Handbook of Industrial Automation

edited by

Richard L. Shell
Ernest L. Hall

University of Cincinnati
Cincinnati, Ohio

MARCEL DEKKER, INC.   NEW YORK • BASEL

ISBN: 0-8247-0373-1

This book is printed on acid-free paper.

Headquarters
Marcel Dekker, Inc.
270 Madison Avenue, New York, NY 10016
tel: 212-696-9000; fax: 212-685-4540

Eastern Hemisphere Distribution
Marcel Dekker AG
Hutgasse 4, Postfach 812, CH-4001 Basel, Switzerland
tel: 41-61-261-8482; fax: 41-61-261-8896

World Wide Web
http://www.dekker.com

The publisher offers discounts on this book when ordered in bulk quantities. For more information, write to Special Sales/Professional Marketing at the headquarters address above.

Current printing (last digit):
10 9 8 7 6 5 4 3 2 1

PRINTED IN THE UNITED STATES OF AMERICA

Preface

This handbook is designed as a comprehensive reference for the industrial automation engineer. Whether in a small or large manufacturing plant, the industrial or manufacturing engineer is usually responsible for using the latest and best technology in the safest, most economic manner to build products. This responsibility requires an enormous knowledge base that, because of changing technology, can never be considered complete. The handbook will provide a handy starting reference covering technical, economic, certain legal standards, and guidelines that should be the first source for solutions to many problems. The book will also be useful to students in the field as it provides a single source for information on industrial automation.

The handbook is also designed to present a related and connected survey of engineering methods useful in a variety of industrial and factory automation applications. Each chapter is arranged to permit review of an entire subject, with illustrations to provide guideposts for the more complex topics. Numerous references are provided to other material for more detailed study.

The mathematical definitions, concepts, equations, principles, and application notes for the practicing industrial automation engineer have been carefully selected to provide broad coverage. Selected subjects from both undergraduate- and graduate-level topics from industrial, electrical, computer, and mechanical engineering as well as material science are included to provide continuity and depth on a variety of topics found useful in our work in teaching thousands of engineers who work in the factory environment. The topics are presented in a tutorial style, without detailed proofs, in order to incorporate a large number of topics in a single volume.

The handbook is organized into ten parts. Each part contains several chapters on important selected topics. Part 1 is devoted to the foundations of mathematical and numerical analysis. The rational thought process developed in the study of mathematics is vital in developing the ability to satisfy every concern in a manufacturing process. Chapters include: an introduction to probability theory, sets and relations, linear algebra, calculus, differential equations, Boolean algebra and algebraic structures and applications. Part 2 provides background information on measurements and control engineering. Unless we measure we cannot control any process. The chapter topics include: an introduction to measurements and control instrumentation, digital motion control, and in-process measurement.

Part 3 provides background on automatic control. Using feedback control in which a desired output is compared to a measured output is essential in automated manufacturing. Chapter topics include distributed control systems, stability, digital signal processing and sampled-data systems. Part 4 introduces modeling and operations research. Given a criterion or goal such as maximizing profit, using an overall model to determine the optimal solution subject to a variety of constraints is the essence of operations research. If an optimal goal cannot be obtained, then continually improving the process is necessary. Chapter topics include: regression, simulation and analysis of manufacturing systems, Petri nets, and decision analysis.

Part 5 deals with sensor systems. Sensors are used to provide the basic measurements necessary to control a manufacturing operation. Human senses are often used but modern systems include important physical sensors. Chapter topics include: sensors for touch, force, and torque, fundamentals of machine vision, low-cost machine vision and three-dimensional vision. Part 6 introduces the topic of manufacturing. Advanced manufacturing processes are continually improved in a search for faster and cheaper ways to produce parts. Chapter topics include: the future of manufacturing, manufacturing systems, intelligent manufacturing systems in industrial automation, measurements, intelligent industrial robots, industrial materials science, forming and shaping processes, and molding processes. Part 7 deals with material handling and storage systems. Material handling is often considered a necessary evil in manufacturing but an efficient material handling system may also be the key to success. Topics include an introduction to material handling and storage systems, automated storage and retrieval systems, containerization, and robotic palletizing of fixed- and variable-size parcels.

Part 8 deals with safety and risk assessment. Safety is vitally important, and government programs monitor the manufacturing process to ensure the safety of the public. Chapter topics include: investigative programs, government regulation and OSHA, and standards. Part 9 introduces ergonomics. Even with advanced automation, humans are a vital part of the manufacturing process. Reducing risks to their safety and health is especially important. Topics include: human interface with automation, workstation design, and physical-strength assessment in ergonomics. Part 10 deals with economic analysis. Returns on investment are a driver to manufacturing systems. Chapter topics include: engineering economy and manufacturing cost recovery and estimating systems.

We believe that this handbook will give the reader an opportunity to quickly and thoroughly scan the field of industrial automation in sufficient depth to provide both specialized knowledge and a broad background of specific information required for industrial automation. Great care was taken to ensure the completeness and topical importance of each chapter.

We are grateful to the many authors, reviewers, readers, and support staff who helped to improve the manuscript. We earnestly solicit comments and suggestions for future improvements.

Richard L. Shell
Ernest L. Hall

Contents

Part 1 Mathematics and Numerical Analysis

Part 2 Measurements and Computer Control

Part 3 Automatic Control

Part 4 Modeling and Operations Research

Part 5 Sensor Systems

Part 6 Manufacturing

Part 10 Economic Analysis

Contributors

Hyder Nihal Agha Research and Development, Motoman, Inc., West Carrollton, Ohio

C. Ray Asfahl University of Arkansas, Fayetteville, Arkansas

William E. Barkman Fabrication Systems Development, Lockheed Martin Energy Systems, Inc., Oak Ridge, Tennessee

Benita M. Beamon Department of Industrial Engineering, University of Washington, Seattle, Washington

Ludwig Benner, Jr. Events Analysis, Inc., Alexandria, Virginia

Amit Bhattacharya Environmental Health Department, University of Cincinnati, Cincinnati, Ohio

Ken Bloemer Ethicon Endo-Surgery Inc., Cincinnati, Ohio

Richard Brook Off Campus Ltd., Palmerston North, New Zealand

William C. Brown Department of Mathematics, Michigan State University, East Lansing, Michigan

Jin Cao Department of Mechanical, Industrial, and Nuclear Engineering, University of Cincinnati, Cincinnati, Ohio

Ming Cao Department of Mechanical, Industrial, and Nuclear Engineering, University of Cincinnati, Cincinnati, Ohio

Enrique Castillo Applied Mathematics and Computational Sciences, University of Cantabria, Santander, Spain

Frank S. Cheng Industrial and Engineering Technology Department, Central Michigan University, Mount Pleasant, Michigan

Ron Collier Department of Mechanical, Industrial, and Nuclear Engineering, University of Cincinnati, Cincinnati, Ohio

Terry R. Collins Department of Industrial Engineering, University of Arkansas, Fayetteville, Arkansas

Jane Cronin Department of Mathematics, Rutgers University, New Brunswick, New Jersey

Richard M. Crowder Department of Electronics and Computer Science, University of Southampton, Southampton, England

Richard B. Darst Department of Mathematics, Colorado State University, Fort Collins, Colorado

William H. DeCamp Motoman, Inc., West Carrollton, Ohio

Steve Dickerson Department of Mechanical Engineering, Georgia Institute of Technology, Atlanta, Georgia

Verna Fitzsimmons Department of Mechanical, Industrial, and Nuclear Engineering, University of Cincinnati, Cincinnati, Ohio

Jeannine Gailey Department of Mechanical, Industrial, and Nuclear Engineering, University of Cincinnati, Cincinnati, Ohio

Sean Gallagher Pittsburgh Research Laboratory, National Institute for Occupational Safety and Health, Pittsburgh, Pennsylvania

Patrick H. Garrett Department of Electrical and Computer Engineering and Computer Science, University of Cincinnati, Cincinnati, Ohio

Ashraf M. Genaidy Department of Mechanical, Industrial, and Nuclear Engineering, University of Cincinnati, Cincinnati, Ohio

Wanek Golnazarian General Dynamics Armament Systems, Burlington, Vermont

Prasanthi Guda Department of Mechanical, Industrial, and Nuclear Engineering, University of Cincinnati, Cincinnati, Ohio

Ali S. Hadi Department of Statistical Sciences, Cornell University, Ithaca, New York

Ernest L. Hall Department of Mechanical, Industrial, and Nuclear Engineering, University of Cincinnati, Cincinnati, Ohio

C. P. Han Department of Mechanical Engineering, Florida Atlantic University, Boca Raton, Florida

Thomas R. Huston Department of Mechanical, Industrial, and Nuclear Engineering, University of Cincinnati, Cincinnati, Ohio

Avraam I. Isayev Department of Polymer Engineering, The University of Akron, Akron, Ohio

Ki Hang Kim Mathematics Research Group, Alabama State University, Montgomery, Alabama

Krishnamohan Kola Department of Mechanical, Industrial, and Nuclear Engineering, University of Cincinnati, Cincinnati, Ohio

Eric M. Malstrom[†] Department of Industrial Engineering, University of Arkansas, Fayetteville, Arkansas

John Mandel[*] National Institute of Standards and Technology, Gaithersburg, Maryland

Jon Marvel Padnos School of Engineering, Grand Valley State University, Grand Rapids, Michigan

A. Kader Mazouz Department of Mechanical Engineering, Florida Atlantic University, Boca Raton, Florida

James D. McGlothlin Purdue University, West Lafayette, Indiana

M. Eugene Merchant Institute of Advanced Manufacturing Sciences, Cincinnati, Ohio

Denny Meyer Institute of Information and Mathematical Sciences, Massey University–Albany, Palmerston North, New Zealand

Angelo B. Mingarelli School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada

Anil Mital Department of Industrial Engineering, University of Cincinnati, Cincinnati, Ohio

J. Steven Moore Department of Occupational and Environmental Medicine, The University of Texas Health Center, Tyler, Texas

[*] Retired.
[†] Deceased.

Diego A. Murio Department of Mathematical Sciences, University of Cincinnati, Cincinnati, Ohio

Lawrence E. Murr Department of Metallurgical and Materials Engineering, The University of Texas at El Paso, El Paso, Texas

Joseph H. Nurre School of Electrical Engineering and Computer Science, Ohio University, Athens, Ohio

Stephen L. Parsley ESKAY Corporation, Salt Lake City, Utah

Arunkumar Pennathur University of Texas at El Paso, El Paso, Texas

Dobrivoje Popovic Institute of Automation Technology, University of Bremen, Bremen, Germany

Shivakumar Raman Department of Industrial Engineering, University of Oklahoma, Norman, Oklahoma

George N. Saridis Professor Emeritus, Electrical, Computer, and Systems Engineering Department, Rensselaer Polytechnic Institute, Troy, New York

Richard L. Shell Department of Mechanical, Industrial, and Nuclear Engineering, University of Cincinnati, Cincinnati, Ohio

Christin Shoaf Department of Mechanical, Industrial, and Nuclear Engineering, University of Cincinnati, Cincinnati, Ohio

J. B. Srivastava Department of Mathematics, Indian Institute of Technology, Delhi, New Delhi, India

Terrence J. Stobbe Industrial Engineering Department, West Virginia University, Morgantown, West Virginia

Allen R. Stubberud Department of Electrical and Computer Engineering, University of California Irvine, Irvine, California

Stephen C. Stubberud ORINCON Corporation, San Diego, California

Hiroyuki Tamura Graduate School of Engineering Science, Osaka University, Toyonaka, Osaka, Japan

Fred J. Taylor Department of Electrical and Computer Engineering and Department of Computer and Information Science Engineering, University of Florida, Gainesville, Florida

Herbert R. Tuttle Graduate Engineering Management, University of Kansas, Lawrence, Kansas

William Wrennall The Leawood Group Ltd., Leawood, Kansas

Chapter 1.1

Some Probability Concepts for Engineers

Enrique Castillo
University of Cantabria, Santander, Spain

Ali S. Hadi
Cornell University, Ithaca, New York

1.1 INTRODUCTION

Many engineering applications involve some element of uncertainty [1]. Probability is one of the most commonly used ways to measure and deal with uncertainty. In this chapter we present some of the most important probability concepts used in engineering applications.

The chapter is organized as follows. Section 1.2 first introduces some elementary concepts, such as random experiments, types of events, and sample spaces. Then it introduces the axioms of probability and some of the most important properties derived from them, as well as the concepts of conditional probability and independence. It also includes the product rule, the total probability theorem, and Bayes' theorem.

Section 1.3 deals with unidimensional random variables and introduces three types of variables (discrete, continuous, and mixed) and the corresponding probability mass, density, and distribution functions. Sections 1.4 and 1.5 describe the most commonly used univariate discrete and continuous models, respectively.

Section 1.6 extends the above concepts of univariate models to the case of bivariate and multivariate models. Special attention is given to joint, marginal, and conditional probability distributions.

Section 1.7 discusses some characteristics of random variables, such as the moment-generating function and the characteristic function.

Section 1.8 treats the techniques of variable transformations, that is, how to obtain the probaiblity distribution function of a set of transformed variables when the probability distribution function of the initial set of variables is known. Section 1.9 uses the transformation techniques of Sec. 1.8 to simulate univariate and multivariate data.

Section 1.10 is devoted to order statistics, giving methods for obtaining the joint distribution of any subset of order statistics. It also deals with the problem of limit or asymptotic distribution of maxima and minima.

Finally, Sec. 1.11 introduces probability plots and how to build and use them in making inferences from data.

1.2 BASIC PROBABILITY CONCEPTS

In this section we introduce some basic probability concepts and definitions. These are easily understood from examples. Classic examples include whether a machine will malfunction at least once during the first month of operation, whether a given structure will last for the next 20 years, or whether a flood will occur during the next year, etc. Other examples include how many cars will cross a given intersection during a given rush hour, how long we will have to wait for a certain event to occur, how much stress level a given structure can withstand, etc. We start our exposition with some definitions in the following subsection.

1.2.1 Random Experiment and Sample Space

Each of the above examples can be described as a *random experiment* because we cannot predict in advance the outcome at the end of the experiment. This leads to the following definition:

Definition 1. Random Experiment and Sample Space: *Any activity that will result in one and only one of several well-defined outcomes, but does not allow us to tell in advance which one will occur is called a random experiment. Each of these possible outcomes is called an elementary event. The set of all possible elementary events of a given random experiment is called the sample space and is usually denoted by* Ω.

Therefore, for each random experiment there is an associated sample space. The following are examples of random experiments and their associated sample spaces:

Rolling a six-sided fair die once yields $\Omega = \{1, 2, 3, 4, 5, 6\}$.
Tossing a fair coin once, yields $\Omega = \{Head, Tail\}$.
Waiting for a machine to malfunction yields $\Omega = \{x : x > 0\}$.
How many cars will cross a given intersection yields $\Omega = \{0, 1, \ldots\}$.

Definition 2. Union and Intersection: *If C is a set containing all elementary events found in A or in B or in both, then write* $C = (A \cup B)$ *to denote the union of A and B, whereas, if C is a set containing all elementary events found in both A and B, then we write* $C = (A \cap B)$ *to denote the intersection of A and B.*

Referring to the six-sided die, for example, if $A = \{1, 3, 5\}$, $B = \{2, 4, 6\}$, and $C = \{1, 2, 3\}$, then $(A \cup B) = \Omega$ and $(A \cup C) = \{1, 2, 3, 5\}$, whereas $(A \cap C) = \{1, 3\}$ and $(A \cap B) = \phi$, where ϕ denotes the *empty* set.

Random events in a sample space associated with a random experiment can be classified into several types:

1. *Elementary vs. composite events.* A subset of Ω which contains more than one elementary event is called a *composite event*. Thus, for example, observing an odd number when rolling a six-sided die once is a composite event because it consists of three elementary events.
2. *Compatible vs. mutually exclusive events.* Two events A and B are said to be *compatible* if they can simultaneously occur, otherwise they are said to be *mutually exclusive* or *incompatible* events. For example, referring to rolling a six-sided die once, the events $A = \{1, 3, 5\}$ and $B = \{2, 4, 6\}$ are incompatible because if one event occurs, the other does not, whereas the events A and $C = \{1, 2, 3\}$ are compatible because if we observe 1 or 3, then both A and C occur.
3. *Collectively exhaustive events.* If the union of several events is the sample space, then the events are said to be *collectively exhaustive*. For example, if $\Omega = \{1, 2, 3, 4, 5, 6\}$, then $A = \{1, 3, 5\}$ and $B = \{2, 4, 6\}$ are collectively exhaustive events but $A = \{1, 3, 5\}$ and $C = \{1, 2, 3\}$ are not.
4. *Complementary events.* Given a sample space Ω and an event $A \in \Omega$, let B be the event consisting of all elements found in Ω but not in A. Then A and B are said to be *complementary events* or B is the *complement* of A (or vice versa). The complement of A is usually denoted by $\bar{A}$. For example, in the six-sided die example, if $A = \{1, 2\}$, $\bar{A} = \{3, 4, 5, 6\}$. Note that an event and its complement are always defined with respect to the sample space Ω. Note also that A and $\bar{A}$ are always mutually exclusive and collectively exhaustive events, hence $(A \cap \bar{A}) = \phi$ and $(A \cup \bar{A}) = \Omega$.

1.2.2 Probability Measure

To measure uncertainty we start with a given sample space Ω, in which all mutually exclusive and collectively exhaustive outcomes of a given experiment are included. Next, we select a class of subsets of Ω which are closed under the union, intersection, complementary and limit operations. Such a class is called a σ-algebra. Then, the aim is to assign to every subset in σ a real value measuring the degree of uncertainty about its occurrence. In order to obtain measures with clear physical and practical meanings, some general and intuitive properties are used to define a class of measures known as *probability measures*.

Definition 3. Probability Measure: *A function p mapping any subset* $A \subseteq \sigma$ *into the interval* $[0, 1]$ *is called a probability measure if it satisfies the following axioms:*

Axiom 1. Boundary: $p(\Omega) = 1$.

Axiom 2. Additivity: *For any (possibly infinite) sequence, $A_1, A_2, \ldots,$ of disjoint subsets of σ, then*

$$p\Big(\bigcup A_i\Big) = \sum p(A_i)$$

Axiom 1 states that despite our degree of uncertainty, at least one element in the universal set Ω will occur (that is, the set Ω is exhaustive). Axiom 2 is an aggregation formula that can be used to compute the probability of a union of disjoint subsets. It states that the uncertainty of a given subset is the sum of the uncertainties of its disjoint parts.

From the above axioms, many interesting properties of the probability measure can be derived. For example:

Property 1. Boundary: $p(\phi) = 0$.

Property 2. Monotonicity: If $A \subseteq B \subseteq \sigma$, then $p(A) \leq p(B)$.

Property 3. Continuity–Consistency: For every increasing sequence $A_1 \subseteq A_2 \subseteq \ldots$ or decreasing sequence $A_1 \supseteq A_2 \supseteq \ldots$ of subsets of σ we have

$$\lim_{i\to\infty} p(A_i) = p(\lim_{i\to\infty} A_i)$$

Property 4. Inclusion–Exclusion: Given any pair of subsets A and B of σ, the following equality always holds:

$$p(A \cup B) = p(A) + p(B) - p(A \cap B) \tag{1}$$

Property 1 states that the evidence associated with a complete lack of information is defined to be zero. Property 2 shows that the evidence of the membership of an element in a set must be at least as great as the evidence that the element belongs to any of its subsets. In other words, the certainty of an element belonging to a given set A must not decrease with the addition of elements to A.

Property 3 can be viewed as a consistency or a continuity property. If we choose two sequences converging to the same subset of σ, we must get the same limit of uncertainty. Property 4 states that the probabilities of the sets A, B, $A \cap B$, and $A \cup B$ are not independent; they are related by Eq. (1).

Note that these properties respond to the intuitive notion of probability that makes the mathematical model valid for dealing with uncertainty. Thus, for example, the fact that probabilities cannot be larger than one is not an axiom but a consequence of Axioms 1 and 2.

Definition 4. Conditional Probability: *Let A and B be two subsets of variables such that $p(B) > 0$. Then, the conditional probability distribution (CPD) of A given B is given by*

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)} \tag{2}$$

Equation (2) implies that the probability of $A \cap B$ can be written as

$$p(A \cap B) = p(B)p(A \mid B) \tag{3}$$

This can be generalized to several events as follows:

$$p(A \mid B_1, \ldots, B_k) = \frac{p(A, B_1, \ldots, B_k)}{p(B_1, \ldots, B_k)} \tag{4}$$

1.2.3 Dependence and Independence

Defintion 5. Independence of Two Events: *Let A and B be two events. Then A is said to be independent of B if and only if*

$$p(A \mid B) = p(A) \tag{5}$$

otherwise A is said to be dependent on B.

Equation (5) means that if A is independent of B, then our knowledge of B does not affect our knowledge about A, that is, B has no information about A. Also, if A is independent of B, we can then combine Eqs. (2) and (5) and obtain

$$p(A \cap B) = p(A)p(B) \tag{6}$$

Equation (6) indicates that if A is independent of B, then the probability of $A \cap B$ is equal to the product of their probabilities. Actually, Eq. (6) provides a definition of independence equivalent to that in Eq. (5).

One important property of the independence relation is its *symmetry*, that is, if A is independent of B, then B is independent of A. This is because

$$p(B \mid A) = \frac{p(A \cap B)}{p(A)} = \frac{p(A)p(B)}{p(A)} = p(B)$$

Because of the symmetry property, we say that A and B are *independent* or *mutually independent*. The practical implication of symmetry is that if knowledge of B is relevant (irrelevant) to A, then knowledge of A is relevant (irrelevant) to B.

The concepts of dependence and independence of two events can be extended to the case of more than two events as follows:

Definition 6. Independence of a Set of Events: *The events $A_1, \ldots, A_m$ are said to be independent if and only if*

$$p(A_1 \cap \ldots \cap A_m) = \prod_{i=1}^{m} p(A_i) \tag{7}$$

otherwise they are said to be dependent.

In other words, $\{A_1, \ldots, A_m\}$ are said to be independent if and only if their intersection probability is equal to the product of their individual probabilities. Note that Eq. (7) is a generalization of Eq. (6).

An important implication of independence is that it is not worthwhile gathering information about independent (irrelevant) events. That is, independence means irrelevance.

From Eq. (3) we get

$$p(A_1 \cap A_2) = p(A_1 \mid A_2)p(A_2) = p(A_2 \mid A_1)p(A_1)$$

This property can be generalized, leading to the so-called *product* or *chain rule*:

$$p(A_1 \cap \ldots \cap A_n) = p(A_1)p(A_2 \mid A_1) \ldots p(A_n \mid A_1 \cap \ldots \cap A_{n-1})$$

1.2.4 Total Probability Theorem

Theorem 1. Total Probability Theorem: *Let $\{A_1, \ldots, A_n\}$ be a class of events which are mutually incompatible and such that $\bigcup_{1 \le i \le n} A_i = \Omega$. Then we have*

$$p(B) = \sum_{1 \le i \le n} p(B \mid A_i)p(A_i)$$

A graphical illustration of this theorem is given in Fig. 1.

1.2.5 Bayes' Theorem

Theorem 2. Bayes' Theorem: *Let $\{A_1, \ldots, A_n\}$ be a class of events which are mutually incompatible and such that $\bigcap_{1 \le i \le n} A_i = \Omega$. Then,*

$$p(A_i \mid B) = \frac{p(B \mid A_i)p(A_i)}{\sum_{1 \le i \le n} p(B \mid A_i)p(A_i)}$$

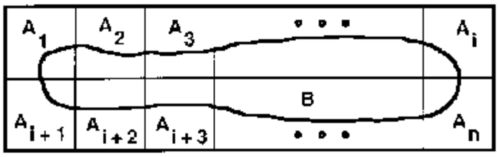


Figure 1 Graphical illustration of the total probability rule.

Probabilities $p(A_i)$ are called *prior probabilities*, because they are the probabilities before knowing the information B. Probabilities $p(A_i \mid B)$, which are the probabilities of A_i after the knowledge of B, are called *posterior probabilities*. Finally, $p(B \mid A_i)$ are called *likelihoods*.

1.3 UNIDIMENSIONAL RANDOM VARIABLES

In this section we define random variables, distinguish among three of their types, and present various ways of presenting their probability distributions.

Definition 7. Random Variable: *A possible vector-valued function $\mathbf{X} : \Omega \to \mathbf{R}^n$, which assigns to each element $\omega \in \Omega$ one and only one vector of real numbers $\mathbf{X}(\omega) = \mathbf{x}$, is called an n-dimensional random variable. The space of $\mathbf{X}$ is $\{\mathbf{x} : \mathbf{x} = \mathbf{X}(\omega, \omega \in \Omega\}$. The space of a random variable $\mathbf{X}$ is also known as the support of $\mathbf{X}$.*

When $n = 1$ in Definition 7, the random variable is said to be *unidimensional* and when $n > 1$, it is said to be *multidimensional*. In this and Secs 1.4 and 1.5, we deal with unidimensional random variables. Multidimensional random variables are treated in Sec. 1.6.

Example 1. *Suppose we roll two dice once. Let A be the outcome of the first die and B be the outcome of the second. Then the sample space $\Omega = \{(1, 1), \ldots (6, 6)\}$ consists of 36 possible pairs (A,B), as shown in Fig. 2. Suppose we define a random variable $X = A + B$, that is, X is the sum of the two numbers observed when we roll two dice once. Then X is a unidimensional random variable. The support of this random variable is the set* $\{2, 3, \ldots, 12\}$ *consisting of 11 elements. This is also shown in Fig. 2.*

1.3.1 Types of Random Variables

Random variables can be classified into three types: discrete, continuous, and mixed. We define and give examples of each type below.

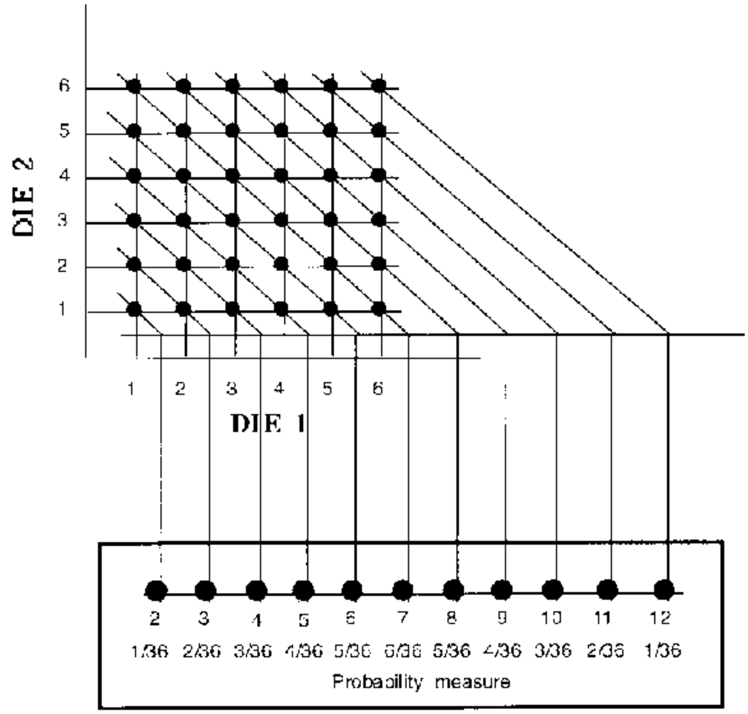


Figure 2 Graphical illustration of an experiment consisting of rolling two dice once and an associated random variable which is defined as the sum of the two numbers observed.

Definition 8. Discrete Random Variables: *A random variable is said to be* discrete *if it can take a finite or countable set of real values.*

As an example of a discrete random variable, let X denote the outcome of rolling a six-sided die once. Since the support of this random variable is the finite set $\{1, 2, 3, 4, 5, 6\}$, then X is discrete random variable. The random variable $X = A + B$ in Fig. 2 is another example of discrete random variables.

Definition 9. Continuous Random Variables: *A random variable is said to be* continuous *if it can take an uncountable set of real values.*

For example, let X denote the weight of an object, then X is a continuous random variable because it can take values in the set $\{x : x > 0\}$, which is an uncountable set.

Definition 10. Mixed Random Variables: *A random variable is said to be mixed if it can take an uncountable set of values and the probability of at least one value of x is positive.*

Mixed random variables are encountered often in engineering applications which involve some type of censoring. Consider, for example, a life-testing situation where n machines are put to work for a given period of time, say 30 days. Let X_i denotes the time at which the ith machine malfunctions. Then X_i is a random variable which can take the values $\{x : 0 < x \leq 30\}$. This is clearly an uncountable set. But at the end of the 30-day period some machines may still be functioning. For each of these machines all what we know is that $X_i \geq 30\}$. Then the probability that $X_i = 30$ is positive. Hence the random variable X_i is of the mixed type. The data in this example is known as *censored* data.

Censoring can be of two types: *right censoring* and *left censoring*. The above example is of the former type. An example of the latter type occurs when we measure say, pollution, using an instrument which cannot detect polution below a certain limit. In this case we have left censoring because only small values are censored. Of course, there are situations where both right and left censoring are present.

1.3.2 Probability Distributions of Random Variables

So far we have defined random variables and their support. In this section we are interested in measuring the probability of each of these values and/or the probability of a subset of these values. We know from Axiom 1 that $p(\Omega) = 1$; the question is then how this probability of 1 is distributed over the elements of Ω. In other words, we are interested in finding the probability distribution of a given random variable. Three equivalent ways of representing the probability distributions of these random variables are: tables, graphs, and mathematical functions (also known as mathematical models).

1.3.3 Probability Distribution Tables

As an example of a probability distribution that can be displayed in a table let us flip a fair coin twice and let X be the number of heads observed. Then the sample space of this random experiment is $\Omega = \{TT, TH, HT, HH\}$, where TH, for example, denotes the outcome: first coin turned up a tail and second a head. The sample space of the random variable X is then $\{0, 1, 2\}$. For example, $X = 0$ occurs when we observe TT. The probability of each of these possible values of X is found simply by counting how many elements of Ω are associated with each value in the support of X. We can see that $X = 0$ occurs when we observe the outcome TT, $X = 1$ occurs when we observe either HT or TH, and $X = 2$ occurs when we observe HH. Since there are four equally likely elementary events in Ω, each element has a probability of 1/4. Hence, $p(X = 0) = 1/4$, $p(X = 1) = 2/4$, and $p(X = 2) = 1/4$. This probability distribution of X can be displayed in a table as in Table 1. For obvious reasons, such tables are called *probability distribution tables*. Note that to denote the random variable itself we use an uppercase letter (e.g., X), but for its realizations we use the corresponding lowercase letter (e.g., x).

Table 1 The Probability Distribution of the Random Variable X Defined as the Number of Heads Resulting from Flipping a Fair Coin Twice

x	$p(x)$
0	0.25
1	0.50
2	0.25

Obviously, it is possible to use tables to display the probability distributions of only discrete random variables. For continuous random variables, we have to use one of the other two means: graphs or mathematical functions. Even in discrete random variables with large number of elements in their support, tables are not the most efficient way of displaying the probability distribution.

1.3.4 Graphical Representation of Probabilities

The probability distribution of a random variable can equivalently be represented graphically by displaying values in the support of X on a horizontal line and erecting a vertical line or bar on top of each of these values. The height of each line or bar represents the probability of the corresponding value of X. For example, Fig. 3 shows the probability distribution of the random variable X defined in Example 1.

For continuous random variables, we have infinitely many possible values in their support, each of which has a probability equal to zero. To avoid this difficulty, we represent the probability of a subset of values by an area under a curve (known as the *probability density curve*) instead of heights of vertical lines on top of each of the values in the subset.

For example, let X represent a number drawn randomly from the interval [0, 10]. The probability distribution of X can be displayed graphically as in Fig. 4. The area under the curve on top of the support of X has to equal 1 because it represents the total probability. Since all values of X are equally likely, the curve is a horizontal line with height equal to 1/10. The height of 1/10 will make the total area under the curve equal to 1. This type of random variable is called a *contin-*

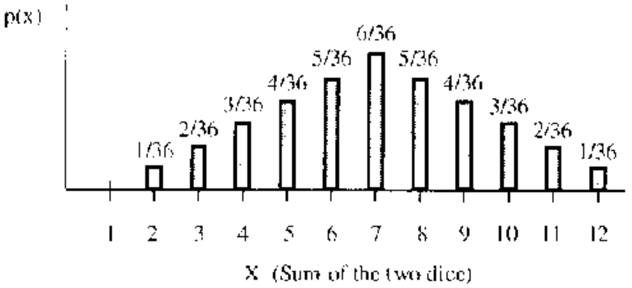


Figure 3 Graphical representation of the probability distribution of the random variable X in Example 1.

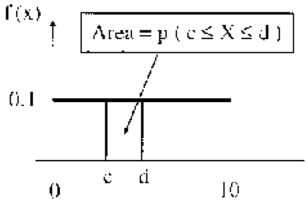


Figure 4 Graphical representation of the pdf of the $U(0, 10)$ random variable X.

uous uniform random variable and is dentoed by $U(a, b)$, where in this example $a = 0$ and $b = 10$.

If we wish, for example, to find the probability that X is between 2 and 6, this probability is represented by the shaded area on top of the interval (2, 6). Note here that the heights of the curve do not represent probabilities as in the discrete case. They represent the *density* of the random variable on top of each value of X.

1.3.5 Probability Mass and Density Functions

Alternatively to tables and graphs, a probability distribution can be displayed using a mathematical function. For example, the probability distribution of the random variable X in Table 1 can be written as

$$p(X = x) = \begin{cases} 0.25 & \text{if } x \in \{0, 2\} \\ 0.50 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

A function like the one in Eq. (8) is known as a *probability mass function* (pmf). Examples of the pmf of other popular discrete random variables are given in Sec. 1.4. Sometimes we write $p(X = x)$ as $p(x)$ for simplicity of notation.

Note that every pmf $p(x)$ must satisfy the following conditions:

$$p(x) > 0, \forall x \in A; \quad p(x) = 0, \forall x \notin A; \quad \sum_{x \in A} p(x) = 1$$

where A is the support of X.

As an example of representing a continuous random variable using a mathematical function, the graph of the continuous random variable X in Fig. 4 can be represented by the function

$$f(x) = \begin{cases} 0.1 & \text{if } 0 \le x \le 10 \\ 0 & \text{otherwise} \end{cases}$$

The pdf for the general uniform random variable $U(a, b)$ is

$$f(x) = \begin{cases} \dfrac{1}{b - a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Functions like the one in Eq. (9) are known as a *probability density function* (pdf). Examples of the pdf of other popular continuous random variables are given in Sec. 1.5. To distinguish between probability mass and density functions, the former is denoted by $p(x)$ (because it represents the probability that $X = x$) and the latter by $f(x)$ (because it represents the height of the curve on top of x).

Note that every pdf $f(x)$ must satisfy the following conditions:

$$f(x) > 0, \forall x \in A; \quad f(x) = 0, \forall x \notin A; \quad \int_{x \in A} f(x) = 1$$

where A is the support of X.

Probability distributions of mixed random variables can also be represented graphically and using *probability mass–density functions* (pmdf). The pmdf of a mixed random variable X is a pair of functions $p(x)$ and $f(x)$ such that they allow determining the probabilities of X to take given values, and X to belong to given intervals, respectively. Thus, the probability of X to take values in the interval (a, b) is given by

$$\sum_{x > a}^{x < b} p(x) + \int_a^b f(x)\, dx$$

The interpretation of each of these functions coincides with that for discrete and continuous random variables. The pmdf has to satisfy the following conditions:

$$p(x) \ge 0, \; f(x) \ge 0, \qquad \sum_{x > -\infty}^{x < \infty} p(x) + \int_{-\infty}^{\infty} f(x)\, dx = 1$$

which are an immediate consequence of their definitions.

1.3.6 Cumulative Distribution Function

An alternative way of defining the probability mass–density function of a random variable is by means of the *cumulative distribution function* (cdf). The cdf of a random variable X is a function that assigns to each real value x the probability of X having values less than or equal to x. Thus, the cdf for the discrete case is

$$P(x) = p(X \le x) = \sum_{a \le x} p(x)$$

and for the continuous case is

$$F(x) = p(X \le x) = \int_{-\infty}^{x} f(x)\,dx$$

Note that the cdfs are denoted by the uppercase letters $P(x)$ and $F(x)$ to distinguish them from the pmf $p(x)$ and the pdf $f(x)$. Note also that since $p(X = x) = 0$ for the continuous case, then $p(X \le x) = p(X < x)$. The cdf has the following properties as a direct consequence of the definitions of cdf and probability:

$F(\infty) = 1$ and $F(-\infty) = 0$.
$F(x)$ is nondecreasing and right continuous.
$f(x) = dF(x)/dx$.
$p(X = x) = F(x) - F(x - 0)$, where $F(x - 0) = \lim_{\varepsilon \to 0} F(x - \varepsilon)$.
$p(a < X \le b) = F(b) - F(a)$.
The set of discontinuity points of $F(x)$ is finite or countable.
Every distribution function can be written as a linear convex combination of continuous distributions and step functions.

1.3.7 Moments of Random Variables

The pmf or pdf of random variables contains all the information about the random variables. For example, given the pmf or the pdf of a given random variable, we can find the mean, the variance, and other moments of the random variable. The results in this section are presented for the continuous random variables using the pdf and cdf, $f(x)$ and $F(x)$, respectively. For the discrete random variables, the results are obtained by replacing $f(x)$, $F(x)$, and the integration symbol by $p(x)$, $P(x)$, and the summation symbol, respectively.

Definition 11. Moments of Order *k*: *Let X be a random variable with pdf $f(x)$, cdf $F(x)$, and support A. Then the kth moment m_k around $a \in A$ is the real number*

$$m_k = \int_A (x - a)^k f(x)\,dx \tag{10}$$

The moments around $a = 0$ are called the central *moments.*

Note that the Stieltjes–Lebesgue integral, Eq. (10), does not always exist. In such a case we say that the corresponding moment does not exist. However, Eq. (10) implies the existence of

$$\int_A |x - a|^k f(x)\,dx$$

which leads to the following theorem:

Theorem 3. Existence of Moments of Lower Order: *If the tth moment around a of a random variable X exists, then the sth moment around a also exists for $0 < s \le t$.*

The first central moment is called the *mean* or the *expected value* of the random variable X, and is denoted by μ or $E[X]$. Let X and Y be random variables, then the expectation operator has the following important properties:

$E[c] = c$, where c is a constant.
$E[aX + bY + c] = aE[X] + bE[Y] + c$; $\forall a, b, c \in \mathbf{A}$.
$a \le Y \le b \Rightarrow a \le E[Y] \le b$.
$|E[Y]| \le E[|y|]$.

The second moment around the mean is called the *variance* of the random variable, and is denoted by $\text{Var}(X)$ or σ^2. The square root of the variance, σ, is called the *standard deviation* of the random variable. The physical meanings of the mean and the variance are similar to the center of gravity and the moment of inertia, used in mechanics. They are the central and dispersion measures, respectively.

Using the above properties we can write

$$\begin{aligned}\sigma^2 &= E[(X - \mu)^2]\\ &= E[X^2 - 2X\mu + \mu^2]\\ &= E[X^2] - 2\mu E[X] + \mu^2 E[1]\\ &= E[X^2] - 2\mu^2 + \mu^2\\ &= E[X^2] - \mu^2\end{aligned} \tag{11}$$

which gives an important relationship between the mean and variance of the random variable. A more general expression can be similarly obtained:

$$E[(X - a)^2] = \sigma^2 + (\mu - a)^2$$

1.4 UNIVARIATE DISCRETE MODELS

In this section we present several important discrete probability distributions that often arise in engineering applications. Table 2 shows the pmf of these distributions. For additional probability distributions, see Christensen [2] and Johnson et al. [3].

Table 2 Some Discrete Probability Mass Functions that Arise in Engineering Applications

Distribution	$p(x)$	Parameters and support
Bernoulli	$\begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \end{cases}$	$0 < p < 1$ $x \in \{0, 1\}$
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$	$n \in \{1, 2, \ldots\}$ $0 < p < 1$ $x \in \{0, 1, \ldots, n\}$
Nonzero binomial	$\dfrac{\binom{n}{x} p^x (1-p)^{n-x}}{1-(1-p)^n}$	$n \in \{1, 2, \ldots\}$ $0 < p < 1$ $x \in \{1, 2, \ldots, n\}$
Geometric	$p(1-p)^{x-1}$	$0 < p < 1$ $x \in \{1, 2, \ldots\}$
Negative binomial	$\binom{x-1}{r-1} p^r (1-p)^{x-r}$	$n \in \{1, 2, \ldots\}$ $0 < p < 1$ $x \in \{0, 1, \ldots, n\}$
Hypergeometric	$\binom{D}{x}\binom{N-D}{n-x} \Big/ \binom{N}{n}$	$(n, N) \in \{1, 2, \ldots\}, n < N$ $\max(0, n-N+D) \leq x \leq \min(n, D)$
Poisson	$\dfrac{e^{-\lambda}\lambda^x}{x!}$	$\lambda > 0$ $x \in \{0, 1, \ldots\}$
Nonzero Poisson	$\dfrac{\lambda^x}{x!(e^{-\lambda}-1)}$	$\lambda > 0$ $x \in \{1, 2, \ldots\}$
Logarithmic series	$\dfrac{-\alpha^x}{x \ln(1-p)}$	$0 < p < 1$ $\alpha > 0$ $x \in \{1, 2, \ldots\}$
Discrete Weibull	$(1-p)^{x^\alpha} - (1-p)^{(x+1)^\alpha}$	$0 < p < 1, \alpha > 0$ $x \in \{0, 1, \ldots\}$
Yule	$\dfrac{n\Gamma(x)\Gamma(n+1)}{\Gamma(n+x+1)}$	$(x, n) \in \{1, 2, \ldots\}$

1.4.1 The Bernoulli Distribution

The *Bernoulli distribution* arises in the following situation. Assume that we have a random experiment with two possible mutually exclusive outcomes: *success*, with probability p, and *failure*, with probability $1-p$. This experiment is called a *Bernoulli trial*. Define a random variable X by

$$X = \begin{cases} 1 & \text{if we obtain success} \\ 0 & \text{if we obtain failure} \end{cases}$$

Then, the pmf of X is as given in Table 2 under the Bernoulli distribution. It can be shown that the corresponding cdf is

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1-p & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

Both the pmf and cdf are presented graphically in Fig. 5.

1.4.2 The Discrete Uniform Distribution

The discrete uniform random variable $U(n)$ is a random variable which takes n equally likely values. These values are given by its support A. Its pmf is

$$p(X = x) = \begin{cases} 1/n & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

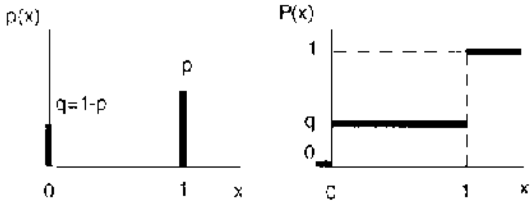


Figure 5 A graph of the pmf and cdf of a Bernoulli distribution.

1.4.3 The Binomial Distribution

Suppose now that we repeat a Bernoulli experiment n times under identical conditions (that is, the outcome of one trial does not affect the outcomes of the others). In this case the trials are said to be independent. Suppose also that the probability of success is p and that we are interested in the number of trials, X in which the outcomes are successes. The random variable giving the number of successes after n realizations of independent Bernoulli experiments is called a *binomial* random variable and is denoted as $B(n, p)$. Its pmf is given in Table 2. Figure 6 shows some examples of pmfs associated with binomial random variables.

In certain situations the event $X = 0$ cannot occur. The pmf of the binomial distribution can be modified to accommodate this case. The resultant random variable is called the *nonzero binomial*. Its pmf is given in Table 2.

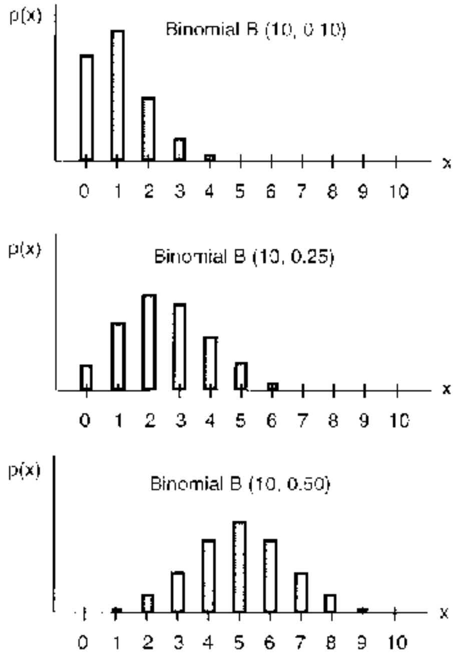


Figure 6 Examples of the pmf of binomial random variables.

1.4.4 The Geometric or Pascal Distribution

Suppose again that we repeat a Bernoulli experiment n times, but now we are interested in the random variable X, defined to be the number of Bernoulli trials that are required until we get the first success. Note that if the first success occurs in the trial number x, then the first $(x - 1)$ trials must be failures (see Fig. 7). Since the probability of a success is p and the probability of the $(x - 1)$ failures is $(1 - p)^{x-1}$ (because the trials are independent), then the $p(X = x) = p(1 - p)^{x-1}$. This random variable is called the *geometric* or *Pascal* random variable and is denoted by $G(p)$.

1.4.5 The Negative Binomial Distribution

The geometric distribution arises when we are interested in the number of Bernoulli trials that are required until we get the first success. Now suppose that we define the random variable X as the number of Bernoulli trials that are required until we get the rth success. For the rth success to occur at the xth trial, we must have $(r - 1)$ successes in the $(x - 1)$ previous trials and one success in the rth trial (see Fig. 8). This random variable is called the *negative binomial* random variable and is denoted by $NB(r, p)$. Its pmf is given in Table 2. Note that the gometric distribution is a special case of the negative binomial distribution obtained by setting $(r = 1)$, that is, $G(p) = NB(1, p)$.

1.4.6 The Hypergeometric Distribution

Consider a set of N items (products, machines, etc.), D items of which are defective and the remaining $(N - D)$ items are acceptable. Obtaining a random sample of size n from this finite population is equivalent to withdrawing the items one by one without replacement.

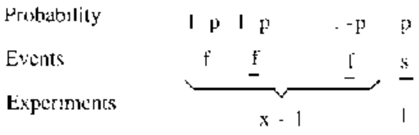


Figure 7 Illustration of the Pascal or geometric random variable, where s denotes success and f denotes failure.

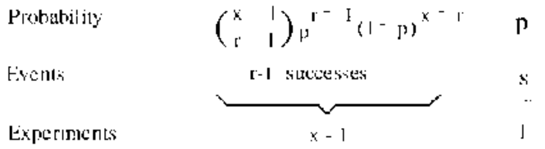


Figure 8 An illustration of the negative binomial random variable.

This yields the hypergeometric random variable, which is defined to be the number of defective items in the sample and is denoted by $HG(N, D, n)$.

Obviously, the number X of defective items in the sample cannot exceed the total number of defective items D nor the sample size n. Similarly, the number $(n - X)$ of acceptable items in the sample cannot be less than zero or exceed n minus the total number of acceptable items $(N - D)$. Thus, we must have $\max(0, n - (N - D)) \leq X \leq \min(n, D)$. This random variable has the *hypergeometric distribution* and its pmf is given in Table 2. Note that the numerator in the pmf is the number of possible samples with x defective and $(n - x)$ acceptable items, and that the denominator is the total number of possible samples.

The mean and variance of the hypergeometric random variable are D and

$$\frac{D(N-n)}{N-1}\left(1 - \frac{D}{N}\right)$$

respectively. When N tends to infinity this distribution tends to the binomial distribution.

1.4.7 The Poisson Distribution

There are events which are not the result of a series of experiments but occur in random time instants or locations. For example, we can be interested in the number of traffic accidents occurring in a time interval, or the number of vehicles arriving at a given intersection.

For these types of random variables we can make the following (Poisson) assumptions:

The probability of occurrence of a single event in an interval of brief duration dt is $\alpha\, dt$, that is, $p_{dt}(1) = \alpha\, dt + o(dt)^2$, where α is a positive constant.

The probability of occurrence of more than one event in the same interval dt, is negligible with respect to the previous one, that is

$$\lim_{dt \to 0} \frac{p_{dt}(x)}{dt} = 0 \qquad (\text{for } x > 1)$$

The number of events occurring in two nonoverlapping intervals are independent random variables.

The probabilities $p_t(x)$ of x events in two time intervals of identical duration, t, are the same.

Based on these assumptions, it can be shown that the pmf of this random variable is:

$$p_t(x) = \frac{e^{-\alpha t}\alpha^x t^x}{x!}$$

Letting $\lambda = \alpha t$, we obtain the pmf of the Poisson random variable as given in Table 2. Thus, the Poisson random variable gives the number of events occurring in period of given duration and is denoted by $P(\lambda)$, where $\lambda = \alpha t$, that is, the intensity α times the duration t.

As in the nonzero binomial case, in certain situations the event $X = 0$ cannot occur. The pmf of the Poisson distribution can be modified to accommodate this case. The resultant random variable is called the *nonzero Poisson*. Its pmf is given in Table 2.

1.5 UNIVARIATE CONTINUOUS MODELS

In this section we give several important continuous probability distributions that often arise in engineering applications. Table 3 shows the pdf and cdf of these distributions. For additional probability distributions, see Christensen [2] and Johnson et al. [4].

1.5.1 The Continuous Uniform Distribution

The uniform random variable $U(a, b)$ has already been introduced in Sec. 1.3.5. Its pdf is given in Eq. (9), from which it follows that the cdf can be written as (see Fig. 9):

$$F(x) = \begin{cases} 0 & \text{if } x < a \\ \dfrac{x-a}{b-a} & \text{if } a \leq x < b \\ 1 & \text{if } x \geq b \end{cases}$$

1.5.2 The Exponential Distribution

The exponential random variable gives the time between two consecutive Poisson events. To obtain its cdf $F(x)$ we consider that the probability of X exceeding x is equal to the probability of no events occurring in a period of duration x. But the probability of the first event is $1 - F(x)$, and the probability of zero events is given by the Poisson probability distribution. Thus, we have

Table 3 Some Continuous Probability Density Functions that Arise in Engineering Applications

Distribution	$p(x)$	Parameters and Support
Uniform	$\dfrac{1}{b-a}$	$a < b$ $a < x < b$
Exponential	$\lambda e^{-\lambda x}$	$\lambda > 0$ $x > 0$
Gamma	$\dfrac{\lambda(\lambda x)^{k-1} e^{-\lambda x}}{\Gamma(k)}$	$\lambda > 0, k \in \{1, 2, \ldots\}$ $x \geq 0$
Beta	$\dfrac{\Gamma(r+t)}{\Gamma(r)\Gamma(t)} x^{r-1}(1-x)^{t-1}$	$r, t > 0$ $0 \leq x \leq 1$
Normal	$\dfrac{1}{\sigma\sqrt{2\pi}} \exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2}\right)$	$-\infty < \mu < \infty$ $\sigma > 0$ $-\infty < x < \infty$
Log–normal	$\dfrac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\dfrac{(\ln x-\mu)^2}{2\sigma^2}\right)$	$-\infty < \mu < \infty$ $\sigma > 0$ $x \geq 0$
Central chi-squared	$\dfrac{e^{-x/2} x^{(n/2)-1}}{2^{n/2}\Gamma(n/2)}$	$n \in \{1, 2, \ldots\}$ $x \geq 0$
Rayleigh	$\dfrac{x}{\sigma^2} \exp\left(-\dfrac{x^2}{2\sigma^2}\right)$	$\sigma > 0$ $x \geq 0$
Central t	$\dfrac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{n\pi}} \left(1+\dfrac{x^2}{n}\right)^{-(n+1)/2}$	$n \in \{1, 2, \ldots\}$ $-\infty < x < \infty$
Central F	$\dfrac{\Gamma((n_1+n_2)/2) n_1^{n_1/2} n_2^{n_2/2} x^{(n_1/2)-1}}{\Gamma(n_1/2)\Gamma(n_2/2)(n_1 x + n_2)^{(n_1+n_2)/2}}$	$(n_1, n_2) \in \{1, 2, \ldots\}$ $x \geq 0$

$$1 - F(x) = p_0(x) = e^{-\lambda x}$$

from which follows the cdf:

$$F(x) = 1 - e^{-\lambda x} \qquad x > 0$$

Taking the derivative of $F(x)$ with respect to x, we obtain the pdf

$$f(x) = \frac{dF(x)}{dx} = \lambda e^{-\lambda x} \qquad x > 0$$

The pdf and cdf for the exponential distribution are drawn in Fig. 10.

1.5.3 The Gamma Distribution

Let Y be a Poisson random variable with parameter λ. Let X be the time up to the kth Poisson event, that is, the time it takes for Y to be equal to k. Thus the probability that X is in the interval $(x, x+dx)$ is $f(x)\,dx$. But this probability is equal to the probability of there having occurred $(k-1)$ Poisson events in a period of duration x times the probability of occurrence of one event in a period of duration dx. Thus, we have

$$f(x)\,dx = \frac{e^{-\lambda x}(\lambda x)^{k-1}}{(k-1)!}\,\lambda\,dx$$

from which we obtain

$$f(x) = \frac{\lambda(\lambda x)^{k-1} e^{-\lambda x}}{(k-1)!} \qquad 0 \leq x < \infty \tag{12}$$

Expression (12), taking into account that the gamma function for an integer k satisfies

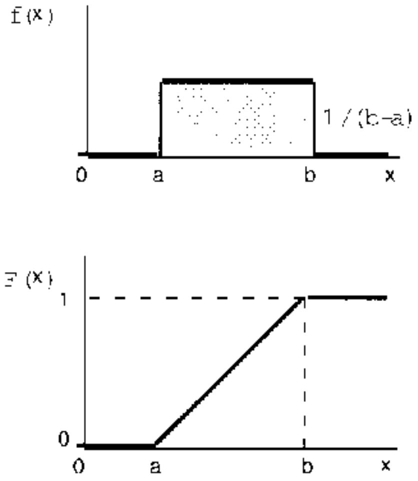


Figure 9 An example of pdf and cdf of the uniform random variable.

$$\Gamma(k) = \int_0^\infty e^{-u} u^{k-1}\, dx = (k-1)! \tag{13}$$

can be written as

$$f(x) = \frac{\lambda(\lambda x)^{k-1} e^{-\lambda x}}{\Gamma(k)} \qquad 0 \le x < \infty \tag{14}$$

which is valid for any real positive k, thus, generalizing the exponential distribution. The pdf in Eq. (14) is known as the gamma distribution with parameters k and λ. The pdf of the gamma random variable is plotted in Fig. 11.

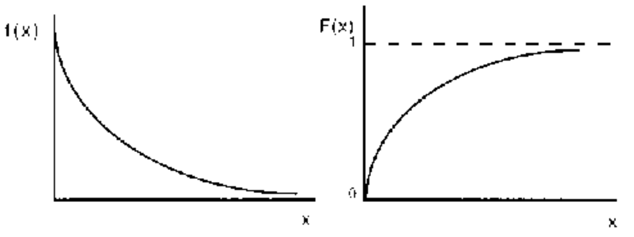


Figure 10 An example of the pdf and cdf of the exponential random variable.

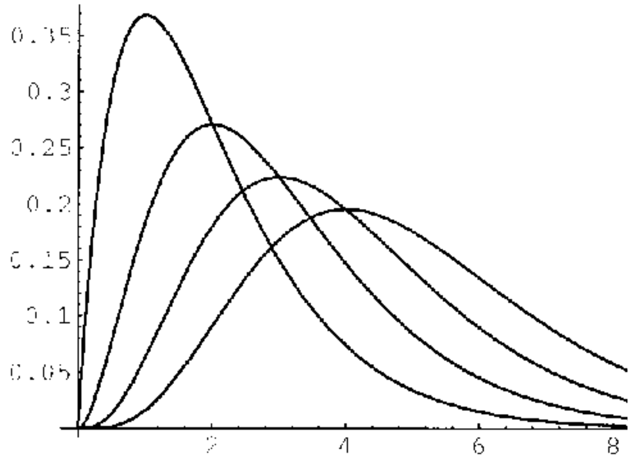


Figure 11 Examples of pdf of some gamma random variables $G(2,1)$, $G(3,1)$, $G(4,1)$, and $G(5,1)$, from left to right.

1.5.4 The Beta Distribution

The beta random variable is denoted as $Beta(r,s)$, where $r > 0$ and $s > 0$. Its name is due to the presence of the beta function

$$\beta(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1} dx \qquad p > 0, q > 0$$

Its pdf is given by

$$\frac{x^{r-1}(1-x)^{s-1}}{\beta(r,s)} \qquad 0 \le x \le 1 \tag{15}$$

Utilizing the relationship between the gamma and the beta functions, Eq. (15) can be expressed as

$$\frac{\Gamma(r+s)}{\Gamma(r)\,\Gamma(s)} x^{r-1}(1-x)^{s-1} \qquad 0 \le x \le 1$$

as given in Table 3. The interest in this variable is based on its flexibility, because it can take many different forms (see Fig. 12), which can fit well many sets of experimental data. Figure 12 shows different examples of the pdf of the beta random variable. Two

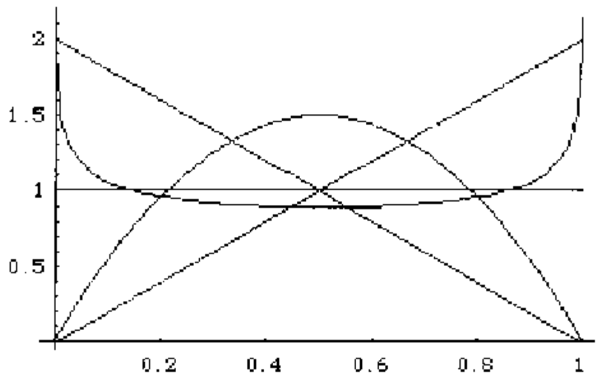


Figure 12 Examples of pdfs of beta random variables.

particular cases of the beta distribution are interesting. Setting $(r=1,\ s=1)$, gives the *standard uniform* $U(0, 1)$ distribution, while setting $(r = 1, s = 2 \text{ or } r = 2, s = 1)$ gives the *triangular* random variable whose cdf is given by $f(x) = 2x$ or $f(x) = 2(1 - x)$, $0 \leq x \leq 1$. The mean and variance of the beta random variable are

$$\frac{r}{r+s} \qquad \text{and} \qquad \frac{rs}{(r+s+1)(r+s)^2}$$

respectively.

1.5.5 The Normal or Gaussian Distribution

One of the most important distributions in probability and statistics is the *normal* distribution (also known as the *Gaussian distribution*), which arises in various applications. For example, consider the random variable, X, which is the sum of n identically and independently distributed (iid) random variables X_i. Then, by the *central limit theorem*, X is asymptotically normal, regardless of the form of the distribution of the random variables X_i.

The normal random variable with parameters μ and σ^2 is denoted by $N(\mu, \sigma^2)$ and its pdf is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad -\infty < x < \infty$$

The change of variable, $Z = (X - \mu)/\sigma$, transforms a normal $N(\mu, \sigma^2)$ random variable X in another random variable Z, which is $N(0.1)$. This variable is called the *standard normal* random variable. The main interest of this change of variable is that we can use tables for the standard normal distribution to calculate probabilities for any other normal distribution. For example, if X is $N(\mu, \sigma^2)$, then

$$\begin{aligned} p(X < x) &= p\left(\frac{X-\mu}{\sigma} < \frac{x-\mu}{\sigma}\right) \\ &= p\left(Z < \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right) \end{aligned}$$

where $\Phi(z)$ is the cdf of the standard normal distribution. The cdf $\Phi(z)$ cannot be given in closed form. However, it has been computed numerically and tables for $\Phi(z)$ are found at the end of probability and statistics textbooks. Thus we can use the tables for the standard normal distribution to calculate probabilities for any other normal distribution.

1.5.6 The Log-Normal Distribution

We have seen in the previous subsection that the sum of iid random variables has given rise to a normal distribution. In some cases, however, some random variables are defined to be the products instead of sums of iid random variables. In these cases, taking the logarithm of the product yields the *log-normal* distribution, because the logarithm of a product is the sum of the logarithms of its components. Thus, we say that a random variable X is log-normal when its logarithm $\ln X$ is normal.

Using Theorem 7, the pdf of the log-normal random variable can be expressed as

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) \qquad x \geq 0$$

where the parameters μ and σ are the mean and the standard deviation of the initial random normal variable. The mean and variance of the log-normal random variable are $e^{\mu+\sigma^2/2}$ and $e^{2\mu}(e^{2\sigma^2} - e^{\sigma^2})$, respectively.

1.5.7 The Chi-Squared and Related Distributions

Let $Y_1, \ldots, Y_n$ be independent random variables, where Y_i is distributed as $N(\mu_i, 1)$. Then, the variable

$$X = \sum_{i=1}^{n} Y_i^2$$

is called a *noncentral chi-squared* random variable with n degrees of freedom, *noncenrality parameter* $\lambda = \sum_{i=1}^{n} \mu_i^2$; and is denoted as $\chi_n^2(\lambda)$. When $\lambda = 0$ we obtain the *central chi-squared* random variable, which is denoted by χ_n^2. The pdf of the central chi-squared random variable with n degrees of freedom is given in Table 3, where $\Gamma(.)$ is the gamma function defined in Eq. (13).

The positive square root of a $\chi_n^2(\lambda)$ random variable is called a *chi* random variable and is denoted by $\chi_n(\lambda)$. An interesting particular case of the $\chi_n(\lambda)$ is the *Rayleigh* random variable, which is obtained for $(n = 2 \text{ and } \lambda = 0)$. The pdf of the Rayleigh random variable is given in Table 3. The Rayleigh distribution is used, for example, to model wave heights [5].

1.5.8 The *t* Distribution

Let Y_1 be a normal $N(\lambda, 1)$ and Y_2 be a χ_n^2 independent random variables. Then, the random variable

$$T = \frac{X_1}{\sqrt{Y_2/n}}$$

is called the *noncentral Student's t* random variable with n degrees of freedom and noncentrality parameter λ and is denoted by $t_n(\lambda)$. When $\lambda = 0$ we obtain the *central Student's t* random variable, which is denoted by t_n and its pdf is given in Table 3. The mean and variance of the central t random variable are 0 and $n/(n-2), n > 2$, respectively.

1.5.9 The *F* Distribution

Let X_1 and X_2 be two independent random variables distributed as $\chi^2_{n_1}(\lambda_1)$ and $\chi^2_{n_2}(\lambda_2)$, respectively. Then, the random variable

$$X = \frac{X_1/n_1}{X_2/n_2}$$

is known as the *noncentral Snedecor F* random variable with n_1 and n_2 degrees of freedom and noncentrality parameters λ_1 and λ_2; and is denoted by $F_{n_1,n_2}(\lambda_1, \lambda_2)$. An interesting particular case is obtained when $\lambda_1 = \lambda_2 = 0$, in which the random variable is called the *noncentral Snedecor F* random variable with n_1 and n_2 degrees of freedom. In this case the pdf is given in Table 3. The mean and variance of the central F random variable are

$$\frac{n_2}{n_2 - 2} \qquad n_2 > 2$$

and

$$\frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 2)^2(n_2 - 4)} \qquad n_2 > 4$$

respectively.

1.6 MULTIDIMENSIONAL RANDOM VARIABLES

In this section we deal with multidimensional random variables, that is, the case where $n > 1$ in Definition 7. In random experiments that yield multidimensional random variables, each outcome gives n real values. The corresponding components are called *marginal* variables. Let $\{X_1, \ldots, X_n\}$ be n-dimensional random variables and $\mathbf{X}$ be the $n \times 1$ vector containing the components $\{X_1, \ldots, X_n\}$. The support of the random variable is also denoted by A, but here A is multidimensional. A realization of the random variable $\mathbf{X}$ is denoted by $\mathbf{x}$, an $n \times 1$ vector containing the components $\{x_1, \ldots, x_n\}$. Note that vectors and matrices are denoted by boldface letters. Sometimes it is also convenient to use the notation $X = \{X_1, \ldots, X_n\}$, which means that X refers to the set of marginals $\{X_1, \ldots, X_n\}$. We present both discrete and continuous multidimensional random variables and study their characteristics. For some interesting engineering multidimensional models see Castillo et al. [6,7].

1.6.1 Multidimensional Discrete Random Variables

A multidimensional random variable is said to be discrete if its marginals are discrete. The pmf of a multidimensional discrete random variable $\mathbf{X}$ is written as $p(\mathbf{x})$ or $p(x_1, \ldots, x_n)$ which means

$$p(\mathbf{x}) = p(x_1, \ldots, x_n) = p(X_1 = x_1, \ldots, X_n = x_n)$$

The pmf of multidimensional random variables can be tabulated in probability distribution tables, but the tables necessarily have to be multidimensional. Also, because of its multidimensional nature, graphs of the pmf are useful only for $n = 2$. The random variable in this case is said to be *two-dimensional*. A graphical representation can be obtained using bars or lines of heights proportional to $p(x_1, x_2)$ as the following example illustrates.

Example 2. *Consider the experiment consisting of rolling two fair dice. Let $X = (X_1, X_2)$ be a two-dimensional random variable such that X_1 is the outcome of the first die and X_2 is the minimum of the two dice. The pmf of X is given in Fig. 13, which also shows the marginal probability of X_2. For example, the probability associated with the pair* $(3, 3)$ *is 4/36, because, according to Table 4, there are four elementary events where $X_1 = X_2 = 3$.*

Table 4 Values of $X_2 = \min(X, Y)$ for Different Outcomes of Two Dice X and Y

	Die 2					
Die 1	1	2	3	4	5	6
1	1	1	1	1	1	1
2	1	2	2	2	2	2
3	1	2	3	3	3	3
4	1	2	3	4	4	4
5	1	2	3	4	5	5
6	1	2	3	4	5	6

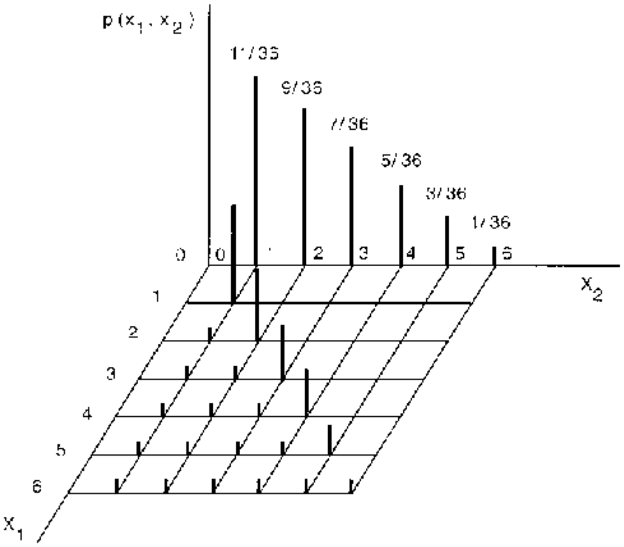


Figure 13 The pmf of the random variable $X = (X_1, X_2)$.

The pmf must satisfy the following properties:

$$p = (x_1, x_2) \geq 0 \qquad \sum_{x_1 \in A} \sum_{x_2 \in A} p(x_1, x_2) = 1 \qquad \text{and}$$

$$P(a_1 \leq X_1 \leq b_1, a_2 \leq X_2 \leq b_2) = \sum_{a_1 \leq x_1 \leq b_1} \sum_{a_2 \leq x_2 \leq b_2} p(x_1, x_2)$$

Example 3. The Multinomial Distribution: *We have seen in Sec. 1.4.3 that the binomial random variable results from random experiments, each one having two possible outcomes. If each random experiment has more than two outcomes, the resultant random variable is called a* multinomial *random variable. Suppose that we perform an experiment with k possible outcomes* $r_1, \ldots, r_k$ *with probabilities* $p_1, \ldots, p_k$*, respectively. Since the outcomes are mutually exclusive and collectively exhaustive, these probabilities must satisfy* $\sum_{i=1}^{k} p_i = 1$*. If we repeat this experiment n times and let* X_i *be the number of times we obtain outcomes* r_i*, for* $i = 1, \ldots, k$*, then* $X = \{X_1, \ldots, X_k\}$ *is a multinomial random variable, which is denoted by* $M(n; p_1, \ldots, p_k)$*. The pmf of* $M(n; p_1, \ldots, p_k)$ *is*

$$p(x_1, x_2, \ldots, x_k; p_1, p_2, \ldots, p_k) = \frac{n!}{x_1!x_2!\ldots, x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

The mean of X_i*, variance of* X_i*, and covariance between* X_i *and* X_j *are*

$$\mu_i = np_i \qquad \sigma_{ii}^2 = np_i(1 - p_i) \quad \text{and} \quad \sigma_{ij}^2 = -mp_i p_j$$

respectively.

1.6.2 Multidimensional Continuous Random Variables

A multidimensional random variable is said to be continuous if its marginals are continuous. The pdf of an n-dimensional continuous random variable $\mathbf{X}$ is written as $f(\mathbf{x})$ or $f(x_1, \ldots, x_n)$. Thus $f(\mathbf{x})$ gives the height of the density at the point $\mathbf{x}$ and $F(\mathbf{x})$ gives the cdf, that is,

$$F(\mathbf{x}) = p(X_1 \leq x_1, \ldots, X_n \leq x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n$$

Similarly, the probability that X_i belongs to a given region, say, $a_i \leq X_i \leq b_i$ for all i is the integral

$$p(a_1 \leq X_1 \leq b_1, \ldots, a_n \leq X_n \leq b_n) = \int_{a_n}^{b_n} \cdots \int_{a_1}^{b_1} f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n$$

The pdf satisfies the following properties:

$$f(x_1, \ldots, x_n) \geq 0$$

$$\int_{a_n}^{b_n} \cdots \int_{a_1}^{b_1} f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n = 1$$

Example 4. *Two-dimensional cumulative distribution function. The cdf of a two-dimensional random variable* (X_1, X_2) *is*

$$F(x_1, x_2) = \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f(x_1, x_2)\, dx_1\, dx_2$$

The relationship between the pdf and cdf is

$$f(x_1, x_2) = \frac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2}$$

Among other properties of two-dimensional cdfs we mention the following:

$F(\infty, \infty) = 1$.
$F(-\infty, x_2) = F(x_1, -\infty) = 0$.
$F(x_1 + a_1, x_2 + a_2) \geq F(x_1, x_2)$, where $a_1, a_2 \geq 0$.
$p(a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2) = F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2)$.
$p(x_1 = x_1, X_2 = x_2) = 0$.

For example, Fig. 14 illustrates the fourth property, showing how the probability that (X_1, X_2) belongs to a given rectangle is obtained from the cdf.

1.6.3 Marginal and Conditional Probability Distributions

We obtain the marginal and conditional distributions for the continuous case. The results are still valid for the discrete case after replacing the pdf and integral symbols by the pmf and the summation symbol, respectively. Let $\{X_1, \ldots, X_n\}$ be n-dimensional continuous random variable with a joint pdf $f(x_1, \ldots, x_n)$. The marginal pdf of the ith component, X_i, is obtained by integrating the joint pdf over all other variables. For example, the marginal pdf of X_1 is

$$f(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n)\, dx_2 \ldots dx_n$$

We define the conditional pdf for the case of two-dimensional random variables. The extension to the n-dimensional case is straightforward. For simplicity of notation we use (X, Y) instead of (X_1, X_2). Let then (Y, X) be a two-dimensional random variable. The random variable Y given $X = x$ is denoted by $(Y \mid X = x)$. The corresponding probability density and distribution functions are called the *conditional* pdf and cdf, respectively.

The following expressions give the conditional pdf for the random variables $(Y \mid X = x)$ and $(X \mid Y = y)$:

$$f_{Y|X=x}(y) = \frac{f_{(X,Y)}(x, y)}{f_X(x)}$$

$$f_{X|Y=y}(x) = \frac{f_{(X,Y)}(x, y)}{f_Y(y)}$$

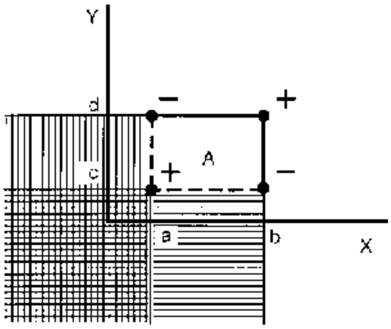


Figure 14 An illustration of how the probability that (X_1, X_2) belongs to a given rectangle is obtained from the cdf.

It may also be of interest to compute the pdf conditioned on events different from $Y = y$. For example, for the event $Y \leq y$, we get

$$F_{X|Y\leq y}(x) = p(X \leq x \mid Y \leq y) = \frac{p(X \leq x, Y \leq y)}{p(Y \leq y)} = \frac{F_{(X,Y)}(x, y)}{F_Y(y)}$$

1.6.4 Moments of Multidimensional Random Variables

The moments of multidimensional random variables are straightforward extensions of the moments for the unidimensional random variables.

Definition 12. Moments of a Multidimensional Random Variable: *The moment* $\mu_{k_1,\ldots,kn;a_1,\ldots,a_n}$ *of order* $(k_1, \ldots, k_n)$, $k_i \in \{0, 1, \ldots\}$ *with respect to the point* $a = (a_1, \ldots, a_n)$ *of the n-dimensional continuous random variable* $X = (X_1, \ldots, X_n)$, *with pdf* $f(x_1, \ldots, x_n)$ *and support A, is defined as the real number*

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (x_1 - a_1)^{k_1}(x_2 - a_2)^{k_2} \ldots (x_n - a_n)^{k_n} \, dF(x_1, \ldots, x_n) \tag{16}$$

For the discrete random variable Eq. (16) becomes

$$\sum_{(x_1,\ldots,x_n)\in A} (x_1 - a_1)^{k_1}(x_2 - a_2)^{k_2} \ldots (x_n - a_n)^{k_n} \, p(x_1, \ldots, x_n)$$

where $f(x_1, \ldots, x_n)$ is the pdf of X.

The moment of first order with respect to the origin is called the *mean vector*, and the moments of second order with respect to the mean vector are called the *variances* and *covariances*. The variances and covariances can conveniently be arranged in a matrix called the *variance–covariance matrix*. For example, in the bivariate case, the variance–covariance matrix is

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{XX} & \sigma_{XY} \\ \sigma_{YX} & \sigma_{YY} \end{pmatrix}$$

where $\sigma_{XX} = \mathrm{Var}(X)$ and $\sigma_{YY} = \mathrm{Var}(Y)$, and

$$\sigma_{XY} = \sigma_{YX} = \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y)\, dF(x, y)$$

is the covariance between X and Y, where μ_X is the mean of the variable X. Note that $\mathbf{\Sigma}$ is necessarily symmetrical.

Figure 15 gives a graphical interpretation of the contribution of each data point to the covariance and its corresponding sign. In fact the contribution term has absolute value equal to the area of the rectangle in Fig. 15(a). Note that such area takes value zero when the corresponding points are on the vertical or the horizontal lines associated with the means, and takes larger values when the point is far from the means.

On the other hand, when the points are in the first and third quadrants (upper-right and lower-left) with respect to the mean, their contributions are positive, and if they are in the second and fourth quadrants (upper-left and lower-right) with respect to the mean, their contributions are negative [see Fig. 15(b)].

Another important property of the variance–covariance matrix is the Cauchy–Schwartz inequality:

$$|\sigma_{XY}| \le \sqrt{\sigma_{XX}\sigma_{YY}} \tag{17}$$

The equality holds only when all the possible pairs (points) are in a straight line.

The pairwise correlation coefficients can also be arranged in a matrix

$$\boldsymbol{\rho} = \begin{pmatrix} \rho_{XX} & \rho_{XY} \\ \rho_{YX} & \rho_{YY} \end{pmatrix}$$

This matrix is called the *correlation matrix*. Its diagonal elements ρ_{XX} and ρ_{YY} are equal to 1, and the off-diagonal elements satisfy $-1 \le \rho_{XY} \le 1$.

1.6.5 Sums and Products of Random Variables

In this section we discuss linear combinations and products of random variables.

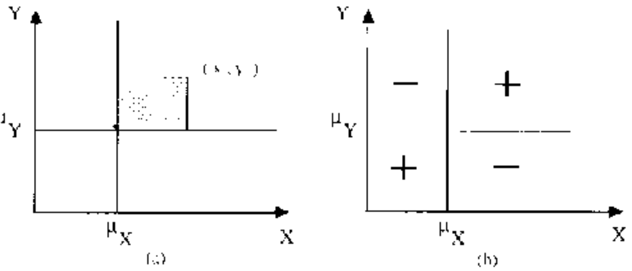


Figure 15 Graphical illustration of the meaning of the covariance.

Theorem 4. Linear Transformations: *Let* $(X_1, \ldots, X_n)$ *be an n-dimensional random variable and* $\boldsymbol{\mu}_X$ *and* $\mathbf{\Sigma}_X$ *be its mean and covariance matrix. Consider the linear transformation*

$$\mathbf{Y} = \mathbf{CX}$$

where $\mathbf{X}$ *is the column vector containing* $(X_1, \ldots, X_n)$ *and* $\mathbf{C}$ *is a matrix of order* $m \times n$. *Then, the mean vector and covariance matrix of the m-dimensional random variable* $\mathbf{Y}$ *are*

$$\boldsymbol{\mu}_Y = \mathbf{C}\boldsymbol{\mu}_X \qquad \text{and} \qquad \Sigma_Y = \mathbf{C}\mathbf{\Sigma}_X\mathbf{C}^T$$

Theorem 5. Expectation of a Product of Independent Random Variables: *If* $X_1, \ldots, X_n$ *are independent random variables with means*

$$E[X_1], \ldots, E[X_n]$$

respectively, then, we have

$$E\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} E[X_i]$$

That is, the expected value of the product of independent random variables is the product of their individual expected values.

1.6.6 Multivariate Moment-Generating Function

Let $X = (X_1, \ldots, X_n)$ be an n-dimensional random variable with cdf $F(x_1, \ldots, x_n)$. The moment-generating function $M_X(t_1, \ldots, t_n)$ of X is

$$M_x(t_1, \ldots, t_n) = \int_{R^n} e^{t_1x_1 + \cdots + t_nx_n}\, dF(x_1, \ldots, x_n)$$

Like in the univariate case, the moment-generating function of a multidimensional random variable may not exist.

The moments with respect to the origin are

$$E[X_1^{\alpha_1} \ldots X_n^{\alpha_n}] = \left.\frac{\partial^{\alpha_1 + \cdots + \alpha_n} M_X(t_1, \ldots, t_n)}{\partial t_1^{\alpha_1} \ldots \partial t_n^{\alpha_n}}\right|_{t_1 = \cdots = t_n = 0}$$

Example 5. *Consider the random variable with pdf*

$$f(x_1, \ldots, x_n) = \begin{cases} \prod_{i=1}^{n} \lambda_i \exp\left(- \sum_{j=1}^{n} \lambda_j x_j \right) & \text{if } 0 \le x_i < \infty \\ 0 & \text{otherwise} \end{cases}$$

$$\lambda_i > 0 \qquad \forall i = 1, \ldots, n$$

Then, the moment-generating function is

$$\begin{aligned} M_X(t_1, \ldots, t_n) &= \int_0^\infty \cdots \int_0^\infty \exp\left(\sum_{j=1}^{n} t_i x_i \right) \prod_{i=1}^{n} \lambda_i \\ &\quad \exp\left(- \sum_{i=1}^{n} \lambda_i x_i \right) dx_1 \ldots dx_n \\ &= \prod_{i=1}^{n} \lambda_i \int_0^\infty \cdots \int_0^\infty \\ &\quad \exp\left[\sum_{i=1}^{n} x_i(t_i - \lambda_i) \right] dx_1 \ldots dx_n \\ &= \prod_{i=1}^{n} \left[\lambda_i \int_0^\infty \exp[x_i(t_i - \lambda_i)]\, dx_i \right] \\ &= \prod_{i=1}^{n} \frac{\lambda_i}{\lambda_i - t_i} \end{aligned}$$

1.6.7 The Multinormal Distribution

Let $\mathbf{X}$ be an n-dimensional normal random variable, which is denoted by $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean vector and covariance matrix, respectively. The pdf of $\mathbf{X}$ is given by

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}[\det(\boldsymbol{\Sigma})]^{1/2}} e^{-0.5(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

The following theorem gives the conditional mean and variance–covariance matrix of any conditional variable, which is normal.

Theorem 6. Conditional Mean and Covariance Matrix: *Let Y and Z be two sets of random variables having a multivariate Gaussian distribution with mean vector and covariance matrix given by*

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_Z \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YZ} \\ \boldsymbol{\Sigma}_{ZY} & \boldsymbol{\Sigma}_{ZZ} \end{pmatrix}$$

where $\boldsymbol{\mu}_Y$ and $\boldsymbol{\Sigma}_{YY}$ are the mean vector and covariance matrix of Y, $\boldsymbol{\mu}_Z$ and $\boldsymbol{\Sigma}_{ZZ}$ are the mean vector and covariance matrix of Z, and $\boldsymbol{\Sigma}_{YZ}$ is the covariance of Y and Z. Then the CPD of Y given $Z = z$ is multivariate Gaussian with mean vector $\boldsymbol{\mu}_{Y|Z=z}$ and covariance matrix $\boldsymbol{\Sigma}_{Y|Z=z}$, where

$$\begin{aligned} \boldsymbol{\mu}_{Y|Z=z} &= \boldsymbol{\mu}_Y + \boldsymbol{\Sigma}_{YZ}\boldsymbol{\Sigma}_{ZZ}^{-1}(z - \boldsymbol{\mu}_z) \\ \boldsymbol{\Sigma}_{Y|Z=z} &= \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YZ}\boldsymbol{\Sigma}_{ZZ}^{-1}\boldsymbol{\Sigma}_{ZY} \end{aligned} \tag{18}$$

For other properties of the multivariate normal distribution, see any multivariate analysis book, such as Rencher [8].

1.6.8 The Marshall–Olkin Distribution

We give two versions of the *Marshall–Olkin* distribution with different interpretations. Consider first a system with two components. Both components are subject to Poissonian processes of fatal shocks, such that if one component is affected by one shock it fails. Component 1 is subject to a Poisson process with parameter λ_1, component 2 is subject to a Poisson process with parameter λ_2, and both are subject to a Poisson process with parameter λ_{12}. This implies that

$$\begin{aligned} \overline{F}(s, t) &= p[X > s, Y > t] \\ &= p\{Z_1(s; \lambda_1) = 0, Z_2(t; \lambda_2) = 0, \} \\ &\quad Z_{12}(\max(s, t); \lambda_{12}) = 0, \} \\ &= \exp[-\lambda_1 s - \lambda_2 t - \lambda_{12} \max(s, t)] \end{aligned}$$

where $Z(s; \lambda)$ represents the number of shocks produced by a Poisson process of intensity λ in a period of duration s and $\overline{F}(s, t)$ is the *survival function.*

This model has another interpretation in terms of nonfatal shocks as follows. Consider the above model of shock occurrence, but now suppose that the shocks are not fatal. Once a shock of intensity λ_1 has occurred, there is a probability p_1 of failure of component 1. Once a shock of intensity λ_2 has occurred, there is a probability p_2 of failure of component 2 and, finally, once a shock of intensity λ_{12} has occurred, there are probabilities p_{00}, p_{01}, p_{10}, and p_{11} of failure of neither of the components, component 1, component 2, or both components, respectively. In this case we have

$$\begin{aligned} \overline{F}(s, t) &= P[X > s, Y > t] \\ &= \exp[-\delta_1 s - \delta_2 t - \delta_{12} \max(s, t)] \end{aligned}$$

where

$$\delta_1 = \lambda_1 p_1 + \lambda_{12} p_{01}; \qquad \delta_2 = \lambda_2 p_2 + \lambda_{12} p_{10};$$
$$\delta_{12} = \lambda_{12} p_{00}$$

This two-dimensional model admits an obvious generalization to n dimensions:

$$\overline{F}(x_1, \ldots, x_n) = \exp\Bigg[-\sum_{i=1}^{n} \lambda_i x_i - \sum_{i<j} \lambda_{ij} \max(x_i x_j)$$
$$- \sum_{i<j<k} \lambda_{ijk} \max(x_i, x_j, x_k) - \cdots$$
$$- \lambda_{12\ldots n} \max(x_1, \ldots, x_n)\Bigg]$$

1.7 CHARACTERISTICS OF RANDOM VARIABLES

The pmf or pdf of random variables contains all the information about the random variables. For example, given the pmf or the pdf of a given random variable, we can find the mean, the variance, and other moments of the random variable. We can also find functions related to the random variables such as the moment-generating function, the characteristic function, and the probability-generating function. These functions are useful in studying the properties of the corresponding probability distribution. In this section we study these characteristics of the random variables.

The results in this section is presented for continuous random variables using the pdf and cdf, $f(x)$ and $F(x)$, respectively. For discrete random variables, the results are obtained by replacing $f(x)$, $F(x)$, and the integration symbol by $p(x)$, $P(x)$, and the summation symbol, respectively.

1.7.1 Moment-Generating Function

Let X be a random variable with a pdf $f(x)$ and cdf $F(x)$. The *moment-generating function* (mgf) of X is defined as

$$M_X(t) = E[e^{tx}] = \int_{-\infty}^{\infty} e^{tx}\, dF_X(x)$$

In some cases the moment-generating function does not exist. But when it exists, it has several very important properties.

The mgf generates the moments of the random variable, hence its name. In fact, the k central moment of the random variable is obtained by evaluating the kth derivative of $M_X(t)$ with respect to t at $t = 0$. That is, if $M^{(k)}(t)$ is the kth derivative of $M_X(t)$ with respect to t, then the kth central moment of X is $m_k = M^{(k)}(0)$.

Example 6. *The moment-generating function of the Bernoulli random variable with pmf*

$$p(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

is

$$M(t) = E[e^{tX}] = e^{t\times 1}p + e^{t\times 0}(1-p) = 1 - p + pe^t$$

For example, to find the first two central moments of X, we first differentiate $M_X(t)$ with respect to t twice and obtain $M^{(1)}(t) = pe^t$ and $M^{(2)}(t) = pe^t$. In fact, $M^{(k)}(t) = pe^t$, for all k. Therefore, $M^{(k)}(0) = p$, which proves that all central moments of X are equal to p.

Example 7. *The moment-generating function of the Poisson random variable with pmf*

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad x \in \{0, 1, \ldots\}$$

is

$$M(t) = E[e^{tX}] = \sum_{x=0}^{\infty} \frac{e^{tx}e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda e^t)^x}{x!}$$
$$= e^{-\lambda}e^{\lambda e^t} = e^{\lambda(e^t - 1)}$$

For example, the first derivative of $M(t)$ with respect to t is $M^{(1)}(t) = \lambda e^{-\lambda}e^t e^{\lambda e^t}$, from which it follows that the mean of the Poisson random variable is $M^{(1)}(0) = \lambda$. The reader can show tht $E[X^2] = M^{(2)}(0) = \lambda + \lambda^2$, from which it follows that $Var(x) = \lambda$, where we have used Eq. (11).

Example 8. *The moment-generating function of the exponential random variable with pdf*

$$f(x) = \lambda e^{-\lambda x} \qquad x \geq 0$$

is

$$M(t) = E[e^{tX}] = \int_0^{\infty} e^{tx}\lambda e^{-\lambda x}\, dx = \lambda \int_0^{\infty} e^{(t-\lambda)x}\, dx$$
$$= \lambda\frac{(-1)}{t-\lambda} = \left(1 - \frac{t}{\lambda}\right)^{-1}$$

from which it follows that $M^{(1)}(t) = \lambda^{-1}(1 - t/\lambda)^2$ and, hence $M^{(1)}(0) = 1/\lambda$, which is the mean of the exponential random variable.

Tables 5 and 6 give the mgf, mean, and variance of several discrete and continuous random variables. The characteristic function $\psi_X(t)$ is discussed in the following subsection.

Table 5 Moment-Generating Functions, Characteristic Functions, Means, and Variances of Some Discrete Probability Distributions

Distribution	$M_X(t)$	$\psi_X(t)$	Mean	Variance
Bernoulli	$1-p+pe^t$	$1-p+pe^{it}$	p	$p(1-p)$
Binomial	$(1-p+pe^t)^n$	$(1-p+pe^{it})^n$	np	$mp(1-p)$
Geometric	$\dfrac{pe^t}{1-e^t(1-p)}$	$\dfrac{pe^{it}}{1-e^{it}(1-p)}$	$1/p$	$\dfrac{(1-p)}{p^2}$
Negative binomial	$\left(\dfrac{pe^t}{1-qe^t}\right)^r$	$\left(\dfrac{pe^{it}}{1-qe^{it}}\right)^r$	$\dfrac{r(1-p)}{p}$	$\dfrac{r(1-p)}{p^2}$
Poisson	$e^{\lambda(e^t-1)}$	$e^{\lambda(e^{it}-1)}$	λ	λ

1.7.2 Characteristic Function

Let X be a univariate random variable with pdf $f(x)$ and cdf $F(x)$. Then, the *characteristic function* (cf) of X is defined by

$$\psi_X(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\, dx \tag{19}$$

where i is the imaginary unit. Like the mgf, the cf is unique and completely characterizes the distribution of the random variable. But, unlike the mgf, the cf always exists.

Note that Eq. (19) shows that $\psi_X(t)$ is the Fourier transform of $f(x)$.

Example 9. *The characteristic function of the discrete uniform random variable $U(n)$ with pmf*

$$p(x) = \frac{1}{n} \qquad x = 1, \ldots, n$$

is

$$\psi_X(t) = \sum_x e^{itx} p(x) = \sum_x e^{itx} \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} e^{itx}$$
$$= \frac{1}{n} \frac{e^{it}(e^{itn}-1)}{(e^{it}-1)}$$

which is obtained using the well-known formula for the sum of the first n terms of a geometric series.

Example 10. *The characteristic function of the continuous uniform random variable $U(0, a)$ with pdf*

$$f(x) = \begin{cases} 1/a & \text{if } 0 \le x \le a \\ 0 & \text{otherwise} \end{cases}$$

Table 6 Moment-Generating Functions, Characteristic Functions, Means, and Variances of Some Continuous Probability Distributions

Distribution	$M_X(t)$	$\psi_X(t)$	Mean	Variance
Uniform	$\dfrac{e^{tb}-e^{ta}}{t(b-a)}$	$\dfrac{e^{itb}-e^{ita}}{it(b-a)}$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$
Exponential	$\left(1-\dfrac{t}{\lambda}\right)^{-1}$	$\left(1-\dfrac{it}{\lambda}\right)^{-1}$	λ	λ^2
Gamma	$\left(1-\dfrac{t}{\lambda}\right)^{-k}$	$\left(1-\dfrac{it}{\lambda}\right)^{-k}$	λk	$k\lambda^2$
Normal	$e^{(\mu t+\sigma^2 t^2/2)}$	$e^{it\mu-\sigma^2 t^2/2}$	μ	σ^2
Central chi-squared	$(1-2t)^{-n/2}$	$(1-2it)^{-n/2}$	n	$2n$

is

$$\psi_X(t) = \int_0^a e^{itx}\frac{1}{a}\,dx = \frac{1}{a}\frac{e^{itx}}{it}\bigg|_0^a = \frac{e^{ita}-1}{iat}$$

Some important properties of the characteristic function are:

$\psi_X = 1$.
$|\psi_X(t)| \le 1$.
$\psi_X(-t) = \overline{\psi_X(t)}$, where $\bar{z}$ is the conjugate of z.
If $Z = aX + b$, where X is a random variable, and a and b are real constants, we have $\psi_Z(t) = e^{itb}\psi_X(at)$, where $\psi_X(t)$ and $\psi_Z(t)$ are the characteristic functions of Z and X, respectively.
The characteristic function of the sum of two independent random variables is the product of their characteristic functions, that is, $\psi_{X+Y}(t) = \psi_X(t)\psi_Y(t)$.
The characteristic function of a linear convex combination of random variables is the linear convex combination of their characteristic functions with the same coefficients: $\psi_{aF_x+bF_y}(t) = a\psi_X(t) + b\psi_Y(t)$.
The characteristic function of the sum of a random number N iid random variables $\{X_1, \ldots, X_n\}$ is given by

$$\psi_S(t) = \psi_N\left(\frac{\log\psi_X(t)}{i}\right)$$

where $\psi_X(t)$, $\psi_N(t)$, and $\psi_S(t)$ are the characteristic functions of X_i, N and $S = \sum_{i=1}^N X_i$, respectively.

One of the main applications of the characteristic function is to obtain the central moments of the corresponding random variable. In fact, in we differentiate the characteristic function k times with respect to t, we get

$$\psi_X^{(k)}(t) = \int_{-\infty}^{\infty} i^k x^k e^{itx} f(x)\,dx$$

which for $t = 0$ gives

$$\psi_X^{(k)}(0) = i^k \int_{-\infty}^{\infty} x^k\,dF(x) = i^k m_k$$

from which we have

$$m_k = \frac{\psi_X^{(k)}(0)}{i^k} \tag{20}$$

where m_k is the kth central moment of X.

Example 11. *The central moments of the Bernoulli random variable are all equal to p. In effect, its characteristic function is*

$$\psi_X(t) = pe^{it} + q$$

and, according to Eq. (20), we get

$$m_k = \frac{\psi_X^{(k)}(0)}{i^k} = \frac{pi^k}{i^k} = p$$

Example 12. *The characteristic function of the gamma random variable Gamma(p, a) is*

$$\psi_X(t) = \left(1 - \frac{it}{a}\right)^{-p}$$

The moments with respect to the origin, according to Eq. (20), are

$$m_k = \frac{\psi_X^{(k)}(0)}{i^k} = \frac{p(p-1)\ldots(p-k+1)}{a^k}$$

Tables 5 and 6 give the cf of several discrete and continuous random variables.

Next, we can extend the characteristic function to multivariate distributions as follows. Let $X = (X_1, \ldots, X_n)$ be an n-dimensional random variable. The characteristic function of X is defined as

$$\psi_X(t) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{i(t_1x_1+t_2x_2+\cdots+t_nx_n)}\, dF_X(x_1, \ldots, x_n)$$

where the integral always exists.

The moments with respect to the origin can be obtained by

$$m_{r_1,\ldots,r_k} = \frac{\psi_X^{(r_1+\cdots+r_k)}(0, \ldots, 0)}{i^{(r_1+\cdots+r_k)}}$$

Example 13. *Consider the random variable with pdf*

$$f(x_1, \ldots, x_n) = \begin{cases} \prod_{i=1}^{n}(\lambda_i e^{-\lambda_i x_i}) & \text{if } 0 \le x_i < \infty; i = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

Its characteristic function is

$$
\begin{aligned}
\psi_X(t) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{i(t_1x_1+\cdots+t_nx_n)}\, dF_X(x_1, \ldots, x_n) \\
&= \int_0^{\infty} \cdots \int_0^{\infty} \exp\left(i \sum_{i=1}^{n} t_i x_i \right) \prod_{i=1}^{n} (\lambda_i e^{-\lambda_i x_i}) \\
&\quad dx_1 \ldots dx_n \\
&= \int_0^{\infty} \cdots \int_0^{\infty} \prod_{i=1}^{n} (\lambda_i e^{x_i(it_i-\lambda_i)})\, dx_1 \ldots dx_n \\
&= \prod_{i=1}^{n} \left(\lambda_i \int_0^{\infty} e^{x_i(it_i-\lambda_i)}\, dx_i \right) \\
&= \prod_{i=1}^{n} \frac{\lambda_i}{\lambda_i - it_i} = \prod_{i=1}^{n} \left(1 - \frac{it_i}{\lambda_i} \right)^{-1}
\end{aligned}
$$

Example 14. *The characteristic function of the multinormal random variable is*

$$
\varphi(t_1, \ldots, t_n) = \exp\left[i \left(\sum_{k=1}^{n} t_k \mu_k - \frac{\sum_{k,j=1}^{n} \sigma_{k_j t_k t_j}}{2} \right) \right]
$$

Example 15. *The characteristic function of the multinominal random variable, $M(n; p_1, \ldots, p_k)$, can be written as*

$$
\begin{aligned}
\varphi(t_1, \ldots, t_k) &= \sum \exp\left(\sum_{j=1}^{k} it_j x_j \right) \\
&\quad p(x_1, x_2, \ldots, x_k) \\
&= \sum \frac{n!}{x_1! x_2! \ldots x_k!} (p_1 e^{it_1})^{x_1} \\
&\quad (p_2 e^{it_2})^{x_2} \ldots (p_k e^{it_k})^{x_k} \\
&= \left(\sum_{j=1}^{k} p_j e^{it_j} \right)^n
\end{aligned}
$$

1.8 TRANSFORMATIONS OF RANDOM VARIABLES

1.8.1 One-to-One Transformations

Theorem 7. Transformations of Continuous Random Variables: *Let $(X_1, \ldots, X_n)$ be an n-dimensional random variable with pdf $f(x_1, \ldots, x_n)$ defined on the set A and let*

$$
\begin{aligned}
Y_1 &= g_1(X_1, \ldots, X_n) \\
Y_2 &= g_2(X_1, \ldots, X_n) \\
&\vdots \\
Y_n &= g_n(X_1, \ldots, X_n)
\end{aligned}
\tag{21}
$$

be a one-to-one continuous transformation from the set A to the set B. Then, the pdf of the random variable $(Y_1, \ldots, Y_n)$ on the set B, is

$$
\begin{aligned}
&f(h_1(y_1, \ldots, y_n), h_2(y_1, \ldots, y_n), \ldots, \\
&\quad h_n(y_1, y_2, \ldots, y_n)) |\det(\mathbf{J})|
\end{aligned}
$$

where

$$
\begin{aligned}
X_1 &= h_1(Y_1, \ldots, Y_n) \\
X_2 &= h_2(Y_1, \ldots, Y_n) \\
&\vdots \\
X_n &= h_n(Y_1, \ldots, Y_n)
\end{aligned}
$$

is the inverse transformation of Eq. (21) and $|\det(\mathbf{J})|$ the absolute value of the determinant of the Jacobian matrix $\mathbf{J}$ of the transformation. The ijth element of $\mathbf{J}$ is given by $\partial X_i / \partial Y_j$.

Example 16. *Let X and Y be two independent normal $N(0, 1)$ random variables. Then the joint pdf is*

$$
f_{X,Y}(x, y) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \qquad -\infty < x, y < \infty
$$

Consider the transformation

$$
\begin{aligned}
U &= X + Y \\
V &= X - Y
\end{aligned}
$$

which implies that

$$
\begin{aligned}
X &= (U + V)/2 \\
Y &= (U - V)/2
\end{aligned}
$$

Then the Jacobian matrix is

$$
\mathbf{J} = \begin{pmatrix} \partial X/\partial U & \partial X/\partial V \\ \partial Y/\partial U & \partial Y/\partial V \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix}
$$

with $|\det(\mathbf{J})| = 1/2$. Thus, the joint density of U and V becomes

$$
\begin{aligned}
g(u, v) &= \frac{1}{4\pi} \exp\left\{ -\frac{1}{2} \left[\left(\frac{u+v}{2} \right)^2 + \left(\frac{u-v}{2} \right)^2 \right] \right\} \\
&= \frac{1}{4\pi} e^{-u^2/4} e^{-v^2/4} \qquad -\infty < u, v < \infty
\end{aligned}
$$

which is the product of a function of u and a function of v defined in a rectangle. Thus, U and V are independent $N(0, 2)$ *random variables.*

1.8.2 Other Transformations

If the transformation Eq. (21) is not one-to-one, the above method is not applicable. Assume that for each point $(x_1, \ldots, x_n)$ in A there is one point in B, but each point in B, has more than one point in A. Assume further that there exists a finite partition $(A_1, \ldots, A_n)$, of A, such that the restriction of the given transformation to each A_i, is a one-to-one transformation. Then, there exist transformations of B in A_i defined by

$$\begin{aligned} X_1 &= h_{1i}(Y_1, \ldots, Y_n) \\ X_2 &= h_{2i}(Y_1, \ldots, Y_n) \\ &\vdots \\ X_n &= h_{ni}(Y_1, \ldots, Y_n) \end{aligned}$$

with jacobians $\mathbf{J}_i \quad i = 1, \ldots, m$. Then, taking into account that the probability of the union of disjoint sets is the sum of the probabilities of the individual sets, we obtain the pdf of the random variable $(Y_1, \ldots, Y_n)$:

$$g(y_1, \ldots, y_n) = \sum_{i=1}^{m} f(h_{1i}, \ldots, h_{ni}) |\det(\mathbf{J}_i)|$$

1.9 SIMULATION OF RANDOM VARIABLES

A very useful application of the change-of-variables technique discussed in the previous section is that it provides a justification of an important method for simulating any random variable using the standard uniform variable $U(0, 1)$.

1.9.1 The Univariate Case

Theorem 8. *Let X be a univariate random variable with cdf* $F(x)$. *Then, the random variable* $U = F(x)$ *is distributed as a standard uniform variable* $U(0, 1)$.

Example 17. Simulating from a Probability Distribution: *To generate a sample from a probability distribution* $f(x)$, *we first compute the cdf,*

$$F(x) = p(X \le x) = \int_{-\infty}^{x} f(x)\, dx$$

We then generate a sequence of random numbers $\{u_1, \ldots, u_n\}$ *from* $U(0, 1)$ *and obtain the corresponding values* $\{x_1, \ldots, x_n\}$ *by solving* $F(x_i) = u_i, i = 1, \ldots n$, *which gives* $x_i = H^{-1}(u_i)$, *where* $H^{-1}(u_i)$ *is the inverse of the cdf evaluated at* u_i. *For example, Fig. 16 shows the cdf* $F(x)$ *and two values* x_1 *and* x_2 *corresponding to the uniform* $U(0, 1)$ *numbers* u_1 *and* u_2.

Theorem 9. Simulating Normal Random Variables: *Let X and Y be independent standard uniform random variables* $U(0, 1)$. *Then, the random variables U and V defined by*

$$\begin{aligned} U &= (-2 \log X)^{1/2} \sin(2\pi Y) \\ V &= (-2 \log X)^{1/2} \cos(2\pi Y) \end{aligned}$$

are independent $N(0, 1)$ *random variables.*

1.9.2 The Multivariate Case

In the multivariate case $(X_1, \ldots, X_n)$, we can simulate using the conditional cdfs:

$$F(x_1), F(x_2|x_1), \ldots, F(x_n|x_1, \ldots, x_{n-1})$$

as follows. First we simulate X_1 with $F(x_1)$ obtaining x_1. Once we have simulated X_{k-1} obtaining x_{k-1}, we simulate X_k using $F(x_k|x_1, \ldots, x_{k-1})$, and we continue the process until we have simulated all X's. We repeat the whole process as many times as desired.

1.10 ORDER STATISTICS AND EXTREMES

Let $(X_1, \ldots, X_n)$ be a random sample coming from a pdf $f(x)$ and cdf $F(x)$. Arrange $(X_1, \ldots, X_n)$ in an

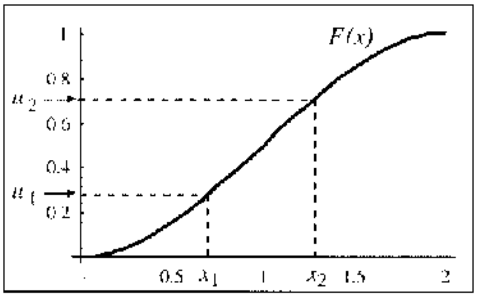


Figure 16 Sampling from a probability distribution $f(x)$ using the corresponding cdf $F(x)$.

increasing order of magnitude and let $X_{1:n} \leq \cdots \leq X_{n:n}$ be the ordered values. Then, the rth element of this new sequence, $X_{r:n}$, is called the rth order statistic of the sample.

Order statistics are very important in practice, especially so for the minimum, $X_{1:n}$ and the maximum, $X_{n:n}$ because they are the critical values which are used in engineering, physics, medicine, etc. (see, e.g., Castillo and Hadi [9–11]). In this section we study the distributions of order statistics.

1.10.1 Distributions of Order Statistics

The cdf of the rth order statistic $X_{r:n}$ is [12, 13]

$$\begin{aligned} F_{r:n}(x) &= P[X_{r:n} \leq x] = 1 - F_{m(x)}(r-1) \\ &= \sum_{k=r}^{n} \binom{n}{k} F^k(x)[1-F(x)]^{n-k} \\ &= r\binom{n}{r} \int_0^{F(x)} u^{r-1}(1-u)^{n-r}\, du \\ &= I_{F(x)}(r, n-r+1) \end{aligned} \tag{22}$$

where $m(x)$ is the number of elements in the sample with value $X_j \leq x$ and $I_p(a,b)$ is the incomplete beta function, which is implicitly defined in Eq. (22).

If the population is absolutely continuous, then the pdf of X_{r_n} is given by the derivative of Eq. (22) with respect to x:

$$\begin{aligned} f_{X_{r:n}}(x) &= r\binom{n}{r} F^{r-1}(x)[1-F(x)]^{n-r} f(x) \\ &= \frac{Fr^{-1}(x)[1-F(x)]^{n-r} f(x)}{\beta(r, n-r+1)} \end{aligned} \tag{23}$$

where $\beta(a,b)$ is the beta function.

Example 18. *Distribution of the minimum order statistic. Letting $r = 1$ in Eqs (22) and (23) we obtain the cdf and pdf of the minimum order statistic:*

$$\begin{aligned} F_{X_{1:n}}(x) &= \sum_{k=1}^{n} \binom{n}{k} F^k(x)[1-F(x)]^{n-k} \\ &= 1 - [1-F(x)]^n \end{aligned}$$

and

$$f_{X_{1:n}}(x) = n[1-F(x)]^{n-1} f(x)$$

Example 19. *Distribution of the maximum order statistic. Letting $r = n$ in Eqs (22) and (23) we obtain the cdf and the pdf of the maximum order statistic: $F_{X_{n:n}}(x) = F(x)^n$ and $f_{X_{n:n}}(x) = nF^{n-1}(x)f(x)$.*

1.10.2 Distributions of Subsets of Order Statistics

Let $X_{r_1:n}, \ldots, X_{r_k:n}$, be the subset of k order statistics of orders $r_1 < \ldots < r_k$, of a random sample of size n coming from a population with pdf $f(x)$ and cdf $F(x)$. With the aim of obtaining the joint distribution of this set, consider the event $x_j \leq X_{r_j:n} < x_j + \Delta x_j$; $1 \leq j \leq k$ for small values of Δx_j, $1 \leq j \leq k$ (see Fig. 17). That is, k values in the sample belong to the intervals $(x_j, x_j + \Delta x_j)$ for $1 \leq j \leq k$ and the rest are distributed in such a way that exactly $(r_j - r_{j-1} - 1)$ belong to the interval $(x_{j-1} + \Delta x_{j-1}, x_j)$ for $1 \leq j \leq k$, where $\Delta x_0 = 0$, $r_0 = 0$, $r_{k+1} = n+1$, $x_0 = -\infty$ and $x_{k+1} = \infty$.

Consider the following multinomial experiment with the $2k+1$ possible outcomes associated with the $2k+1$ intervals illustrated in Fig. 17. We obtain a sample of size n from the population and determine to which of the intervals they belong. Since we assume independence and replacement, the numbers of elements in each interval is a multinomial random variable with parameters

$$\begin{aligned} &\{n; f(x_1)\Delta x_1, \ldots, f(x_k)\Delta x_k, [F(x_1) - F(x_0)], \\ &[F(x_2) - F(x_1)], \ldots, [F(x_{k+1}) - F(x_k)]\} \end{aligned}$$

where the parameters are n (the sample size) and the probabilities associated with the $2k+1$ intervals. Consequently, we can use the results for multinomial random variables to obtain the joint pdf of the k order statistics and obtain

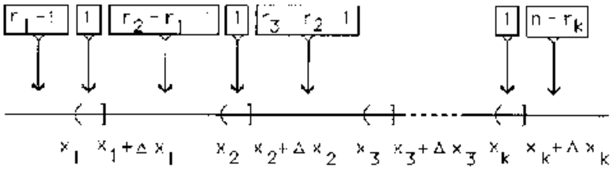


Figure 17 An illustration of the multinomial experiment used to determine the joint pdf of a subset of k order statistics.

$$f_{r_1,\ldots,r_k:n}(x_1,\ldots,x_k) = n! \prod_{i=1}^{k} f(x_i) \prod_{j=1}^{k+1} \frac{[F(x_j) - F(x_{j-1})]^{r_j - r_{j-1} - 1}}{(r_j - r_{j-1} - 1)!} \qquad (24)$$

1.10.3 Distributions of Particular Order Statistics

1.10.3.1 Joint Distribution of Maximum and Minimum

Setting $k = 2$, $r_1 = 1$ and $r_2 = n$ in Eq. (24), we obtain the joint distribution of the maximum and the minimum of a sample of size n, which becomes

$$f_{1,n:n}(x_1, x_2) = n(n-1)f(x_1)f(x_2)[F(x_2) - F(x_1)]^{n-2} \qquad x_1 \le x_2$$

1.10.3.2 Joint Distribution of Two Consecutive Order Statistics

Setting $k = 2$, $r_1 = i$ and $r_2 = i + 1$ in Eq. (24), we get the joint density of the statistics of orders i and $i + 1$:

$$f_{i,i+1:n}(x_1,\ x_2) = \frac{n! f(x_1)f(x_2)F^{i-1}(x_1)[1 - F(x_2)]^{n-i-1}}{(i-1)!(n-i-1)!} \qquad x_1 \le x_2$$

1.10.3.3 Joint Distribution of Any Two Order Statistics

The joint distribution of the statistics of orders r and s $(r < s)$ is

$$f_{(r,s:n}(x_r, x_s) = \frac{n! f(x_r)f(x_s)F^{r-1}(x_r)[F(x_s) - F(x_r)]^{s-r-1}[1 - F(x_s)]^{n-s}}{(r-1)!(s-r-1)!(n-s)!} \qquad x_r \le x_s$$

1.10.3.4 Joint Distribution of all Order Statistics

The joint density of all order statistics can be obtained from Eq. (24) setting $k = n$ and obtain

$$f_{1,\ldots,n:n}(x_1,\ldots,x_n) = n! \prod_{i=1}^{n} f(x_i) \qquad x_1 \le \cdots \le x_n$$

1.10.4 Limiting Distributions of Order Statistics

We have seen that the cdf of the maximum Z_n and minimum W_n of a sample of size n coming from a population with cdf $F(x)$ are $H_n(x) = P[Z_n \le x] = F^n(x)$ and $L_n(x) = P[W_n \le x] = 1 - [1 - F(x)]^n$. When n tends to infinity we have

$$\lim_{n\to\infty} H_n(x) = \lim_{n\to\infty} F^n(x) = \begin{cases} 1 & \text{if } F(x) = 1 \\ 0 & \text{if } F(x) < 1 \end{cases}$$

and

$$\lim_{n\to\infty} L_n(x) = \lim_{n\to\infty} 1 - [1 - F(x)]^n = \begin{cases} 0 & \text{if } F(x) = 0 \\ 1 & \text{if } F(x) > 0 \end{cases}$$

which means that the limit distributions are degenerate.

With the aim of avoiding degeneracy, we look for linear transformations $Y = a_n + b_n x$, where a_n and b_n are constants, depending on n, such that the limit distributions

$$\lim_{n\to\infty} H_n(a_n + b_n x) = \lim_{n\to\infty} F^n(a_n + b_n x) = H(x) \quad \forall x \qquad (25)$$

and

$$\lim_{n\to\infty} L_n(c_n + d_n x) = \lim_{n\to\infty} 1 - [1 - F(c_n + d_n x)]^n = L(x) \qquad \forall x \qquad (26)$$

are not degenerate.

Definition 13. Domain of Attraction of a Given Distribution: *A given distribution, $F(x)$, is said to belong to the domain of attraction for maxima of $H(x)$, if Eq. (25) holds for at least one pair of sequences $\{a_n\}$ and $\{b_n > 0\}$. Similarly, when $F(x)$ satisfies (26) we say that it belongs to the domain of attraction for minima of $L(x)$.*

The problem of limit distribution can then be stated as:

1. Find conditions under which Eqs (25) and (26) are satisfied.
2. Give rules for building the sequences $\{a_n\}$, $\{b_n\}$, $\{c_n\}$, and $\{d_n\}$.
3. Find what distributions can occur as $H(x)$ and $L(x)$.

The answer of the third problem is given by the following theorem [14–16].

Theorem 10. Feasible Limit Distribution for Maxima: *The only nondegenerate distributions $H(x)$ satisfying Eq. (25) are*

$$\text{Frechet}: H_{1,g}(x) = \begin{cases} \exp(-x^{-g}) & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Weibull}: H_{2,g}(x) = \begin{cases} 1 & \text{if } \le 0 \\ \exp[-(-x)^{g}] & \text{otherwise} \end{cases}$$

and

$$\text{Gumbel}: H_{3,0}(x) = \exp[-\exp(-x)] \qquad -\infty < x < \infty$$

Theorem 11. Feasible Limit Distribution for Minima: *The only nondegenerate distributions $L(x)$ satisfying Eq. (26) are*

$$\text{Frechet}: L_{1,g}(x) = \begin{cases} 1 - \exp[-(-x)^{-g}] & \text{if } x < 0 \\ 1 & \text{otherwise} \end{cases}$$

$$\text{Weibull}: L_{2,g}(x) = \begin{cases} 1 - \exp(-x^{g}) & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

and

$$\text{Gumbel}: L_{3,0}(x) = 1 - \exp(-\exp x) \qquad -\infty < x < \infty$$

To know the domains of attraction of a given distribution and the associated sequences, the reader is referred to Galambos [16].

Some important implications of his theorems are:

1. Only three distributions (Frechet, Weibull, and Gumbel) can occur as limit distributions for maxima and minima.
2. Rules for determining if a given distribution $F(x)$ belongs to the domain of attraction of these three distributions can be obtained.
3. Rules for obtaining the corresponding sequences $\{a_n\}$ and $\{b_n\}$ or $\{c_n\}$ and $\{d_n\}$ $(i = 1, \ldots)$ can be obtained.
4. A distribution with no finite end in the associated tail cannot belong to the Weibull domain of attraction.
5. A distribution with finite end in the associated tail cannot belong to the Frechet domain of attraction.

Next we give another more efficient alternative to solve the same problem. We give two theorems [13, 17] that allow this problem to be solved. The main advantage is that we use a single rule for the three cases.

Theorem 12. Domain of Attraction for Maxima of a Given Distribution: *A necessary and sufficient condition for the continuous cdf $F(x)$ to belong to the domain of attraction for maxima of $H_c(x)$ is that*

$$\lim_{\varepsilon\to 0} \frac{F^{-1}(1-\varepsilon) - F^{-1}(1-2\varepsilon)}{F^{-1}(1-2\varepsilon) - F^{-1}(1-4\varepsilon)} = 2^{c}$$

where c is a constant. This implies that:

If $c < 0$, $F(x)$ belongs to the Weibull domain of attraction for maxima.
If $c = 0$, $F(x)$ belongs to the Gumbel domain of attraction for maxima.
If $c > 0$, $F(x)$ belongs to the Frechet domain of attraction for maxima.

Theorem 13. Domain of Attraction for Minima of a Given Distribution: *A necessary and sufficient condition for the continuous cdf $F(x)$ to belong to the domain of attraction for minima of $L_c(x)$ is that*

$$\lim_{\varepsilon\to 0} \frac{F^{-1}(\varepsilon) - F^{-1}(2\varepsilon)}{F^{-1}(2\varepsilon) - F^{-1}(4\varepsilon)} = 2^{c}$$

This implies that:

If $c < 0$, $F(x)$ belongs to the Weibull domain of attraction for minima.
If $c = 0$, $F(x)$ belongs to the Gumbel domain of attraction for minima.
If $c > 0$, $F(x)$ belongs to the Frechet domain of attraction for minima.

Table 7 shows the domains of attraction for maxima and minima of some common distributions.

1.11 PROBABILITY PLOTS

One of the graphical methods commonly used by engineers is the probability plot. The basic idea of probability plots, of a biparametric family of distributions, consists of modifying the random variable and the probability drawing scales in such a manner that the cdfs become a family of straight lines. In this way, when the cdf is drawn a linear trend is an indication of the sample coming from the corresponding family.

In addition, probability plots can be used to estimate the parameters of the family, once we have checked that the cdf belongs to the family.

However, in practice we do not usually know the exact cdf. We, therefore, use the empirical cdf as an approximation to the true cdf. Due to the random character of samples, even in the case of the sample

Table 7 Domains of Attraction of the Most Common Distributions

	Domain of attraction	
Distribution[a]	for maxima	for minima
Normal	Gumbel	Gumbel
Exponential	Gumbel	Weibull
Log-normal	Gumbel	Gumbel
Gamma	Gumbel	Weibull
Gumbel$_M$	Gumbel	Gumbel
Gumbel$_m$	Gumbel	Gumbel
Rayleigh	Gumbel	Weibull
Uniform	Weibull	Weibull
Weibull$_M$	Weibull	Gumbel
Weibull$_m$	Gumbel	Weibull
Cauchy	Frechet	Frechet
Pareto	Frechet	Weibull
Frechet$_M$	Frechet	Gumbel
Frechet$_m$	Gumbel	Frechet

[a] M = for maxima; m = for minima.

coming from the given family, the corresponding graph will not be an exact straight line. This complicates things a little bit, but if the trend approximates linearity, we can say that the sample comes from the associated family.

In this section we start by discussing the empirical cdf and define the probability graph, then give examples of the probability graph for some distributions useful for engineering applications.

1.11.1 Empirical Cumulative Distribution Function

Let $x_{i:n}$ denote the ith observed order statistic in a random sample of size n. Then the *empirical cumulative distribution function* (ecdf) is defined as

$$p(X = x) = \begin{cases} 0 & \text{if } x < x_{1:n} \\ i/n & \text{if } x_{i:n} \le x \le x_{i+1:n} \quad i = 1, \ldots, n-1 \\ 1 & \text{if } x > x_{n:n} \end{cases}$$

This is a jump (step) function. However, there exist several methods that can be used to smooth this function, such as linear interpolation methods [18].

1.11.2 Fundamentals of Probability Plots

A probability plot is simply a scatter plot with transformed scales for the two-dimensional family to become the set of straight lines with positive slope (see Castillo [19], pp. 131–173).

Let $F(x; a, b)$ be a biparametric family of cdfs, where a and b are the parameters. We look for a transformation

$$\xi = g(x) \qquad \eta = h(y) \tag{27}$$

such that the family of curves $y = F(x; a, b)$ after transformation (27) becomes a family of straight lines.

Note that this implies

$$h(y) = h[F(x; a, b)] = ag(x) + b \quad \Leftrightarrow \quad \eta = a\xi + b \tag{28}$$

where the variable η is called the reduced variable.

Thus, for the existence of a probabilistic plot associated with a given family of cdfs $F(x; a, b)$ it is necessary to have $F(x; a, b) = h^{-1}[ag(x) + b]$.

As we mentioned above, in cases where the true cdf is unknown we estimate the cdf by the ecdf. But the ecdf has steps $0, 1/n, 2/n, \ldots, 1$. However, the two extremes 0 and 1, when we apply the scale transformation become $-\infty$ and ∞, respectively, in the case of many families. Thus, they cannot be drawn.

Due to the fact that in the order statistic $x_{i:n}$ the probability jumps from $(i-1)/n$ to i/n, one solution, which has been proposed by Hazen [20], consists of using the value $(i - 1/2)/n$; thus, we draw on the probability plot the points

$$(x_{i:n}, (i - 0.5)/n) \qquad i = 1, \ldots, n$$

Other alternative plotting positions are given in Table 8. (For a justification of these formulas see Castillo [13], pp. 161–166.)

In the following subsection we give examples of probability plots for some commonly used random variables.

Table 8 Plotting Positions Formulas

Formula	Source
$(x_{i:n}, i/(n+1))$	—
$(x_{i:n}, (i - 0.375)/(n + 0.25))$	Blom [21]
$(x_{i:n}, (i - 0.5)/n)$	Hazen [20]
$(x_{i:n}, (i - 0.44)/(n + 0.12))$	Gringorten [22]

1.11.3 The Normal Probability Plot

The cdf $F(x; \mu, \sigma)$ of a normal random variable can be written as

$$F(x; \mu, \sigma) = \Phi\left(\frac{x-\mu}{\sigma}\right) \tag{29}$$

where μ and σ are the mean and the standard deviation, respectively, and $\Phi(x)$ is the cdf of the standard normal variable $N(0, 1)$. Then, according to Eqs (27) and (28), Eq. (29) gives

$$\xi = g(x) = x \quad \eta = h(y) = \Phi^{-1}(y) \quad a = \frac{1}{\sigma} \quad b = \frac{-\mu}{\sigma}$$

and the family of straight lines becomes

$$\eta = a\xi + b = \frac{\xi - \mu}{\sigma} \tag{30}$$

Once the normality assumption has been checked, estimation of the parameters μ and σ is straightforward. In fact, setting $\eta = 0$ and $\eta = 1$ in Eq. (30), we obtain

$$\begin{aligned} \eta = 0 \quad &\Rightarrow \quad 0 = (\xi - \mu)/\sigma \quad \Rightarrow \quad \xi = \mu \\ \eta = 1 \quad &\Rightarrow \quad 1 = (\xi - \mu)/\sigma \quad \Rightarrow \quad \xi = \mu + \sigma \end{aligned} \tag{31}$$

Figure 18 shows a normal probability plot, where the ordinate axis has been transformed by $\eta = \Phi^{-1}(y)$, whereas the abscissa axis remains untransformed. Note that we show the probability scale Y and the reduced scale η.

1.11.4 The Log-Normal Probability Plot

The case of the log-normal probability plot can be reduced to the case of the normal plot if we take into account that X is log-normal iff $Y = \log(X)$ is normal. Consequently, we transform X into $\log(x)$ and obtain a normal plot. Thus, the only change consists of transforming the X scale to a logarithmic scale (see Fig. 19). The mean μ^* and the standard deviation σ^* of the log-normal distribution can then be estimated by

$$\mu^* = e^{\mu + \sigma^2/2} \qquad \sigma^{*2} = e^{2\mu}\left(e^{2^2} - e^{\sigma^2}\right)$$

where μ and σ are the values obtained according to Eq. (31).

1.11.5 The Gumbel Probability Plot

The Gumbel cdf for maxima is

$$F(x; \lambda, \delta) = \exp\left[-\exp\left(-\frac{x-\lambda}{\delta}\right)\right] \qquad -\infty < x < \infty \tag{32}$$

Let $p = F(x; \lambda, \delta)$. Then taking logarithms of $1/p$ twice we get

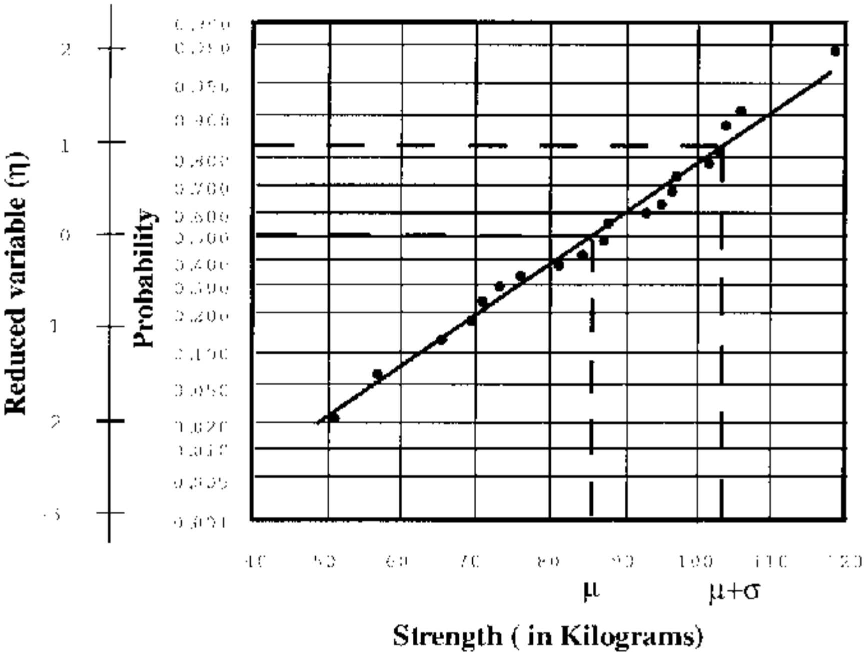


Figure 18 An example of a normal probability plot.

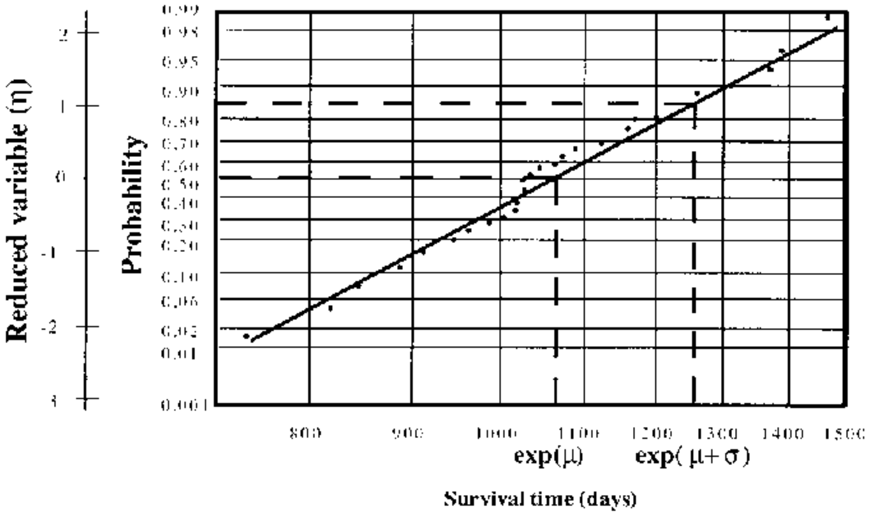


Figure 19 An example of a log-normal probability plot.

$$-\log\left[\log\left(\frac{1}{p}\right)\right] = \frac{x-\lambda}{\delta}$$

Upon comparison with Eqs (27) and (28), we get

$$\begin{aligned}\xi &= g(x) = x\\ \eta &= h(p) = -\log\left[\log\left(\frac{1}{p}\right)\right]\\ &= -\log(-\log p) \qquad a = 1/\delta \qquad b = -\lambda/\delta\end{aligned} \tag{33}$$

which shows that the transformation Eq. (33) transforms Eq. (32) to a family of straight lines

$$\eta = a\xi + b = \frac{\xi - \lambda}{\delta}$$

Estimation of the two parameters λ and δ can be done by noting that for $\eta = 0$ and $\eta = 1$. Therefore,

$$\begin{aligned}\eta = 0 = (\xi - \lambda)/\delta \quad &\Rightarrow \quad \xi = \lambda\\ \eta = 1 = (\xi - \lambda)/\delta \quad &\Rightarrow \quad \xi = \lambda + \delta\end{aligned}$$

Thus, once we have fitted a straight line to the data, the abscissas associated with the reduced variable, η, namely 0 and 1, are the values λ and $\lambda + \delta$, respectively. Figure 20 shows a Gumbel probability plot for

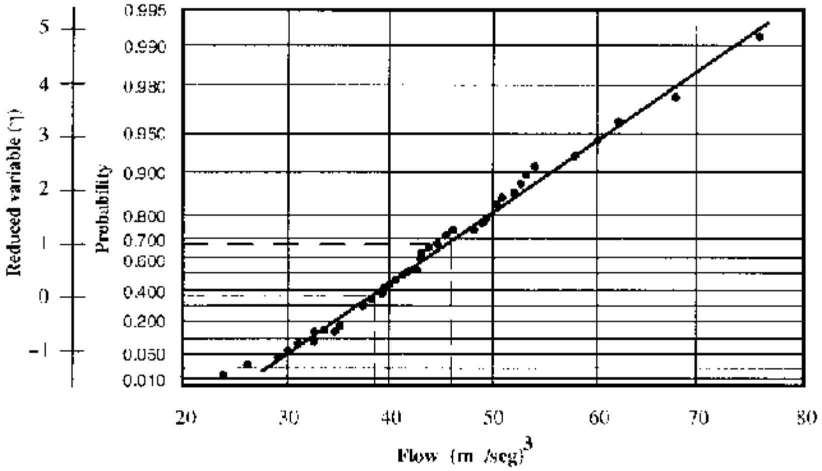


Figure 20 An example of a Gumbel probability plot for maxima.

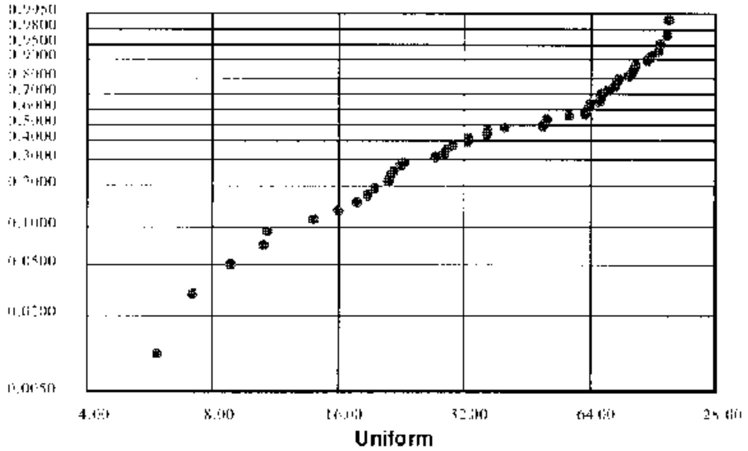


Figure 21 An example of a Weibull probability plot for minima.

maxima in which the ordinate axis has been transformed according to Eq. (33) and the abscissa axis remains unchanged.

1.11.6 The Weibull Probability Plot

The Weibull cdf for maxima is

$$y = F(x; \lambda, \beta, \delta) = \exp\left[-\left(\frac{\lambda - x}{\delta}\right)^{\beta}\right] \qquad (34)$$

$$-\infty < x \le \lambda$$

Letting $p = F(x; \lambda, \beta, \delta)$ and taking logarithms twice we get

$$\begin{aligned} -\log(-\log y) &= -\beta \log\left(\frac{\lambda - x}{\delta}\right) \\ &= -\beta \log(\lambda - x) + \beta \log \delta \end{aligned}$$

Comparison with Eqs (27) and (28) gives

$$\begin{aligned} \xi &= g(x) = -\log(\lambda - x) \\ \eta &= h(y) = -\log(-\log y) \end{aligned} \qquad (35)$$

and

$$a = \beta \qquad b = \beta \log \delta \qquad (36)$$

This shows that the transformation Eq. (35) transforms Eq. (34) to a family of straight lines

$$\eta = a\xi + b = \beta(\xi + \log \delta)$$

Note that the η scale coincides with that for the Gumbel plot, but now the ξ scale is logarithmic.

Since now the parameter λ is unknown, we must proceeds by successive approximations until we get a straight line and then we can proceed to estimate the remaining parameters β and δ noting that for $\eta = 0$ and $\eta = 1$ we obtain

$$\eta = 0 = \beta(\xi + \log \delta) \quad \Rightarrow \quad \xi = -\log \delta$$

$$\eta = 1 = \beta(\xi + \log \delta) \quad \Rightarrow \quad \xi = \frac{1}{\beta} - \log \delta$$

Figure 21 shows a Weibull probability plot for minima.

REFERENCES

1. JR Benjamin, CA Cornell. *Probability, Statistics and Decision for Civil Engineers*. New York: McGraw Hill Book Company, 1970.
2. R Christensen. *Data Distributions: A Statistical Handbook*. Lincoln, MA: Entropy Limited, 1984.
3. N Johnson, S Kotz, AW Kemp. *Discrete Univariate Distributions*. New York: John Wiley & Sons, 1997.
4. N Johnson, S Kotz, N Balakrishnan. *Continuous Univariate Distributions*. New York: John Wiley & Sons, 1994.
5. E Castillo, JM Sarabia. Extreme value analysis of wave heights. *J Res Nat Inst Stand Technol*, 99: 445–454, 1994.
6. E Castillo, JM Gutiérrez, AS Hadi. *Expert Systems and Probabilistic Network Models*. New York: Springer-Verlag, 1992.
7. E Castillo, JM Gutiérrez, AS Hadi. *Sistemas Expertos y Redes Probabilísticas*. Madrid: Academia de Ingeniería, 1997.

8. AC Rencher. *Methods of Multivariate Analysis*. New York: John Wiley & Sons, 1995.
9. E Castillo, AS Hadi. Parameter and quantile estimation for the generalized extreme-value distribution. *Environmetrics* 5: 417–432, 1994.
10. E Castillo, AS Hadi. Modeling life-time data with application to fatigue models. *J Am Statist Assoc* 90: 1041–1054, 1995.
11. E Castillo, AS Hadi. Fitting the generalized Pareto distribution to data. *J Am Statist Assoc* 92: 1609–1620, 1997.
12. HA David. *Order Statistics*. New York: John Wiley & Sons, 1981.
13. E Castillo. *Extreme Value Theory in Engineering*. New York: Academic Press, 1988.
14. RA Fisher, LHC Tippett. Limiting forms of the frequency distributions of the largest or smallest member of a sample. *Proc Cambridge Philos Soc* 24 180–190, 1928.
15. J Tiago de Oliveira. Extremal distributions. *Rev Fac Cienc* A6: 121–146, 1958.
16. J Galambos. *The Asymptotic Theory of Extreme Order Statistics*. Malabar, FL: Krieger, 1987.
17. E Castillo, J Galambos, JM Sarabia. The selection of the domain of attraction of an extreme value distribution from a set of data. *Proceedings, Oberwolfach, Extreme Value Theory, Lecture Notes in Statistics* 51: 181–190, 1987.
18. JS Simonoff. *Smoothing Methods in Statistics*. New York: Springer-Verlag, 1996.
19. E Castillo *Introduccion a la estadistica aplicada*. Madrid: Paraninfo, 1978.
20. A Hazen. *Flood flows: A Study of Frequencies and magnitudes*. New York: John Wiley & Sons, 1930.
21. G Blom. *Statistical Estimates and Transformed Beta-Variables*. Uppsala, Sweden: Almqvist and Wiksell; New York: John Wiley & Sons, 1958.
22. II Gringorten. A plotting rule for extreme probability paper. *J Geophys Res* 68: 813–814, 1963.

Chapter 1.2

Introduction to Sets and Relations

Diego A. Murio
University of Cincinnati, Cincinnati, Ohio

2.1 SETS

The concept of sets is basic to all mathematics and mathematical applications. A *set* is simply a collection of objects and we assume that this notion is intuitively clear. Note, however, that the word "collection" is as undefined, in this setting, as is the word "set." Throughout this chapter, we shall denote sets by capital letters, such as A, B, X, and elements of these sets by lowercase letters, such as a, b, x. If A is a set and x is an object that belongs to A, we say "x is an element of A," "x belongs to A," or "x is a member of A" and write

$$x \in A$$

If x is not an element of A, we write

$$x \notin A$$

It is convenient to have several ways of describing sets. If a set does not have too many elements, we can describe it by *listing* its elements between a pair of curly brackets. For example, the set consisting of the whole numbers 1 to 3 can be written as

$$A = \{1, 2, 3\}$$

Alternatively, if B is the set of all elements from some collection X that satisfy some property P, we write this as

$$B = \{x \in X \mid x \text{ satisfies the property } P\}$$

or

$$B = \{x \in X : x \text{ satisfies the property } P\}$$

These are read as "the set of x elements that belong to X such that property P is true." If X is understood, we simply write

$$B = \{x : x \text{ satisfies the property } P\}$$

For example, the set

$$D = \{x : x \text{ is a positive even integer}\}$$

describes the set made up of all positive, even integers: 2, 4, 6, . . .

The set

$$C = \{x : x^2 - 1 = 0\}$$

is a fancy way to describe the set

$$C = \{-1, 1\}$$

It can happen that there is no $x \in X$ that satisfies property P. In such a case the set will contain no elements at all. This set is called the *empty* (or *null* or *void*) set and it is denoted $\emptyset$. For instance, if X represents the set of real numbers,

$$\emptyset = \{x \in X : x^2 + 1 = 0\}$$

because the square of any real number is nonnegative and, consequently, it is always true that $x^2 + 1 \geq 1$. This last statement implies that $x^2 + 1 = 0$ is *never true*.

If A and B are sets such that each member of A is also a member of B, then A is a *subset* of B or, equivalently, A is *contained* in B, and we write $A \subseteq B$.

The set A is said to be a *proper subset* of B if $A \subseteq B$ and there is some element $x \in B$ such that $x \notin A$; that is, every element of A is in B but B contains at least one element that is not in A. The corresponding notation is $A \subset B$ and we also say that the set A is *strictly contained* in B.

Two sets A and B are *equal*, written $A = B$, if every element of A is an element of B and every element of B is an element of A. In other words, $A = B$ if and only if we have both $A \subseteq B$ and $B \subseteq A$. This statement has a double implication and it should be understood as follows:

1. If $A = B$ then $A \subseteq B$ and $B \subseteq A$.
2. If $A \subseteq B$ and $B \subseteq A$ then $A = B$.

Notice that the order in which the elements of a set are listed does not matter, nor does the repetition of elements. For example, the sets $A = \{a, b, c\}$ and $B = \{c, b, a, a, b\}$ are both the same set: $A = B$. Notice also that for any given set A,

$$A \subseteq A$$

and

$$\emptyset \subseteq A$$

Additional sets can be formed by using the elements of a given set. For instance, if $A = \{a, b, c\}$, the set whose elements are all the subsets of A, the *power set* $P(A)$, is given by

$$P(A) = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}$$

Note that all the members of $P(A)$, except A itself, are proper subsets of A.

If A is a finite set, the cardinality of A, written $|A|$, indicates the number of elements in A. For our previous example, $|A| = 3$ and $|P(A)| = 8$. In general, it can be shown that if $|A| = n$, then $|P(A)| = 2^n$.

2.1.1 Set Operations

The *union* of the sets A and B, written $A \cup B$, is the set of all elements that belong to either A or B or both. In symbols,

$$A \cup B = \{x : x \in A \text{ or } x \in B\}$$

The *intersection* of the sets A and B, written $A \cap B$, is the set of all elements that belong to both A and B. In other words,

$$A \cap B = \{x : x \in A \text{ and } x \in B\}$$

Sometimes we have to work with sets all of which are subsets of a given set U, called a *universal set*. This particular set of reference must be explicitly given or clearly inferred from the context.

The *complement* of the set A *relative* to the set B, denoted $B - A$, is defined as the set of all elements belonging to B but not to A, i.e.,

$$B - A = \{x : x \in B \text{ and } x \notin A\}$$

A very common situation occurs when the set B is the universal set U. In this case the complement of the set A relative to U, $U - A$, is simply called the *complement* of the set A and denoted A^c.

To help us understand the new terms and definitions, we now look at some simple examples.

Let $A = \{a, b, c, 1, 2, 3\}$, $B = \{1, a, b, c\}$, and $C = \{3, 2\}$. Then

$$A \cup B = \{a, b, c, 1, 2, 3\}$$
$$A \cap B = \{1, a, b, c\}$$
$$B \cup C = \{a, b, c, 1, 2, 3\}$$
$$A - B = \{2, 3\}$$
$$B - A = \emptyset$$
$$B \cap C = \emptyset$$

If two sets have an empty intersection, we say that they are *disjoint*.

If the universal set U is given by $U = \{1, 2, 3, 4, 5, 6, 7, 8\}$ and we consider the sets $A = \{2, 6, 8\}$, $B = \{1, 2, 3, 4\}$, and $C = \{1, 3, 5, 7\}$, then

$$A^c = \{1, 3, 4, 5, 7\}$$
$$B^c = \{5, 6, 7, 8\}$$
$$C^c = \{2, 4, 6, 8\}$$
$$A^c - (B - C) = \{1, 3, 4, 5, 7\} - \{2, 4\} = \{1, 3, 5, 7\}$$

The following properties can be derived from the previous definitions. If A, B, and C are sets, then

$$A \cap B = B \cap A \tag{1}$$
$$A \cup B = B \cup A \tag{2}$$
$$(A \cap B) \cap C = A \cap (B \cap C) \tag{3}$$
$$(A \cup B) \cup C = A \cup (B \cup C) \tag{4}$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \tag{5}$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \tag{6}$$

They represent, in order, commutativity of intersection and union, associativity of interesection and union, and distributivity of intersection with respect to union and union with respect to intersection.

Properties (7) and (8) below are known as De Morgan's laws. They relate the operations of intersection, union, and complementation:

$$(A \cap B)^c = A^c \cup B^c \tag{7}$$
$$(A \cup B)^c = A^c \cap B^c \tag{8}$$

Other useful relationships involving the universal set, the empty set, and complements are

$$A \cap \emptyset = \emptyset \tag{9}$$
$$A \cup \emptyset = A \tag{10}$$
$$A \cap B \subseteq A \subseteq A \cup B \tag{11}$$
$$A \cap U = A \tag{12}$$
$$A \cup U = U \tag{13}$$

and

$$(A^c)^c = A \tag{14}$$
$$U^c = \emptyset \tag{15}$$
$$A \cap A^c = \emptyset \tag{16}$$
$$A \cup A^c = U \tag{17}$$

2.1.1.1 Venn Diagrams

Before supplying the proof of some of the previous propositions, we point out that it is possible to illustrate properties of sets graphically by means of Venn diagrams. The diagrams are very simple and start by drawing a rectangle whose interior points represent the elements of the universal set (Fig. 1).

Any subset of the universal set is now represented by a circle within the rectangle. The elements of the set are represented by the points inside the circle. For example, if we want to represent the set A, we draw the diagram shown in Fig. 2.

Notice that the points of the rectangle, outside the circle, represent the elements of the set A^c (Fig. 3).

The concept of union of two sets, A and B, is represented as shown in Fig. 4.

For the interesection of two sets, A and B, the corresponding Venn diagram is as shown in Fig. 5.

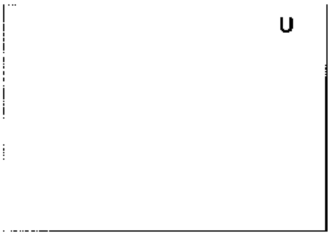


Figure 1 Universal set U.

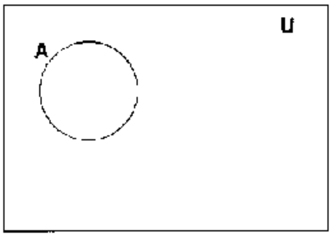


Figure 2 Set A.

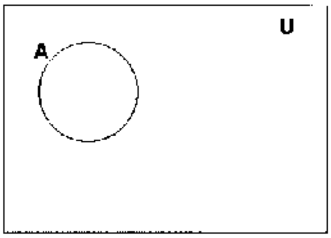


Figure 3 Complement set A^c.

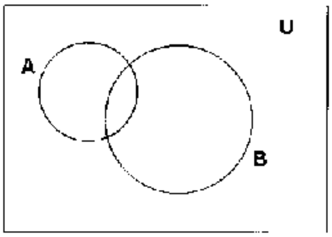


Figure 4 Union set $A \cup B$.

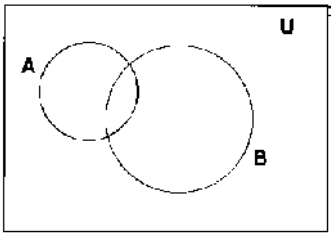


Figure 5 Intersection set $A \cap B$.

Diagrams get a little more complicated if we want to visualize some other expressions. For example, to illustrate De Morgan's law, Eq. (7), we must proceed in stages. First, we generate a Venn diagram for the set represented by the left-hand side of the equation. Second, we generate another Venn diagram for the set represented by the right-hand side of the equation. Finally, we must compare both diagrams (Fig. 6).

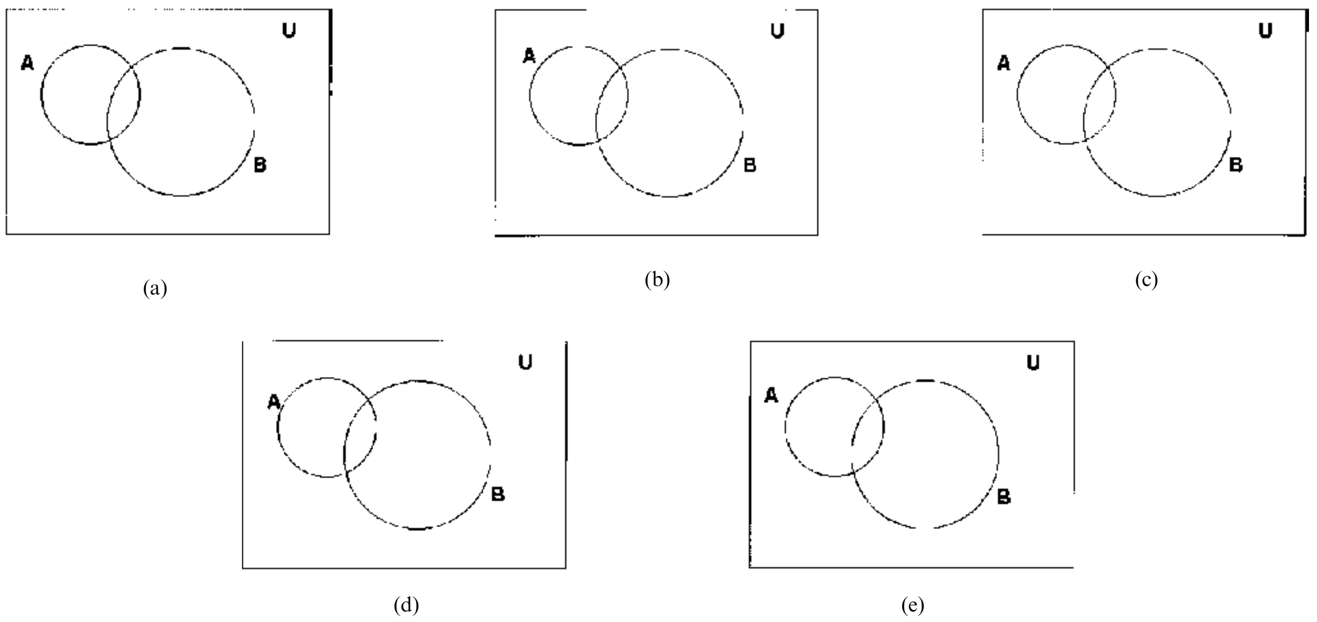


Figure 6 Venn diagrams for De Morgan's law. Left- and right-hand sides of Eq. (7).

Notice that Venn diagrams do not prove De Morgan's law. They simply validate the property for the particular arrangement of sets in the diagrams.

2.1.1.2 Rigorous Proofs

To rigorously prove property (7), according to our definition of equality of two sets, we must show the double inclusion

$$(A \cap B)^c \subseteq A^c \cup B^c \tag{18}$$

and

$$A^c \cap B^c \subseteq (A \cap B)^c \tag{19}$$

Proof. *In proving Eq. (18), we consider an arbitrary element $x \in (A \cap B)^c$. Then since x is an element that belongs to the complement of the set $A \cap B$, $x \notin A \cap B$. If $x \notin A \cap B$, either $x \notin A$ or $x \notin B$. This means that $x \in A^c$ or $x \in B^c$. That is, $x \in A^c \cup B^c$. We have shown that if $x \in (A \cap B)^c$ then $x \in A^c \cap B^c$. In other words, $(A \cap B)^c \subseteq A^c \cap B^c$.*

In proving Eq. (19), we consider an arbitrary element $x \in A^c \cup B^c$. Then either $x \in A^c$ or $x \in B^c$. If $x \in A^c$, then $x \notin A$. If $x \notin A$, then $x \notin A \cap B$. Analogously, if $x \in B^c$, then $x \notin A \cap B$. All together, if $x \in A^c \cup B^c$, $x \notin A \cap B$. In other words, if $x \in A^c \cup B^c$, $x \in (A \cap B)^c$. Since x is arbitrary, this property is true for every element in $A^c \cup B^c$ and this implies that $A^c \cup B^c \subseteq (A \cap B)^c$.

Finally, from (1) and (2), it follows that $(A \cap B)^c = A^c \cup B^c$.

As a second example, let us prove proposition (14). Again, the method of proof consists of showing that the set on the left-hand side of the proprosed equality is a subset of the set on the right-hand side and vice versa. We have to show that

$$(A^c)^c \subseteq A \tag{20}$$

and

$$A \subseteq (A^c)^c \tag{21}$$

Proof. *In order to demonstrate Eq. (20), we consider an arbitrary element $x \in (A^c)^c$. Then $x \notin A^c$ and this implies that $x \in A$. We have shown that $(A^c)^c \subseteq A$.*

To prove Eq. (21), we consider an arbitrary element $x \in A$. Then $x \notin A^c$ which means that x belongs to the complement of A^c, i.e., $x \in (A^c)^c$. This shows that $A \subseteq (A^c)^c$.

Finally, from Eqs (20) and (21), it follows that $A = (A^c)^c$.

2.1.1.3 Indexed Family of Sets

The notions of union and intersection of two sets can be generalized to unions and intersections of more than two sets after introducing the notion of indexed family of sets. Let Δ be a set and assume that with each element $\gamma \in \Delta$ there is associated a subset A_γ of a given set S. The collection of all such sets A_γ is called an *indexed family* of subsets of S with Δ as the index set and it is denoted

$$\{A_\gamma\}_{\gamma\in\Delta}$$

Given an indexed family of sets, we define

$$\bigcup_{\gamma\in\Delta} A_\gamma = \{x : x \in A_\gamma \text{ for some } \gamma \in \Delta\}$$

and

$$\bigcap_{\gamma\in\Delta} A_\gamma = \{x : x \in A_\gamma \text{ for all } \gamma \in \Delta\}$$

Virtually all the notation used for sets applies to families of sets as well. In the event that the index set is a subset of the set N of positive integers, it is customary to write, if $S = N$,

$$\bigcap_{\gamma\in N} A_\gamma = \bigcap_{n=1}^{\infty} A_n$$

or, if $S = \{1, 2, 3, 4, 5, \ldots, k\}$,

$$\bigcap_{n=1}^{k} A_n$$

instead of

$$\bigcap_{\gamma\in S} A_\gamma$$

For instance, the collection of sets $\{1, 3\}$, $\{2, 4\}$, $\{3, 5\}, \ldots, \{n, n+2\}, \ldots$ may be considered as an indexed family of sets with index set N. We can write $A_n = \{n, 2n\}$ and the family of sets become $\{A_n : n \in N\}$.

We borrow two more examples, without proof, from real analysis. If R denotes the set of real numbers, given a and b in R, $a < b$, we define the open interval of real numbers (a, b) as the set

$$(a, b) = \{x : x \in R \text{ and } a < x < b\}$$

Now consider the sets $A_n = (0, 1/n)$ and $B_n = (1/n, 1)$ for each $n \in N$. Then,

$$\bigcap_{n=1}^{\infty} A_n = \emptyset$$

and

$$\bigcap_{n=1}^{\infty} B_n = (0, 1)$$

2.2 RELATIONS

2.2.1 Ordered Pairs

Given any two objects x and y, it is possible to form the set whose only members are x and y. We write $\{x, y\}$ or $\{y, x\}$. In this section we are interested in another object that can be constructed from two elements: an *ordered pair* (x, y), where x is called the first coordinate and y is the second coordinate. The "ordered" clause emphasizes that the order in which the objects x and y appear is essential. The way of defining the ordered pair (x, y) as a set is as follows:

$$(x, y) = \{\{x, \}, \{x, y\}\}$$

Next, we prove a natural property of ordered pairs; that is, $(x, y) = (u, v)$ if and only if $x = u$ and $y = v$.

Proof. *If* $x = u$ *and* $y = v$, *then* $(x, y) = \{\{x\}, \{x, y\}\} = \{\{u\}, \{u, v\}\} = (u, v)$.

Suppose now that $(x, y) = (u, v)$. *There are two possibilities: (1)* $x = y$ *and (2)* $x \neq y$.

1. *If* $x = y$, *then* $\{\{x\}\} = \{\{u\}, \{u, v\}\}$. *Since the set on the left-hand side has only one member, the set on the right must also have one member. This can only happen if* $\{u\} = \{u, v\}$, *which is true only if* $u = v$. *Consequently, we have* $\{\{x\}\} = \{\{u\}\}$. *Thus,* $\{x\} = \{u\}$ *and* $x = u$. *Altogether, we obtain* $x = y = u = v$.
2. *If* $x \neq y$, $\{x\} \neq \{x, y\}$. *Since* $\{\{x\}, \{x, y\}\} = \{\{u\}, \{u, v\}\}$, $\{x\} \in \{\{u\}, \{u, v\}\}$ *and we conclude that* $\{x\} = \{u\}$. *Notice that* $\{x\} = \{u, v\}$ *is impossible since u and v are distinct. Thus, it follows that* $x = u$. *Similarly, from* $\{x, y\} \in \{\{u\}, \{u, v\}\}$, *the only remaining element that can be equal to* $\{x, y\}$ *is* $\{u, v\}$. *Finally, since we already know that* $x = u$, *the equality* $\{x, y\} = \{u, v\}$ *implies* $y = v$.

The notion of ordered pairs can be generalized very naturally to sets involving three, four, or more elements. For example, if x, y, and z are three objects, the *ordered triplet* (x, y, z) is defined as the ordered pair $((x, y), z)$, etc.

2.2.2 Cartesian Product

By restricting the choice of coordinates of ordered pairs to elements of given sets, we arrived at the concept of Cartesian products. The name is derived from the classical method of determining the coordinates of a point in the plane by the French mathematician René Descartes (1596–1650). Given two sets A and B, the set of all ordered pairs (x, y) with $x \in A$ and $y \in B$, is called the *Cartesian product of A and B* and it is denoted $A \times B$. In symbols,

$$A \times B = \{(x, y) : x \in A \text{ and } y \in B\}$$

As an example, let $A = \{1, 2, 3\}$ and $B = \{a, b\}$. By the above definition we have

$$A \times B = \{(1, a), (1, b), (2, a), (2, b), (3, a), (3, b)\}$$

and

$$B \times A = \{(a, 1), (a, 2), (a, 3), (b, 1), (b, 2), (b, 3)\}$$

We notice that in general the Cartesian product of two sets is not commutative. Also, if $|A| = m$ and $|B| = n$, then $|A \times B| = |B \times A| = mn$.

The Cartesian product can be represented pictorially as the set of points in Fig. 7.

We leave as an interesting exercise the task to show that the Cartesian product distributes with respect to intersection and union of sets. In symbols, if A, B, and C are any three sets, then

$$A \times (B \cap C) = (A \times B) \cap (A \times C)$$
$$A \times (B \cup C) = (A \times B) \cup (A \times C)$$

and

$$A \times (B - C) = (A \times B) - (A \times C)$$

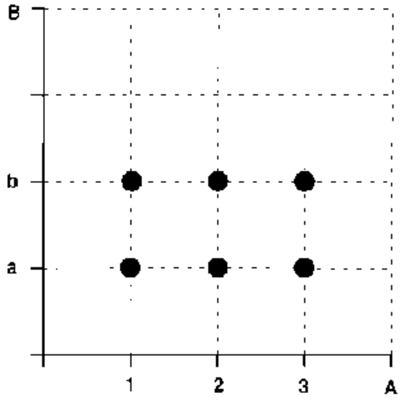


Figure 7 Cartesian product $A \times B$.

2.2.3 Relations

A *relation* is a set of ordered pairs. More precisely, given two sets A and B, a *binary relation* $\Re$ *from* a set A to a set B is a subset of the Cartesian product $A \times B$. To indicate that the ordered pair $(x, y) \in \Re$ we write $x\Re y$ and we say that x is $\Re$ related to y. In the particular case of having $A = B$, we say that $\Re$ is a relation *in A* (or B) instead of *from A* to A.

The set of all x which are in relation $\Re$ with some y is called the *domain* of $\Re$ and it is denoted Dom $\Re$:

$$\text{Dom } \Re = \{x : \text{ there exists } y \text{ such that } x\Re y\}$$

The set of all y such that, for some x, x is in relation $\Re$ with y is called the *range* of $\Re$ and it is denoted Ran $\Re$:

$$\text{Ran } \Re = \{y : \text{ there exists } x \text{ such that } x\Re y\}$$

Notice that the sets Dom $\Re$ and Ran $\Re$ represent the sets of first and second coordinates, respectively, of all ordered pairs in $\Re$.

Given a relation $\Re$ from A to B, the *inverse of the relation* $\Re$, denoted $\Re^{-1}$, is the relation from B to A such that $y\Re^{-1}x$ if and only if $x\Re y$. In symbols,

$$\Re^{-1} = \{(y, x) : (x, y) \in \Re\}$$

An example will help to understand this concept. Suppose that we are given the following relation:

$$\Re = \{(x, y) : x, y \in \{0, 1, 2\} \text{ and } x < y\}$$

We can easily list the elements of $\Re$,

$$\Re = \{(0, 1), (0.2), (1, 2)\}$$

This is shown graphically in Fig. 8.

By inspection, it follows that

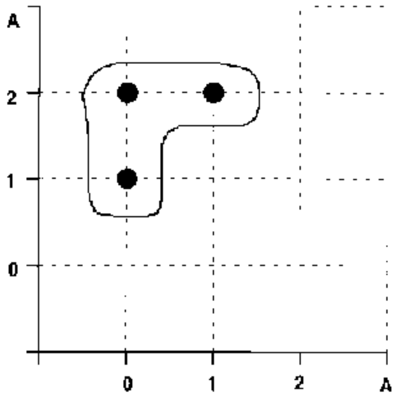


Figure 8 Relation $\Re$.

$$\text{Dom}\ \Re = \{0, 1\}$$
$$\text{Ran}\ \Re = \{1, 2\}$$

and

$$\Re^{-1} = \{(1, 0), (2, 0), (2, 1)\}$$

Consider now the relation

$$\Re_1 = \{(x, y)x, y \in \{0, 1, 2\} \text{ and } x \leq y\}$$

Then

$$\Re_1 = \{(0, 0), (0, 1), (0, 2), (1, 1), (1, 2), (2, 2)\}$$

As a subset of the Cartesian product of the set {0, 1, 2} by itself, the relation $\Re_1$ can be visualized as shown in Fig. 9.

Also,

$$\text{Dom}\ \Re_1 = \{0, 1, 2\}$$
$$\text{Ran}\ \Re_1 = \{0, 1, 2\}$$

and

$$\Re_1^{-1} = \{(0, 0), (1, 0), (1, 1), (2, 0), (2, 1), (2, 2)\}$$

Finally, consider the *identity* relation given by

$$\Re_2 = \{(x, y) : x, y \in \{0, 1, 2\} \text{ and } x = y\}$$

Then

$$\Re_2 = \{(0, 0), (1, 1), (2, 2)\}$$

Pictorially we obtain the diagonal of the graph shown in Fig. 10.

In this example, we have

$$\text{Dom}\ \Re_2 = \{0, 1, 2\}$$
$$\text{Ran}\ \Re_2 = \{0, 1, 2\}$$

and

$$\Re_2^{-1} = \{0, 0), (1, 1), (2, 2)\}$$

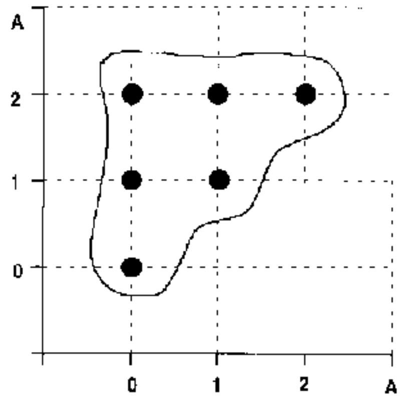


Figure 9 Relation $\Re_1$.

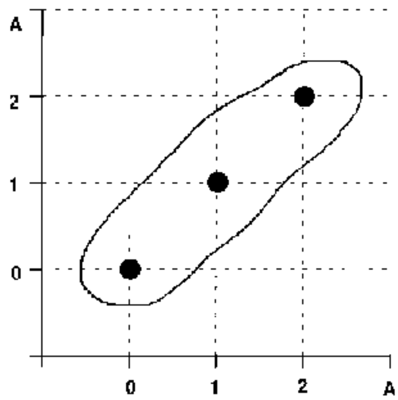


Figure 10 Relation $\Re_2$.

We close this section indicating a general procedure to combine two binary relations to obtain a new binary relation. If $\Re_1$ and $\Re_2$ are two binary relations, we define the *composition* of $\Re_1$ and $\Re_2$ by the relation

$$\Re_2 \circ \Re_1 = \{(x, z) : \text{there exists } y \text{ for which } x\Re_1 y \text{ and } y\Re_2 z\}$$

Notice that $\Re_1$ is applied first and $\Re_2$ second and also that, in general, the composition of relations is not commutative, i.e.,

$$\Re_2 \circ \Re_1 \neq \Re_1 \circ \Re_2$$

For example, let R denote the set of real numbers and consider the binary relations

$$\Re_1 = \{(x, y) : x, y \in R \text{ and } y = x^2\}$$

and

$$\Re_2 = \{(x, y) : x, y \in R \text{ and } y = x + 1\}$$

Then

$$\Re_2 \circ \Re_1 = \{(x, z) : x, z \in R \text{ and } z = x^2 + 1\}$$

and

$$\Re_1 \circ \Re_2 = \{(x, z) : x, z \in R \text{ and } z = (x + 1)^2\}$$

Note that for $\Re_2 \circ \Re_1$ we have $x\Re_1 x^2$ and $x^2\Re_2(x^2 + 1)$, while for $\Re_1 \circ \Re_2$ we have $x\Re_2(x + 1)$ and $(x + 1)\Re_1(x + 1)^2$.

In general, if $\Re_1$, $\Re_2$, and $\Re_3$ are binary relations, the following properties hold:

$$(\Re_1 \circ \Re_2) \circ \Re_3 = \Re_1 \circ (\Re_2 \circ \Re_3)$$
$$(\Re_1 \circ \Re_2)^{-1} = \Re_2^{-1} \circ \Re_1^{-1}$$

and

$$(\Re_1^{-1})^{-1} = \Re_1$$

2.2.4 Equivalence Relations

Now we concentrate our attention on the properties of a binary relation $\mathfrak{R}$ defined in a set X.

1. $\mathfrak{R}$ is called *reflexive* in X, if and only if, for all $x \in X$, $x\mathfrak{R}x$.
2. $\mathfrak{R}$ is called *symmetrical* in X, if and only if, for all $x, y \in X$, $x\mathfrak{R}y$ implies $y\mathfrak{R}x$.
3. $\mathfrak{R}$ is called *transitive* in X, if and only if, for all $x, y, z \in X$, $x\mathfrak{R}y$ and $y\mathfrak{R}z$ implies $x\mathfrak{R}z$.

A binary relation $\mathfrak{R}$ is called an *equivalence relation* on X if it is reflexive, symmetrical and transitive.

As an example, consider the set Z of integer numbers and let n be an arbitrary positive integer. The congruence relation modulo n on the set Z is defined by $x \equiv y$ (modulo n) if and only if $x - y = kn$ for some $k \in Z$. The congruence relation is an equivalence relation on Z.

Proof

1. *For each $x \in Z$, $x - x = 0n$. This means that $x \equiv x$ (modulo n) which implies that the congruence relation is reflexive.*
2. *If $x \equiv y$ (modulo n), $x - y = kn$ for some $k \in Z$. Multiplying both sides of the last equality by -1, we get $y - x = -kn$ which implies that $y \equiv x$ (modulo n). Thus, the congruence relation is symmetrical.*
3. *If $x \equiv y$ (modulo n) and $y \equiv z$ (modulo n), we have $x - y = k_1 n$ and $y - z = k_2 n$ for some k_1 and k_2 in Z. Writing $x - z = x - y + y - z$, we get $x - z = (k_1 + k_2)n$. Since $k_1 + k_2 \in Z$, we conclude that $x \equiv z$ (modulo n). This shows that the congruence relation is transitive.*

From 1–3 it follows that the congruence relation (modulo n) is an equivalence relation on the set Z of integer numbers.

In particular, we observe that if we choose $n = 2$, then $x \equiv y$ (modulo 2) means that $x - y = 2k$ for some integer k. This is equivalent to saying that either x and y are both even or both x and y are odd. In other words, any two even integers are equivalent, any two odd integers are equivalent but an even integer can not be equivalent to an odd one. The set Z has been divided into two disjoint subsets whose union gives Z. One such proper subset is the set of even integers and the other one is the set of odd integers.

2.2.4.1 Partitions and Equivalence Relations

The situation described in the last example is quite general. To study equivalence relations in more detail we need to introduce the concepts of *partition* and *equivalence class*.

Given a nonempty set X, a *partition* S of X is a collection of nonempty subsets of X such that

1. If $A, B \in S, A \neq B$, then $A \cap B = \emptyset$.
2. $\bigcup_{A \in S} A = X$.

If $\mathfrak{R}$ is an equivalence relation on a nonempty set X, for each member $x \in X$ the *equivalence class* associated with x, denoted $x/\mathfrak{R}$, is given by

$$x/\mathfrak{R} = \{z \in X : x\mathfrak{R}x\}$$

The set $x/\mathfrak{R}$ is a subset of X and, consequently, an element of the power set $P(X)$. Thus, the set

$$X/\mathfrak{R} = \{y \in P(x) : y = x/\mathfrak{R} \text{ for some } y \in X\}$$

is also a well-defined subset of $P(x)$ called the *quotient set* of X by $\mathfrak{R}$.

The correspondence between the partition of a nonempty set and the equivalence relation determined by it is established in the following propositions.

The quotient set $x/\mathfrak{R}$ of a set X by an equivalence relation $\mathfrak{R}$ is a partition of the set X.

The converse of this statement also holds; that is, each partition of X generates an equivalence relation on X. In fact, if S is a partition of a nonempty set X, we can define the relation

$$X/S = \{(x, y) \in X \times X : x \in s \text{ and } y \in s \text{ for some } s \in S\}$$

This is an equivalence relation on X, and the equivalent classes induced by it are precisely the elements of the partition S, i.e.,

$$X/(X/S) = S$$

Intuitively, equivalence relations and partitions are two different ways to describe the same collection of subsets.

2.2.5 Order Relations

Order relations constitute another common type of relations. Once again, we begin by introducing several definitions.

A binary relation $\mathfrak{R}$ in X is said to be *antisymmetrical* if for all $x, y \in X, x\mathfrak{R}y$ and $y\mathfrak{R}x$ imply $x = y$.

A binary relation $\Re$ in X is *asymmetrical* if for any $x, y \in X$, $x\Re y$ implies that $y\Re x$ does not hold. In other words, we can not have $x\Re y$ an $y\Re x$ both true.

A binary relation $\Re$ in X is a *partial ordering* of X if and only if it is reflexive, antisymmetrical, and transitive. The pair $(X, \Re)$ is called and *ordered set*.

A binary relation in X is a *strict (or total) ordering* of X if and only if it is asymmetrical and transitive.

For example, consider the set of integers

$$X = \{1, 3, 2\}$$

and the binary relation in X given by

$$\Re_1 = \{(x, y) : x, y \in X \text{ and } x \leq y\}$$

This gives explicitly

$$\Re_1 = \{(1, 1), (2, 2), (3, 3), (1, 2), (1, 3), (2, 3)\}$$

It is a simple task to check that $\Re_1$ is a partial ordering of the set X. It requires a little extra thinking to realize that now the least and the greatest elements of X have been identified.

On the same set X, the binary relation defined by

$$\Re_1 = \{(x, y) : x, y \in X \text{ and } x < y\}$$
$$=\{(1, 2), (1, 3), (2, 3)\}$$

is an example of a strict ordering of X.

It is also possible to establish a correspondence between partial orderings and strict orderings of a set:

If $\Re_1$ is a partial ordering of X, then the binary relation $\Re_2$ defined in X by

$$x\Re_2 y \text{ if and only if } x\Re_1 y \text{ and } x \neq y$$

is a strict ordering of X.

Finally, if $\Re_2$ is a strict ordering of X, then the relation $\Re_1$ defined in X by

$$x\Re_1 y \text{ if and only if } x\Re_2 y \text{ or } x = y$$

is a partial ordering of X.

GENERAL REFERENCES

1. PR Halmos. *Naive Set Theory*. New York: Van Nostrand Réinhold, 1960.
2. K Hrbacek, T Jech. *Introduction to Set Theory*. New York: Marcel Dekker, 1978.

Chapter 1.3

Linear Algebra

William C. Brown
Michigan State University, East Lansing, Michigan

3.1 MATRICES

3.1.1 Shapes and Sizes

Throughout this chapter, F will denote a field. The four most commonly used fields in linear algebra are $\mathbb{Q}$ = rationals, $\mathbb{R}$ = reals, $\mathbb{C}$ = complex numbers and $\mathbb{Z}_p$ = the integers modulo a prime p. We will also let $\mathbb{N} = \{1, 2, \ldots\}$, the set of natural numbers.

Definition 1. *Let $m, n \in \mathbb{N}$. An $m \times n$ matrix A with entries from F is a rectangular array of m rows and n columns of numbers from F.*

The most common notation used to represent an $m \times n$ (read "m by n") matrix A is displayed in Eq. (1):

$$A = \begin{pmatrix} a_{11}, & a_{12} & ,\ldots, & a_{1n} \\ \vdots & & & \vdots \\ a_{m1}, & a_{m2} & ,\ldots, & a_{mn} \end{pmatrix} \tag{1}$$

If A is the $m \times n$ matrix displayed in Eq. (1), then the field elements a_{ij} $(i = 1, \ldots, m; j = 1, \ldots, n)$ are called the entries of A. We will also use $[A]_{ij}$ to denote the i, jth entry of A. Thus, $a_{ij} = [A]_{ij}$ is the element of F which lies in the ith row and jth column of A. By the size of A, we will mean the expression $m \times n$. Thus, size $(A) = m \times n$ if A has m rows and n columns. Notice that the size of a matrix is a pair of positive integers with a "×" put between them. Negative numbers and zero are not allowed to appear in the size of a matrix.

Definition 2. *The set of all $m \times n$ matrices with entries from F will be denoted by $M_{m\times n}(F)$.*

Matrices of various shapes are given special names in linear algebra. Here is a brief list of some of the more famous shapes and some pictures to illustrate the definitions.

1. A matrix is square if $m = n$.

$$(a), \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}, \ldots$$

$$\text{size} = 1 \times 1, \quad 2 \times 2, \qquad 3 \times 3, \ldots \tag{2a}$$

2. An $m \times n$ matrix is called a column vctor if $n = 1$.

$$(a), \quad \begin{pmatrix} a \\ b \end{pmatrix}, \ldots \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$$

$$\text{size} = 1 \times 1, 2 \times 1, \ldots, n \times 1 \tag{2b}$$

3. An $m \times n$ matrix is called a row vector if $m = 1$.

$$(a), \quad (a, b), \ldots, (a_1, \ldots, a_n)$$

$$\text{size} = 1 \times 1, 1 \times 2, \ldots, \quad 1 \times n \tag{2c}$$

4. An $m \times n$ matrix A is upper triangular if $[A]_{ij} = 0$ whenever $i > j$.

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1m} & \dots & a_{1n} \\ 0 & a_{22} & a_{23} & \dots & a_{2m} & \dots & a_{2n} \\ 0 & 0 & a_{33} & \dots & a_{3m} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & a_{mm} & \dots & a_{mn} \end{pmatrix}$$

$$\text{if } m \le n \tag{2d}$$

5. An $m \times n$ matrix A is lower triangular if $[A]_{ij} = 0$ whenever $i < j$.

$$A = \begin{pmatrix} a_{11} & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ a_{21} & a_{22} & 0 & \dots & 0 & 0 & \dots & 0 \\ a_{31} & a_{32} & a_{33} & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mm} & 0 & \dots & 0 \end{pmatrix}$$

$$\text{if } m \le n \tag{2e}$$

6. An $m \times n$ matrix A is diagonal if $[A]_{ij} = 0$ whenever $i \neq j$.

$$\begin{pmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & a_{nn} \end{pmatrix} \quad \text{if } m = n \tag{2f}$$

7. A square matrix is symmetric (skew-symmetric) if $[A]_{ij} = [A]_{ji}(-[A]_{ji})$ for all $i, j = 1, \dots, n$.

$$(a), \begin{pmatrix} a & b \\ b & c \end{pmatrix}, \begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix}, \dots \quad \text{symmetric} \tag{2g}$$

size $= 1 \times 1, 2 \times 2, 3 \times 3$

$$(0), \begin{pmatrix} 0 & b \\ -b & 0 \end{pmatrix}, \begin{pmatrix} 0 & b & c \\ -b & 0 & e \\ -c & -e & 0 \end{pmatrix}, \dots \tag{2h}$$

skew-symmetric

Definition 3. *A submatrix of A is a matrix obtained from A by deleting certain rows and/or columns of A. A partition of A is a series of horizontal and vertical lines drawn in A which divide A into various submatrices.*

Example 1. *Suppose* $A = \begin{pmatrix} x & y \\ z & w \end{pmatrix}$. *Then*

$$(x), (y), (z), (w), \begin{pmatrix} x \\ z \end{pmatrix}, \begin{pmatrix} y \\ w \end{pmatrix}, (x, y), (z, w), \begin{pmatrix} x & y \\ z & w \end{pmatrix} \tag{3}$$

is a complete list of the submatrices of A.

$$\left(\begin{array}{c|c} x & y \\ z & w \end{array}\right), \left(\begin{array}{cc} x & y \\ \hline z & w \end{array}\right), \left(\begin{array}{c|c} x & z \\ \hline y & w \end{array}\right), \begin{pmatrix} x & y \\ z & w \end{pmatrix} \tag{4}$$

are all partitions of A.

The most important partitions of a matrix A are its column and row partitions.

Definition 4. *Let*

$$A = \begin{pmatrix} a_{11}, & \dots, & a_{1n} \\ \vdots & & \vdots \\ a_{m1}, & \dots, & a_{mn} \end{pmatrix} \in M_{m \times n}(F)$$

1. *For each $j = 1, \dots, n$, the $m \times 1$ submatrix*

$$\mathrm{Col}_j(A) = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{pmatrix}$$

of A is called the jth column of A.

2. *$A = (\mathrm{Col}_1(A) \mid \mathrm{Col}_2(A) \mid \dots \mid \mathrm{Col}_n(A))$ is called the column partition of A.*
3. *For each $i = 1, \dots, m$, the $1 \times n$ submatrix $Row_i(A) = (a_{i1}, \dots, a_{in})$ of A is called the ith row of A.*
4.

$$A = \left(\begin{array}{c} \mathrm{Row}_1(A) \\ \hline \vdots \\ \hline \mathrm{Row}_m(A) \end{array}\right)$$

is called the row partition of A.

We will cut down on the amount of space required to show a column or row partition by employing the following notation. In Definition 4, let $\xi_j = \mathrm{Col}_j(A)$ for $j = 1, \dots, n$ and let $\alpha_i = \mathrm{Row}_i(A)$ for $i = 1, \dots, m$. Then the column partition of A will be written $A = (\xi_1 \mid \xi_2 \mid \dots \mid \xi_n)$ and the row partition of A will be written $A = (\alpha_1; \alpha_2; \dots; \alpha_m)$.

3.1.2 Matrix Arithmetic

Definition 5. *Two matrices A and B with entries from F are said to be equal if size $(A) =$ size (B) and $[A]_{ij} = [B]_{ij}$ for all $i = 1, \dots, m$; $j = 1, \dots, n$. Here $m \times n =$ size (A).*

If A and B are equal, we will write $A = B$. Notice that two matrices which are equal have the same size. Thus, the 1×1 matrix (0) is not equal to the 1×2 matrix $(0,0)$. Matrix addition, scalar multiplication, and multiplication of matrices are defined as follows.

Definition 6

1. *Let A, $B \in M_{m\times n}(F)$. Then $A + B$ is the $m \times n$ matrix whose i,jth entry is given by $[A + B]_{ij} = [A]_{ij} + [B]_{ij}$ for all $i = 1, \dots, m$ and $j = 1, \dots, n$.*
2. *If $A \in M_{m\times n}(F)$ and $x \in F$, then xA is the $m \times n$ matrix whose i,jth entry is given by $[xA]_{ij} = x[A]_{ij}$ for all $i = 1, \dots, m$ and $j = 1, \dots, n$.*
3. *Let $A \in M_{m\times n}(F)$ and $C \in M_{n\times p}(F)$. Then AC is the $m \times p$ matrix whose i,jth entry is given by*

$$[AC]_{ij} = \sum_{k=1}^{n} [A]_{ik}[C]_{kj} \qquad \text{for } i = 1, \dots, m; j = 1, \dots, p$$

Example 2. *Let*

$$A = \begin{pmatrix} 1 & 0 & 3 \\ 1 & 1 & 2 \end{pmatrix} \quad B = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 2 \end{pmatrix} \in M_{2\times 3}(\mathbb{Q})$$

and let

$$C = \begin{pmatrix} 0 & 2 \\ -1 & 1 \\ 4 & 0 \end{pmatrix} \in M_{3\times 2}(\mathbb{Q})$$

Then

$$A + B = \begin{pmatrix} 2 & 0 & 4 \\ 2 & 1 & 4 \end{pmatrix} \quad 6A = \begin{pmatrix} 6 & 0 & 18 \\ 6 & 6 & 12 \end{pmatrix} \quad AC = \begin{pmatrix} 12 & 2 \\ 7 & 3 \end{pmatrix} \tag{5}$$

Notice that addition is defined only for matrices of the same size. Multiplication is defined only when the number of columns of the first matrix is equal to the number of rows of the second matrix.

The rules for matrix addition and scalar multiplication are summarized in the following theorem:

Theorem 1. *Let A, B, $C \in M_{m\times n}(F)$. Let x, $y \in F$. Then*

1. $A + B = B + A$.
2. $(A + B) + C = A + (B + C)$.
3. $A + 0 = A$.
4. $A + (-1)A = 0$.
5. $(xy)A = x(yA)$.
6. $x(A + B) = xA + xB$.
7. $(x + y)A = xA + yA$.
8. $1A = A$.

When no explicit reference is given, a proof of the quoted theorem can be found in Brown [1]. The number zero appearing in 3 and 4 above denotes the $m \times n$ matrix all of whose entries are zero. The eight statements given in Theorem 1 imply $M_{m\times n}(F)$ is a vector space over F (see definition 16 in Sec. 3.2) when vector addition and scalar multiplication are given as in 1 and 2 of Definition 6.

Theorem 1(2) implies matrix addition is associative. It follows from this statement that expressions of the form $x_1A_1 + \cdots + x_rA_r$ ($A_i \in M_{m\times n}(F)$ and $x_i \in F$) can be used unambiguously. Any placement of parentheses in this expression will result in the same answer. The sum $x_1A_1 + \cdots + x_rA_r$ is called a linear combination of $A_1, \dots, A_r$. The numbers $x_1, \dots, x_r$ are called the scalars of the linear combination.

Example 3. *Let*

$$A = \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix} \quad B = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \quad C = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$$

We view A, B, $C \in M_{2\times 2}(\mathbb{Z}_3)$. Then

$$2A + B + 2C = \begin{pmatrix} 0 & 0 \\ 1 & 2 \end{pmatrix} \tag{6}$$

The rules for matrix multiplication are as follows:

Theorem 2. *Let A, $D \in M_{m\times n}(F)$, $B \in M_{n\times p}(F)$, $C \in M_{p\times q}(F)$ and $E \in M_{r\times m}(F)$. Let $x \in F$. Then*

1. $(AB)C = A(BC)$.
2. $(A + D)B = AB + DB$.
3. $E(A + D) = EA + ED$.
4. $0A = 0$ *and* $A0 = 0$.
5. $I_mA = A$ *and* $AI_n = A$.
6. $x(AB) = (xA)B = A(xB)$.

In Theorem 2(4), the zero denotes the zero matrix of various sizes. In Theorem 2(5), I_n denotes the $n \times n$ identity matrix. This is the diagonal matrix given by $[I_n]_{jj} = 1$ for all $j = 1, \dots, n$. Theorem 2 implies $M_{n\times n}(F)$ is an associative algebra with identity [2, p. 36] over the field F.

Consider the following system of m equations in unknowns $x_1, \dots, x_n$:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ \vdots \qquad\qquad\qquad & \quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned} \tag{7}$$

In Eq. (7), the a_{ij}'s and b_i's are constants in F. Set

$$A = \begin{pmatrix} a_{11}, & \ldots, & a_{1n} \\ \vdots & & \vdots \\ a_{m1}, & \ldots, & a_{mn} \end{pmatrix} \quad B = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \; X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \tag{8}$$

Using matrix multiplication, the system of linear equations in Eq. (7) can be written succinctly as

$$AX = B \tag{9}$$

We will let F^n denote the set of all column vectors of size n. Thus, $F^n = M_{n\times 1}(F)$. A column vector $\xi \in F^n$ is called a solution to Eq. (9) if $A\xi = B$. The $m \times n$ matrix $A = (a_{ij}) \in M_{m\times n}(F)$ is called the coefficient matrix of Eq. (7). The partitioned matrix $(A \mid B) \in M_{m\times(n+1)}(F)$ is called the augmented matrix of Eq. (7). Matrix multiplication was invented to handle linear substitutions of variables in Eq. (7). Suppose $y_1, \ldots, y_p$ are new variables which are related to $x_1, \ldots, x_n$ by the following set of linear equations:

$$\begin{aligned} x_1 &= c_{11}y_1 + \cdots + c_{1p}y_p \\ &\;\vdots \qquad\qquad\qquad\qquad \text{(here } c_{uv} \in F \text{ for all } u, v) \\ x_n &= c_{n1}y_1 + \cdots + c_{np}y_p \end{aligned} \tag{10}$$

Set

$$C = \begin{pmatrix} c_{11}, & \ldots, & c_{1p} \\ \vdots & & \vdots \\ c_{n1}, & \ldots, & c_{np} \end{pmatrix} \in M_{n\times p}(F)$$

Substituting the expressions in Eq. (10) into Eq. (7) produces m equations in $y_1, \ldots, y_p$. The coefficient matrix of the new system is AC, the matrix product of A and C.

Definition 7. *A square matrix $A \in M_{n\times n}(F)$ is said to be invertible (or nonsingular) if there exists a square matrix $B \in M_{n\times n}(F)$ such that $AB = BA = I_n$.*

If $A \in M_{n\times n}(F)$ is invertible and $AB = BA = I_n$ for some $B \in M_{n\times n}(F)$, then B is unique and will be denoted by A^{-1}. A^{-1} is called the inverse of A.

Example 4. *Let*

$$A = \begin{pmatrix} x & y \\ z & w \end{pmatrix} \in M_{2\times 2}(F)$$

and assume $\Delta = xw - yz \neq 0$. Then A is invertible with inverse given by

$$A^{-1} = \begin{pmatrix} w/\Delta & -y/\Delta \\ -z/\Delta & x/\Delta \end{pmatrix} \tag{11}$$

If $m = n$ in Eq. (7) and the coefficient matrix A is invertible, then $AX = B$ has the unique solution $A^{-1}b$.

Definition 8. *Let $A \in M_{m\times n}(F)$. The transpose of A is denoted by A^t. A^t is the $n \times m$ matrix whose entries are given by $[A^t]_{ij} = [A]_{ji}$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$.*

A square matrix is symmetric (skew-symmetric) if and only if $A = A^t(-A^t)$. When the field $F = \mathbb{C}$, the complex numbers, the Hermitian conjugate (or conjugate transpose) of A is more useful than the transpose.

Definition 9. *Let $A \in M_{m\times n}(\mathbb{C})$. The Hermitian conjugate of A is denoted by A^*. A^* is the $n \times m$ matrix whose entries are given by $[A^*]_{ij} = [\bar{A}]_{ji}$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$.*

In Definition 9, the bar over $[A]_{ji}$ denotes the conjugate of the complex number $[A]_{ji}$. For example,

$$\begin{aligned} &\begin{pmatrix} 1+i & 2 \\ 2-i & i \end{pmatrix}^* = \begin{pmatrix} 1-i & 2+i \\ 2 & -i \end{pmatrix} \text{ and} \\ &\begin{pmatrix} 1+i & 2 \\ 2-i & i \end{pmatrix}^t = \begin{pmatrix} 1+i & 2-i \\ 2 & i \end{pmatrix} \end{aligned} \tag{12}$$

The following facts about transposes and Hermitian conjugates are easy to prove.

Theorem 3. *Let A, $C \in M_{m\times n}(F)$ and $B \in M_{n\times p}(F)$. Then*

1. $(A + C)^t = A^t + C^t$.
2. $(AB)^t = B^t A^t$.
3. $(A^t)^t = A$.
4. *If $m = n$ and A is invertible, then A^t is also invertible. In this case, $(A^t)^{-1} = (A^{-1})^t$.*

If $F = \mathbb{C}$, then we also have

5. $(A + C)^* = A^* + C^*$.
6. $(AB)^* = B^* A^*$.
7. $(A^*)^* = A$.
8. *If A is invertible, so is A^* and $(A^*)^{-1} = (A^{-1})^*$.*

3.1.3 Block Multiplication of Matrices

Theorem 4. *Let $A \in M_{m\times n}(F)$ and $B \in M_{n\times p}(F)$. Suppose*

$$A = \left(\begin{array}{c|c|c} A_{11} & \dots & A_{1k} \\ \hline \vdots & & \vdots \\ \hline A_{r1} & \dots & A_{rk} \end{array}\right) \quad \text{and}$$

$$B = \left(\begin{array}{c|c|c} B_{11} & \dots & B_{1t} \\ \hline \vdots & & \vdots \\ \hline B_{k1} & \dots & B_{kt} \end{array}\right)$$

are partitions of A and B such that size $(A_{ij}) = m_i \times n_j$ and size $(B_{jl}) = n_j \times p_l$. Thus, $m_1 + \cdots + m_r = m$, $n_1 + \cdots + n_k = n$, and $p_1 + \cdots + p_t = p$. For each $i = 1, \dots, r$ and $j = 1, \dots, t$, set

$$C_{ij} = \sum_{q=1}^{k} A_{iq}B_{qj} \qquad \text{(multiplication of blocks)}$$

Then

$$AB = \left(\begin{array}{c|c|c} C_{11} & \dots & C_{1t} \\ \hline \vdots & & \vdots \\ \hline C_{r1} & \dots & C_{rk} \end{array}\right)$$

Notice that the only hypothesis in Theorem 4 is that every vertical line drawn in A must be matched with the corresponding horizontal line in B. There are four special cases of Theorem 4 which are very useful. We collect these in the next theorem.

Theorem 5. *Let $A \in M_{m\times n}(F)$.*

1. *If $\xi = (x_1, \dots, x_n)^t \in F^n$, then $A\xi = \sum_{i=1}^{n} x_i$ $Col_i(A)$.*
2. *If $B = (\xi_1 \mid \dots \mid \xi_p) \in M_{n\times p}(F)$, then $AB = (A\xi_1 \mid \dots \mid A\xi_p)$.*
3. *If $\lambda = (y_1, \dots, y_m) \in M_{1\times m}(F)$, then $\lambda A = \sum_{i=1}^{m} y_i\,\mathrm{Row}_i(A)$.*
4. *If $C = (\lambda_1; \dots; \lambda_r) \in M_{r\times m}(F)$, $CA = (\lambda_1 A; \dots; \lambda_r, A)$.*

Definition 10. *Let $A \in M_{m\times n}(F)$.*

1. *$CS(A) = \{A\xi \mid \xi \in F^n\}$ is called the column space of A.*
2. *$RS(A) = \{\lambda A \mid \lambda \in M_{1\times m}(F)\}$ is called the row space of A.*

Theorem 5 implies that the column space of A consists of all linear combinations of the columns of A. $RS(A)$ is all linear combinations of the rows of A. Using all four parts of Theorem 5, we have

$$\begin{aligned} &CS(AB) \subseteq CS(A) \\ &RS(AB) \subseteq RS(B) \end{aligned} \tag{13}$$

The column space of A is particularly important for the theory of linear equations. Suppose $A \in M_{m\times n}(F)$ and $B \in F^m$. Theorem 5 implies that $AX = B$ has a solution if and only if $B \in CS(A)$.

3.1.4 Gaussian Elimination

The three elementary row operations that can be performed on a given matrix A are as follows:

(α) Interchange two rows of A
(β) Add a scalar times one row of A to another row of A
(δ) Multiply a row of A by a nonzero scalar
(14)

There are three corresponding elementary column operations which can be preformed on A as well.

Definition 11. *Let $A_1, A_2 \in M_{m\times n}(F)$. A_1 and A_2 are said to be row (column) equivalent if A_2 can be obtained from A_1 by applying finitely many elementary row (column) operations to A_1.*

If A_1 and A_2 are row (column) equivalent, we will write $A_1 \tilde{r} A_2$ ($A_1 \tilde{c} A_2$). Either one of these relations is an equivalence relation on $M_{m\times n}(F)$. By this, we mean

$$\begin{array}{ll} A_1 \tilde{r} A_1 & (\tilde{r} \text{ is reflexive}) \\ A_1 \tilde{r} A_2 \Leftrightarrow A_2 \tilde{r} A_1 & (\tilde{r} \text{ is symmetric}) \\ A_1 \tilde{r} A_2, A_2 \tilde{r} A_3 \Rightarrow A_1 \tilde{r} A_3 & (\tilde{r} \text{ is transitive}) \end{array} \tag{15}$$

Theorem 6. *Let A, $C \in M_{m\times n}(F)$ and B, $D \in F^m$. Suppose $(A \mid B)\, \tilde{r}\, (C \mid D)$. Then the two linear systems of equations $AX = B$ and $CX = D$ have precisely the same solutions.*

Gaussian elimination is a strategy for solving a system of linear equations. To find all solutions to the linear system of equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned} \tag{16}$$

carry out the following three steps.

1. Set up the augmented matrix of Eq. (16):

$$(A \mid B) = \left(\begin{array}{ccc|c} a_{11}, & \dots, & a_{1n} & b_1 \\ \vdots & & \vdots & \vdots \\ a_{m1}, & \dots, & a_{mn} & b_m \end{array}\right)$$

2. Apply elementary row operations to $(A \mid B)$ to obtain a matrix $(C \mid D)$ in upper triangular form.
3. Solve $CX = D$ by back substitution.

By Theorem 6, this algorithm yields a complete set of solutions to $AX = B$.

Example 5. *Solve*

$$\begin{aligned} 2x+3y+4z&=10 \\ x-y-z&=2 \qquad (F=\mathbb{Q}) \qquad\qquad (*) \\ x+z&=3 \end{aligned}$$

Following steps 1–3, we have

1.

$$\left(\begin{array}{ccc|c} 2 & 3 & 4 & 10 \\ 1 & -1 & -1 & 2 \\ 1 & 0 & 1 & 3 \end{array}\right)$$

is the augmented matrix of $(*)$

2.

$$\left(\begin{array}{ccc|c} 2 & 3 & 4 & 10 \\ 1 & -1 & -1 & 2 \\ 1 & 0 & 1 & 3 \end{array}\right) \underset{(\alpha)}{\longrightarrow} \left(\begin{array}{ccc|c} 1 & 0 & 1 & 3 \\ 1 & -1 & -1 & 2 \\ 2 & 3 & 4 & 10 \end{array}\right)$$

$$\underset{(\beta)}{\longrightarrow} \left(\begin{array}{ccc|c} 1 & 0 & 1 & 3 \\ 0 & -1 & -2 & -1 \\ 0 & 3 & 2 & 4 \end{array}\right) \underset{(\delta)}{\longrightarrow}$$

$$\left(\begin{array}{ccc|c} 1 & 0 & 1 & 3 \\ 0 & 1 & 2 & 1 \\ 0 & 3 & 2 & 4 \end{array}\right) \underset{(\beta)}{\longrightarrow} \left(\begin{array}{ccc|c} 1 & 0 & 1 & 3 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & -4 & 1 \end{array}\right)$$

(upper triangular).

The letters below the arrows indicate which type of elementary row operation was used on the matrix on the left to get the matrix on the right.

3. Solve

$$\begin{aligned} x+z&=3 & && x&=13/4 \\ y+2z&=1 \Rightarrow & & y=3/2 & \\ -4z&=1 & z=-1/4 & & \end{aligned}$$

Thus, $x = 13/4$, $y = 3/2$ and $z = -1/4$ is the (unique) solution to $(*)$.

Definition 12. *Let $A \in M_{m\times n}(F)$. A system of equations of the form $AX = 0$ is called a homogeneous system of equations. A nonzero, column vector $\xi \in F^n$ is called a nontrivial solution of $AX = 0$ if $A\xi = 0$.*

Using Gaussian elimination, we can prove the following theorem.

Theorem 7. *Let $A \in M_{m\times n}(F)$.*

1. *The homogeneous system of equations $AX = 0$ has a nontrivial solution if $m < n$.*
2. *Suppose $m = n$. The homogeneous system of equations $AX = 0$ has only $X = 0$ as a solution if and only if $A \,\tilde{r}\, I_n$.*

3.1.5 Elementary Matrices and Inverses

Definition 13. *An elementary matrix (of size $m \times m$) is a matrix obtained from I_m by performing a single elementary row operation on I_m.*

Pictures of the three types of elementary matrices are as follows:

1. E_{ij} will denote the matrix obtained from I_m by interchanging rows i and j. These matrices are called transpositions.

$$E_{ij}=\left[\begin{array}{ccccccccccc}
1&&&&&&&&&&\\
&\ddots&&&&&&&&&\\
&&1&&&&&&&&\\
&&&0&\cdots&\cdots&0&1&&&\\
&&&\vdots&1&&&&&&\\
&&&\vdots&&\vdots&&\vdots&&&\\
&&&0&&&1&\vdots&&&\\
&&&1&0&\vdots&\vdots&0&&&\\
&&&&&&&&1&&\\
&&&&&&&&&\ddots&\\
&&&&&&&&&&1
\end{array}\right]\begin{array}{l}\\ \\ \\ i\\ \\ \\ \\ j\\ \\ \\ \\ \end{array}$$

$$\qquad\qquad i \qquad\qquad\qquad j \qquad\qquad (17a)$$

2. $E_{ij}(c)$ will denote the matrix obtained from I_m by adding c times row j of I_m to row i of I_m.

$$E_{ij}(c) = \begin{array}{c} \left[\begin{array}{ccccccccccc} 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & & & & & & & & & & \vdots \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & c & 0 & \cdots & 0 \\ \vdots & & & & & & & & & & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 1 & 0 & \cdots & 0 \\ \vdots & & & & & & & & & & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 1 \end{array}\right] \begin{array}{c} \\ \\ i \\ \\ j \\ \\ \\ \end{array} \\ \begin{array}{ccccccccccc} & & & & & & & j & & & \end{array} \end{array} \tag{17b}$$

(here $i < j$)

3. $E_i(C)$ will denote the elementary matrix obtained from I_m by multiplying the ith row of I_m by $c \neq 0$.

$$E_i(c) = \left[\begin{array}{cccccc} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & & & & & \vdots \\ 0 & \cdots & c & \cdots & & 0 \\ \vdots & & & & & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \end{array}\right] \begin{array}{c} \\ \\ i \\ \\ \\ \end{array} \tag{17c}$$

Each elementary matrix is invertible. Multiplying on the left by elementary matrices performs row operations on A.

Theorem 8

1. $E_{ij}^{-1} = E_{ij}$.
2. $E_{ij}(c)^{-1} = E_{ij}(-c)$.
3. $E_i(c)^{-1} = E_i(1/c)$.
4. *For any $A \in M_{m\times n}(F)$,*
   a. *$E_{ij}A$ is the $m \times n$ matrix obtained from A by interchanging rows i and j of A.*
   b. *$E_{ij}(c)A$ is the $m \times n$ matrix obtained from A by adding c times row j of A to row i of A.*
   c. *$E_i(c)A$ is the $m \times n$ matrix obtained from A by multiplying row i of A by c.*

Thus, two $m \times n$ matrices A and B are row equivalent if and only if there exist a finite number of elementary matrices $E_1, \ldots, E_k$ such that $E_kE_{k-1}\cdots E_2E_1A = B$.

Example 6. *In Example 5, we showed that*

$$(A \mid B) = \left(\begin{array}{ccc|c} 2 & 3 & 4 & 10 \\ 1 & -1 & -1 & 2 \\ 1 & 0 & 1 & 3 \end{array}\right) \tilde{r} \left(\begin{array}{ccc|c} 1 & 0 & 1 & 3 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & -4 & 1 \end{array}\right)$$
$$= (C \mid D).$$

The sequence of elementary matrices used there are as follows:

$$E_{32}(-3)E_2(-1)E_{31}(-2)E_{21}(-1)E_{13}(A \mid B) = (C \mid D)$$

We can also multiply a given matrix A on the right by elementary matrices. Multiplying on the right performs elementary column operations on A.

Theorem 9. *Let $A \in M_{m\times n}(F)$ be a nonzero matrix. Then there exist elementary matrices $E_1, \ldots, E_k$ and $E'_1, \ldots, E'_\ell$ such that*

$$E_k \cdots E_2E_1AE'_1E'_2 \cdots E'_\ell = \left(\begin{array}{c|c} I_t & 0 \\ \hline 0 & 0 \end{array}\right) \tag{18}$$

The positive integer t appearing in Theorem 9 is called the rank of A.

Elementary matrices can be used to characterize invertible matrices.

Theorem 10. *Let $A \in M_{n\times n}(F)$. Then the following conditions are equivalent:*

1. *A is invertible.*
2. *A has a left inverse, i.e., $BA = I_n$ for some $B \in M_{n\times n}(F)$.*
3. *A has a right inverse, i.e., $AC = I_n$ for some $C \in M_{n\times n}(F)$.*
4. *The homogeneous system of equations $AX = 0$ has no nontrivial solution.*
5. *A is a finite product of elementary matrices.*

The proof of Theorem 10 incorporates an algorithm for computing A^{-1}, which is effective if n is small and the entries of A are reasonable:

Suppose A is invertible.

1. Form the $n \times 2n$ partitioned matrix $(A \mid I_n)$.
2. Apply row operations to $(A \mid I_n)$ so that A on the left in $(A \mid I_n)$ is changed to I_n. Then I_n on the right will change to A^{-1}.

In symbols,

$$(A \mid I_n)\, \tilde{r}\, (I_n \mid A^{-1}) \tag{19}$$

See Brown [1; Ex. 5.15, Chap. I] for a concrete example.

One nice application of this material involves the Vandermonde matrix:

$$V = \begin{pmatrix} 1 & a_0 & a_0^2 & \cdots & a_0^n \\ 1 & a_1 & a_1^2 & & a_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & a_n & a_n^2 & & a_n^n \end{pmatrix} \in M_{(n+1)\times(n+1)}(\mathbb{R}) \tag{20}$$

This matrix is invertible if $a_0, a_1, \dots, a_n$ are distinct real numbers. V is used to prove the following interpolation theorem:

Theorem 11. *Suppose $a_0, \dots, a_n$ are $n+1$ distinct real numbers. Let $b_0, \dots, b_n \in \mathbb{R}$. There exists a polynomial (with real coefficients) $p(t)$ such that the degree of $p(t)$ is at most n and $p(a_i) = b_i$ for all $i = 0, 1, \dots, n$.*

3.1.6 *LU*-Factorizations

LU-Factorizations are refinements of Gaussian elimination.

Definition 14. *Let $A \in M_{n\times n}(F)$. A has an LU-factorization if there exists a unit, lower triangular matrix*

$$L = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \ell_{21} & 1 & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \dots & 1 \end{pmatrix} \in M_{n\times n}(F)$$

and an upper triangular matrix

$$U = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ 0 & u_{22} & \dots & u_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots 0 & u_{nn} \end{pmatrix} \in M_{n\times n}(F)$$

such that $A = LU$.

A given matrix $A \in M_{n\times n}(F)$ may not have an LU-factorization. The deciding factor is whether any transpositions are needed to row reduce A to an upper triangular matrix.

Definition 15. *A square matrix of the form*

$$L = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & c_{i+1} & 1 & & \\ & & \vdots & & \ddots & \\ & & c_n & \dots & \dots & 1 \end{pmatrix}_{n\times n}$$

is called a Frobenius matrix (L has ones on its diagonal, constants $c_{i+1}, \dots, c_n$ in positions $(i+1, i) \dots (n, i)$ and zeros elsewhere).

Frobenius matrices are invertible. The inverse of L is obtained from L by changing the signs of $c_{i+1}, \dots, c_n$. If R_i denotes the ith row of A and L is the Frobenius matrix pictured above, then

$$LA = \left(\begin{array}{c} R_1 \\ \hline \vdots \\ \hline R_i \\ \hline R_{i+1} + c_{i+1}R_i \\ \hline \vdots \\ \hline R_n + c_nR_i \end{array}\right) \tag{21}$$

Thus, multiplying A by L on the left performs elementary row operations of type (β) on A below row i. Hence, the process which row reduces A to an upper triangular matrix can be stated as follows:

Theorem 12. *Let $A \in M_{n\times n}(F)$. There exist a finite number of Frobenius matrices $L_1, \dots, L_k$ and a finite number of transpositions $E_1, \dots, E_k$ such that $L_kE_k \cdots L_1E_1A = U$ an upper triangular matrix.*

Example 7. *Suppose*

$$A = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 0 & 1 \end{pmatrix} \in M_{3\times 3}(\mathbb{Q})$$

Then

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 4 & 1 \end{pmatrix}}_{L_2}\underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{E_2}\underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix}}_{L_1} \underbrace{\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{E_1}\underbrace{\begin{pmatrix} 0 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 0 & 1 \end{pmatrix}}_{A} = \underbrace{\begin{pmatrix} 1 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 7 \end{pmatrix}}_{U} \tag{22}$$

If no transpositions are required in Theorem 12, i.e., $E_i = I_n$ for all $i = 1, \dots, k$, then $L_kL_{k-1} \dots L_1A = U$. Consequently, $A = (L_1^{-1}L_2^{-1} \dots L_k^{-1})U$. It is easy to see that $L = L_1^{-1}L_2^{1} \dots L_k^{-1}$ is a unit, lower triangular matrix and hence A has an LU-factorization. This proves part of our next theorem.

Theorem 13. *Let $A \in M_{n\times n}(F)$.*

1. *A has an LU-factorization if and only if no row interchanges are required in the Gaussian reduction of A to upper triangular form.*
2. *Suppose A has an LU-factorization. If A is invertible, then the factorization is unique.*
3. *For any $A \in M_{n\times n}(F)$, there exists a permutation matrix P, i.e., a finite product of transposition, such that PA has an LU-factorization.*

There are several important applications of LU-factorizations. We will give one application now and another in the section on determinants. LU-Factorizations are used to solve systems of equations. To solve $AX = B$:

1. Find an LU-factorization $PA = LU$.
2. Replace $AX = B$ with $(PA)X = PB$, i.e., $L(UX) = PB$.
3. Solve $LY = PB$ by forward substitution.
4. Solve $UX = Y$ by back substitution. (23)

Thus, replacing A by an LU-factorization converts $AX = B$ into two simpler problems $LY = PB$ and $UX = Y$. These last two systems are lower triangular and upper triangular respectively. These systems are usually easier to solve than the original.

3.2 VECTOR SPACES

3.2.1 Definitions and Examples

Definition 16. *A vector space over F is a nonempty set V together with two functions $(\alpha, \beta) \to \alpha + \beta$ from $V \times V \to V$ and $(x, \alpha) \to x\alpha$ from $F \times V \to V$ which satisfy the following conditions:*

V1. $\alpha + \beta = \beta + \alpha$ *for all* $\alpha, \beta \in V$.
V2. $\alpha + (\beta + \delta) = (\alpha + \beta) + \delta$ *for all* $\alpha, \beta, \delta \in V$.
V3. *There exists an element* $0 \in V$ *such that* $\alpha + 0 = \alpha$ *for all* $\alpha \in V$.
V4. *For any* $\alpha \in V$, *there exists a* $\beta \in V$ *such that* $\alpha + \beta = 0$.
V5. $(xy)\alpha = x(y\alpha)$ *for all* $\alpha \in V$ *and* $x, y \in F$.
V6. $x(\alpha + \beta) = x\alpha + x\beta$ for all $\alpha, \beta \in V$ and $x \in F$.
V7. $(x + y)\alpha = x\alpha + y\alpha$ *for all* $\alpha \in V$ *and* $x, y \in F$.
V8. $1\alpha = \alpha$ *for all* $\alpha \in V$.

Suppose $(V, (\alpha, \beta) \to \alpha + \beta, (x, \alpha) \to x\alpha)$ is a vector space over F. The elements in V are called vectors and will usually be denoted by Greek letters. The elements in F are called scalars and will be represented by lowercase English letters. The function $(\alpha, \beta) \to \alpha + \beta$ is called vector addition. The function $(x, \alpha) \to x\alpha$ is called scalar multiplication. Notice that a vector space is actually an ordered triple consisting of a set of vectors, the function vector addition and the function scalar multiplication. It is possible that a given set V can be made into a vector space over F in many different ways by specifying different vector additions or scalar multiplications on V. Thus, when defining a vector space, all three pieces of information (the vectors, vector addition and scalar multiplication) must be given.

Suppose $(V, (\alpha, \beta) \to \alpha + \beta, (x, \alpha) \to x\alpha)$ is a vector space over F. When the two functions $(\alpha, \beta) \to \alpha + \beta$, $(x, \alpha) \to x\alpha$ are understood from the context or when it is not important to know the exact forms of these functions, we will drop them from our notation and simply call the vector space V. Axioms V1 and V2 say that vector addition is commutative and associative. There is only one vector $0 \in V$ which satisfies V3. This vector is called the zero vector of V. If $\alpha \in V$, there is only one vector $\beta \in V$ such that $\alpha + \beta = 0$. The vector β is called the inverse of α and written $-\alpha$. The following facts about addition and scalar multiplication are true in any vector space.

Theorem 14. *Let V be a vector space over F. Then*

1. *Any parentheses placed in* $\alpha_1 + \cdots + \alpha_n$ *result in the same vector.*
2. $x0 = 0$ *for all* $x \in F$.
3. $0\alpha = 0$ *for all* $\alpha \in V$.
4. $(-1)\alpha = -\alpha$ *for all* $\alpha \in V$.
5. *If* $x\alpha = 0$ *then,* $x = 0$ *or* $\alpha = 0$.

The reader will notice that we use 0 to represent the zero vector in V as well as the zero scalar in F. This will cause no real confusion in what follows. Theorem 14(1) implies linear combinations of vectors in V, i.e., sums of the form $x_1\alpha_1 + \cdots + x_n\alpha_n$, can be written unambiguously with no parentheses.

The notation for various vector spaces students encounter when studying linear algebra is becoming standard throughout most modern textbooks. Here is a short list of some of the more important vector spaces. If the reader is in doubt as to what addition or scalar multiplication is in the given example, consult Brown [1, 2].

1. F^S = all functions from a set S to the field F.
2. $M_{m\times n}(F)$ = the set of all $m \times n$ matrices with entries from F.
3. $F[X]$ = the set of all polynomials in X with coefficients from F.
4. $CS(A)$, $RS(A)$, and $NS(A) = \{\xi \in F^n \mid A\xi = 0\}$ for any $A \in M_{m\times n}(F)$. ($NS(A)$ is called the null space of A.)
5. $C^k(I) = \{f \in \mathbb{R}^I \mid f$ is k times differentiable on $I\}$. (I here is usually some open or closed set contained in $\mathbb{R}$).
6. $\mathcal{R}([a, b]) = \{f \in \mathbb{R}^{[a,b]} \mid f$ is Riemann integrable on $[a, b]\}$. (24)

Definition 17. *Let W be a nonempty subset of a vector space V. W is a subspace of V if $\alpha + \beta \in W$ and $x\alpha \in W$ for all $\alpha, \beta \in W$ and $x \in F$.*

Thus, a subset is a subspace if it is closed under vector addition and scalar multiplication. $\mathbb{R}[X]$, $C^k([a, b])$ and $\mathcal{R}([a, b])$ are all subspaces of $\mathbb{R}^{[a,b]}$. If $A \in M_{m\times n}(F)$, then $NS(A)$ is a subspace of F^n, $CS(A)$ is a subspace of F^m and $RS(A)$ is a subspace of $M_{1\times n}(F)$. One of the most important sources of subspaces are linear spans.

Definition 18. *Let S be a subset of a vector space V. The set of all linear combinations of vectors from S is called the linear span of S. We will let $L(S)$ denote the linear span of S.*

If $S = \emptyset$, i.e., S is empty, then we set $L(S) = (0)$. Notice, $\alpha \in L(S)$ if $\alpha = x_1\beta_1 + \cdots + x_n\beta_n$ for some $\beta_1, \ldots, \beta_n \in S$ and $x_1, \ldots, x_n \in F$. If S is finite, say $S = \{\gamma_1, \ldots, \gamma_r\}$, then we often write $L(\gamma_1, \ldots, \gamma_r)$ for $L(S)$. For example, if $A = (\xi_1 | \ldots | \xi_n) \in M_{m\times n}(F)$, then $L(\xi_1, \ldots, \xi_n) = CS(A)$.

Theorem 15. *Let V be a vector space over F.*

1. *For any subset $S \subseteq V$, $L(S)$ is a subspace of V.*
2. *If $S_1 \subseteq S_2 \subseteq V$, then $L(S_1) \subseteq L(S_2) \subseteq V$.*
3. *If $\alpha \in L(S)$, then $\alpha \in L(S_1)$ for some finite subset $S_1 \subseteq S$.*
4. $L(L(S)) = L(S)$.
5. *Exchange principle: If $\beta \in L(S \cup \{\alpha\})$ and $\beta \notin L(S)$, then $\alpha \in L(S \cup \{\beta\})$.*

The exchange principle is used to argue any two bases of V have the same cardinality.

Definition 19. *A vector space V is finite dimensional if $V = L(S)$ for some finite subset S of V.*

If V is not finite dimensional, we say V is infinite dimensional. $M_{m\times n}(F)$, $CS(A)$, $RS(A)$ and $NS(A)$ are all examples of finite-dimensional vector spaces over F. $\mathbb{R}[X]$, $C^k((0, 1))$ and $\mathcal{R}([0, 1])$ are all infinite-dimensional vector spaces over $\mathbb{R}$.

3.2.2 Bases

Definition 20. *Let S be a subset of a vector space V.*

1. *The set S is linearly dependent (over F) if there exist distinct vectors $\alpha_1, \ldots, \alpha_n \in S$ and nonzero scalars $x_1, \ldots, x_n \in F$ such that $x_1\alpha_1 + \cdots + x_n\alpha_n = 0$.*
2. *The set S is linearly independent (over F) if S is not linearly dependent.*
3. *S is a basis of V if S is linearly independent and $L(S) = V$.*

Suppose S is a basis of V. Then every vector in V is a linear combination of vectors from S. To be more precise, if $\beta \in V$ and $\beta \neq 0$, then there exist $\alpha_1, \ldots, \alpha_n \in S$ (all distinct) and nonzero scalars $x_1, \ldots, x_n \in F$ such that $\beta = x_1\alpha + \cdots + x_n\alpha_n$. Furthermore, this representation is unique. By this, we mean if $\beta = y_1\gamma_1 + \cdots + y_t\gamma_t$ with $y_1, \ldots, y_t$ nonzero scalars and $\gamma_1, \ldots, \gamma_t$ distinct vectors in S, then $n = t$ and after a suitable permutation of the symbols, $\alpha_1 = \gamma_1, \ldots, \alpha_n = \gamma_n$ and $x_1 = y_1, \ldots, x_n = y_n$.

The four basic theorems about bases are listed in our next theorem.

Theorem 16

1. *Every vector space has a basis.*
2. *If S is a linearly independent subset of V, then $S \subseteq B$ for some basis B of V.*
3. *If $L(S) = V$, then S contains a basis of V.*
4. *Any two bases of V have the some cardinality.*

A proof of Theorem 16 can be found in Brown [2]. A much simpler proof of Theorem 16 when V is finite dimensional can be found in Brown [1]. The common cardinality of the bases of V will be denoted by $\dim(V)$ and called the dimension of V. $\text{Dim}(V)$ is a finite cardinal number, i.e., $0, 1, 2, \ldots$ if and only if V is finite dimensional. If V is infinite dimensional, then $\dim(V)$ is an infinite cardinal number. In this case, we will simply write $\dim(V) = \infty$. In our next example, we list the dimensions of some of the more important finite-dimensional vector spaces.

Example 8

1. $\dim(M_{m\times n}(F)) = mn$. *A basis of $M_{m\times n}(F)$ is given by the matrix units $B = \{A_{ij} \mid i = 1, \ldots, m; j = 1, \ldots, n\}$ of $M_{m\times n}(F)$. Here A_{ij} is the $m \times n$ matrix having a 1 in its i,jth entry and* 0 *elsewhere.*
2. $\dim(F^n) = n$. *A basis of F^n is given by $B = \{\varepsilon_1, \ldots, \varepsilon_n\}$ where $I_n = (\varepsilon_1 \mid \ldots \mid \varepsilon_n)$. B is usually called the canonical basis of F^n.*
3. *Let* $\mathcal{P}_n(\mathbb{R}) = \{p(X) \in \mathbb{R}[X] \mid \text{degree}(p) \leq n\}$. *Then* $\dim(\mathcal{P}_n(\mathbb{R})) = n + 1$. $B = \{1, X, \ldots, X^n\}$ *is a basis of* $\mathcal{P}_n(\mathbb{R})$.

4. *Let $A \in M_{m\times n}(F)$. The dimension of $CS(A)$ is called the rank of A and written* $\mathrm{rk}(A)$. *The dimension of $NS(A)$ is called the nullity of A and written $\nu(A)$. It follows from Theorem 17 that* $0 \le \mathrm{rk}(A) \le m$ *and* $0 \le \nu(A) \le n$.

Theorem 17. *Suppose W is a subspace of V and* $\dim(V) < \infty$. *Then* $\dim(W) \le \dim(V)$ *with equality if and only if $W = V$.*

There are many standard theorems about $\mathrm{rk}(A)$, i.e., $\dim(CS(A))$. Here are some of the more important ones.

Theorem 18. *Let $A \in M_{m\times n}(F)$ and $B \in M_{n\times p}(F)$.*

1. $\dim(CS(A)) = \dim(RS(A))$.
2. $0 \le \mathrm{rk}(A) \le \min\{m, n\}$.
3. $\mathrm{rk}(A) = \mathrm{rk}(A^t)$ $(= \mathrm{rk}(A^*)$ *if* $F = \mathbb{C})$.
4. $\mathrm{rk}(A) = \mathrm{rk}(PAQ)$ *for any invertible matrices P, Q.*
5. $\mathrm{rk}(AB) \le \min\{\mathrm{rk}(A), \mathrm{rk}(B)\}$.
6. $\mathrm{rk}(A) + \nu(A) = n$.

Theorem 18(1) implies that the rank of A can be computed from the row space of A as well as the column space of A. Theorem 18(4) implies that the integer t appearing in Theorem 9 is the rank of A. Theorem 18(6) is called Sylvester's law of nullity. We also have

Theorem 19. *Let $A \in M_{n\times n}(F)$. Then the following statements are equivalent:*

1. *A is invertible.*
2. $\mathrm{rk}(A) = n$.
3. *The columns of A are linearly independent.*
4. *The rows of A are linearly independent.*
5. $CS(A) = F^n$.
6. $RS(A) = M_{1\times n}(F)$.
7. $\nu(A) = 0$.

For systems of linear equations, we have:

Theorem 20. *Let $A \in M_{m\times n}(F)$ and $B \in F^m$. The linear system of equations $AX = B$ has a solution if and only if* $\mathrm{rk}(A \mid B) = \mathrm{rk}(A)$.

3.2.3 Coordinate Maps and Change-of-Basis Matrices

One of the most important applications of bases is to convert abstract problems into concrete matrix problems which machines can then solve. To see how this is done, suppose V is a finite-dimensional vector space over F. Suppose $\dim(V) = n$. Choose a basis $\{\alpha_1, \ldots, \alpha_n\}$ of V and form the n-tuple $B = (\alpha_1, \ldots, \alpha_n)$. Then B determines a function $[*]_B : V \to F^n$ which is defined as follows:

Definition 21

$$[\gamma]_B = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad \text{if } \gamma = x_1\alpha_1 + \cdots + x_n\alpha_n \text{ in } V$$

The definition makes sense because the entries in B form a basis of V. Every vector $\gamma \in V$ can be written in one and only one way as a linear combination of $\alpha_1, \ldots, \alpha_n$. The function $[*]_B$ is called a coordinate map on V. If we permute the entries of B or choose a new basis altogether, we get a different coordinate map. Every coordinate map satisfies the following conditions:

Theorem 21. *Let $\{\alpha_1, \ldots, \alpha_n\}$ be a basis of V and set $B = (\alpha_1, \ldots, \alpha_n)$. Then,*

1. $[x\lambda + y\delta]_B = x[\lambda]_B + y[\delta]_B$ *for all $\lambda, \delta \in V$ and x, $y \in F$.*
2. $[*]_B : V \to F^n$ *is one-to-one and onto.*
3. *$\gamma_1, \ldots, \gamma_t$ are linearly independent in V if and only if $[\gamma_1]_B, \ldots, [\gamma_t]_B$ are linearly independent in F^n.*
4. *$\gamma \in L(\gamma_1, \ldots, \gamma_t)$ in V if and only if $[\gamma]_B \in L([\gamma_1]_B, \ldots, [\gamma_t]_B)$ in F^n.*
5. *If $W_i, i = 1, 2, 3$, are subspaces of V, then $W_1 + W_2 = W_3$ $(W_1 \cap W_2 = W_3)$ if and only if $[W_1]_B + [W_2]_B = [W_3]_B$ $([W_1)_B \cap [W_2]_B = [W_3]_B)$ in F^n.*

Let us give one example which illustrates the power of Theorem 21.

Example 9. *Let $V = \mathcal{P}_4(\mathbb{R})$. Suppose $f_1(X) = 1 + X + X^4$, $f_2(X) = 1 - X + X^3 - X^4$ and $f_3(X) = 1 + 3X - X^3 + 3X^4$. Are f_1, f_2, f_3 linearly independent in V? To answer this question, we use Theorem 21(3).*

Let $B = (1, X, X^2, X^3, X^4)$. B is an ordered basis of V. The matrix

$$A = ([f_1]_B | [f_2]_B | [f_3]_B)$$
$$= \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 3 \\ 0 & 0 & 0 \\ 0 & 1 & -1 \\ 1 & -1 & 3 \end{pmatrix} \tilde{r} \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

has rank 2. Thus f_1, f_2, f_3 are linearly dependent.

Suppose $B = (\alpha_1, \dots, \alpha_n)$ and $C = (\beta_1, \dots, \beta_n)$ are two ordered bases of V. How are the coordinate maps $[*]_B$ and $[*]_C$ related?

Definition 22. *Let $B = (\alpha_1, \dots, \alpha_n)$ and $C = (\beta_1, \dots, \beta_n)$ be two ordered bases of V. Set*

$$M(C, B) = ([\beta_1]_B | [\beta_2]_B | \dots | [\beta_n]_B)$$

$M(C, B)$ is an $n \times n$ matrix called the change-of-basis matrix from C to B.

Theorem 22. *Let $B = (\alpha_1, \dots, \alpha_n)$ and $C = (\beta_1, \dots, \beta_n)$ be two ordered basis of V. Then for all $\gamma \in V$, $M(C, B)[\gamma]_C = [\gamma]_B$.*

Theorem 22 implies each change-of-basis matrix $M(C, B)$ is invertible with inverse $M(B, C)$.

3.2.4 Linear Transformations

Definition 23. *Let V and W be vector spaces over F. A function $T : V \to W$ is called a linear transformation if $T(x\alpha + y\beta) = xT(\alpha) + yT(\beta)$ for all $\alpha, \beta \in V$ and $x, y \in F$.*

We will let $\mathrm{Hom}(V, W)$ denote the set of all linear transformations from V to W. In algebra, a linear transformation is also called a homomorphism. The symbols denoting the complete set of homomorphisms from V to W comes from the word homomorphism. The function $T : V \to W$ given by $T(\alpha) = 0$ for all $\alpha \in V$ is clearly a linear transformation (called the zero map). Hence, $\mathrm{Hom}(V, W)$ contains at least one map. Here are some standard examples of linear transformations.

Example 10

1. *Coordinate maps $[*]_B : V \to F^n$ are linear transformations by Theorem 21(1).*
2. *Let $V = W$ and define $T : V \to V$ by $T(\alpha) = \alpha$ for all $\alpha \in V$. T is called the identity map on V and will be denoted by 1_V.*
3. *$T : M_{m\times n}(F) \to M_{n\times m}(F)$ given by $T(A) = A^t$ is a linear transformation (notice $S : M_{m\times n}(\mathbb{C}) \to M_{n\times m}(\mathbb{C})$ given by $S(A) = A^*$ is not a linear transformation).*
4. *Let $A \in M_{m\times n}(F)$. A defines a linear transformation $\mu_A : F^n \to F^m$ given by $\mu_A(\xi) = A\xi$ for all $\xi \in V$. The map μ_A is called multiplication by A.*
5. *Let I be a nonempty subset of $\mathbb{R}$ and set $V = \mathbb{R}^I$. Let $a \in I$. The map $E_a : \mathbb{R}^I \to \mathbb{R}$ given by $E_A(f) = f(a)$ is a linear transformation called evaluation at a.*
6. *Let I be an open interval in $\mathbb{R}$. The map $D : C^1(I) \to \mathbb{R}^I$ given by $D(f) = f'$ (the derivative of f) is a linear transformation.*
7. *The map $S : \mathcal{R}([a, b]) \to \mathbb{R}$ given by $S(f) = \int_a^b f(t)\, dt$ is a linear transformation.*

Definition 24. *Let $T \in \mathrm{Hom}(V, W)$.*

1. $\mathrm{Ker}(T) = \{\alpha \in V | T(\alpha) = 0\}$.
2. $\mathrm{Im}(T) = \{T(\alpha) | \alpha \in V\}$.
3. *T is injective (one-to-one, monomorphism) if* $\mathrm{Ker}(T) = (0)$.
4. *T is surjective (onto, epimorphism) if* $\mathrm{Im}(T) = W$.
5. *T is an isomorphism if T is both injective and surjective.*

The linear transformations in Example 10(1–3) are all isomorphisms. μ_A is an isomorphism if and only if A is invertible. The set $\mathrm{Ker}(T)$ is a subspace of V and is called the kernel of T. The set $\mathrm{Im}(T)$ is a subspace of W and is called the image (or range) of T. If T is an isomorphism, then T is a bijective map (one-to-one and onto) from V to W. In this case, there is a well defined inverse map T^{-1}: $W \to V$ given by $T^{-1}(\beta) = \alpha$ if $T(\alpha) = \beta$. It is easy to prove $T^{-1} \in \mathrm{Hom}(W, V)$.

Definition 25. *Two vector spaces V and W over F are said to be isomorphic if there exists a linear transformation $T : V \to W$ which is an isomorphism.*

If V and W are isomorphic, we will write $V \cong W$. Our remarks before Definition 25 imply $V \cong W$ if and only if $W \cong V$. Isomorphic vector spaces are virtually identical. Only the names of the vectors are being changed by the isomorphism. Example 10(1) implies the following statement: if $\dim(V) = n$, then $V \cong F^n$.

Thus, up to isomorphism, F^n is the only finite-dimensional vector space over F of dimension n.

The construction of linear transformations is facilitated by the following existence theorem. We do not assume V is finite dimensional here.

Theorem 23. *Let V be a vector space over F and suppose $B = \{\alpha_i \mid i \in \Delta\}$ is a basis of V. Let W be another vector space over F and suppose $\{\beta_i \mid i \in \Delta\}$ is any subset of W. Then*

1. *A linear transformation from V to W is completely determined by its values on B. Thus, if T, $S \in \mathrm{Hom}(V, W)$ and $T(\alpha_i) = S(\alpha_i)$ for all $i \in \Delta$, then $T = S$.*
2. *There exists a unique linear transformation $T : V \to W$ such that $T(\alpha_i) = \beta_i$ for all $i \in \Delta$.*

A proof of Theorem 23 can be found in Brown [2]. The four basic theorems connecting linear transformations and dimensions are as follows.

Theorem 24. *Let V be a finite-dimensional vector space over F. Let $T \in \mathrm{Hom}(V, W)$.*

1. *If T is surjective, then W is finite dimensional. In this case,* $\dim(V) \geq \dim(W)$.
2. *Suppose* $\dim(V) = \dim(W)$. *Then T is an isomorphism if and only if T is injective.*
3. *Suppose* $\dim(V) = \dim(W)$. *Then T is an isomorphism if and only if T is surjective.*
4. $\dim(\mathrm{Ker}(T)) + \dim(\mathrm{Im}(T)) = \dim(V)$.

Finally, $\mathrm{Hom}(V, W)$ is itself a vector space over F when addition and scalar multiplication of linear transformations are defined as follows: For $T, S \in \mathrm{Hom}(V, W)$, set

$$\begin{aligned} (T + S)(\alpha) &= T(\alpha) + S(\alpha) \quad \text{for all } \alpha \in V \\ (xT)(\alpha) &= x(T(\alpha)) \quad \text{for all } x \in F \text{ and } \alpha \in V \end{aligned} \tag{25}$$

3.2.5 Matrix Representations of Linear Transformations

Suppose V and W are both finite-dimensional vector spaces over F. Say $\dim(V) = n$ and $\dim(W) = m$. Let $B = (\alpha_1, \ldots, \alpha_n)$ be an ordered basis of V and $C = (\beta_1, \ldots, \beta_m)$ be an ordered basis of W. Let $T \in \mathrm{Hom}(V, W)$. We can then define an $m \times n$ matrix as follows:

Definition 26. $M(T; B, C) = ([T(\alpha_1)]_C | [T(\alpha_2)]_C | \ldots | [T(\alpha_n)]_C)$.

$M(T; B, C)$ is called the matrix representation of T with respect to B and C. The ith column of $M(T; B, C)$ is just the coordinates of $T(\alpha_i)$ with respect to C.

Example 11. *Let $V = \mathcal{P}_3(\mathbb{R})$, $W = \mathcal{P}_2(\mathbb{R})$ and $D : \mathcal{P}_3(\mathbb{R}) \to \mathcal{P}_2(\mathbb{R})$ be ordinary differentiation ($D(f) = f'$). Let $B = (1, X, X^2, X^3)$ and $C = (1, X, X^2)$. Then*

$$M(D; B, C) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix} \in M_{3\times 4}(\mathbb{R}) \tag{26}$$

The following diagram is certainly one of the most important diagrams in linear algebra:

$$\begin{tikzcd} V \arrow[r, "T"] \arrow[d, "{[*]_B}"'] & W \arrow[d, "{[*]_C}"] \\ F^n \arrow[r, "\mu_A"] & F^m \end{tikzcd} \tag{27}$$

Here $A = M(T; B, C)$.

Theorem 25. $[T(\alpha)]_C = A[\alpha]_B$ *for all $\alpha \in V$.*

Theorem 25 implies that the diagram (27) commutes, i.e., the composite maps $[*]_C \circ T$ and $\mu_A \circ [*]_B$ are the same. The vertical maps in (27) are isomorphisms which translate the abstract situation $V \xrightarrow{T} W$ into the concrete situation $F^n \xrightarrow{\mu_A} F^m$. Machines do computations with the bottom row of (27).

Theorem 26. *In diagram (27),*

1. $\mathrm{Ker}(T) \cong NS(A)$.
2. $\mathrm{Im}(T) \cong CS(A)$.

The rank and nullity of a linear transformation T are defined to be the dimensions of $\mathrm{Im}(T)$ and $\mathrm{Ker}(T)$ respectively. Let us return to Example 11 for an illustration of how Theorem 26 works. Suppose we want to compute the rank and nullity of $D : \mathcal{P}_3(\mathbb{R}) \to \mathcal{P}_2(\mathbb{R})$. Since

$$M(D; B, C) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix} \in M_{3\times 4}(\mathbb{R}) \tag{28}$$

we conclude $\mathrm{rk}(D) = 3$ and $\nu(D) = 1$.

If we vary $T \in \mathrm{Hom}(V, W)$, we get a function $M(*; B, C) : \mathrm{Hom}(V, W) \to M_{m\times n}(F)$. This map is an isomorphism of vector spaces.

Theorem 27. *$M(*;B,C) : \mathrm{Hom}(V,W) \to M_{m\times n}(F)$ is an isomorphism.*

Thus,

1. $M(xT + yS; B, C) = xM(T; B, C) + yM(S; B, C)$.
2. $M(T; B, C) = 0$ if and only if T is the zero map.
3. Given $A \in M_{m\times n}(F)$ there exists a $T \in \mathrm{Hom}(V, W)$ such that $M(T; B, C) = A$. (29)

Suppose V, W, and Z are finite dimensional vector spaces over F. If $T \in \mathrm{Hom}(V,W)$ and $S \in \mathrm{Hom}(W,Z)$, then the composite map $S \circ T : V \to Z$ is a linear transformation. If D is an ordered basis of Z, then

$$M(S \circ T; B, D) = M(S; C, D)M(T; B, C) \tag{30}$$

If $T : V \to W$ is an isomorphism, then

$$M(T^{-1}; C, B) = M(T; B, C)^{-1} \tag{31}$$

Suppose we change bases in V and W. Let $B' = (\alpha_1', \ldots, \alpha_n')$ and $C' = (\beta_1', \ldots, \beta_m')$ be two new, ordered bases of V and W respectively. Then we have two matrix representations of $T: M(T; B, C)$ and $M(T; B', C')$. Recall that $M(B, B')$ and $M(C, C')$ denote change-of-basis matrices in V and W. The relation between the matrix representations of T is as follows:

$$M(T; B', C') = M(C, C')M(T; B, C)M(B, B')^{-1} \tag{32}$$

Since change of bases matrices are invertible, a simple translation of Theorem 9 gives us the following theorem:

Theorem 28. *Suppose V and W are finite-dimensional vector spaces over F. Let $T \in \mathrm{Hom}(V, W)$ and suppose the rank of T is t. Then there exist ordered bases B and C of V and W respectively such that*

$$M(T; B, C) = \left(\begin{array}{c|c} I_t & 0 \\ \hline 0 & 0 \end{array}\right)$$

There is a special case of Eq. (32) which is worth mentioning here. Suppose $V = W$ and $T \in \mathrm{Hom}(V, V)$. If B and B' are two ordered bases of V, then $M(T; B, B)$ and $M(T; B', B')$ are two $n \times n$ matrices representing T. If $U = M(B, B')$, then Eq. (32) becomes

$$M(T; B', B') = UM(T; B, B)U^{-1} \tag{33}$$

Definition 27. *Let $A_1, A_2 \in M_{n\times n}(F)$. A_1 is similar to A_2 if $A_1 = UAU^{-1}$ for some invertible matrix $U \in M_{n\times n}(F)$.*

The relation in Eq. (33) implies that any two matrix representations of $T \in \mathrm{Hom}(V, V)$ are similar. We then have the following questions: What is the simplest matrix representation of T? In other words, what is the simplest analog of Theorem 28? In terms of matrices, the question becomes: What is the simplest matrix B which is similar to a given matrix A? The answers to these questions are called canonical forms theorems. The important canonical forms (e.g., the Jordan canonical form, rational canonical form, etc.) are discussed in Brown [2].

3.3 DETERMINANTS AND EIGENVALUES

3.3.1 Determinants

Let $n \in \mathbb{N}$ and set $\Delta(n) = \{1, 2, \ldots, n\}$. A permutation (on n letters) is a bijective function from $\Delta(n)$ to $\Delta(n)$. We will let S_n denote the set of all permutations on n letters. If $\sigma \in S_n$, then σ is represented by a $2 \times n$ matrix

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & \ldots & n \\ i_1 & i_2 & i_3 & \ldots & i_n \end{pmatrix} \tag{34}$$

Here $\sigma(1) = i_1, \sigma(2) = i_2, \ldots, \sigma(n) = i_n$. Thus, $i_1, \ldots, i_n$ are the numbers $1, 2, \ldots, n$ in some different order. Obviously, S_n is a finite set of cardinality $n!$. Permutations can be composed as functions on $\Delta(n)$. Composition determines a binary operation $S_n \times S_n \to S_n[(\sigma, \tau) \to \sigma \circ \tau]$ which endows S_n with the structure of a group. See Brown [1, Chap. III] for more details.

Definition 28. *A permutation $\sigma \in S_n$ is called a cycle of length r if there exist distinct integers $i_1, \ldots, i_r \in \Delta(n)$ such that:*

1. *$\sigma(i_1) = i_2, \sigma(i_2) = i_3, \ldots, \sigma(i_{r-1}) = i_r$, and $\sigma(i_r) = i_1$.*
2. *$\sigma(j) = j$ for all $j \in \Delta(n)\backslash\{i_1, \ldots, i_r\}$.*

If σ is a cycle of length r, we will write $\sigma = (i_1, i_2, \ldots, i_r)$. A two-cycle (a, b) interchanges a and b and leaves all other elements of $\Delta(n)$ invariant. Two-cycles are also called transpositions. Every permutation in S_n is a product of disjoint cycles (i.e., cycles having no entries in common). Every cycle $\sigma = (i_1, \ldots, i_r)$ is a product of transpositions: $(i_1, \ldots, i_r) = (i_1, i_r)(i_1, i_{r-1}) \ldots (i_1, i_2)$. Thus, every permutation is a finite product of transpositions.

Example 12. *Let*

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 2 & 3 & 4 & 1 & 6 & 5 & 8 & 9 & 7 \end{pmatrix} \in S_9$$

Then

$$\begin{aligned} \sigma &= (1,2,3,4)(5,6)(7,8,9) \\ &= (1,4)(1,3)(1,2)(5,6)(7,9)(7,8) \end{aligned} \tag{35}$$

If $\sigma \in S_n$ is a product of an even (odd) number of transpositions, then any factorization of σ into a product of transpositions must contain an even (odd) number of terms. A permutation σ is said to be even (odd) if σ is a product of an even (odd) number of transpositions. We can now define a function $\operatorname{sgn}(*) : S_n \rightarrow \{-1, 1\}$ by the following rules:

$$\operatorname{sgn}(\sigma) = \begin{cases} 1 & \text{if } \sigma \text{ is even} \\ -1 & \text{if}\sigma \text{ is odd} \end{cases} \tag{36}$$

The number $\operatorname{sgn}(\sigma)$ is called the sign of σ. If e denotes the identity map on $\Delta(n)$, then $e = (a,b)(b,a)$ and, hence, $\operatorname{sgn}(e) = 1$. Any transposition is odd. Hence, $\operatorname{sgn}((a,b)) = -1$. If σ is the permutation given in Eq. (35), then σ is even and $\operatorname{sgn}(\sigma) = 1$.

We can now define the determinant, $\det(A)$, of an $n \times n$ matrix A.

Definition 29. *Let $A = (a_{ij}) \in M_{n\times n}(F)$. The determinant of A is defined to be the following element of F*

$$\det(A) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) a_{1\sigma(1)} a_{2\sigma(2)} \ldots a_{n\sigma(n)} \tag{37}$$

The symbols in Eq. (37) mean add all possible products $\operatorname{sgn}(\sigma) a_{1\sigma(1)} a_{2\sigma(2)} \ldots a_{n\sigma(n)}$ as σ ranges over S_n. If we let A vary, the determinant defines a function $\det(*) : M_{n\times n}(F) \rightarrow F$.

The value of $\det(A)$ is numerically hard to compute. For instance, if $n = 5$, there are 120 products to compute, store, and add together in Eq. (37). Fortunately, we have many theorems which help us compute the value of $\det(A)$. A summary of the more elementary properties of the determinant is given in Theorem 29 below.

Theorem 29. *Let $A, B \in M_{n\times n}(F)$. Then*

1. *If* $\text{Row}_i(A) = 0$ (*or* $\text{Col}_i(A) = 0$), *then* $\det(A) = 0$.
2. *If* $\text{Row}_i(A) = \text{Row}_j(A)$ *for some* $i \neq j$ [*or* $\text{Col}_i(A) = \text{Col}_j(A)$], *then* $\det A = 0$.
3. *If A is upper or lower triangular, then* $\det(A) = \prod_{i=1}^{n} [A]_{ii}$.
4. *If $A = (\alpha_1; \ldots; \alpha_n)$ is the row partition of A and $\alpha_i = \beta + \delta$ for some $i = 1, \ldots, n$ and some $\beta, \delta \in M_{1\times n}(F)$, then*
   a. $\det(\alpha_1; \ldots; x\alpha_i; \ldots; \alpha_n) = x\det(\alpha_1; \ldots; \alpha_i; \ldots; \alpha_n)$.
   b. $\det(\alpha_1; \ldots; \beta + \delta; \ldots; \alpha_n) = \det(\alpha_1; \ldots; \beta; \ldots; \alpha_n) + \det(\alpha_1; \ldots; \delta; \ldots \alpha_n)$.
5. *If $A = (\alpha_1; \ldots; \alpha_n)$ is the row partition of A and $\sigma \in S_n$, then* $\det(\alpha_{\sigma(1)}; \ldots; \alpha_{\sigma(n)}) = \operatorname{sgn}(\sigma) \det\,(\alpha_1; \ldots; \alpha_n)$.
6. $\det(AB) = \det(A)\det(B)$.
7. a. $\det(E_{ij}A) = -\det(A) \quad (i \neq j)$.
   b. $\det(E_{ij}(c)A) = \det(A) \quad (i \neq j)$.
   c. $\det(E_j(c)A) = c\det(A)$.
8. *If $PA = LU$ is an LU-factorization of PA, then* $\det(A) = \operatorname{sgn}(P)(\prod_{i=1}^{n} [U]_{ii})$.
9. $\det(A) = \det(A^t)$.
10. *A is invertible if and only if* $\det(A) \neq 0$.

The corresponding statements for columns in Theorem 29(4–6) are also true. The matrix P in (8) is a permutation matrix and, consequently, has the form $P = (\varepsilon_{\sigma(1)} \mid \ldots \mid \varepsilon_{\sigma(n)})$ where $(\varepsilon_1 \mid \ldots \mid \varepsilon_n) = I_n$ and $\sigma \in S_n$. Then $\operatorname{sgn}(P)$ is defined to be $\operatorname{sgn}(\sigma)$. Theorem 29(8) is an important application of LU-factorizations: To compute $\det(A)$, factor PA for a suitable permutation matrix P, compute the product of the diagonal elements of U and then $\det(A) = \operatorname{sgn}(P)(\prod_{i=1}^{n} [U]_{ii})$. For example, Eq. (22) implies $\det(A) = -7$. Theorem 29(10) is one of the most important properties of the determinant. A is singular (i.e., not invertible) if and only if $\det(A) = 0$.

Definition 30. *Let $A \in M_{n\times n}(F)$. Assume $n \geq 2$.*

1. *For $i, j = 1, \ldots, n$, $M_{ij}(A)$ will denote the $(n-1) \times (n-1)$ submatrix of A obtained by deleting row i and column j of A.*
2. $\operatorname{cof}_{ij}(A) = (-1)^{i+j} \det(M_{ij}(A))$.
3. $\operatorname{adj}(A)$ *is the $n \times n$ matrix whose i, jth entry is given by* $[\operatorname{adj}(A)]_{ij} = \operatorname{cof}_{ji}(A)$.

The determinant of the $(n-1) \times (n-1)$ submatrix $M_{ij}(A)$ is called the i, jth minor of A. The i, jth minor of A with sign $(-1)^{i+j}$ is $\operatorname{cof}_{ij}(A)$ and is called the i, jth cofactor of A. The matrix defined in 3 is called the adjoint of A.

Theorem 30. Laplace Expansion: *Let $A \in M_{n\times n}(F)$. Then* $\operatorname{adj}(A)A = A\operatorname{adj}(A) = (\det A)I_n$.

Analyzing the entries in Theorem 30 gives us the following identities:

$$\begin{aligned}&\sum_{j=1}^{n}[A]_{ij}\operatorname{cof}_{kj}(A)=\delta_{ik}\det(A)\\&\qquad\qquad\text{for all } i,k=1,\ldots,n\\&\sum_{i=1}^{n}[A]_{ij}\operatorname{cof}_{ik}(A)=\delta_{jk}\det(A)\\&\qquad\qquad\text{for all } j,k=1,\ldots,n\end{aligned}\tag{38}$$

here

$$\delta_{uv}=\begin{cases}1 & \text{if } u=v\\ 0 & \text{if } u\neq v\end{cases}$$

is Kronecker's delta function.

If A is invertible, then Theorem 30 implies $A^{-1}=(\det(A))^{-1}\operatorname{adj}(A)$. Equation (11) is just Laplace's theorem when $n=2$. The last elementary result we will give concerning determinants is Cramer's rule.

Theorem 31. Cramer's Rule: *Let $A=(\xi_1\mid\ldots\mid\xi_n)\in M_{n\times n}(F)$. Let $B\in F^n$. Suppose A is invertible. Then the unique solution $(x_1,\ldots,x_n)^t\in F^n$ to the system of equations $AX=B$ is given by*

$$\begin{aligned}x_i=(\det(A))^{-1}\det(\xi_1\mid\ldots\mid\xi_{i-1}\mid B\mid\xi_{i+1}\mid\ldots\mid\xi_n)\\ \text{for all } i=1,\ldots,n\end{aligned}\tag{39}$$

3.3.2 Eigenvalues

Definition 31

1. *Let $A\in M_{n\times n}(F)$. A scalar $d\in F$ is called an eigenvalue (or characteristic value) of A if there is a nonzero vector $\xi\in F^n$ such that $A\xi=d\xi$.*
2. *Let V be a vector space over F and $T\in\operatorname{Hom}(V,V)$. A scalar $d\in F$ is an eigenvalue (or characteristic value) of T if there is a nonzero vector $\lambda\in V$ such that $T(\lambda)=d\lambda$.*

Eigenvalues of matrices and linear transformations are related to each other by diagram (27). If A is any matrix representation of T, then d is an eigenvalue of T if and only if d is an eigenvalue of A. For this reason, we will present only the theory for matrices. The reader can translate the results given here into the corresponding theorems about linear transformations (on finite-dimensional vector spaces) by using (27).

Definition 32. *Let $A\in M_{n\times n}(F)$. Then $\mathcal{S}_F(A)=\{d\in F\mid d$ is an eigenvalue of $A\}$ is called the spectrum of A.*

The spectrum of A could very well be empty. Consider the following well-known example.

Example 13. *Let*

$$A=\begin{pmatrix}0&-1\\1&0\end{pmatrix}\in M_{2\times2}(\mathbb{R})$$

The matrix A represents a rotation (in the counterclockwise direction) of 90° in the plane $\mathbb{R}^2$. It is easy to see $A\xi=d\xi$ for some $d\in\mathbb{R}$ implies $\xi=0$. Thus, $\mathcal{S}_{\mathbb{R}}(A)=\emptyset$.

If we view A as a complex matrix, i.e., $A\in M_{2\times2}(\mathbb{C})$, then

$$A\begin{pmatrix}1\\-i\end{pmatrix}=i\begin{pmatrix}1\\-i\end{pmatrix}\qquad A\begin{pmatrix}-1\\-i\end{pmatrix}=-i\begin{pmatrix}-1\\-i\end{pmatrix}\tag{40}$$

Here $i=\sqrt{-1}$. It is easy to show $\mathcal{S}_{\mathbb{C}}(A)=\{-i,i\}$.

Example 13 shows that the base field F is important when computing eigenvalues. Thus, the notation $\mathcal{S}_F(A)$ for the spectrum of A includes the field F in the symbols.

Definition 33. *Let $A\in M_{n\times n}(F)$ and let X denote an indeterminate over F. The polynomial $C_A(X)=\det(XI_n-A)$ is called the characteristic polynomial of A.*

For any matrix $A\in M_{n\times n}(F)$, the characteristic polynomial has the following form:

$$C_A(X)=X^n+a_1X^{n-1}+\cdots+a_{n-1}X+a_n.\tag{41}$$

In Eq. (41), $a_1,\ldots,a_n\in F$. The coefficients $a_1,\ldots,a_n$ appearing in $C_A(X)$ all have various interpretations which are related to A. For example,

$$a_1=-\sum_{i=1}^{n}[A]_{ii}\qquad a_n=(-1)^n\det(A)\tag{42}$$

At any rate, $C_A(X)$ is always a nonzero polynomial of degree n whose leading term is 1. The connection between $\mathcal{S}_F(A)$ and $C_A(X)$ is given in our next theorem.

Theorem 32. $\mathcal{S}_F(A)=\{d\in F\mid C_A(d)=0\}$.

Thus, the zeros of $C_A(X)$ in F are precisely the eigenvalues of A. In Example 13, $C_A(X)=X^2+1$. The zeros of X^2+1 are $\{-i,i\}\subseteq\mathbb{C}$. Hence, $\mathcal{S}_{\mathbb{R}}(A)=\emptyset$ and $\mathcal{S}_{\mathbb{C}}(A)=\{\pm i\}$.

Although Theorem 32 is simple, it is only useful when $C_A(X)$ (an $n\times n$ determinant) can be computed and the roots of $C_A(X)$ in F can be computed. For large n or "bad" matrices A, more sophisticated methods (such as the power method or inverse power method when $F = \mathbb{R}$ or $\mathbb{C}$) must be employed. One of the central problems of numerical linear algebra is to devise iterative methods for computing eigenvalues. A good elementary reference for these techniques is Cullen [4].

Notice that Theorem 32 implies A has at most n distinct eigenvalues in F. Also, Theorem 32 implies if A is similar to B, then $\mathcal{S}_F(A) = \mathcal{S}_F(B)$.

Definition 34. *Let $d \in \mathcal{S}_F(A)$. The subspace $NS(dI_n - A)$ is called the eigenspace of A associated with d. The nonzero vectors in $NS(dI_n - A)$ are called eigenvectors (or characteirstic vectors) of A associated with d.*

We will let $\mathcal{E}_A(d)$ denote the eigenspace of A associated with the eigenvalue d. If $d_1, \dots, d_r$ are distinct eigenvalues in $\mathcal{S}_F(A)$ and ξ_i is a nonzero vector in $\mathcal{E}_A(d_i)$, then $\xi_1, \dots, \xi_r$ are linearly independent over F. This leads immediately to our next theorem.

Theorem 33. *Let $A \in M_{n\times n}(F)$. A is similar to a diagonal matrix if and only if F^n has a basis consisting of eigenvectors of A.*

There are many applications of Theorem 33.

Example 14. *Suppose $A \in M_{n\times n}(F)$. How do we compute A^k for all $k \geq 2$? If F^n has a basis $\{\xi_1, \dots, \xi_n\}$ consisting of eigenvectors of A, then the problem is easily solved. Suppose $A\xi_i = d_i\xi_i$ for $i = 1, \dots, n$. Set $P = (\xi_1 \mid \dots \mid \xi_n)$. Since $\mathrm{rk}(P) = n$, P is invertible.*

$$\begin{aligned} AP &= A(\xi_1 \mid \dots \mid \xi_n) \\ &= (A\xi_1 \mid \dots \mid A\xi_n) = (d_1\xi_1 \mid \dots \mid d_n\xi_n) = PD \end{aligned} \tag{43}$$

Here

$$D = \begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_n \end{pmatrix}$$

Thus, $A = PDP^{-1}$ and

$$A^k = (PDP^{-1})^k = PD^kP^{-1} = P\begin{pmatrix} d_1^k & & 0 \\ & \ddots & \\ 0 & & d_n^k \end{pmatrix}P^{-1} \tag{44}$$

There are many iteration-type problems in which A^k must be computed for all $k \geq 1$. These problems are easily solved if A has enough eigenvectors to span the space. The reader is urged to consult Brown [1] for other applications of eigenvalues.

3.4 INNER-PRODUCT SPACES

3.4.1 Real and Complex Definitions

Inner products are defined on vector spaces defined over $\mathbb{R}$ or $\mathbb{C}$. A vector space V whose field of scalars $F = \mathbb{R}(\mathbb{C})$ is called a real (complex) vector space. The definition of an inner product is slightly different for the two cases.

Definition 35. *Let V be a real vector space, i.e., $F = \mathbb{R}$. An inner product on V is a function $\langle *, * \rangle : V \times V \to \mathbb{R}$ which satisfies the following conditions:*

1. *$\langle \alpha, \alpha \rangle$ is positive for every nonzero $\alpha \in V$.*
2. *$\langle x\alpha + y\beta, \gamma \rangle = x\langle \alpha, \gamma \rangle + y\langle \beta, \gamma \rangle$ for all $\alpha, \beta, \gamma \in V$ and $x, y \in \mathbb{R}$.*
3. $\langle \alpha, \beta \rangle = \langle \beta, \alpha \rangle$ for all $\alpha, \beta \in V$.

A real vector space V together with an inner product $\langle *, * \rangle$ on V is called an inner-product space. We will denote an inner-product space by the ordered pair $(V, \langle *, * \rangle)$.

If $(V, \langle *, * \rangle)$ is an inner-product space, then $\langle 0, \alpha \rangle = 0$ for all $\alpha \in V$. Also, $\langle \gamma, x\alpha + y\beta \rangle = x\langle \gamma, \alpha \rangle + y\langle \gamma, \beta \rangle$ by 2 and 3. Hence, $\langle *, * \rangle$ is a bilinear function on $V \times V$.

Example 15

1. *Let $V = \mathbb{R}^n$. Define $\langle \alpha, \beta \rangle = \alpha^t\beta$. Here we identify the 1×1 matrix $\alpha^t\beta$ with its single entry in $\mathbb{R}$. Then $(\mathbb{R}^n, \langle \alpha, \beta \rangle = \alpha^t\beta)$ is an inner-product space. The function $\langle \alpha, \beta \rangle = \alpha^t\beta$ is called the standard inner product (in calculus, the dot product) on $\mathbb{R}^n$.*
2. *Let $V = \mathbb{R}^n$. Let $c_1, \dots, c_n$ be any positive numbers in $\mathbb{R}$. If $\alpha = (x_1, \dots, x_n)^t$ and $\beta = (y_1, \dots, y_n)^t$, set $\langle \alpha, \beta \rangle' = \sum_{i=1}^n c_i x_i y_i$. then $(\mathbb{R}^n, \langle *, * \rangle')$ is an inner-product space.*

Thus, a given real vector space V can have many different inner products on it.

3. *Let $V = C([a, b])$, the continuous real-valued functions on a closed interval $[a, b] \subseteq \mathbb{R}$. If we define $\langle f, g \rangle = \int_a^b f(x)g(x)dx$, then $(V, \langle *, * \rangle)$ is an inner-product space.*

Definition 36. *Let V be a complex vector space, i.e., $F = \mathbb{C}$. An inner product on V is a function $\langle *, * \rangle : V \times V \to \mathbb{C}$ which satisfies the following conditions:*

1. $\langle \alpha, \alpha \rangle$ *is a positive real number for every nonzero* $\alpha \in V$.
2. $\langle x\alpha + y\beta, \gamma \rangle = x\langle \alpha, \gamma \rangle + y\langle \beta, \gamma \rangle$ for all $\alpha, \beta, \gamma \in V$ and $x, y \in \mathbb{C}$.
3. $\langle \alpha, \beta \rangle = \overline{\langle \beta, \alpha \rangle}$ *for all* $\alpha, \beta \in V$.

In 3, $\overline{\langle \beta, \alpha \rangle}$ denotes the conjugate of the complex number $\langle \beta, \alpha \rangle$. A complex vector space V together with an inner product $\langle *, * \rangle$ on V will be called a complex inner-product space $(V, \langle *, * \rangle)$.

If $(V, \langle *, * \rangle)$ is a complex inner-product space, then $\langle 0, \alpha \rangle = \langle \alpha, 0 \rangle = 0$ for all $\alpha \in V$ and $\langle \gamma, x\alpha + y\beta \rangle = \bar{x}\langle \gamma, \alpha \rangle + \bar{y}\langle \gamma, \beta \rangle$. Thus $\langle *, * \rangle$ is linear in its first variable and conjugate linear in its second variable. The inner product given on $\mathbb{C}^n$ by $\langle \alpha, \beta \rangle = \alpha^t \bar{\beta}$ is called the standard inner product on $\mathbb{C}^n$.

The theorems for real and complex inner products are very similar. Usually if one erases the conjugation symbol in a complex proof, one gets the real proof. For this reason, we will state all future results for complex inner-product spaces and leave the corresponding results for real vector spaces to the reader.

Let $(V, \langle *, * \rangle)$ be a (complex) inner-product space. We can define a length function $\| * \| : V \to \mathbb{R}$ on V for by setting

$$\|\alpha\| = \sqrt{\langle \alpha, \alpha \rangle} \qquad \text{for all } \alpha \in V \tag{45}$$

By Definition 36(1) [or 35(1)], $\langle \alpha, \alpha \rangle \geq 0$ for any $\alpha \in V$. $\sqrt{\langle \alpha, \alpha \rangle}$ denotes the nonnegative real number whose square is $\langle \alpha, \alpha \rangle$. Thus, $\|\alpha\|^2 = \langle \alpha, \alpha \rangle$. One of the most important inequalities in mathematics is the Cauchy–Schwarz inequality:

$$|\langle \alpha, \beta \rangle| \leq \|\alpha\| \|\beta\| \qquad \text{for all } \alpha, \beta \in V \tag{46}$$

In Eq. (46), $|\langle \alpha, \beta \rangle|$ denotes the modulus of the complex number $\langle \alpha, \beta \rangle$. The function $\| * \|$ defined in Eq. (45) is called the norm associated with the inner product $\langle *, * \rangle$. The norm satisfies the following inequalities:

$$\|\alpha\| > 0 \qquad \text{if } \alpha \neq 0 \tag{47a}$$

$$\|0\| = 0 \tag{47b}$$

$$\|x\alpha\| = |x| \|\alpha\| \qquad \text{for all } \alpha \in V \text{ and } x \in \mathbb{C} \tag{47c}$$

$$\|\alpha + \beta\| \leq \|\alpha\| + \|\beta\| \qquad \text{for all } \alpha, \beta \in V \tag{47d}$$

The inequality in Eq. (47d) is called the triangle inequality. Its proof follows immediately from the Cauchy–Schwarz inequality.

The norm associated with the inner product $\langle *, * \rangle$ defines a distance function $d : V \times V \to \mathbb{R}$ given by the following equation

$$d(\alpha, \beta) = \|\alpha - \beta\| \qquad \text{for all } \alpha, \beta \in V \tag{48}$$

The distance function satisfies the following inequalities:

$$d(\alpha, \beta) \geq 0 \qquad \text{for all } \alpha, \beta \in V \tag{49a}$$

$$d(\alpha, \beta) = 0 \qquad \text{if and only if } \alpha = \beta \tag{49b}$$

$$d(\alpha, \beta) = d(\beta, \alpha) \qquad \text{for all } \alpha, \beta \in V \tag{49c}$$

$$d(\alpha, \beta) \leq d(\alpha, \gamma) + d(\gamma, \beta) \qquad \text{for all } \alpha, \beta, \gamma \in V \tag{49d}$$

Thus, any inner-product space $(V, \langle *, * \rangle)$ is a normed vector space $(V, \| * \|)$ with norm given by Eq. (45) and a metric space (V, d) with metric (i.e., distance function) d given by Eq. (48). Since we have a distance function on $(V, \langle *, * \rangle)$, we can extend many results from the calculus to $(V, \langle *, * \rangle)$. For more details, the reader is referred to Brown [2].

3.4.2 Orthogonality

Definition 37. *Let $(V, \langle *, * \rangle)$ be an inner-product space.*

1. *Two vectors $\alpha, \beta \in V$ are said to be orthogonal if $\langle \alpha, \beta \rangle = 0$.*
2. *A set of vectors $\{\alpha_i \mid i \in \Delta\} \subseteq V$ is said to be pairwise orthogonal if α_i and α_j are orthogonal whenever $i \neq j$.*

Notice that $\langle \alpha, \beta \rangle = 0$ if and only if $\langle \beta, \alpha \rangle = 0$. Thus, α and β are orthogonal if and only if β and α are orthogonal.

Theorem 34. *Let $\alpha_1, \ldots, \alpha_n$ be pairwise orthogonal, nonzero vectors in $(V, \langle *, * \rangle)$. Then*

1. *$\alpha_1, \ldots, \alpha_n$ are linearly independent.*
2. *If $\gamma \in L(\alpha_1, \ldots, \alpha_n)$, then*

$$\gamma = \sum_{j=1}^{n} \left(\frac{\langle \gamma, \alpha_j \rangle}{\langle \alpha_j, \alpha_j \rangle} \right) \alpha_j$$

3. *If $\gamma \in L(\alpha_1, \ldots, \alpha_n)$, then*

$$\|\gamma\| = \left\{ \sum_{j=1}^{n} \frac{|\langle \gamma, \alpha_j \rangle|^2}{\langle \alpha_j, \alpha_j \rangle} \right\}^{1/2}$$

A set of vectors $\alpha_1, \dots, \alpha_n \in (V, \langle *, * \rangle)$ is said to be orthonormal if $\langle \alpha_i, \alpha_j \rangle = 0$ whenever $i \neq j$ and $\|\alpha_i\| = 1$ for all $i = 1, \dots, n$. If $\alpha_1, \dots, \alpha_n$ are orthonormal, then Theorem 34 implies that $B = (\alpha_1, \dots, \alpha_n)$ is an ordered basis of $W = L(\alpha_1, \dots, \alpha_n)$. In this case, the coordinate map $[*]_B : W \to \mathbb{C}^n$ is particularly easy to compute. By 2 and 3, we have

$$[\gamma]_B = \begin{pmatrix} \langle \gamma, \alpha_1 \rangle \\ \vdots \\ \langle \gamma, \alpha_n \rangle \end{pmatrix} \quad \text{for any } \gamma \in W = L(\alpha_1, \dots, \alpha_n) \tag{50a}$$

$$\|\gamma\| = \left\{ \sum_{j=1}^{n} |\langle \gamma, \alpha_j \rangle|^2 \right\}^{1/2} \quad \text{for any } \gamma \in L(\alpha_1, \dots, \alpha_n) \tag{50b}$$

The Gram–Schmidt process allows us to construct an orthonormal basis of any finite-dimensional subspace W of $(V, \langle *, * \rangle)$.

Theorem 35. Gram–Schmidt: *Let $\alpha_1, \dots, \alpha_n$ be linearly independent vectors in $(V, \langle *, * \rangle)$. Then there exist pairwise orthogonal vectors $\lambda_1, \dots, \lambda_n$ such that $L(\alpha_1, \dots, \alpha_j) = L(\gamma_1, \dots, \gamma_j)$ for $j = 1, \dots, n$.*

The vectors $\lambda_1, \dots, \lambda_n$ in Theorem 35 are defined inductively as follows: $\lambda_1 = \alpha_1$. Having defined $\lambda_1, \dots, \lambda_r$, λ_{r+1} is defined by the following equation:

$$\lambda_{r+1} = \alpha_{r+1} - \sum_{j=1}^{r} \left(\frac{\langle \alpha_{r+1}, \lambda_j \rangle}{\langle \lambda_j, \lambda_j \rangle} \right) \lambda_j \tag{51}$$

To produce an orthonormal basis for $L(\alpha_1, \dots, \alpha_n)$, replace $\lambda_1, \dots, \lambda_n$ by $\lambda_1/\|\lambda_1\|, \dots, \lambda_n/\|\lambda_n\|$.

Theorem 35 can be used to construct the orthogonal complement of a subspace W.

Theorem 36. *Let $(V, \langle *, * \rangle)$ be a finite-dimensional inner-product space. Let W be a subspace of V. Then there exists a unique subspace $W' \subseteq V$ such that*

1. $W + W' = V$.
2. $W \cap W' = (0)$.
3. *Every vector in W is orthogonal to every vector in W'.*

The unique subspace W' given in Theorem 36 is called the orthogonal complement of W and written $W^{\perp}$. Clearly, $\dim(W) + \dim(W^{\perp}) = \dim V$.

3.4.3 Least-Squares Problems

There are three main problems in numerical linear algebra:

1. Find effective methods for solving linear systems of equations $AX = B$.
2. Find methods for computing eigenvalues of a square matrix A.
3. Find effective methods for solving least-squares problems.

We have already talked about the first two problems. We will now consider the third problem.

Suppose W is a subspace of some inner-product space $(V, \langle *, * \rangle)$. Let $\alpha \in V$. Is there a vector $P(\alpha) \in W$ which is closest to α? In other words, is there a vector $P(\alpha) \in W$ such that $\|\alpha - P(\alpha)\| = \min\{\|\alpha - \beta\| \mid \beta \in W\}$? If $V = \mathbb{R}^n$ and $\langle \alpha, \beta \rangle = \alpha^t \beta$, then $\|\alpha - \beta\|^2 = \sum_{i=1}^{n} (a_i - x_i)^2$. Here $\alpha = (a_1, \dots, a_n)^t$ and $\beta = (x_1, \dots, x_n)^t \in W$. Finding a vector $P(\alpha)$ in W which is closest to α is equivalent to finding $(x_1, \dots, x_n)^t \in W$ such that $(a_1 - x_1)^2 + \dots + (a_n - x_n)^2$ is as small as possible. Thus, we are trying to minimize a sum of squares. This is where the name "least-squares problem" originates.

If $\dim(W) = \infty$, there may be no vector in W which is closest to α. For a concrete example, see Brown [2, p. 212]. If W is finite dimensional, then there is a unique vector $P(\alpha)$ in W closest to α.

Theorem 37. *Let $(V, \langle *, * \rangle)$ be an inner-product space and let W be a finite-dimensional subspace of V. Let $\alpha \in V$. Then there exists a unique vector $P(\alpha) \in W$ such that $\|\alpha - P(\alpha)\| = \min\{\|\alpha - \beta\| \mid \beta \in W\}$. Furthermore, if $\{\alpha_1, \dots, \alpha_n\}$ is any pairwise, orthogonal basis of W, then*

$$P(\alpha) = \sum_{j=1}^{n} \left(\frac{\langle \alpha, \alpha_j \rangle}{\langle \alpha_j, \alpha_j \rangle} \right) \alpha_j$$

The unique vector $P(\alpha)$ satisfying Theorem 37 is called the orthogonal projection of α onto W. The map $P_W(*) : V \to V$ given by $P_W(\alpha) = P(\alpha)$ for all $\alpha \in V$ is called the orthogonal projection of V onto W. This map satisfies the following properties:

1. $P_W \in \mathrm{Hom}(V, V)$.
2. $\alpha - P_W(\alpha)$ is orthogonal to W for every $\alpha \in V$.
3. $\mathrm{Im}(P_W) = W$, $\mathrm{Ker}(P_W) = W^{\perp}$.
4. $P_W^2 = P_W$. (52)

Theorem 37 has important applications in the theory of linear equations. Suppose $A \in M_{m\times n}(\mathbb{C})$ and $B \in \mathbb{C}^m$.

Definition 38. *A vector $\xi \in \mathbb{C}^n$ is called a least-squares solution to $AX = B$ if $\|A\xi - B\| \le ||A\lambda - B||$ for all $\lambda \in \mathbb{C}^n$.*

Here $\| * \|$ is the induced norm from the standard inner product $\langle \alpha, \beta \rangle = \alpha^t \bar{\beta}$ on $\mathbb{C}^n$. Thus, ξ is a least-squares solution to $AX = B$ if and only if $A\xi = P_{CS(A)}(B)$. In particular, Theorem 37 guarantees least-squares solutions always exist. If $B \in CS(A)$, then $AX = B$ is consistent, i.e., there exists a vector $\lambda = \mathbb{C}^n$ such that $A\lambda = B$. In this case, any least-squares solution to $AX = B$ is an ordinary solution to the system.

Theorem 38. *Let $A \in M_{m\times n}(\mathbb{C})$ and $B \in \mathbb{C}^m$. A vector $\xi \in \mathbb{C}^n$ is a least-squares solution to $AX = B$ if and only if $\bar{\xi}$ is a solution to $(A^t\bar{A})X = A^t\bar{B}$. The least-squares solution is unique if* $\mathrm{rk}(A) = n$.

The equations $(A^t\bar{A})X = A^t\bar{B}$ are called the normal equations of A. Theorem 38 implies the solutions of the normal equations determine the least-squares solutions to $AX = B$. Solutions to the normal equations when $\mathrm{rk}(A) < n$ have an extensive literature. For applications to curve fitting, see Brown [1].

3.4.4 Normal Matrices

In this section, $F = \mathbb{R}$ or $\mathbb{C}$. $\langle *, * \rangle$ will always denote the standard inner product on F^n. Thus, $\langle \alpha, \beta \rangle = \alpha^t \beta$ if $F = \mathbb{R}$ and $\langle \alpha, \beta \rangle = \alpha^t \bar{\beta}$ if $F = \mathbb{C}$. If $A \in M_{n\times n}(F)$, then A^* will denote the Hermitian conjugate of A. Thus, $A^* = A^t$ if the entries of A are all real numbers and, in general, $A^* = (\bar{A})^t$. There is an important relationship between the standard inner product and A and A^*.

Theorem 39. *Let $A \in M_{n\times n}(F)$. Then $\langle A\alpha, \beta \rangle = \langle \alpha, A^*\beta \rangle$ for all $\alpha, \beta \in F^n$.*

Definition 39. *Let $A \in M_{n\times n}(\mathbb{C})$. A is unitary if $AA^* = A^*A = I_n$.*

If the entries of A in Definition 39 are all real [i.e., $A \in M_{n\times n}(\mathbb{R})$] and A is unitary, then $AA^t = A^tA = I_n$. In this case, A is called an orthogonal matrix. The following theorem characterizes unitary and orthogonal matrices.

Theorem 40. *Suppose $A = (\xi_1 \mid \ldots \mid \xi_n) \in M_{n\times n}(\mathbb{C})$. Then the following statements are equivalent:*

1. *A is unitary.*
2. *$\langle A\alpha, A\beta \rangle = \langle \alpha, \beta \rangle$ for all $\alpha, \beta \in \mathbb{C}^n$.*
3. *$\{\xi_1, \ldots, \xi_n\}$ is an orthonormal basis of $\mathbb{C}^n$.*
4. *$A = M(B, C)$, a change-of-basis matrix between two orthonormal bases B and C of $\mathbb{C}^n$.*

The same theorem is true with $\mathbb{C}$ replaced by $\mathbb{R}$ and unitary replaced by orthogonal. An important corollary to Theorem 40 is the following observation. If A is unitary, then

$$\mathcal{S}_{\mathbb{C}}(A) \subseteq \{z \in \mathbb{C} \mid |z| = 1\} \tag{53}$$

Definition 40. *Let $A \in M_{n\times n}(\mathbb{C})$. A is Hermitian if $A = A^*$.*

Theorem 41. *Let $A \in M_{n\times n}(\mathbb{C})$. If A is Hermitian, then $\mathcal{S}_{\mathbb{C}}(A) \subseteq \mathbb{R}$.*

If the entries in A are real, then $A = A^*$ if and only if A is symmetric. Theorem 41 implies any real, symmetric matrix has all of its eigenvalues in $\mathbb{R}$. Here is a handy chart of the complex and real names of some important types of matrices:

$M_{n\times n}(\mathbb{C})$	$M_{n\times n}(\mathbb{R})$
A unitary: $A^*A = I_n$	A orthogonal: $A^tA = I_n$
A Hermitian: $A = A^*$	A symmetric: $A = A^t$
A skew-Hermitian: $A^* = -A$	A skew-symmetric: $A^t = -A$

These are all special cases of normal matrices.

Definition 41. *Let $A \in M_{n\times n}(\mathbb{C})$. A is normal if $AA^* = A^*A$.*

Theorem 42. Schur:

1. *Let $A \in M_{n\times n}(\mathbb{R})$ such that $\mathcal{S}_{\mathbb{C}}(A) \subseteq \mathbb{R}$. Then there exists an orthogonal matrix P such that $P^t AP$ is upper triangular.*
2. *Let $A \in M_{n\times n}(\mathbb{C})$. There exists a unitary matrix P such that P^*AP is upper triangular.*

Notice the difference between the two theorems. If $F = \mathbb{C}$, there are no hypotheses on A. Any (square) matrix is unitarily similar to an upper-triangular matrix. The corresponding theorem for real matrices cannot be true. The matrix

$$A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \in M_{2\times 2}(\mathbb{R})$$

is not similar to any upper-triangular matrix since $\mathcal{S}_{\mathbb{R}}(A) = \emptyset$. However, if all eigenvalues of A (in $\mathbb{C}$) in fact are real numbers, then 1 implies A is orthogonally similar to an upper-triangular matrix. For example, Theorems 41 and 42(1) imply that any symmetric matrix $A \in M_{n\times n}(\mathbb{R})$ is orthogonally similar to a diagonal matrix. In fact, more is true.

Theorem 43. *Let $A \in M_{n\times n}(\mathbb{C})$. A is normal if and only if there exists a unitary matrix P such that P^*AP is diagonal.*

In particular, Hermitian and skew-Hermitian matrices are unitarily similar to diagonal matrices.

We conclude this section with an easy application of Theorem 43.

Theorem 44. *Let $A \in M_{n\times n}(\mathbb{C})$ be a normal matrix.*

1. *A is Hermitian if and only if $\mathcal{S}_{\mathbb{C}}(A) \subseteq \mathbb{R}$.*
2. *A is unitary if and only if $\mathcal{S}_{\mathbb{C}}(A) \subseteq \{z \in \mathbb{C} \mid |z| = 1\}$.*

3.5 FURTHER READING

This chapter consists of definitions and theorems that would normally be found in a junior level course in linear algebra. For more advanced courses the reader could try Brown [2] or Greub [5]. For an introduction to the theory of matrices over arbitrary commutative rings, see Brown [3]. For a basic treatment of numerical results, see Cullen [4]. For a more advanced level treatment of numerical results, see Demmel [6].

REFERENCES

1. WC Brown. Matrices and Vector Spaces. Pure and Applied Mathematics, vol 145. New York: Marcel Dekker, 1991.
2. WC Brown. A Second Course In Linear Algebra. New York: John Wiley & Sons, 1988.
3. WC Brown. Matrices Over Commutative Rings. Pure and Applied Mathematics, vol 169. New York: Marcel Dekker, 1993.
4. CG Cullen. An Introduction to Numerical Linear Algebra. Boston: PWS Publishing, 1994.
5. W Greub. Linear Algebra. Graduate Texts in Mathematics, vol 23, 4th ed. New York: Springer-Verlag, 1981.
6. J Demmel. Numerical Linear Algebra. Berkeley Mathematics Lecture Notes, vol 1, University of California, Berkeley, CA, 1993.

Chapter 1.4

A Review of Calculus

Angelo B. Mingarelli
School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada

4.1 FUNCTIONS, LIMITS, AND CONTINUITY

4.1.1 Functions and Their Properties

A *function* is a rule which associates with each object of one set, called the *domain* [denoted by the symbol Dom(f)], a single object $f(x)$ from a second set called the *range* [denoted by the symbol, Ran(f)]. All functions will be *real valued* in this chapter. This means that their range is always a subset of the set of all real numbers, while their domain is always some interval. We recall the notation for intervals; the symbol (a, b) denotes the set of points $\{x : a < x < b\}$, and this is called an *open interval*, while $[a, b]$ represents the set $\{x : a \leq x \leq b\}$, which is called a *closed interval*. On the other hand, the symbols $(a, b], [a, b)$ each denote the sets $\{x : a < x \leq b\}$ and $\{x : a \leq x < b\}$, respectively (either one of these is called a *semiopen interval*). The rules $f(x) = x^3, g(x) = \cos x, h(x) = \sqrt{x}$ are various examples of functions, with $h(x)$ being defined only when $x \geq 0$. The *sum of two functions*, f, g, say, is defined by the rule $(f + g)(x) = f(x) + g(x)$ with a similar definition being applied to the difference. The operation known as the *product of two functions*, f, g, say, is now defined by the rule $(fg)(x) = f(x)\,g(x)$. For example, with f, g as above, their sum, $(f + g)(x) = x^3 + \cos x$, whereas their product $(fg)(x) = x^3 \cos x$. The *quotient of two functions* is only defined when the denominator is nonzero. In general, $(f/g)(x) = f(x)/g(x)$ represents the quotient of f, g, while in our case, $(f/g)(x) = x^3 \sec x$, which is only defined when $\cos x \neq 0$. When c is a constant (a real number), the symbol cf is defined by $(cf)(x) = cf(x)$. In particular, the *identity function*, denoted by the symbol "1," is defined by the rule $1(x) = x$. An important function in calculus is the so-called *absolute value* function; it is defined by the rule: $|x| = x, x \geq 0$, while, if $x < 0$, $|x| = -x$. In either case, the absolute value of a number is that same number (if it is positive) or the original *unsigned* number (with its minus sign changed to a plus sign). Thus, $|-5| = -(-5) = 5$, while $|3.45| = 3.45$. When using square roots we will always take it that $\sqrt{x^2} = |x|$, for any x.

Another operation which is available on two specified functions is that of *composition*. We recall this notion here: given two functions, f, g where the range of g is contained in the domain of f, we define the *composition of f and g*, denoted by the symbol $f \circ g$, whose values are given by $(f \circ g)(x) = f(g(x))$. As an example, let $f(x) = x^2 + 1, g(x) = x - 1$. then $(f \circ g)(x) = f(g(x)) = g(x)^2 + 1 = (x - 1)^2 + 1$. On the other hand, $(g \circ f)(x) = g(f(x)) = f(x) - 1 = x^2$ and this shows that the operation of *composition is not commutative*, that is, $(g \circ f)(x) \neq (f \circ g)(x)$, in general.

Let f, F be two given function with domains, Dom(f), Dom(F), and ranges, Ran(f), Ran(F). We say that *f (resp. F) is the inverse function of F (resp. f)* if both their compositions give the identity function, that is, if $(f \circ F)(x) = (F \circ f)(x) = x$ [and, as is usual, Dom(f) = Ran(F) and Dom(F) = Ran(f)].

Sometimes this relation is written as $(f \circ f^{-1})(x) = (f^{-1} \circ f)(x) = x$. For instance, the functions f, F defined by the rules $f(x) = x^2$ and $F(x) = \sqrt{x}$ are inverses of one another because their composition is the identity function. In order that two functions f, F be inverses of one another it is necessary that each function be *one-to-one* on their respective domains. This means that the *only* solution of the equation $f(x) = f(y)$ [resp. $F(x) = F(y)$] is the solution $x = y$, whenever x, y are in Dom(f), [resp. Dom(F)]. The simplest geometrical test for deciding whether a given function is one-to-one is the so-called *horizontal line test*. Basically, one looks at the graph of the given function on the xy-plane, and if every horizontal line through the range of the function intersects the graph at only one point, then the function is one-to-one and so it has an inverse function. The *graph of the inverse function* is obtained by reflecting the graph of the original function in the xy-plane about the line $y = x$.

At this point we introduce the notion of the *inverse of a trigonometric function*. The graphical properties of the sine function indicate that it has an inverse when $\mathrm{Dom}(\sin) = [-\pi/2, \pi/2]$. Its inverse is called the *arcsine function* and it is defined for $-1 \leq x \leq 1$ by the rule that $y = \arcsin x$ means that y is an angle whose sine is x. Thus $\arcsin(1) = \pi/2$, since $\sin(\pi/2) = 1$. The cosine function with $\mathrm{Dom}(\cos) = [0, \pi]$ has an inverse called the *arccosine function*, also defined for $-1 \leq x \leq 1$, whose rule is given by $y = \arccos x$ which means tht y is an angle whose cosine is x. Thus, $\arccos(1) = 0$, since $\cos(0) = 1$. Finally, the tangent function defined on ($-\pi/2, \pi/2)$ has an inverse called the *arctangent function* defined on the interval $(-\infty, +\infty)$ by the statement that $y = \arctan x$ only when y is an angle in $(-\pi/2, \pi/2)$ whose tangent is x. In particular, $\arctan(1) = \pi/4$, since $\tan(\pi/4) = 1$. The remaining inverse trigonometric functions can be defined by the relations $y = \operatorname{arccot} x$, the *arccotangent function*, only when y is an angle in $(0, \pi)$ whose cotangent is x (and x is in $(-\infty, +\infty)$). In particular, $\operatorname{arccot}(0) = \pi/2$, since $\cot(\pi/2) = 0$. Furthermore, $y = \operatorname{arcsec} x$, the *arcsecant function*, only when y is an angle in $[0, \pi]$, different from $\pi/2$, whose secant is x (and x is outside the closed interval $[-1, 1]$). In particular, $\operatorname{arcsec}(1) = 0$, since $\sec 0 = 1$. Finally, $y = \operatorname{arccsc}(1)x$, the *arccosecant function*, only when y is an angle in $[-\pi/2, \pi/2]$, different from 0, whose cosecant is x (and x is outside the closed interval $[-1, 1]$). In particular, $\operatorname{arccsc}(1) = \pi/2$, since $\csc(\pi/2) = 1$. Moreover,

$$\sin(\arcsin x) = x, \quad -1 \leq x \leq 1 \quad \arcsin(\sin x) = x, \quad -\pi/2 \leq x \leq \pi/2 \tag{1}$$

$$\cos(\arccos x) = x, \quad -1 \leq x \leq 1 \quad \arccos(\cos x) = x \quad 0 \leq x \leq \pi \tag{2}$$

$$\tan(\arctan x) = x, \quad -\infty < x < +\infty \quad \arctan(\tan x) = x, \quad -\pi/2 < x < \pi/2 \tag{3}$$

$$\cot(\operatorname{arccot} x) = x, \quad -\infty < x < +\infty \quad \operatorname{arccot}(\cot x) = x, \quad 0 < x < \pi \tag{4}$$

$$\sec(\operatorname{arcsec} x) = x \quad |x| \geq 1 \quad \operatorname{arcsec}(\sec x) = x, \quad 0 \leq x \leq \pi, x \neq \pi/2 \tag{5}$$

$$\csc(\operatorname{arccsc} x) = x, \quad |x| \geq 1 \quad \operatorname{arccsc}(\csc x) = x, \quad -\pi/2 \leq x \leq \pi/2, x \neq 0 \tag{6}$$

$$\arccos x + \arcsin x = \pi/2, \quad -1 \leq x \leq 1 \tag{7}$$

$$\operatorname{arccot} x + \arctan x = \pi/2, \quad -\infty < x < +\infty \tag{8}$$

$$\operatorname{arcsec} x + \operatorname{arccsc} x = \pi/2, \quad |x| \geq 1 \tag{9}$$

$$\sin(\arccos x) = \cos(\arcsin x) = \sqrt{1 - x^2}, \quad -1 \leq x \leq 1 \tag{10}$$

Note that *other notations for an inverse function include the symbol f^{-1} for the inverse function of f*, whenever it exists. This is not to be confused with the *reciprocal function*. We used F in this section, and $\arcsin(x)$ instead of $\sin^{-1} x$, in order to avoid this possible confusion.

A relation between two variables say, x, y, is said to be an *implicit relation* if there is an equation connecting the two variables which forms the locus of a set of points on the xy-plane which may be a self-intersecting curve. For example, the locus of points defined by the implicit relation $x^2 + y^2 - 9 = 0$ forms a circle of radius equal to 3. We can then isolate one of the variables x or y, say x, call it an *independent variable* and then have, in some cases, y being a function of x (y then is called a *dependent variable*, because the value of y depends on the actual value of x chosen). When this happens we say that y is defined implicitly as a function of x or *y is an implicit function of x*. In Sec. 4.2.1 we will use the chain rule for derivatives to find the derivative of an implicit function.

4.1.2 Finite Limits

Let f be a given real-valued function whose domain is an interval I of the real line. Most of calculus may be reduced to the notion of *limits*. Let a, L be real numbers. We say that *the function f has the limit L as x approaches a* (or *the limit of f as x approaches a exists and is equal to L*) if, for any given $\varepsilon > 0$, no matter how small, one can find a corresponding number $\delta > 0$ with the property that whenever $|x - a| < \delta$ we have $|f(x) - L| < \varepsilon$. In the same spirit, we say that f has a *limit, L, from the right (resp. left) as x approaches a* if for any given $\varepsilon > 0$, no matter how small, one can find a corresponding number $\delta > 0$ with the property that whenever $0 < x - a < \delta$ (resp. $-\delta < x - a < 0$), we have $|f(x) - L| < \varepsilon$. The symbols for the *limit, limit from the right*, and *limit from the left* at a are denoted respectively by

$$\lim_{x \to a} f(x) = L \qquad \lim_{x \to a+} f(x) = L \qquad \lim_{x \to a-} f(x) = L$$

Fundamental in the theory of limits is the fact that a function f has the limit L as x approaches a, if and only if each of the left- and right-hand limits exist and are each equal to L. Furthermore, the limit of a sum (resp. difference, product) of two limits is the sum (resp. difference, product) of the individual limits. In the case of a quotient, the limit of a quotient of two functions is the quotient of the limits of each function if the limit of the denominator is nonzero.

For example, if H denotes the *Heaviside function* (Fig. 1) where $H(x) = +1$ when x is in $[0, \infty)$ and $H(x) = -1$ when x is in $[-\infty, 0)$, then

$$\lim_{x \to 0+} H(x) = +1 \qquad \lim_{x \to 0-} H(x) = -1$$

but the actual limit as x approaches 0 does not exist (since the left- and right-hand limits are unequal). On the other hand, the limit of $xH(x)$ as x approaches 0 is 0. Note that $xH(x) = |x|$, the absolute value of x.

Fundamental limits:

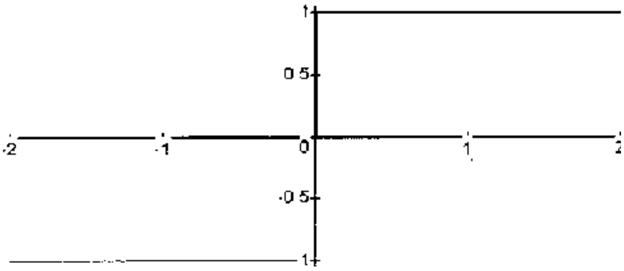


Figure 1 The Heaviside function.

$$\lim_{x \to 0} \frac{1 - \cos x}{x} = 0 \qquad \lim_{x \to 0} \frac{\sin x}{x} = 1 \tag{11}$$

More generally, if the function f has the limit 0 as x approaches 0, then

$$\lim_{x \to 0} \frac{1 - \cos(f(x))}{f(x)} = 0 \qquad \lim_{x \to 0} \frac{\sin(f(x))}{f(x)} = 1 \tag{12}$$

For example, it follows that

$$\lim_{x \to 0+} \frac{1 - \cos(\sqrt{x})}{\sqrt{x}} = 0 \qquad \lim_{x \to 0} \frac{\sin(x^3)}{x^3} = 1$$

If a is any number and $b \neq 0$ then

$$\lim_{x \to 0} \frac{\sin(ax)}{bx} = \frac{a}{b} \qquad \lim_{x \to 0} \frac{\sin(ax)}{\sin(bx)} = \frac{a}{b} \tag{13}$$

4.1.3 Infinite Limits and Limits at Infinity

Let f be a function defined to the left (resp. right) of $x = a$. We say that *the function f has the limit* $+\infty$ *as x approaches a from the left (resp. right)* [or *the limit of f as x approaches a from the left (resp. right) exists and is equal to* $+\infty$] if for every given $X > 0$, there is a $\delta > 0$ such that whenever $-\delta < x - a < 0$ (resp. $0 < x - a < \delta$) we have $f(x) > X$. A similar definition applies to the case where the limit of f as x approaches a is $-\infty$. Explicitly, we say that *the function f has the limit* $-\infty$ *as x approaches a from the left (resp. right)* [or *the limit of f as x approaches a from the left (resp. right) exists and is equal to* $-\infty$] if, for every given $X > 0$, there is a $\delta > 0$ such that whenever $-\delta < x - a < 0$ (resp. $0 < x - a < \delta$) we have $f(x) < -X$. The symbols used to denote each one of these limits are, respectively,

$$\lim_{x \to a-} f(x) = +\infty \qquad \lim_{x \to a+} f(x) = +\infty$$
$$\lim_{x \to a-} f(x) = -\infty \qquad \lim_{x \to a+} f(x) = -\infty$$

If any one (or more) of the above limits exists, we call the line $x = a$ a *vertical asymptote* of the graph of f. Thus, the function f defined by $f(x) = 1/x$ has a vertical asymptote at $x = 0$, while $g(x) = (x - 3)/(x^2 - 4)$ has two vertical asymptotes (at $x = \pm 2$). In the preceding example, the limit of g as x approaches -2 from the left is $+\infty$ while, if x approaches -2 from the right, its limit is $-\infty$.

Now, let f be a real-valued function whose domain is an interval I of the form $(-\infty, a)$ or $(a, +\infty)$ where a is unspecified, and let L be a real number. We say that *the function f has the limit L as x approaches* $+\infty$ *(resp.* $-\infty$*)* [or *the limit of f as x approaches* ∞ *(resp.* $-\infty$*)*

exists and is equal to L] if, for any given $\varepsilon > 0$, there is a value of x, say X, such that whenever $x > X > 0$ (resp. $x < X < 0$) we have $|f(x) - L| < \varepsilon$. The symbols used to denote these limits are respectively,

$$\lim_{x\to\infty} f(x) = L \qquad \lim_{x\to-\infty} f(x) = L$$

If either one (or both) of the above limits exists, and $y = f(x)$, we call the line $y = L$ a *horizontal asymptote* of the graph of f. Thus, once again, the function f defined by $f(x) = 1/x$ has the line $y = 0$ as a horizontal asymptote, while if $f(x) = (x^2 + 4)/(x^2 - 4)$ then the graph of f has the two vertical asymptotes at $x = \pm 2$ and the line $y = 1$ as a horizontal asymptote.

The *Euler exponential function*, $\exp(x)$, or e^x, may be defined by means of the following limits:

$$\begin{aligned}\lim_{h\to 0}(1 + xh)^{1/h} &= \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n \\ &= e^x \qquad -\infty < x < \infty\end{aligned} \tag{14}$$

All other exponential functions of the form $f(x) = a^x$ may be defined in terms of Euler's exponential function via the relation $a^x = e^{x \ln a}$, where $\ln x$ or $\log x$ is the inverse function of the exponential function, called the *natural logarithm*. Note that $\mathrm{Dom}(e^x) = -\infty < x < +\infty$, while its range is $0 < x < +\infty$. It follows that $\mathrm{Dom}(\log x) = 0 < x < +\infty$, while its range is $-\infty < x < +\infty$. To convert a logarithm from a given base $a > 0$ to base e, where $e = 2.7182818284590\ldots$ and vice versa, we use the *change-of-base formula*,

$$\log_a x = \frac{\ln x}{\ln a} \tag{15}$$

Figure 2 shows graphs of a function of type a^x.

Some limits:

$$\lim_{x\to\infty} e^{\alpha x} x^{-\beta} = 0 \qquad \text{if } \alpha \le 0, \beta > 0 \tag{16}$$

$$\lim_{x\to\infty} \frac{\log x}{x^\beta} = 0 \qquad \text{if } \beta > 0 \tag{17}$$

$$\lim_{x\to 0} x^\beta \log x = 0 \qquad \text{if } \beta > 0 \tag{18}$$

4.1.4 Continuity

If f is defined on an interval including the number a, then we say that f *is continuous* at a if the limit of f as x approaches a exists and is equal to L and, in addition, $L = f(a)$. Intuitively, continuity at a means that the *graph of f has no "break" or is "not infinite" at* $x = a$. A function f is called *discontinuous* at a if f is not continuous at a. The notions of *right- and left-continuity* of a function f at a point a are defined using the analogous right- and left-hand limits discussed in Sec. 4.1.2. For example, the Heaviside function, H, introduced in Sec. 4.1.2, is continuous whenever $x < 0$ and $x > 0$ but H is not continuous at $x = 0$ since its right-

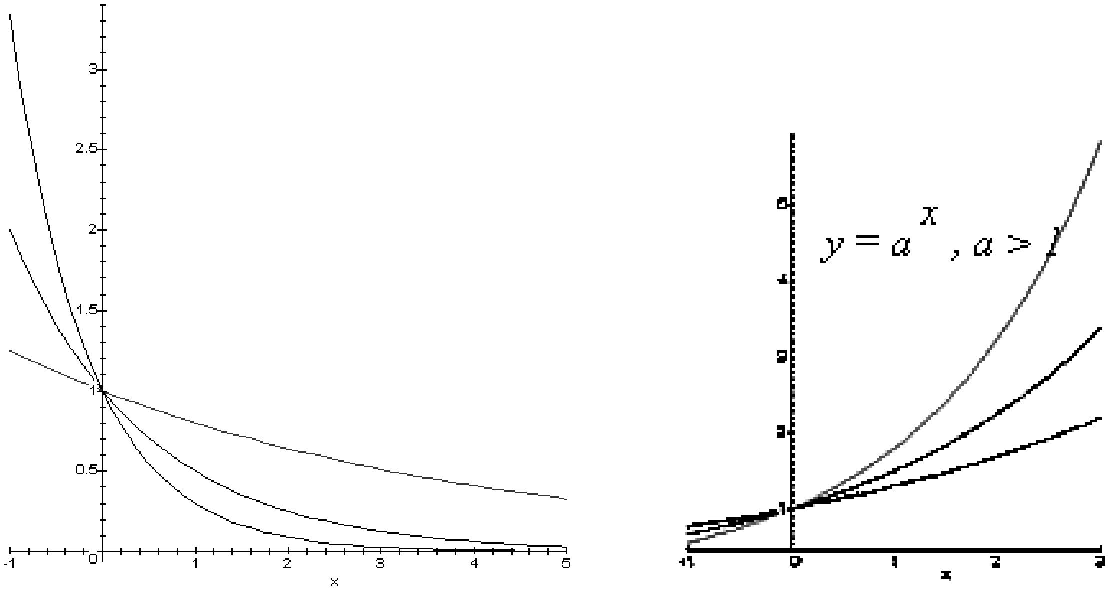


Figure 2 The function $y = a^x$: left, $a < 1$; right, $a > 1$.

and left-hand limits differ there. One crucial property of continuous functions is the so-called *intermediate-value property* which states that a function f which is continuous on $[a, b]$ takes on every value in its range, or, in particular, between the points $f(a)$ and $f(b)$. Use of this result shows *Balzano's theorem*, namely, that if $f(a) > 0$ and $f(b) < 0$ then there must be a root of f, say c, inside (a, b), that is, if $f(a) > 0$ and $f(b) < 0$ then there is a point c such that $f(c) = 0$, a result which is very useful in the practical problem of finding the roots of various functions (see Sec. 4.2.3).

Every *polynomial* of degree $n \geq 0$ with real coefficients, that is, every expression of the form

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

where $a_0, a_1, \ldots, a_n$ are real numbers, is continuous on the real line (i.e., at *every point* of the real line). Sums, differences, and products/quotients (with a nonzero denominator) of continuous functions give continuous functions, while every *rational function* (a quotient of any two polynomials) is continuous at every point where the denominator is nonzero. Thus, $f(x) = (x - 3)/(x^2 - 4)$ is continuous at every point x except when $x = \pm 2$. On the other hand, the slightly modified function $f(x) = (x - 3)/(x^2 + 4)$ is continuous at every point x or, more simply put, *continuous everywhere*.

The *composition of two continuous functions (see Sec. 4.1.1) is also continuous*, so that, for instance $h(x) = \sin(\cos x)$ is continuous for each x, since $h(x) = f(g(x))$ where $f(x) = \sin x$ and $g(x) = \cos x$. Euler's exponential function, e^x, is continuous on $(-\infty, +\infty)$, while its inverse function, the natural logarithm $\ln x$, is continuous on $(0, +\infty)$. The same is true of all other exponential functions of the form a^x where $a > 0$. Figure 3 is a generic graph of points of continuity and discontinuity.

4.2 DIFFERENTIABILITY AND THE CHAIN RULE

4.2.1 The Derivative

One of the most important definitions involving limits is that of the *derivative*. The derivative of a function f at the point $x = a$, denoted by $f'(a)$, or $df/dx(a)$, is defined by the two equivalent definitions

$$\begin{aligned} \frac{df}{dx}(a) = f'(a) &= \lim_{h \to 0} \frac{f(a+h) - f(a)}{h} \\ &= \lim_{x \to a} \frac{f(x) - f(a)}{x - a} \end{aligned} \tag{19}$$

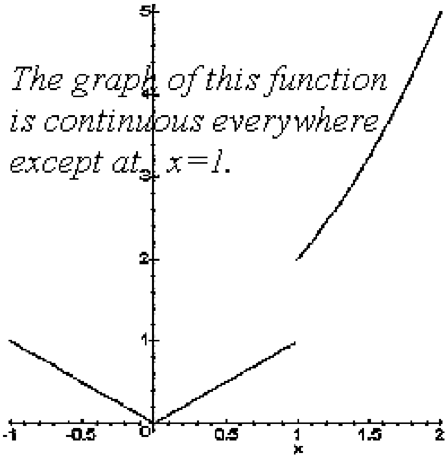

The graph of this function is continuous everywhere except at x=1.

Figure 3 A discontinuous function.

whenever either limit exists (in which case so does the other). The *right derivative* (resp. *left derivative*) is defined by the right-hand (resp. left-hand) limits

$$f'_+(a) = \lim_{h \to 0+} \frac{f(a+h) - f(a)}{h} = \lim_{x \to a+} \frac{f(x) - f(a)}{x - a} \tag{20}$$

and

$$f'_-(a) = \lim_{h \to 0-} \frac{f(a+h) - f(a)}{h} = \lim_{x \to a-} \frac{f(x) - f(a)}{x - a} \tag{21}$$

A function f is said to be *differentiable at the point a* if its derivative $f'(a)$ exists there. This is equivalent to saying that both the left- and right-hand derivatives exist at a and are equal. A function f is said to be *differentiable everywhere* if it is differentiable at every point a of the real line. For example, the function f defined by the absolute value of x, namely $f(x) = |x|$, is differentiable at every point except at $x = 0$ where $f'_-(0) = -1$ and $f'_+(0) = 1$. On the other hand, the function g defined by $g(x) = x|x|$ is differentiable everywhere.

The derivative of the derivative of a given function f at $x = a$ is called the *second derivative* of f at $x = a$ and is denoted by $f''(a)$. The derivative of the second derivative is called the *third derivative* [denoted by $f'''(a)$] and so on. The function g defined above by $g(x) = x|x|$ does not have a second derivative at $x = 0$ [i.e., $f''(0)$ does not exist] even though it is differentiable there. It is a fundamental fact that *if $f'(a)$ exists then f is continuous at a.* First derivatives may be thought of as the *velocity* or as the slope of the tangent line to the graph of the function $y = f(x)$ at the point $x = a$, while second derivatives appear in physical applications under the name of *acceleration*.

The *binomial theorem* states that if n is a positive integer,

$$(x+y)^n + \sum_{r=0}^{n} C_{n,r} x^{n-r} y^r \tag{22}$$

where $C_{n,r}$ denotes the *binomial coefficients* defined by $C_{n,0} = 1$ and, for $r > 0$,

$$C_{n,r} = \frac{n!}{r!(n-r)!} \tag{23}$$

As usual, $r!$ denotes the *factorial symbol*, that is, it is the product of the first r numbers, $1, 2, \ldots, (r-1)$. The binomial theorem allows one to prove the *power rule*, namely, that if k is any real number,

$$\frac{d}{dx} x^k = k x^{k-1} \tag{24}$$

The derivative has the following properties (whenever it exists). Let f, g be any two given differentiable functions at the point x and let k be any real number; then,

$$(f \pm g)'(x) = f'(x) \pm g'(x) \quad \text{(sum/difference rule)} \tag{25}$$

$$(kf)'(x) = kf'(x) \tag{26}$$

$$(fg)'(x) = f'(x)\,g(x) + f(x)\,g'(x) \quad \text{(product rule)} \tag{27}$$

$$\left(\frac{f}{g}\right)'(x) = \frac{f'(x)\,g(x) - f(x)\,g'(x)}{g(x)^2} \quad \text{(quotient rule)} \tag{28}$$

The most useful of all the rules is the *chain rule*, which is used to find the derivative of the composition (see Sec. 4.1.1) of two or more functions. It states that if f, g are two given functions with f differentiable at the point $g(a)$ and g itself differentiable at $x = a$, then their composition $(f \circ g)$ is also differentiable at a and its value there is given by

$$(f \circ g)'(a) = f'(g(a))\,g'(a) \tag{29}$$

It is sometimes written in the form

$$\frac{d}{dx} f(g(x)) = f'(g(x))\,g'(x) \quad \text{(chain rule)} \tag{30}$$

For example, if $f(x) = \sin u(x)$ where u is differentiable at x, then $f'(x) = \cos u(x)\, u'(x)$.

The *generalized power rule* states that for any differentiable function u,

$$\frac{d}{dx} u(x)^k = k\, u(x)^{k-1} \frac{du}{dx}, \qquad k \text{ constant} \tag{31}$$

If *f and F are inverse functions* then the relation $f(F(x)) = x$ and the chain rule shows that

$$F'(x) = \frac{1}{f'(F(x))}$$

In the next set of logarithmic and exponential expressions let u, v each be differentiable functions and a a constant; then

$$\begin{aligned} \frac{d}{dx} e^{u(x)} &= e^{u(x)} \frac{du}{dx} \\ \frac{d}{dx} \ln u(x) &= \frac{1}{u(x)} \frac{du}{dx} \qquad \text{if } u(x) > 0 \end{aligned} \tag{32}$$

$$\begin{aligned} \frac{d}{dx} a^{u(x)} &= a^{u(x)} \ln(a) \frac{du}{dx} \\ \frac{d}{dx} \log_a u(x) &= \frac{1}{u(x)\ln(a)} \frac{du}{dx} \qquad \text{if } u(x) > 0 \end{aligned} \tag{33}$$

$$\frac{d}{dx} v(x)^{u(x)} = v(x)^{u(x)} \left(\frac{u(x)}{v(x)} \frac{dv}{dx} + \ln(v(x)) \frac{du}{dx} \right) \quad \text{if } v(x) > 0 \tag{34}$$

If u is a differentiable function then

$$\begin{aligned} \frac{d}{dx} \sin u(x) &= \cos u(x) \frac{du}{dx} \\ \frac{d}{dx} \cos u(x) &= -\sin u(x) \frac{du}{dx} \end{aligned} \tag{35}$$

$$\begin{aligned} \frac{d}{dx} \tan u(x) &= \sec^2 u(x) \frac{du}{dx} \\ \frac{d}{dx} \cot u(x) &= -\csc^2 u(x) \frac{du}{dx} \end{aligned} \tag{36}$$

$$\begin{aligned} \frac{d}{dx} \sec u(x) &= \sec u(x) \tan u(x) \frac{du}{dx} \\ \frac{d}{dx} \csc u(x) &= -\csc u(x) \cot u(x) \frac{du}{dx} \end{aligned} \tag{37}$$

while, if $|u(x)| < 1$ then

$$\begin{aligned} \frac{d}{dx} \arcsin u(x) &= \frac{1}{\sqrt{1 - u(x)^2}} \frac{du}{dx} \\ \frac{d}{dx} \arccos u(x) &= -\frac{1}{\sqrt{1 - u(x)^2}} \frac{du}{dx} \end{aligned} \tag{38}$$

or, for any function u,

$$\begin{aligned}\frac{d}{dx}\arctan u(x) &= \frac{1}{1+u(x)^2}\frac{du}{dx}\\ \frac{d}{dx}\operatorname{arccot} u(x) &= -\frac{1}{1+u(x)^2}\frac{du}{dx}\end{aligned}\tag{39}$$

and if $|u(x)| > 1$,

$$\begin{aligned}\frac{d}{dx}\operatorname{arcsec} u(x) &= \frac{1}{|u(x)|\sqrt{u(x)^2-1}}\frac{du}{dx}\\ \frac{d}{dx}\operatorname{arccsc} u(x) &= -\frac{1}{|u(x)|\sqrt{u(x)^2-1}}\frac{du}{dx}\end{aligned}\tag{40}$$

We define the *hyperbolic functions* by setting

$$\sinh x = \frac{e^x - e^{-x}}{2} \qquad \cosh x = \frac{e^x + e^{-x}}{2}\tag{41}$$

with the remaining functions being defined by rules similar to the circular (trigonometric) functions; for example, $\operatorname{sech} x = 1/\cosh x$, $\tanh x = \sinh x/\cosh x$, etc. For these functions we have the following differentiation formulae:

$$\begin{aligned}\frac{d}{dx}\sinh u(x) &= \cosh u(x)\,\frac{du}{dx}\\ \frac{d}{dx}\cosh u(x) &= \sinh u(x)\,\frac{du}{dx}\end{aligned}\tag{42}$$

$$\begin{aligned}\frac{d}{dx}\tanh u(x) &= \operatorname{sech}^2 u(x)\frac{du}{dx}\\ \frac{d}{dx}\coth u(x) &= -\operatorname{csch}^2 u(x)\frac{du}{dx}\end{aligned}\tag{43}$$

$$\begin{aligned}\frac{d}{dx}\operatorname{sech} u(x) &= -\operatorname{sech} u(x)\tanh u(x)\frac{du}{dx}\\ \frac{d}{dx}\operatorname{csch} u(x) &= -\operatorname{csch} u(x)\coth u(x)\frac{du}{dx}\end{aligned}\tag{44}$$

The graphs of the hyperbolic sine and cosine functions are shown in Fig. 4, and those of the hyperbolic cotangent and cosecant functions in Fig. 5.

4.2.2 L'Hospital's Rule

We begin by defining the notion of an indeterminate form. A limit problem of the form

$$\lim_{x\to a}\frac{f(x)}{g(x)}$$

is called an *indeterminate form* if the expression $f(a)/g(a)$ is one of the following types: $\pm\infty/\infty$ or $0/0$. In either case it is sometimes possible to determine the limit by appealing to *L'Hospital's rule*. Before describing this rule, we define the notion of a *neighborhood of a point a*. Briefly stated, if a is finite, a neighborhood of a consists of an open interval (see Sec. 4.1.1) containing a. In the same vein, a *left-neighborhood* of $x = a$ consists of an open interval with a has its right endpoint [or an interval of the form $(a-\delta, a)$ where $\delta > 0$]. Similarly, a *right-neighborhood* of $x = a$ consists of an open interval with a has its left endpoint (or an interval of the form $(a, a+\delta)$ where $\delta > 0$). A *punctured neighborhood of a* is a set of points which is the union of two open intervals of the form $(a-\delta, a)$ and $(a, a+\mu)$

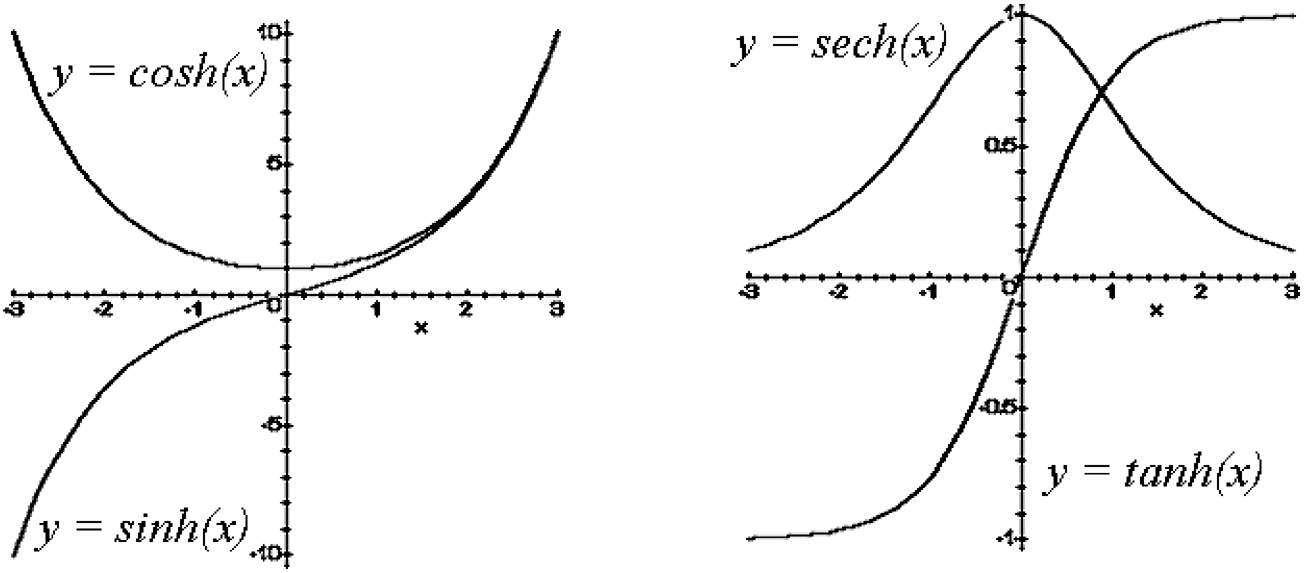


Figure 4 The hyperbolic sine and cosine functions.

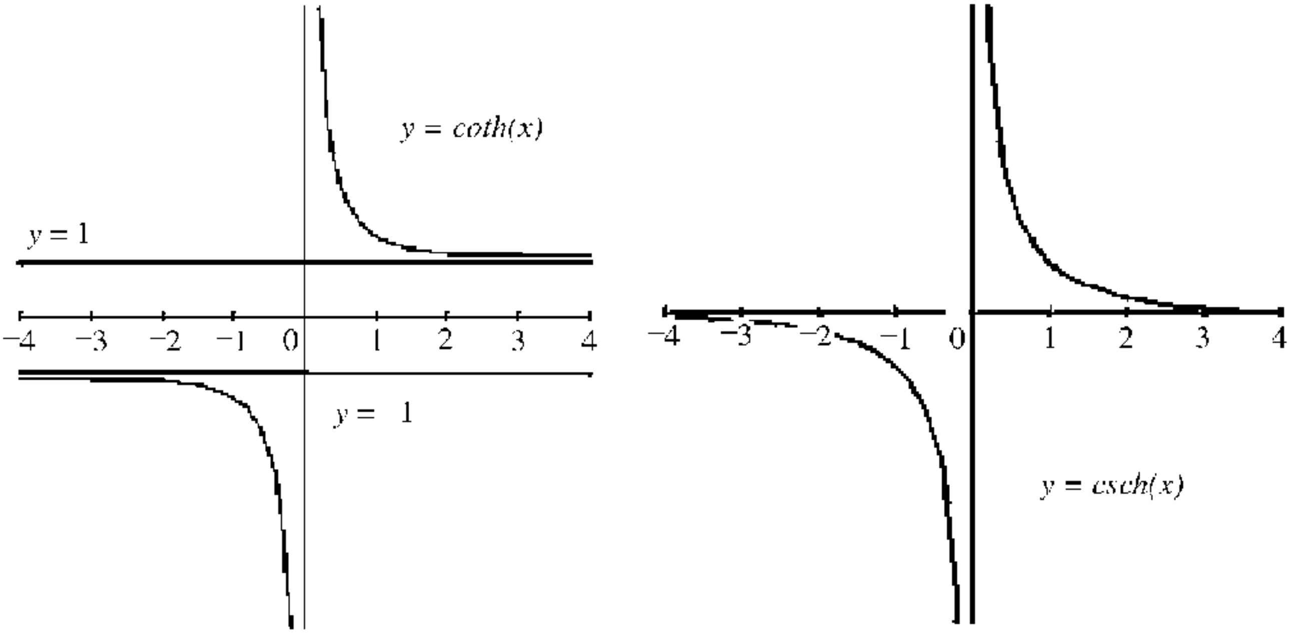


Figure 5 The hyperbolic contangent and cosecant.

where $\delta > 0, \mu > 0$ are not necessarily given. Notice that a punctured neighborhood does not include the point itself. Thus, for example, the union of the intervals $(-0.5, 0)$ and $(0, 0.2)$ is a punctured neighborhood of the point 0 while the interval $(-0.5, 0.2)$ is a neighborhood of 0.

Now, *L'Hospital's rule* may be stated as follows: let f, g be two functions defined and differentiable in a punctured neighborhood of a, where a is finite. If $g'(x) \neq 0$ in this punctured neighborhood of a and $f(a)/g(a)$ is one of the following types: $\pm\infty/\infty$ or $0/0$, then

$$\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{x \to a} \frac{f'(x)}{g'(x)} \tag{45}$$

provided the limit on the right of Eq. (45) exists (or is $+\infty$, or $-\infty$). The rule also holds if, *instead* of assuming that $f(a)/g(a)$ is of the type $\pm\infty/\infty$ or $0/0$ we only have that $\lim_{x\to a} f(x) = 0$ and $\lim_{x\to a} g(x) = 0$ or $\lim_{x\to a} f(x) = \pm\infty$ and $\lim_{x\to a} g(x) = \pm\infty$. The rule is also valid when the quantity a is replaced by $\pm\infty$, or even if the limits are *one-sided limits* (i.e., limit as x approaches a from the right or left, see Sec. 4.1.2). For example, the limits (11)–(13) can all be found using this rule. Other indeterminate forms such as 0^∞, 1^∞, and $\infty - \infty$ can sometimes be converted to indeterminate forms of the type $\pm\infty/\infty$ or $0/0$ by algebraic manipulation, taking logarithms, etc. In addition,

$$\lim_{x \to \infty} \left(\sqrt{ax+b} - \sqrt{ax+d} \right) = 0 \qquad \text{if } a > 0 \tag{46}$$

$$\lim_{x \to 0} \frac{e^x - 1}{x} = 1 \tag{47}$$

$$\lim_{x \to \infty} \left(1 + \frac{a}{x} \right)^{bx} = e^{ab} \tag{48}$$

$$\lim_{x \to \infty} \left(\frac{ax+b}{cx+d} \right) = \frac{a}{c} \qquad \text{if } c \neq 0 \tag{49}$$

$$\lim_{x \to 0+} x^{1/x} = 1 \tag{50}$$

$$\lim_{x \to 0} \frac{\tan(\alpha x)}{x} = \alpha \tag{51}$$

$$\lim_{n \to \infty} \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} - \ln n \right) = 0.57721\ldots \tag{52}$$

For example, in the case of Eq. (51),

$$\lim_{x \to 0} \frac{\tan(\alpha x)}{x} = \lim_{x \to 0} \frac{\alpha \sec^2(\alpha x)}{1} = \alpha$$

since the derivative of $\tan(\alpha x)$ is $\alpha \sec^2(\alpha x)$, the derivative of x is 1, and $\sec 0 = 1$.

4.2.3 Newton's Method for Finding Roots

In the event that one wants to find the roots of an equation of the form $f(x) = 0$, where f is given and differentiable in an open interval containing the root sought, there is a powerful technique which approximates the value of the root(s) to arbitrary precision. It is easily programmable and many subroutines exist on the market which do this for you. The idea is as follows: choose a point x_0 as a starting point (hopefully it is *close* to the desired root). With this value of x_0 define x_1 by setrting $x_1 = x_0 - f(x_0)/f'(x_0)$. We can now define x_2 by setting $x_2 = x_1 - f(x_1)/f'(x_1)$. This gives us the three values x_0, x_1, x_2, the last of which (namely, x_2) is closer to the desired root than the first (i.e., x_0). We define x_n, the nth term of this sequence of numbers, by

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} \qquad \text{for } n \geq 1 \tag{53}$$

If the sequence, x_n of numbers *converges* to a limit, say L, then *the limit, L, of the sequence defined by Eq. (53) is a root of the equation $f(x) = 0$ in the required interval*, that is, $f(L) = 0$. This is the basic idea of *Newton's method.*

For example, if $f(x) = x^3 - 2x - 1$ and we want to find a root of the equation $f(x) = 0$ near the point 1.5, then we find $f'(x)$ set up the iteration Eq. (53) for this function, and then check for convergence of the resulting sequence $x_n, n = 1, 2, 3, \ldots$ So we set $x_0 = 1.5$, from which the form of the iterative procedure given by Eq. (53) can be derived, namely,

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} = x_{n-1} - \frac{x_{n-1}^3 - 2x_{n-1} - 1}{3x_{n-1}^2 - 2},$$
$$\text{for } n \geq 1$$

In this case, $x_1 = 1.6315789\ldots, x_2 = 1.6181835\ldots$, $x_3 = 1.6180340\ldots, x_4 = 1.618033989\ldots$, $x_5 = 1.618033989\ldots$, with rapid convergence to the root closest to the initial value, 1.5, namely, the root whose value is approximately 1.618033989. On the other hand, had we chosen $x_0 = 1.2$, then $x_1, = 1.9206896\ldots$, $x_2 = 1.6731874\ldots$, $x_3 = 1.6203940\ldots$, $x_4 = 1.6180385\ldots$, $x_5 = 1.618033989\ldots$. Even in this case we get very good convergence after only a few terms.

4.2.4 Curve Sketching

We outline here the basic steps required in sketching a given planar curve defined by a function. A function f is said to be *increasing* (resp. *decreasing*) if given any pair of points $x_1 < x_2$ in its domain, we have $f(x_1) < f(x_2)$ [resp. $f(x_1) > f(x_2)$]. If f is differentiable on its domain then f is increasing on a given interval if its derivative is positive there, i.e., f is increasing (resp. decreasing) whenever $f'(x) > 0$ [resp. $f'(x) < 0$]. In each of these cases the slope of the tangent line at any point on the graph of $y = f(x)$ is positive (resp. negative). For example, if f is defined by $f(x) = x^2$ then f is increasing when $x > 0$ and decreasing when $x < 0$.

The graph of a differentiable function f is said to be *concave up* (resp. *concave down*) if given any pair of points $x_1 < x_2$ in the domain of its derivative, we have $f'(x_1) < f'(x_2)$ [resp. $f'(x_1) > f'(x_2)$]. If f is twice differentiable on its domain then f is concave up on a given interval if its second derivative is positive there, i.e., f is concave up (resp. concave down) whenever $f''(x) > 0$ [resp. $f''(x) < 0$]. For example, the graph of the function defined by $f(x) = x^3$ is concave up whenever $x > 0$ and concave down when $x < 0$. A point c is called a *critical point* of a function f defined on an interval I (containing c) if either $f'(c) = 0$ or f' (c) does not exist (either as a finite number, or as a two-sided limit). Examples of critical points are furnished by the following two examples: $f(x) = 1/x$ at $x = 0$ and $f(x) = x^2$ at $x = 0$. The continuous function defined by $f(x) = |x|$ has a critical point at $x = 0$ since it is not differentiable there (i.e., no two-sided limit of f' exists at $x = 0$).

A function f is said to have a *local maximum* at a point $x = a$ if there is a neighborhood of a in which $f(x) < f(a)$. In this case, the value of $f(a)$ is called the *local maximum value*. It is said to have a *global maximum* at $x = a$ if $f(x) < f(a)$ for every x in the domain of f. In this case the value of $f(a)$ is called the *global maximum value*. For example, if $f(x) = -x^2$ then f has a global maximum at $x = 0$ and this global maximum value is equal to 0. If we set $f(x) = (x-1)(x-2)(x-3)$ and $\text{Dom}(f) = [0, 5]$, then f has a local maximum at $x = 2 - 1/\sqrt{3}$, which is not a global maximum, since this occurs at $x = 5$. It f is differentiable we can check the nature of a critical point, a, of f by using the *first derivative test for a maximum*; that is, if $f'(x) > 0$ for x in a left neighborhood (Sec. 4.2.2) of a and $f'(x) < 0$ for x in a right neighborhood (Sec. 4.2.2) of a, then f has a local maximum at $x = a$. In the event that f is twice differentiable on its domain, there is the *second derivative test for a maximum*, which states that if $x = a$ is a critical point of f and $f''(a) < 0$ then it is a local maximum. The global maximum (and its value) is determined by taking that critical point c where $f(c)$ has the *largest* maximum value.

The function f is said to have a *local minimum* at a point $x = a$ if there is a neighborhood of a in which $f(x) > f(a)$. In this case, the value of $f(a)$ is called the *local minimum value*. It is said to have a *global minimum* at $x = a$ if $f(x) > f(a)$ for every x in the domain of f. In this case, the value of $f(a)$ is called the *global minimum value*. For example, if $f(x) = x^2$ then f has a global minimum at $x = 0$ and this global minimum value is equal to 0. If we set $f(x) = (x-1)(x-2)(x-3)$ and $\mathrm{Dom}(f) = [0, 5]$, then f has a local minimum at $x = 2 + 1/\sqrt{3}$ which is not a global minimum since this occurs at $x = 0$. If f is differentiable we can check the nature of a critical point, a, of f by using the *first derivative test for a minimum*; that is, if $f'(x) < 0$ for x in a left neighborhood (Sec. 4.2.2) of a and $f'(x) > 0$ for x in a right neighborhood of a, then f has a local minimum at $x = a$. In the event that f is twice differentiable on its domain, there is the *second derivative test for a minimum* which states that if $x = a$ is a critical point of f and $f''(a) > 0$ then it is a local minimum. The global minimum (and its value) is determined by taking that critical point c where $f(c)$ has the *smallest* minimum value.

A function f is said to have a *point of inflection* at $x = a$ if it changes its concavity around $x = a$, that is, if there is a left neighborhood of $x = a$ in which the graph of f is concave down and a right neighborhood of $x = a$ in which the graph of f is concave up, *or* if there is a left neighborhood of $x = a$ in which the graph of f is concave up and a right neighborhood of $x = a$ in which the graph of f is concave down. For example, if $f(x) = x^3$, the graph of f has a point of inflection at $x = 0$ but the graph of $f(x) = x^2$ does not (because the graph is *always* concave up around $x = 0$). If f is twice differentiable, a necessary (but not sufficient) condition for $x = a$ to be a point of inflection is that $f''(a) = 0$. In this case, we must then check around $x = a$ for a change in concavity. We recall the definitions of asymptotes as presented in Sec. 1.3. The usual rules for finding the graph of a function f now follow.

Find the intervals where f is increasing and decreasing.

Determine the critical points.

Find all local maxima and minima and points of inflection.

Find the roots of f (use may be made of Newton's method, Sec. 4.2.3).

Locate the intervals where the graph of f is concave up or down.

Find the vertical and horizontal asymptotes of f, if any (one may have to use L'Hospital's rule, Sec. 4.2.2). We give an example of a typical graph in Fig. 6, this one for the function f defined by $f(x) = 4x/(1 + x^2)$.

4.2.5 Implicit Differentiation

Implicit relations (Sec. 4.1.1) are useful because they define a *curve* in the xy-plane, a curve which is not, generally speaking, the graph of a function. For example, the circle of radius equal to 2 defined by the implicit relation $x^2 + y^2 = 4$ is not the graph of a unique function. In this case, assuming that y is some function of x, the derivative is found by repeated applications of the chain rule, and possibly all the other rules in this section as well. The basic idea is best described by an example. Assume that y can be written as a differentiable function of x and that there is some implicit relation like

$$y^3 + 7y = x^3$$

We take the derivative of both sides. We see that

$$3y^2 \frac{dy}{dx} + 7\frac{dy}{dx} = 3x^2$$

since

$$\frac{d}{dx}y^3 = 3y^2 \frac{dy}{dx}$$

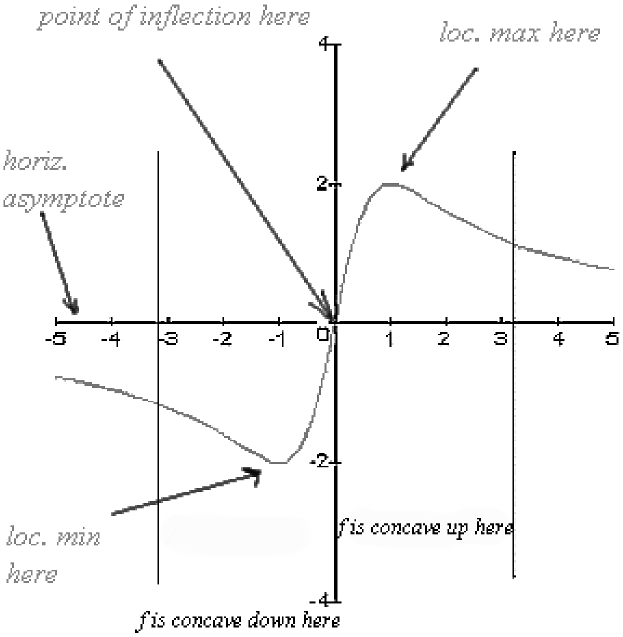


Figure 6 The function $f(x) = 4x/(1 + x^2)$.

by the generalized power rule. We can now solve for the expression dy/dx and find a formula for the derivative, namely,

$$\frac{dy}{dx} = \frac{3x^2}{3y^2 + 7}$$

Now we can find the derivative easily at any point (x, y) on the curve $y^3 + 7y = x^3$. For instance, the derivative at the point $(2, 1)$ on this curve is given by substituting the values $x = 2, y = 1$ in the formula for the derivative just found, so that $dy/dx = 6/5$. As usual, this represents the value of the slope of the tangent line to the curve $y^3 + 7y = x^3$ at the point $(2, 1)$. The graph of an implicit relation is then found by solving the relation for all possible pairs (x, y) satisfying the equation defining the relation, along with the derivative information gathered through implicit differentiation.

4.3 ANTIDERIVATIVES AND INTEGRATION

4.3.1 Antiderivatives

One can think of the antiderivative of a given function as a kind of *inverse* to the operation of differentiation, which we saw in Sec. 4.2. This notion is motivated by the so-called *fundamental theorem of calculus* which we will see below. Given a function f, whose domain is an interval $I = [a, b]$, we define its *antiderivative* as another function, F, also defined and continuous on $[a, b]$, differentiable on (a, b) and with the property that $F'(x) = f(x)$ for every point x in I (except possibly the endpoints). Not every function has an antiderivative. For example, it is known that the function f defined by $f(x) = e^{x^2}$ has no antiderivative that can be written as a sum of a *finite* number of algebraic expressions. Its antiderivative is, in fact given by an *infinite series*. Basic to this section is the fact that any continuous function or, more generally, any function which is continuous except for countably many points does have an antiderivative. For example, the function $F(x) = x^4$ is an antiderivative of $f(x) = 4x^3$, since $F'(x) = f(x)$ by the power rule. Although the derivative of a given function is unique when it is defined, this is not the case for antiderivatives. *All antiderivatives differ by a constant*. Thus if F is an antiderivative of f, then $F + c$ is also an antiderivative of f if c is a constant, by definition.

Since the antiderivative of a function f depends on f its symbol is denoted universally by

$$\text{An antiderivative of } f = F(x) = \int f(x)\,dx$$
$$= \int_a^x f(x)\,dx$$

where a is some point of the domain of f. The difference between two values of an antiderivative, say, $F(b) - F(a)$, is denoted by the symbol

$$\int_a^b f(x)\,dx = F(b) - F(a) \tag{54}$$

Thus, for example, $\int_a^b x\,dx = (b^2 - a^2)/2$, since $F(x) = x^2/2$ for $f(x) = x$. Various antiderivatives are displayed below. Note that, in every case, C denotes a constant.

$$\int e^x\,dx = e^x + C \qquad \int \frac{dx}{x} = \ln|x| + C \tag{55}$$

$$\int \sin x\,dx = -\cos x + C \qquad \int \cos x\,dx = \sin x + C \tag{56}$$

Use of the chain rule, Eq. (30), gives the more general formulae, analogous to those in Sec. 4.2.1,

$$\begin{aligned} &\int e^{u(x)} \frac{du}{dx}\,dx = e^{u(x)} + C \\ &\int \frac{1}{u(x)} \frac{du}{dx}\,dx = \ln u(x) + C, \quad \text{if } u(x) > 0 \end{aligned} \tag{57}$$

$$\begin{aligned} &\int \cos u(x) \frac{du}{dx}\,dx = \sin u(x) + C \\ &\int \sin u(x) \frac{du}{dx}\,dx = -\cos u(x) + C \end{aligned} \tag{58}$$

$$\begin{aligned} &\int \tan u(x) \frac{du}{dx}\,dx = -\ln|\cos u(x)| + C \\ &\int \sec u(x) \frac{du}{dx}\,dx = \ln|\sec u(x) + \tan u(x)| + C \end{aligned} \tag{59}$$

$$\begin{aligned} &\int \frac{1}{\sqrt{1 - x^2}} \frac{du}{dx}\,dx = \arcsin u(x) + C \\ &\int \frac{1}{1 + u(x)^2} \frac{du}{dx}\,dx = \arctan u(x) + C \end{aligned} \tag{60}$$

$$\int a^{u(x)} \frac{du}{dx}\,dx = \frac{a^{u(x)}}{\ln a} + C \qquad \text{if } a > 0 \tag{61}$$

with similar formulae for the rest of the functions in Sec. 4.2.1.

4.3.2 The Integral and Its Properties

Let $I = [a, b]$ be an interval. By a *partition, denoted by* Π, *of the interval* I we mean a subdivision of I into n subintervals

$$\Pi\colon [a, x_1], [x_1, x_2], [x_2, x_3], \ldots, [x_{n-1}, b]$$

where, by definition, $x_0 = a$, $x_n = b$. The points x_i can be chosen in any fashion whatsoever and do not have to be equally spaced within I. The *norm of the partition* Π, denoted by $\|\Pi\|$, is defined to be the length of the largest subinterval making up the partition Π. For example, if we let $I = [0, 1]$ and define Π to be the partition $[0, 1/5]$, $[1/5, 1/3]$, $[1/3, 1/2]$, $[1/2, 7/8]$, $[7/8, 1]$ then the norm of this partition is equal to $7/8 - 1/2 = 3/8 = 0.375$, which is the length of the largest subinterval contained within Π.

Given a partition Π of I and f a function defined on I, we define a *Riemann sum* as follows. Let ξ_i be a point in the subinterval $[x_{i-1}, x_i]$ and consider the sum

$$f(\xi_1)(x_1 - x_0) + f(\xi_2)(x_2 - x_1) + \ldots + f(\xi_n)(x_n - x_{n-1}) = \sum_{i=1}^{n} f(\xi_i)(x_i - x_{i-1})$$

Geometrically, this value can be thought of as representing an *approximation of the area under the graph of the curve* $y = f(x)$, *if* $f(x) \geq 0$, between the vertical lines $x = a$ and $x = b$. We now define the notion of a limit of the Riemann sum as the norm of Π approaches 0, denoted by the symbol

$$\lim_{\|\Pi\| \to 0} \sum_{i=1}^{n} f(\xi_i)(x_i - x_{i-1})$$

as follows. If given any number $\varepsilon > 0$ we can find a $\delta > 0$ such that whenever Π is *any* partition with $\|\Pi\| < \delta$, it follows that

$$\left| \sum_{i=1}^{n} f(\xi_i)(x_i - x_{i-1}) - L \right| < \varepsilon$$

then we say that the *limit of the Riemann sum as the norm of the partition* Π *approaches* 0 *is* L.

Note that the numerical value of the limit, L, just defined generally depends on the choice of f and of the quantities ξ_i and the partition Π. *If this limit,* L, *is independent of the choice of the partition* Π *and the choice of the* ξ_i *within each subdivision of* Π, then we call this value of L the *definite integral of* f *from* a *to* b. When this happens we simply say that f *is integrable over* I and f is called *the integrand* of this integral. One of the consequences of this definition is that the definite integral of a function f which is continuous on $I = [a, b]$ exists. This limit also exists under much more general conditions but it is not necessary to delve into these matters here.

The *mean value theorem* states that if f is differentiable on (a, b) and continuous on $[a, b]$ then there exists some point c inside (a, b) with the property that $f(b) - f(a) = f'(c)(b - a)$. Using this theorem it is not difficult to show that, in fact,

The definite integral of f from a to $b = F(b) - F(a)$

where F is any antiderivative of f. So, we can write

$$\int_a^b f(x)\,dx = \lim_{\|\Pi\| \to 0} \sum_{i=1}^{n} f(\xi_i)(x_i - x_{i-1}) = F(b) - F(a) \tag{62}$$

and, in particular,

$$\int_a^b 1\,dx = b - a$$

Now, if $f(x) \geq 0$, we can also *define the area under the graph of the curve* $y = f(x)$ *between the lines* $x = a$ *and* $x = b$ by the definite integral of f from a to b, that is Eq. (62).

Let x be a generally unspecified point in a given interval $[a, b]$ on which we define f. Assuming that the definite integral of f from a to b exists, one can then write the equality

$$\int_a^x f(t)\,dt = F(x) - F(a) \tag{63}$$

where, as usual, F is some antiderivative of f. The fact that we changed the symbol x within the integral to a t is of no consequence to the value of the integral. These changes reflect the fact that these *inner variables* can be denoted by *any* symbol you want. This means that we can think of the quantity on the left of Eq. (63) as a function of x, and this equality is valid for every x in $[a, b]$. The quantity on the left of Eq. (63) is also called an *indefinite integral of* f. This identifies the notions of an indefinite integral with that of an antiderivative. These two notions are *equivalent* on account of Eqs. (62) and (63). The following properties of the *integral* now follow easily:

$$\int_a^b f(t)\,dt = \int_a^c f(t)\,dt + \int_c^b f(t)\,dt \qquad a \leq c \leq b$$

If k is any constant, then

$$\int_a^b k f(t)\,dt = k \int_a^b f(t)\,dt$$

from which follows the fact that

$$\int_a^b k\,dt = k\int_a^b 1\,dt = k(b-a)$$

Generally, if f, g are both integrable over I then

$$\int_a^b (f(t) \pm g(t))\,dt = \int_a^b f(t)\,dt \pm \int_a^b g(t)\,dt$$

Other properties of the integral that follow directly from its definition include

$$\text{If } f(x) \geq g(x) \text{ over } [a,b] \text{ then } \int_a^b f(x)\,dx \geq \int_a^b g(x)\,dx \qquad \text{(monotonicity property)}$$

from which we easily deduce that

$$\text{If } f(x) \geq 0 \text{ over } [a,b] \text{ then } \int_a^b f(x)\,dx \geq 0$$

and

$$\left|\int_a^b f(x)\,dx\right| \leq \int_a^b |f(x)|\,dx$$

(triangle inequality for integrals)

A consequence of the definition of antiderivative and Eq. (63) is the *fundamental theorem of calculus* which states that if f is continuous over $[a,b]$ then f has an indefinite integral and

$$\frac{d}{dx}\int_a^x f(t)\,dt = f(x) \tag{64}$$

and, if f is differentiable over (a,b) and continuous over $[a,b]$, then

$$\int_a^b f'(t)\,dt = f(b) - f(a) \tag{64$'$}$$

More generally, there is *Leibniz's formula* which follows from the chain rule and Eq. (64). If f is continuous over the real line and a, b are differentiable functions there, then

$$\frac{d}{dx}\int_{a(x)}^{b(x)} f(t)\,dt = f(b(x))\frac{db}{dx} - f(a(x))\frac{da}{dx} \tag{65}$$

(Leibniz's formula)

For example, it follows from Eq. (63) that

$$\frac{d}{dx}\int_a^x e^{-t^2}\,dt = \frac{d}{dx}\int_a^x e^{-s^2}\,ds = e^{-x^2}$$

and, from Eq. (65), that

$$\begin{aligned}\frac{d}{dx}\int_{x+1}^{x^2} e^{-t^2}\,dt &= e^{-(x^2)^2}\frac{d}{dx}(x^2) - e^{-((x+1)^2)^2}\frac{d}{dx}(x+1)\\ &= 2xe^{-x^4} - e^{-(x+1)^4}\end{aligned}$$

4.4 TECHNIQUES OF INTEGRATION

4.4.1 Integration by Substitution

The evaluation of indefinite and corresponding definite integrals is of major importance in calculus. In this section we introduce the method of substitution as a possible rule to be used in the evaluation of indefinite or definite integrals. It is based on a change-of-variable formula, Eq. (67) below, for integrals which we now describe. Given a definite integral of f over $I = [a,b]$ we can write [see Eq. (62)],

$$\int_a^b f(x)\,dx = F(b) - F(a) \tag{66}$$

The substitution $u(x) = t$, where we assume that u has a differentiable inverse function $x = u^{-1}(t)$, inside the integral corresponds to the *change-of-variable formula*

$$F(b) - F(a) = \int_{u(a)}^{u(b)} f(u^{-1}(t))\left(\frac{d}{dt}u^{-1}(t)\right)dt \tag{67}$$

which is itself equivalent to the relation

$$\begin{aligned}\frac{d}{dt}F(u^{-1}(t)) &= F'(u^{-1}(t))\frac{d}{dt}u^{-1}(t)\\ &= f(u^{-1}(t))\frac{d}{dt}u^{-1}(t)\end{aligned} \tag{68}$$

if F is an antiderivative of f, by the chain rule, Eq. (30). Integrating both sides of Eq. (68) over the interval $u(a), u(b)$ and using the fundamental theorem of calculus, we obtain Eq. (67). In practice, we proceed as follows. In order to evaluate the integral

$$\int_a^b f(x)\,dx = \int_0^2 2xe^{x^2}\,dx$$

we make the substitution $u(x) = x^2 = t$, whose inverse is given by $x = \sqrt{t} = u^{-1}(t)$. Using this substitution we see that $u(0) = 0$, and $u(2) = 4$. Since $f(u^{-1}(t)) = f(\sqrt{t}) = 2\sqrt{t}\,e^t$, and the derivative of u^{-1} (t) is $1/(2\sqrt{t})$, Eq. (67) becomes, in this case,

$$\begin{aligned}\int_0^2 2xe^{x^2}\,dx &= \int_0^4 2\sqrt{t}\,e^t\frac{1}{2\sqrt{t}}\,dt\\ &= \int_0^4 e^t\,dt = (e^4 - 1)\end{aligned}$$

The *shortcut to integration by substitution*, which amounts to the same thing as an answer can be summarized by setting $t = x^2$, $dt = 2x\,dx$ with the limits being changed according to the rule $t = 0$ when $x = 0$, $t = 4$ when $x = 2$. We find

$$\int_0^2 2xe^{x^2}\,dx = \int_0^4 e^t\,dt$$

as before, but more directly.

4.4.2 Integration by Parts

When every other technique fails try using *integration by parts*. This method is based on the *product rule* for derivatives (Sec. 4.2.1); in fact it is a sort of *inverse* of this method. Starting with the ordinary product rule, Eq. (27), namely,

$$(fg)'(x) = f'(x)\,g(x) + f(x)\,g'(x)$$

we can integrate both sides, say, from a to b. Use of Eq. (64′) and some adjustments show that, when used with a definite integral,

$$\int_a^b f(x)\,g'(x)\,dx = (fg)(b) - (fg)(a) - \int_a^b f'(x)\,g(x)\,dx \tag{69}$$

However, it is more commonly written in the following indefinite integral form:

$$\int u\,dv = uv - \int v\,du \tag{70}$$

For example, in order to evaluate

$$\int xe^x\,dx$$

we set $u = x, dv = e^x$. From this we obtain, $du = dx, v = e^x$. Substituting these values into Eq. (70) we get

$$\int xe^x\,dx = xe^x - \int e^x\,dx = xe^x - e^x + C$$

where C is a constant of integration. In the event where one of the terms is a simple trigonometric function (such as sin or cos), the method needs to be adapted, as the following shows. For example, the evaluation of

$$\int e^x \sin x\,dx$$

requires that we set $u = e^x$, $dv = \sin x\,dx$. Then $du = e^x\,dx$, $v = -\cos x$. The method then gives us

$$\int e^x \sin x\,dx = -e^x \cos x + \int e^x \cos x\,dx$$

We apply the same technique once again, except that now we set $u = e^x$, $dv = \cos x\,dx$ in the integral on the right of the last display. From this, $du = e^x\,dx$, $v = \sin$ x and we now find

$$\int e^x \sin x\,dx = -e^x \cos x + e^x \sin x - \int e^x \sin x\,dx$$

It follows that

$$\int e^x \sin x\,dx = \frac{-e^x \cos x + e^x \sin x}{2}$$

This method can always be used when one of the factors is either a sine or cosine and the other is an exponential.

4.4.3 Trigonometric Integrals

A trigonometric integral is an integral whose integrand contains only trigonometric functions and their powers. These are best handled with the repeated use of *trigonometric identities*. Among those which are most commonly used we find (here u or x is in radians):

$$\cos^2 u = \frac{1 + \cos(2u)}{2} \qquad \sin^2 u = \frac{1 - \cos(2u)}{2} \tag{71}$$

$$\begin{matrix} \cos^2 u + \sin^2 u = 1 & \sec^2 u - \tan^2 u = 1 \\ \csc^2 u - \cot^2 u = 1 & \end{matrix} \tag{72}$$

$$\sin(2u) = 2 \sin u \cos u \tag{73}$$

As an example, we consider the problem of finding an antiderivative of the function $f(x) = \cos^4 x$. This problem is tackled by writing $f(x) = \cos^4 x = \cos^2 x \cos^2 x$ $= (1 - \sin^2 x)\cos^2 x$ and then using the first of Eqs (71), (73), and the second of Eq. (71) along with a simple change of variable. The details for evaluating integrals of the form

$$\int \cos^m x \sin^n x\,dx \tag{74}$$

where m, n are positive integers are given here. Similar ideas apply in the case where the integral Eq. (74) involves other trigonometric functions.

m is odd, n is even. Solve the first of Eqs (72) for $\cos^2$ x and substitute the remaining in lieu of the cosine expression leaving one cosine term to the side. Follow this with a substitution of variable, namely, $u = \sin x$, $du = \cos x\,dx$, which now reduces the integrand to a polynomial in u and this is easily integrated.

m is odd, n is odd. Factor out a copy of each of $\sin x$, $\cos x$ leaving behind even powers of both $\sin x$, $\cos x$. Convert either one of these even powers in terms of the other using Eq. (72), and then perform a simple substitution, as before.

m is even, n is odd. Proceed as in the case where m is odd and n is even with the words *sine* and *cosine* interchanged.

m is even, n is even. Remove all even powers of the sine and cosine by applying Eq. (71) repeatedly. In addition to Eqs (71)–(73) there are a few other formulae which may be useful as they *untangle* the products. For any two angles, A, B, these are

$$\sin(A)\sin(B) = \frac{\cos(A-B)-\cos(A+B)}{2} \tag{75}$$

$$\sin(A)\cos(B) = \frac{\sin(A-B)+\sin(A+B)}{2} \tag{76}$$

$$\cos(A)\cos(B) = \frac{\cos(A-B)+\cos(A+B)}{2} \tag{77}$$

For example,

$$\begin{aligned}\int \sin^4 x \cos^2 x\,dx &= \frac{1}{8}\int (1-\cos(2x))^2(1+\cos(2x))\,dx \\ &= \frac{1}{8}\int (1-\cos(2x)-\cos^2(2x) \\ &\qquad + \cos^3(2x))\,dx\end{aligned}$$

where the first three integrals may be evaluated without much difficulty. The last integral above reduces to the case where m is odd and n is even (actually, $n = 0$ here).

4.4.4 Trigonometric Substitutions

A trigonometric substitution is particularly useful when the integrand has a particular form, namely, if it is the sum or difference of two squares, one of which is a constant. The substitutions can be summarized as follows. If the integrand contains a term of the form:

$\sqrt{a^2-x^2}$, where $a > 0$ is a constant: set $x = a\sin\theta$, $dx = a\cos\theta\,d\theta$, $\sqrt{a^2-x^2} = a\cos\theta$, if $-\pi/2 < \theta < \pi/2$.

$\sqrt{a^2+x^2}$, where $a > 0$ is a constant: set $x = a\tan\theta$, $dx = a\sec^2\theta\,d\theta$, $\sqrt{a^2+x^2} = a\sec\theta$, if $-\pi/2 < \theta < \pi/2$.

$\sqrt{x^2-a^2}$, where $a > 0$ is a constant: set $x = a\sec\theta$, $dx = a\sec\theta.\ \ \tan\theta\,d\theta$, $\sqrt{x^2-a^2} = a\tan\theta$, if $0 < \theta < \pi/2$.

For example,

$$\int \frac{x^2}{\sqrt{9-x^2}}\,dx = \int \frac{(3\sin\theta)^2 3\cos\theta\,d\theta}{3\cos\theta} = 9\int \sin^2\theta\,d\theta$$

and the last one is handled by means of Eq. (71) and a substitution. Thus,

$$9\int \sin^2\theta\,d\theta = \frac{9}{2}\int (1-\cos(2\theta))\,d\theta = \frac{9}{2}\theta - \frac{9}{4}\sin(2\theta) + C$$

But $x = 3\sin\theta$ means that $\theta = \arcsin(x/3)$. Moreover, by trigonometry, $(9/4)\sin(2\theta) = (9/4)\sin(2\arcsin(x/3)) = (1/2)x\sqrt{9-x^2}$. Hence,

$$\int \frac{x^2}{\sqrt{9-x^2}}\,dx = \frac{9}{2}\arcsin\left(\frac{x}{3}\right) - \frac{x\sqrt{9-x^2}}{2} + C$$

4.4.5 Partial Fractions

The method of *partial fractions* applies to the case where the integrand is a *rational function* (see Sec. 4.1.4). It is known from algebra that every polynomial with real coefficients can be factored into a product of linear factors [e.g., products of factors of the form $(x-r)^p$], and a product of quadratic factors called *quadratic irreducibles* (e.g., ax^2+bx+c where $b^2-4ac < 0$, i.e., it has no real roots). For example, $x^4-1 = (x^2+1)(x-1)(x+1)$. It follows that the numerator and denominator of every rational function can also be factored in this way. In order to factor a given polynomial in this way one can use Newton's method (Sec. 4.2.3) in order to find all its real roots successively.

Now, in order to evaluate an expression of the form

$$\int \frac{a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0}{b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0}\,dx \tag{78}$$

where m, n are integers and the coefficients are assumed real, there are two basic cases.

$n \geq m$. In this case we apply the classical *method of long division* which indicates that we divide the numerator into the denominator resulting in a polynomial and a remainder term which is a rational function whose numerator has degree one (or more) less than the denominator. For example, long division gives us that

$$\frac{x^4}{x^2-1} = x^2 + 1 + \frac{1}{x^2-1}$$

Here, the remainder is the rational function on the right of the last display (whose numerator has degree 0 and whose denominator has degree 2).

The remainder may be integrated using the idea in the next case.

$n < m$. We factor the denominator completely into a product of linear and irreducible factors. Next, we decompose this quotient into a *partial fraction* in the following sense. To each factor (of the denominator) of the form $(x-r)^p$, there corresponds a sum of terms of the form

$$\frac{A_1}{x-r}+\frac{A_2}{(x-r)^2}+\frac{A_3}{(x-r)^3}+\cdots+\frac{A_p}{(x-r)^p}$$

where the A's are to be found. To every quadratic irreducible factor (of the denominator) of the form $(ax^2+bx+c)^q$ where $b^2-4ac<0$, there corresponds, in its *partial fraction decomposition*, a sum of terms of the form

$$\frac{B_1x+C_1}{ax^2+bx+c}+\frac{B_2x+C_2}{(ax^2+bx+c)^2}+\cdots+\frac{B_qx+C_q}{(ax^2+bx+c)^q}$$

where the B's and C's are to be found, as well. The method for finding the A's, B's, and C's is best described using an example. In order to evaluate

$$\int \frac{x}{x^4-1}\,dx$$

which is a rational function, we find its partial fraction decomposition, which looks like

$$\frac{x}{x^4-1}=\frac{A_1}{x-1}+\frac{A_2}{x+1}+\frac{B_1x+C_1}{x^2+1}$$

since the factors of the denominator are $(x-1)(x+1)(x^2+1)$ and each such factor is simple (i.e., $p=1, q=1$). We *multiply both sides of the last display by the denominator*, x^4-1 and we proceed formally, *canceling out as many factors as possible* in the process. In this case, we get

$$x=A_1(x+1)(x^2+1)+A_2(x-1)(x^2+1)+(B_1x+C_1)(x-1)(x+1)$$

Since the last relation must be true for every value of x, we can set $x=1,-1$ and then any two other values of x, say, $x=0,2$ in order to *get a system of equations* (four of them) in the four given unknowns, A_1, A_2, B_1, C_1. Solving this system, we get the values $A_1=1/4, A_2=1/4, B_1=-1/2, C_1=0$ so that

$$\frac{x}{x^4-1}=\frac{1/4}{x-1}+\frac{1/4}{x+1}-\frac{(1/2)x}{x^2+1}$$

It follows that

$$\begin{aligned}\int\frac{x}{x^4-1}\,dx &= \frac{1}{4}\int\frac{dx}{x-1}+\frac{1}{4}\int\frac{dx}{x+1}-\frac{1}{2}\int\frac{x\,dx}{x^2+1}\\ &=\frac{1}{4}\ln|x-1|+\frac{1}{4}\ln|x+1|-\frac{1}{4}\ln(x^2+1)\end{aligned}$$

4.4.6 Numerical Integration

When the evaluation of a definite integral is required and every possible method of finding an antiderivative fails, one resorts to *numerical integration*, that is, the numerical approximation of the value of the definite integral. The two principal techniques here are the *trapezoidal rule* and *Simpson's rule*. Many other methods (midpoint rule, quadrature formulae, etc.) exist as well and the reader may consult any manual in numerical analysis for further details. In the *trapezoidal rule* the value of the definite integral of a given integrable function f on an interval $[a,b]$ is approximated by

$$\int_a^b f(x)\,dx \approx T_n=\frac{b-a}{2n}(f(x_0)+2f(x_1)+2f(x_2)+\cdots+2f(x_{n-1})+f(x_n))$$

where $x_0=a$, $x_n=b$ and $x_i=a+i(b-a)/n$, if $i=0,1,2,3,\ldots,n$. This method uses *line segments* to mimic the *curvature* of the graph of the function f. The larger the value of n the better the approximation and one is limited only by computing power. For example, if $n=30$,

$$\int_0^1 e^{x^2}\,dx \approx 1.4631550$$

whereas if we choose $n=40$ the value is approximately 1.46293487, while for $n=50$ the value is 1.462832952 with accuracy to three decimal places. In general, the error obtained in using the trapezoidal rule on a twice differentiable function f which is continuous on $[a,b]$ is given by

$$\left|\int_a^b f(x)\,dx - T_n\right| \le \frac{K(b-a)^3}{12n^2}$$

if $|f''(x)|\le K$ for x in $[a,b]$.

Simpson's rule states that, if we choose n to be an *even number*,

$$\int_a^b f(x)\,dx \approx S_n = \frac{b-a}{3n}(f(x_0)+4f(x_1)+2f(x_2) + 4f(x_3)+\cdots+4f(x_{n-1})+f(x_n))$$

where the coefficients on the right alternate between 4 and 2 except in the initial and final positions. This particular method uses *parabolic segments* to mimic the *curvature* of the graph of the function f and usually results in a better approximation (in contrast to the trapezoidal rule) for small values of n. For example, if we choose $n = 30$ as before, we find

$$\int_0^1 e^{x^2}\,dx \approx 1.462652118$$

with accuracy to five decimal places already. In this case, the error obtained in using Simpson's rule on a *four-times* differentiable function f which is continuous on $[a, b]$ is given by

$$\left|\int_a^b f(x)\,dx - S_n\right| \le \frac{K(b-a)^3}{180n^4}$$

if $|f^{(4)}(x)| \le K$ for x in $[a, b]$.

4.4.7 Improper Integrals

In some cases a definite integral may have one or both of its limits infinite, in which case we need to define the meaning of the integral. The natural definition involves interpreting the definite integral as a limit of a definite integral with finite limits. In the evaluation of the resulting limit use may be made of L'Hospital's rule (Sec. 4.2.2) in conjunction with the various techniques presented in Sec. 4.3. Given a function f defined and integrable on every finite interval on the real line, we define an *improper integral with infinite limit(s)* in a limiting sense, as follows:

$$\int_a^\infty f(x)\,dx = \lim_{T\to\infty}\int_a^T f(x)\,dx$$
$$\int_{-\infty}^a f(x)\,dx = \lim_{T\to-\infty}\int_T^a f(x)\,dx$$

whenever this limit exists and is finite, in which case we say that the *improper integral converges*. In the event that the limit does not exist as a finite number, we say the *improper integral diverges*. A similar definition applies when both limits are infinite, e.g.,

$$\int_{-\infty}^\infty f(x)\,dx = \int_a^\infty f(x)\,dx + \int_{-\infty}^a f(x)\,dx$$

provided this limit exists and is finite. For example, the evaluation of

$$\int_0^\infty x\,e^{-x}\,dx = \lim_{T\to\infty}\int_0^T xe^{-x}\,dx$$

requires integration by parts after which

$$\lim_{T\to\infty}\int_0^T x\,e^{-x}\,dx = \lim_{T\to\infty}(-(T+1)\,e^{-T}+1) = 1$$

by L'Hospital's rule. If an antiderivative cannot be found using any method, one resorts to numerical integration (Sec. 4.4.6).

4.5 APPLICATIONS OF THE INTEGRAL

4.5.1 The Area Between Two Curves

In this section we outline the main applications of the integral and its main interpretation, namely, as *the area under two given curves*. Let $y = f(x)$, $y = g(x)$ denote the graph of two curves defined on a common domain $[a, b]$ and we assume each function f, g is integrable over $[a, b]$. The *area between the two curves* is defined to be the expression

$$\text{The area between the two curves} = \int_a^b |f(x)-g(x)|\,dx \tag{79}$$

where the absolute value function was defined in Sec. 4.1.1. In the event that $f(x) \ge 0$ the *area under that part of the graph of f lying above the x-axis and between a, b* is given simply by the definite integral,

$$\int_a^b f(x)\,dx$$

For example, the area between the curves $y = x^2 - 1$, $y = x^2 + 1$ above the interval $[-1, 1]$ is given by $\int_{-1}^1 |(x^2-1)-(x^2+1)|dx = 4$. This area is depicted graphically in Fig. 7.

4.5.2 Volume and Surface Area of a Solid of Revolution

Next, let $y = f(x), y = g(x)$ denote the graph of two curves defined on a common domain $[a, b]$ and assume each function f, g is integrable over $[a, b]$, as above. In addition, we assume that $f(x) \le g(x)$, $a \le x \le b$. Now, consider the planar region defined by the curves $x = a$, $x = b$, $y = f(x)$, $y = g(x)$. This region is a closed region and it can be rotated (out of the plane) about an arbitrary line $x = L$ where $L < a$, or $L > b$, thus forming a

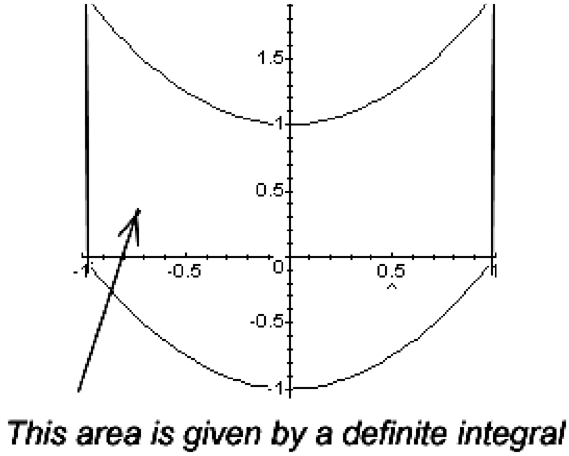


Figure 7 The area between $y = x^2 - 1$ and $y = x^2 + 1$ above the interval $[-1, 1]$.

solid of revolution. The *volume of the solid of revolution obtained by revolving this region about the line* $x = L$ is given by

$$2\pi \int_a^b |L - x|(f(x) - g(x))\,dx \tag{80}$$

On the other hand, if we revolve this region about the line $y = M$ where M exceeds the largest value of $f(x)$, $a \le x \le b$ or is smaller than the smallest value of $g(x)$, $a \le x \le b$, then the *volume of the solid of revolution thus obtained by revolving the region about the line* $y = M$ is given by

$$2\pi \int_a^b \left|\frac{f(x) + g(x)}{2} - M\right|(f(x) - g(x))\,dx$$

Similar formulae may be derived in case the planar region under discussion is bounded by curves of the form $x = h(y)$, $x = k(y)$ where $c \le y \le d$ and we are revolving about an arbitrary vertical or horizontal line. We point out a *theorem of Pappus* which states that if a closed region in the plane is rotated about a line which does not intersect the region, then the volume of the resulting solid is the product of the area of the region and the distance traveled by the *center of mass* (Sec. 4.5.4). If we set $g(x) = 0$, then the *surface area of the solid of revolution obtained by revolving the region about the x-axis*, also called a *surface of revolution* is given by the expression

$$\int_a^b 2\pi f(x)\sqrt{1 + f'(x)^2}\,dx$$

provided $f(x) \ge 0$ and f is differentiable. For example, if we rotate the region bounded by the two curves $y = 1 - x^2$, $y = x^2 - 1$ about the line $x = 2$, say, we get (from Eq. (81)), the integral

$$2\pi \int_{-1}^{1} |2 - x|[(1 - x^2) - (x^2 - 1)]\,dx = \frac{32\pi}{3}$$

The surface area of the solid of revolution obtained by revolving that part of the curve $y = 1 - x^2$ with $y \ge 0$ about the x-axis is given by

$$2\pi \int_{-1}^{1} (1 - x^2)\sqrt{1 + 4x^2}dx = \frac{7\pi\sqrt{5}}{8} - \frac{17\pi \ln(\sqrt{5} - 2)}{16}$$

4.5.3 The Length of a Curve

The *length of a segment of a curve* given by the graph of a differentiable function f on the interval $[a, b]$ is given by

$$\int_a^b \sqrt{1 + f'(x)^2}\,dx \tag{81}$$

where, as usual, $f'(x)$ denotes the derivative. The methods of Sec. 4.4 (trigonometric substitutions) may prove useful in the evaluation of an integral of the type Eq. (81). If these should fail, one can always resort to numerical integration (Sec. 4.4.6). For example, the length of the arc of the curve given by the function f where $f(x) = (a^{2/3} - x^{2/3})^{3/2}$, between the points $x = 0$ and $x = a$, is given by

$$\begin{aligned}\int_0^a \sqrt{1 + f'(x)^2}\,dx &= \int_0^a \sqrt{1 + (a^{2/3}x^{-2/3} - 1)}\,dx \\ &= \int_0^a a^{1/3}x^{-1/3}\,dx = \frac{3a}{2}\end{aligned}$$

If the differentiable curve is given *parametrically* by a set of points (x, y) where $x = x(t)$, $y = y(t)$, $a \le t \le b$, are each differentiable functions of t, then its length is given by

$$\int_a^b \sqrt{x'(t)^2 + y'(t)^2}\,dt \tag{82}$$

For example, the parametric equations of a circle of radius R centered at a point $P = (x_0, y_0)$, are given by $x = x_0 + R\cos t$, $y = y_0 + R\sin t$ where $0 \le t \le 2\pi$. In this case, the circumference of the circle is given by Eq. (82) where $x'(t)^2 + y'(t)^2 = R^2$ resulting in the value $2\pi R$, as expected.

4.5.4 Moments and Centers of Mass

Let $y = f(x)$, $y = g(x)$ denote the graph of two curves defined on a common domain $[a, b]$ and assume each function f, g is integrable over $[a, b]$, as above. In addition, we assume that $f(x) \geq g(x)$, $a \leq x \leq b$. Then the *center of mass* or *centroid* of the region of uniform mass density defined by the curves $x = a, x = b$, $y = f(x)$, $y = g(x)$ is given by the point $(\bar{x}, \bar{y})$ where

$$\bar{x} = \frac{M_y}{m} = \frac{1}{m}\int_a^b x(f(x) - g(x))\,dx$$

$$m = \int_a^b (f(x) - g(x))\,dx$$

$$\bar{y} = \frac{M_x}{m} = \frac{1}{m}\int_a^b \frac{(f(x)^2 - g(x)^2)}{2}\,dx$$

where M_x, M_y represent the *moment about the x-axis (resp. y-axis)*. For example, the centroid of the region bounded by the intersection of the line $y = f(x) = x$ and the parabola $y = g(x) = x^2$ has mass $m = \int_0^1 (x - x^2)\,dx = 1/6$. In this case,

$$\bar{x} = 6\int_0^1 (x^2 - x^3)\,dx = \frac{1}{2}$$

$$\bar{y} = 6\int_0^1 \frac{(x^2 - x^4)}{2} \qquad dx = \frac{2}{5}$$

4.6 DIFFERENTIAL EQUATIONS

4.6.1 First-Order Equations

The essence of most practical physical applications of calculus includes differential equations. By a *differential equation of order n* we mean an equation involving some unknown function say, $y(x)$, and its derivatives up to order n. By a *classical solution* of a differential equation of order n we mean a function, y, which has continuous derivatives up to and including that of order n and whose values satisfy the equation at every point x under consideration. For example, the function $y(x) = e^x$ is a solution of the differential equation of order 1, $y'(x) = y(x)$. By the *general solution* of a differential equation of order n is meant a solution which has the property that it contains every other solution of the same equation for particular choices of parameters appearing in it. For example the general solution of the equation $y'(x) = y(x)$ is given by $y(x) = ce^x$ where c is an arbitrary constant. Every other solution of this equation must agree with a particular choice of the parameter, c, in the general solution just written. It is a general fact that the general solution of a differential equation of order n must contain n parameters.

The simplest form of all differential equations of the first order is known as a *separable equation*. It has the simple form

$$\frac{dy}{dx} = f(x)\,g(y) \tag{83}$$

Dividing both sides by $g(y)$, assumed nonzero, and integrating both sides with respect to x gives the general solution in *implicit form*

$$\int_{y(a)}^{y(x)} \frac{1}{g(t)}\,dt - \int_a^x f(t)\,dt = C \tag{84}$$

where C is a constant, and a is a prescribed point. For example, the general solution of the separable equation

$$\frac{dy}{dx} = \frac{x^2}{y\,e^y}$$

is given by

$$\int_{y(a)}^{y(x)} t\,e^t\,dt - \int_a^x t^2\,dt = C$$

or, upon integration,

$$y(x)\,e^{y(x)} - e^{y(x)} - \frac{x^3}{3} = C$$

where C includes all the quantities, $y(a), a$ etc., being constant, they can all be absorbed in some new constant, denoted again by C. This equation can also be written in the form

$$ye^y - e^y - \frac{x^3}{3} = C$$

where the dependence of y on the independent variable x is suppressed. As such it becomes an implicit relation and it defines y as a function of x under some conditions (derived from a general result called the *implicit function theorem*).

A *linear differential equation* of the first order has the special form

$$\frac{dy}{dx} + P(x)y = Q(x) \tag{85}$$

and its general solution may be written explicitly as

$$y(x) = \frac{1}{\exp\left(\int_a^x P(t)\,dt\right)} \left(\int_a^x Q(t)\exp\left(\int_a^t P(u)\,du\right)dt + C\right) \tag{86}$$

where C is a parameter (constant). For example, the general solution of the equation

$$\frac{dy}{dx} - y = x^2$$

is given by

$$\begin{aligned} y(x) &= \frac{1}{e^{-x}}\left(\int_a^x t^2 e^{-t}\,dt + C\right) \\ &= e^x(e^{-x}(-x^2 - 2x - 2) + C), \\ &= Ce^x - x^2 - 2x - 2 \end{aligned}$$

and the *particular solution* for which $y = 1$ when $x = 0$ is given by $y(x) = 3e^x - x^2 - 2x - 2$.

4.6.2 Partial Derivatives of Functions of Two Variables and Exact Equations

We turn briefly to a definition of a *partial derivative*, that is, the extension of the notion of *derivative*, also called an *ordinary derivative*, to functions of two or more variables. A *function of two variables* is a function, f, whose domain is a set of points in a plane, usually considered as the xy-plane. Its values are denoted by $f(x, y)$ and it acts on two arguments, instead of one. For example, the function defined by $f(x, y) = xe^{xy}$ is such a function and $f(1, 0) = 1$. The values $f(x + h, y)$ are defined as usual by replacing every occurrence of the symbol x in the expression for $f(x, y)$ by the new symbol $x + h$ and then simplifying this new expression. For example, with f defined earlier, $f(x + h, y) = (x + h)e^{(x+h)y} = (x + h)e^{xy}e^{hy}$. In a similar way one can define the meaning of, say, $f(x, y + k)$. In our case, this implies that, for example, $f(x, y + k) = xe^{x(y+k)}$. This then enables us to define the notion of a partial derivative as a limit, as we did in Sec. 4.2. For functions of two variables, we define the *partial derivative of f with respect to x at the point (a,b)* as

$$\frac{\partial f}{\partial x} = \lim_{h \to 0} \frac{f(a + h, b) - f(a, b)}{h} \tag{87}$$

whenever this limit exists and is finite. In the same way we can define the *partial derivative of f with respect to y at the point (a,b)* as

$$\frac{\partial f}{\partial y} = \lim_{k \to 0} \frac{f(a, b + k) - f(a, b)}{k} \tag{88}$$

These two quantities represent the rate of change of the function f in the direction of the two principal axes, the x- and y-axes. In practice, these partial derivatives are found by thinking of one variable as a constant and taking the ordinary derivative of the f with respect to the other variable. The operation of taking a partial derivative can be thought of as being an operation on a function of *one* variable, and so all the rules and properties that we know of in Sec. 2 apply in this case as well. For example, for f defined earlier,

$$\frac{\partial f}{\partial x} = \frac{\partial}{\partial x}(xe^{xy}) = xye^{xy} + e^{xy}$$

by the product rule, since every occurrence of the variable y in the value of $f(x, y)$ triggers the rule that it *be thought of as a constant*. A function of two variables is called *differentiable at* (a, b) if each one of its two partial derivatives exists at (a, b) and is finite. As usual it is *differentiable in a region* if it is differentiable at every point of the region. The notion of continuity for a function of two variables is similar to the one presented in Sec. 4.1.4, except that the *notion of a limit needs to be updated* to take into account the fact that there are now two variables, (x, y), approaching a specified point, (a, b). In this case, we say that a *function f of two variables is continuous at* (a, b) if it is defined at (a, b), i.e., $f(a, b)$ is a finite number, and if

$$\lim_{(x,y)\to(a,b)} f(x, y) = f(a, b)$$

where, by the symbol on the left, we mean the following. For each given $\varepsilon > 0$, we can find a number δ, generally depending on $\varepsilon > 0$, with the property that whenever

$$\sqrt{(x - a)^2 + (y - b)^2} < \delta$$

we also have

$$|f(x, y) - f(a, b)| < \varepsilon$$

Every polynomial function in two variables is continuous, for example, $f(x, y) = -1 + x - x^2y + 3x^2y^3$ is such a function. More generally, the product of any two polynomials of one variable, say, $p(x)$, $q(y)$ gives a polynomial of two variables. As in Secs. 4.1 and 4.2, the composition of continuous functions is also continuous, and so forth.

We can now turn to the solution of a so-called *exact differential equation*. We assume that the functions $P(x, y)$, $Q(x, y)$ appearing below are each differentiable functions of two variables, in the sense above. A first-order (ordinary) differential equation of the form

$$P(x, y) + Q(x, y)\frac{dy}{dx} = 0 \tag{89}$$

is called an *exact differential equation* if the functions of two variables, P, Q satisfy the equations

$$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x} \tag{90}$$

in the region which is the intersection of the domains of each function. In this case the differential equation Eq. (89) can be solved and a general solution may be found as an implicit relation. However, before writing down the general solution, we need to have some working knowledge of the meaning of an expression like

$$\int P(x, y)\, dx$$

By this we mean that we integrate the function $P(x, y)$ with respect to x and we *think* of every occurrence of the symbol y as being a constant (as before). For example,

$$\int e^{xy}\, dx = \frac{e^{xy}}{y} + (\text{some function of } y)$$

Another example is furnished by

$$\int e^{xy}\, dy = \frac{e^{xy}}{x} + (\text{some function of } x)$$

The last two terms which are *functions of x and or y* are the two-dimensional equivalent of the *constant* of integration which appears after we evaluate an indefinite integral for a function of one variable. In the next formula, we will set them both to zero.

The *general solution of the exact equation* Eq. (89) is given implicitly by the relation

$$f(x, y) = c$$

where c is an ordinary constant, and

$$f(x, y) = \int P(x, y)\, dx + \int \left[Q(x, y) - \frac{\partial}{\partial y}\left(\int P(x, y)\, dx \right) \right] dy$$

For example, we solve the equation

$$(\ln(y) + 3y^2) + \left(\frac{x}{y} + 6xy \right) \frac{dy}{dx} = 0$$

by noting that here, $P(x, y) = \ln(y) + 3y^2$ and $Q(x, y) = x/y + 6xy$ with the exactness criterion Eq. (90) being verified, since

$$\frac{\partial}{\partial y} (\ln(y) + 3y^2) = \frac{\partial}{\partial x} \left(\frac{x}{y} + 6xy \right) = \frac{1}{y} + 6y$$

Next,

$$\int P(x, y)\, dx = \int (\ln(y) + 3y^2)\, dx = x(\ln(y) + 3y^2)$$

and so

$$\frac{\partial}{\partial y} \int P(x, y)\, dx = x\left(\frac{1}{y} + 6y \right)$$

Finally we note that

$$Q(x, y) - \frac{\partial}{\partial y} \int P(x, y)\, dx = 0$$

and it follows that

$$f(x, y) = x(\ln(y) + 3y^2)$$

or the general solution is given implicitly by

$$x(\ln(y) + 3y^2) = c$$

where c is an arbitrary constant. It is difficult to isolate the y variable in the last expression but, nevertheless, this *does* give a general solution and one which is practical, since, given any initial condition, we can determine c, and therefore the locus of all points in the xy-plane making up the graph of the required solution.

4.6.3 Integrating Factors

By an *integrating factor* of a first-order differential equation of the form Eq. (89) is meant a function of *one* variable, call it I, with the property that

$$I\, P(x, y) + I\, Q(x, y) \frac{dy}{dx} = 0$$

is exact. Of course the original equation is not assumed to be exact, but, in some cases, it can be turned into an exact equation by multiplying throughout by this integrating factor. We describe two cases in which Eq. (89) can be transformed into an exact equation. If the quotient

$$\frac{1}{Q(x, y)} \left(\frac{\partial P}{\partial y} - \frac{\partial Q}{\partial x} \right) = \text{a function of } x \text{ alone} \tag{91}$$

then an integrating factor is given by the exponential function

$$I(x) = \exp\left(\int \frac{1}{Q(x, y)} \left(\frac{\partial P}{\partial y} - \frac{\partial Q}{\partial x} \right) dx \right)$$

On the other hand, if

$$\frac{1}{P(x, y)} \left(\frac{\partial P}{\partial y} - \frac{\partial Q}{\partial x} \right) = \text{a function of } y \text{ alone} \tag{92}$$

then an integrating factor is given by the exponential function (note the *minus* sign),

$$I(y) = \exp\left(-\int \frac{1}{P(x, y)} \left(\frac{\partial P}{\partial y} - \frac{\partial Q}{\partial x} \right) dy \right)$$

In both cases the general solution may be written as

$$f(x, y) = c$$

where c is a constant and, in the case where $I(x)$ is a function of x alone, f is given by

$$f(x, y) = \int I(x)\,P(x, y)\,dx + \int\left[I(x)\,Q(x, y) - \frac{\partial}{\partial y}\left(\int I(x)\,P(x, y)\,dx\right)\right]dy \tag{93}$$

while in the case where $I(y)$ is a function of y alone, f is given by

$$f(x, y) = \int I(y)\,P(x, y)\,dx + \int\left[I(y)\,Q(x, y) - \frac{\partial}{\partial y}\left(\int I(y)\,P(x, y)\,dx\right)\right]dy \tag{94}$$

For example, if we wish to solve the differential equation

$$(1 - xy) + x(y - x)\frac{dy}{dx} = 0$$

we note that $P(x, y) = 1 - xy$, $Q(x, y) = x(y - x)$ and

$$\frac{1}{Q(x, y)}\left(\frac{\partial P}{\partial y} - \frac{\partial Q}{\partial x}\right) = -\frac{1}{x}$$

so that $I(x) = 1/x$, if $x > 0$. In this case,

$$\int I(x)\,P(x, y)\,dx = \ln x - xy$$

$$I(x)\,Q(x, y) - \frac{\partial}{\partial y}\left(\int I(x)\,P(x, y)\,dx\right) = y$$

and it follows that the general solution is given by $f(x, y) = c$ where

$$f(x, y) = \ln x - xy + \frac{y^2}{2}$$

In particular, the solution which passes through the point $(1, 0)$ is given implicitly by

$$\ln x - xy + \frac{y^2}{2} = 0$$

since $c = 0$ in this case.

More techniques for finding the solutions of various first-order differential equations and those of higher order as well, may be found in Refs. 1 and 2.

REFERENCES

1. D Zwillinger. Handbook of Differential Equations. New York: Academic Press, 1989.
2. M Abramowitz, I Stegun. Handbook of Mathematical Functions. New York: Dover, 1965.

Chapter 1.5

Ordinary Differential Equations

Jane Cronin
Rutgers University, New Brunswick, New Jersey

5.1 INTRODUCTION

A differential equation is an equation which involves an unknown function and its derivatives. We will consider *ordinary* differential equations which concern only functions of one independent variable. (If the unknown function is a function of two or more variables and partial derivatives with respect to two variables occur, the differential equation is called a *partial* differential equation.) Solving the differential equation means determining a function which satisfies it. For example, suppose we are given the differential equation

$$\frac{dy}{dx} = \sin 2x \qquad (\mathcal{E})$$

then a solution of the differential equation is

$$y(x) = -\tfrac{1}{2}\cos 2x + C$$

where C is an arbitrary constant. This equation is particularly easy to solve because on the left, we have only dy/dx and on the right we have an expression involving only the independent variable x. However, even an equation as simple as $(\mathcal{E})$ requires finding an integral or antiderivative of $\sin 2x$, and if the right-hand side were more complicated, say,

$$x^{15}\sin(x^2)$$

then finding the antiderivative or integral might present more serious problems. It is for this reason that we must often resort to using a table of integrals. Short tables can be found in most calculus textbooks. For more extensive tables, see Refs 1, 2, or 3. There are also integral tables in some computer software systems (see, e.g., Ref. 4).

Few of the differential equations encountered in physics, engineering, or chemistry are as simple as $(\mathcal{E})$. Many involve second or higher derivatives of the unknown function and the expression on the right usually involves y as well as x. Thus it is clear from the start that there are serious problems in solving differential equations.

Differential equations have been studied since Newton invented the calculus, which means that people have worked on them for more than 300 years. Our purpose here is to describe, in brief form, some techniques for studying solutions of differential equations. Before proceeding to this description, we mention some general properties of solutions.

In solving equation $(\mathcal{E})$ we obtained actually an infinite set of solutions because C is an arbitrary constant. Very often we are concerned with finding a solution which satisfies an *initial condition*, that is, a solution which has a given value for a given value of the independent value. For example, to find the solution of $(\mathcal{E})$ which equals 0 if $x = 2\pi$, we write

$$-\tfrac{1}{2}\cos 2(2\pi) + C = 0$$

Thus

$$C = \tfrac{1}{2}$$

and the desired solution is

$$y(x) = -\tfrac{1}{2}\cos 2x + \tfrac{1}{2}$$

Under quite general conditions, the solution which satisfies an initial condition is unique, and this has important practical and theoretical consequences.

In most of the discussion which follows, we concern ourselves with the question of how to find or to approximate the solution. Thus we are assuming implicitly that there is a solution to be found. This is not always true. (For examples of nonuniqueness and nonexistence, see Ref. 5, p. 27 and p. 29). Moreover, there is the allied question of the domain of the solution, i.e., the values of the independent value for which the solution is defined. As we shall see, even simple-looking equations present difficulties along this line. Finally, there is the question of relationships among the various solutions of an equation. Important aspects of this question are linear independence (especially for solutions of linear equations) and stability (for solutions of all classes of equations). Roughly speaking, a given solution is stable if solutions which are near it for some value of the independent variable stay close to the given solution for all larger values of the independent value. We shall return later to a description of these properties.

In describing methods for solving differential equations, we will use informal descriptions and exhibit examples. We will not be concerned with proofs, but will simply give references for further development or more rigorous treatment. The references given are intended to provide only a minimal sampling of available material. The literature on differential equations is huge. A few words about notation and numbering: first, in the examples during the discussion of calculations it is sometimes necessary to refer just once or twice to an immediately preceding equation. Instead of using an unnecessary numbering system for these equations we refer to such equations with (*). Second, the examples are numbered consecutively, independent of the section in which they appear.

We begin by describing some of the classical techniques which are sometimes very effective but apply only to special classes of equations. Then we proceed to an account of second-order linear equations, with constant coefficients and also with nonconstant coefficients. (An equation of nth order is one in which $d^n y/dx^n$ appears but no derivative or order higher than n appears.) Then we treat first-order linear systems with constant coefficients. After that, we describe briefly a couple of the major topics in nonlinear equations. Because these topics are large and because of limitations of space, our treatment is indeed brief.

Finally, we consider the very important question of whether to use numerical analysis, i.e., whether to 'put the differential equation on the computer' rather than try to solve it by 'pencil-and-paper' methods.

5.2 SOME CLASSICAL TECHNIQUES

First we will describe some techniques which were developed mainly in the 1700s and 1800s but which are still often useful.

5.2.1 Separable Equations

Regard dy/dx as the ratio of two differentials and multiply through by dx. If the equation can be written so that the variable x appears only on one side and the variable y on the other, then integrate the two sides separately.

Example 1

$$(1 + x^2)\frac{dy}{dx} = xy$$

$$\frac{dy}{y} = \frac{x\,dx}{1 + x^2}$$

$$\ln|y| = \tfrac{1}{2}\ln(1 + x^2) + C = \ln\sqrt{1 + x^2} + C$$

where C is a constant of integration.

$$|y| = e^C\sqrt{1 + x^2}$$

(Since C is real, then $e^C > 0$.) Thus we obtain two solutions: $y = e^C\sqrt{1 + x^2}$ and $y = -e^C\sqrt{1 + x^2}$.

Example 2

$$y^3\frac{dy}{dx} = (y^4 + 2)\cos x$$

$$\frac{y^3}{y^4 + 2}\,dy = \cos x\,dx$$

$$\tfrac{1}{4}\ln(y^4 + 2) = \sin x + C$$

$$\ln(y^4 + 2) = 4\sin x + C \qquad (*)$$

A solution of the differerntial equation is obtained by solving the equation $(*)$ for y as a function of x. We say that $(*)$ yields an implicit solution of the differential equation. In this case, we can solve $(*)$ as follows:

$$e^{\ln(y^4+2)} = e^{(4\sin x + C)} = e^{4\sin x}e^C$$

$$y^4 + 2 = Ke^{4\sin x}$$

where K is a positive constant.

$$y^4 = Ke^{4\sin x} - 2$$

Since

$$-1 \le \sin x \le 1$$

then

$$e^{-4} \le e^{4\sin x} \le e^4$$

If $Ke^{-4} > 2$, then

$$Ke^{4\sin x} - 2 > 0$$

and

$$y = \pm(Ke^{4\sin x} - 2)^{1/4}$$

(If $Ke^{4\sin x} < 2$, then y would not be real.)

5.2.2 Linear Equation of First Order

A *linear differential equation* is an equation in which the dependent variable and its derivatives are all of degree 1 (have exponent 1).

The general formula for the solution of the linear equation of first order

$$\frac{dy}{dx} + P(x)y = Q(x)$$

is

$$y = e^{-\int P(x)dx}\left\{\int [Q(x)]e^{\int P(x)dx}\,dx + C\right\}$$

Example 3

$$\frac{dy}{dx} - 4y = e^x \qquad y(0) = 1$$

$$P(x) = -4$$

$$\int P(x)dx = -4x \qquad -\int P(x)dx = 4x$$

$$Q(x) = e^x$$

$$\begin{aligned} y &= e^{4x}\left\{\int e^x e^{-4x}dx + C\right\} \\ &= e^{4x}\left\{\int e^{-3x}dx + C\right\} \\ &= e^{4x}\left\{-\frac{1}{3}e^{-3x} + C\right\} \\ &= -\frac{1}{3}e^x + Ce^{4x} \end{aligned}$$

$$y(0) = -\frac{1}{3} + C = 1$$

$$C = \frac{4}{3}$$

$$y(x) = -\frac{1}{3}e^x + \frac{4}{3}e^{4x}$$

5.2.3 Exact Differential Equations

The differential equation

$$M(x, y) + N(x, y)\frac{dy}{dx} = 0$$

is *exact* if there is a function $F(x, y)$ such that

$$\frac{\partial F}{\partial x} = M \qquad \text{and} \qquad \frac{\partial F}{\partial y} = N$$

Then the differential equation can be written

$$\frac{\partial F}{\partial x} + \frac{\partial F}{\partial y}\frac{dy}{dx} = 0$$

But if $y(x)$ is a solution of the differential equation, we have

$$\frac{d}{dx}F[x, y(x)] = \frac{\partial F}{\partial x}[x, y(x)] + \frac{\partial F}{\partial y}[x, y(x)]\frac{dy}{dx} = 0$$

Integrating, we have

$$F[x, y(x)] = 0$$

where we have, for convenience, chosen the constant of integration to be zero. It can be proved that the differential equation

$$M(x, y) + N(x, y)\frac{dy}{dx} = 0$$

is *exact* if and only if

$$\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x}$$

Example 4

$$\frac{2xy+1}{y} + \frac{y-x}{y^2}\frac{dy}{dx} = 0$$

$$M = \frac{2xy+1}{y} \qquad N = \frac{y-x}{y^2}$$

The equation is exact because

$$\frac{\partial M}{\partial y} = \frac{\partial}{\partial y}\left(2x + \frac{1}{y}\right) = -\frac{1}{y^2} = \frac{\partial N}{\partial x}$$

Let $\int^{(x)} M(x,y)\,dx$ *be the antiderivative of* M *regarded as a function of* x.

We seek a function $F(x,y)$ *such that*

$$\frac{\partial F}{\partial x} = M \qquad \text{and} \qquad \frac{\partial F}{\partial y} = N$$

Since $\partial F/\partial x = M$, *then integrating with respect to* x, *we obtain*

$$F(x,y) = \int^{(x)} M\,dx + g(y)$$

where $g(y)$ *is an arbitrary function of* y. *(Function* g *is arbitrary because if we differentiate with respect to* x, *any function of* y *only can be regarded as a constant during the differentiation.) Thus we require that*

$$\frac{\partial F}{\partial y} = \frac{\partial}{\partial y}\int^{(x)} M\,dx + g'(y) = N$$

or

$$\frac{\partial}{\partial y}\int^{(x)} M\,dx - N(x,y) = g'(y) \qquad (*)$$

Since the right-hand side of $(*)$ *is a function of* y *only, then the left-hand side of* $(*)$ *must be a function of* y *only. This is, in fact, true because taking the partial derivative of the left-hand side with respect to* x, *we obtain*

$$\begin{aligned}\frac{\partial}{\partial x}\frac{\partial}{\partial y}\int^{(x)} M\,dx - \frac{\partial N}{\partial x} &= \frac{\partial}{\partial y}\frac{\partial}{\partial x}\int^{(x)} M\,dx - \frac{\partial N}{\partial x}\\ &= \frac{\partial M}{\partial y} - \frac{\partial N}{\partial x} = 0\end{aligned}$$

In this problem,

$$\begin{aligned}g'(y) &= \frac{\partial}{\partial y}\int^{(x)} M\,dx - N = \frac{\partial}{\partial y}\left[x^2 + \frac{x}{y}\right] - \frac{y-x}{y^2}\\ &= -\frac{x}{y^2} - \frac{1}{y} + \frac{x}{y^2} = -\frac{1}{y}\end{aligned}$$

Hence

$$g(y) = -\ln|y|$$

and the implicit solution is

$$F(x,y) - x^2 + \frac{x}{y} - \ln|y| = 0$$

Sometimes even if the equation

$$M(x,y) + N(x,y)\frac{dy}{dx} = 0$$

is not exact, there exists a function $\mu(x,y)$ *such that if the equation is multiplied by* $\mu(x,y)$, *the result is an exact differential equation. Such a function* $\mu(x,y)$ *is called an integrating factor.*

Example 5

$$x\frac{dy}{dx} - y - x^3 = 0$$

The equation is not exact since

$$M = -y - x^3 \qquad \text{and} \qquad N = x$$

so that

$$\frac{\partial M}{\partial y} = -1 \qquad \text{and} \qquad \frac{\partial N}{\partial x} = 1$$

If we multiply by the integrating factor $\mu(x,y) = 1/x^2$, the equation becomes

$$-\frac{y}{x^2} - x + \frac{1}{x}\left(\frac{dy}{dx}\right) = 0 \qquad (*)$$

in which $M = -(y/x^2) - x$, $N = 1/x$ and

$$\frac{\partial M}{\partial y} = -\frac{1}{x^2} \qquad \frac{\partial N}{\partial x} = -\frac{1}{x^2}$$

Thus we have obtained an exact equation which can be solved as follows:

$$\int^{(x)} M\,dx = \frac{y}{x} - \frac{x^2}{2}$$

$$\frac{\partial}{\partial y}\left[\frac{y}{x} - \frac{x^2}{2}\right] = \frac{1}{x}$$

$$g'(y) = \frac{\partial}{\partial y}\int^{(x)} M\,dx - N = \frac{1}{x} - \left(\frac{1}{x}\right) = 0$$

$$g(y) = C$$

$$F(x,y) = \frac{y}{x} - \frac{x^2}{2} + C = 0$$

$$y = \frac{x^3}{2} - Cx$$

We may also solve $(*)$ as follows. Rewrite $(*)$ as

$$-\frac{y}{x^2}\,dx - x\,dx + \frac{1}{x}\,dy = 0$$

or

$$\frac{xdy - y\,dx}{x^2} - x\,dx = 0$$

or

$$d\left(\frac{y}{x}\right) = x\,dx$$

Integrating each side, we obtain

$$\frac{y}{x} = \frac{x^2}{2} + C$$

or

$$y = \frac{x^3}{2} + Cx$$

(Since C is an arbitrary constant, its sign has no significance.) A strategic choice of an integrating factor can simplify significantly the problem of solving a differential equation, but finding an integrating factor may require skill, experience, and luck.

5.2.4 Substitution Methods

5.2.4.1 Homogeneous Equations

A *homogeneous first-order equation* is an equation which can be written in the form

$$\frac{dy}{dx} = F\left(\frac{y}{x}\right)$$

Example 6

$$(xe^{y/x} + y) - x\frac{dy}{dx} = 0$$

Divide by x and obtain the homogeneous equation

$$\frac{dy}{dx} = e^{y/x} + \frac{y}{x} \tag{1}$$

Let

$$v = \frac{y}{x} \text{ or } y = vx \tag{2}$$

Then

$$\frac{dy}{dx} = v + x\,\frac{dv}{dx} \tag{3}$$

Substituting (2) and (3) in (1) yields

$$v + x\frac{dv}{dx} = e^v + v$$

$$+x\frac{dv}{dx} = e^v$$

This is a separable equation.

$$e^{-v}dv = \frac{dx}{x}$$

or

$$e^{-v} = -\ln|x| + C$$

If $C = \ln\ K$, then this equation becomes

$$e^{-\frac{y}{x}} = -\ln|x| + \ln\ K = \ln(K/|x|)$$

or

$$\ln e^{-y/x} = -\frac{y}{x} = \ln[\ln(K/|x|]$$

Therefore

$$y = -x\ln[\ln(K/|x|)]$$

5.2.4.2 Bernoulli's Equation

This is

$$\frac{dy}{dx} + P(x)y = [Q(x)]y^k$$

where P, Q are functions of x only and k is any fixed real number.

Divide by y^k:

$$y^{-k}\frac{dy}{dx} + P(x)[y^{-k+1}] = Q(x)$$

Let $v = y^{-k+1}$. Then

$$\frac{dv}{dx} = (-k+1)y^{-k}\frac{dy}{dx}$$

Substitution yields

$$\frac{1}{1-k}\,\frac{dv}{dx} + [P(x)]v = Q(x)$$

or

$$\frac{dv}{dx} + (1-k)[P(x)]v = (1-k)Q(x)$$

This is a linear first-order equation and we have already described a general formula for its solution.

Example 7. $y' - y = -y^2$. *(We use sometimes y' and y'' to denote dy/dx and d^2y/dx^2, respectively.)*

Divide by y^2:

$$\frac{y'}{y^2} - \frac{1}{y} = -1 \tag{$*$}$$

Let $v = y^{-2+1} = 1/y$. Then

$$\frac{dv}{dx} = -y^{-2}\frac{dy}{dx}$$

and $(*)$ *becomes*

$$-\frac{dv}{dx} - v = -1$$

or

$$\frac{dv}{dx} + v = 1$$

This is a first-order linear equation for which we have a general formula for the solution. We have here $P(x) = 1$ *and* $Q(x) = 1$. *Hence*

$$\begin{aligned} y^{-1} = v = e^{-x}&\left\{\int e^x\,dx + C\right\} \\ &= e^{-x}(e^x) + Ce^{-x} = 1 + Ce^{-x} \\ \frac{1}{y} &= 1 + Ce^{-x} \\ y &= \frac{1}{1 + Ce^{-x}} = \frac{e^x}{e^x + C} \end{aligned}$$

5.2.4.3 Dependent Variable Absent

Example 8

$$xy'' + (x^2 - 1)(y' - 1) = 0$$

The dependent variable y *is absent.*

Let $p = y'$; *then* $p' = y''$ *and the equation becomes*

$$xp' + (x^2 - 1)(p - 1) = 0$$

$$\frac{p'}{p - 1} + \frac{x^2 - 1}{x} = 0$$

$$\frac{p'}{p - 1} + x - \frac{1}{x} = 0 \quad \text{or} \quad \frac{dp}{p - 1} + \left(x - \frac{1}{x}\right) dx = 0$$

$$\ln|p - 1| + \frac{x^2}{2} - \ln|x| = C$$

$$\ln\frac{|p - 1|}{|x|} + \frac{x^2}{2} = C$$

$$\frac{|p - 1|}{|x|} = e^{-(x^2/2)+C} = C_1 e^{-x^2/2} \quad \text{where } C_1 > 0$$

$$|p - 1| = C_1|x|e^{-x^2/2}$$

If $p > 1$,

$$\frac{dy}{dx} = C_1|x|e^{-x^2/2} + 1$$

If $x > 0$,

$$y = -C_1 e^{-x^2/2} + x + C_2$$

If $x < 0$,

$$\frac{dy}{dx} = C_1(-x)e^{-x^2/2} + 1$$

and

$$y = C_1 e^{-x^2/2} + x + C_2$$

If $p < 1$,

$$-\frac{dy}{dx} = C_1|x|e^{-x^2/2} - 1$$

and if $x > 0$,

$$y = C_1 e^{-x^2/2} + x + C_2$$

If $x < 0$,

$$y = -C_1 e^{-x^2/2} + x + C_2$$

5.2.4.4 Independent Variable Absent

The procedure is to take y to be the independent variable and let y' be the new dependent variable.

Example 9. $yy'' + (y')^2 + 1 = 0$. *Let* $y' = p$. *Then*

$$y'' = \frac{dp}{dx} = \frac{dp}{dy}\frac{dy}{dx} = p\frac{dp}{dy}$$

The equation becomes

$$yp\frac{dp}{dy} + p^2 + 1 = 0$$

$$y\frac{dp}{dy} + p + \frac{1}{p} = 0$$

$$\frac{dp}{p + 1/p} + \frac{dy}{y} = 0$$

$$\frac{p\,dp}{p^2 + 1} + \frac{dy}{y} = 0$$

$$\tfrac{1}{2}\ln(p^2 + 1) + \ln|y| = C$$

$$\ln(p^2 + 1) + \ln y^2 = C$$

$$\ln[(p^2 + 1)y^2] = C$$

$$(p^2 + 1) = \frac{K}{y^2} \qquad \text{where } K = e^c > 0$$

$$(p^2) = \frac{K}{y^2} - 1 = \frac{K - y^2}{y^2}$$

$$\frac{dy}{dx} = p = \frac{\sqrt{K - y^2}}{y}$$

$$\frac{y\,dy}{\sqrt{K - y^2}} = dx$$

$$-(K - y^2)^{1/2} = x + C$$

$$K - y^2 = (x + C)^2$$
$$y^2 = K - (x + C)^2$$

In the preceding pages, we have given a sampling of the many ingenious techniques which have been developed to study particular classes of first-order differential equations. More complete discussions including extensions of the techniques we have described and other techniques can be found in standard textbooks. There is an excellent discussion in Ref. 5, Chap. 1. Very extensive and thorough treatments are given in Refs 6–8. Each of these references presents a large number of ordinary differential equations and then solutions. Reference 8 is a more extensive compilation than Refs 6 or 7, but Refs 6 and 7 contain more theory and references.

5.3 A GEOMETRICAL APPROACH

We have been describing methods for obtaining explicit formulas for solutions of differential equations. There are, however, other ways of obtaining useful information about solutions. A geometrical approach yields a good qualitative (nonnumerical) understanding not only of particular solutions but also of the relationships among the solutions. If we consider a first-order equation

$$\frac{dy}{dx} = f(x, y)$$

then each solution $y(x)$ represents a curve in the xy-plane. If this curve passes through a point (x_0, y_0) the slope of the curve at (x_0, y_0) is $f(x_0, y_0)$. Thus to get an idea of how the solutions behave, we indicate by an arrow with initial point (x_0, y_0) the slope $f(x_0, y_0)$ at that point. A solution which passes through (x_0, y_0) must be tangent at (x_0, y_0) to the arrow with initial point (x_0, y_0). If we look at the entire collection of arrows, this will, in many cases, give considerable information about the solutions even if we do not calculate any of the solutions.

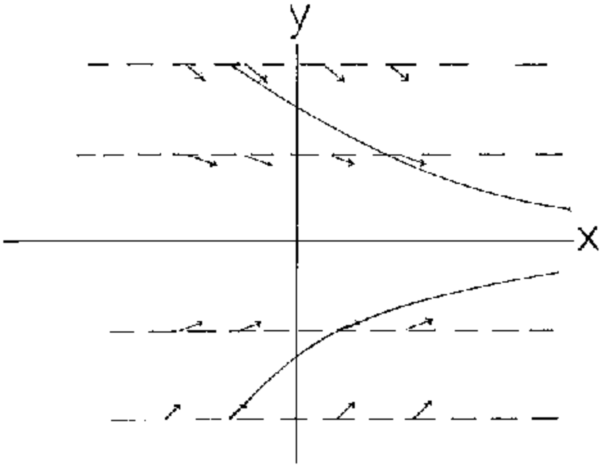


Figure 1

Example 10. *$dy/dx = -2y$. As shown in Fig. 1, all the arrows on a horizontal line $y = k$, a constant, have the same slope. If $k > 0$, then as k increases the slope becomes more and more negative. If $k < 0$ then as k decreases, the slope becomes increasingly positive.*

Thus a reasonable guess at the appearance of two typical solutions are the curves sketched in Fig. 1. Notice that $y(x) = 0$ for all x is a solution of the differential equation and that the sketch suggests that as $x \to \infty$, i.e. as x increases without bound, then each solution $y(x)$ approaches the solution $y(x) = 0$ for all x. (If this occurs then we say that the solution $y(x) = 0$ is asymptotically stable.) Of course, it is easy to verify the conjectures we have made about the solutions because the differential equation in this case is easy to solve:

$$\frac{dy}{y} = -2\,dx$$
$$\ln|y| = -2x + C$$
$$|y| = e^c e^{-2x}$$

Thus all the solutions have the form

$$y(x) = Ke^{-2x}$$

where K is a positive constant or a negative constant. However, the geometrical description that we have given is often enlightening if the problem of solving the differential equation is more difficult.

5.4 SECOND-ORDER LINEAR EQUATIONS WITH CONSTANT COEFFICIENTS

The techniques described so far are, in the main, unrelated. They are effective with special classes of equations, but there is no logical structure which relates them. Now we begin a more systematic study. We have already obtained a formula for the solution of a first-order linear equation. Now we consider *second-order linear equations with constant coefficients*, i.e., equations of the form

$$y'' + by' + cy = g(x) \tag{L}$$

where b, c are constants, and g is a given function of x. If $g(x) = 0$, for all x, the equation is *homogeneous*. (Note that this is a different definition of the word homogeneous from the definition used in Example 6.) The procedure for solving the homogeneous equation is as follows. Find the roots r_1, r_2 of the quadratic equation

$$r^2 + br + c = 0$$

Then the general solution has the form

$$Ae^{r_1 x} + be^{r_2 x}$$

where A, B are constants. The only exception is the case $r_1 = r_2$, which will be discussed below.

Example 11. $y'' - 3y' + 2y = 0$

$$r^2 - 3r + 2 = (r-2)(r-1) = 0$$

$$r_1 = 2 \qquad r_2 = 1$$

The general solution is

$$Ae^{2x} + Be^{x}$$

Example 12. $y'' - 6y' + 25y = 0$

$$r^2 - 6r + 25 = 0$$

By the quadratic formula

$$r_1, r_2 \quad \text{are} \quad \frac{6 \pm \sqrt{36 - 100}}{2} = 3 \pm 4i$$

The general solution is

$$\begin{aligned} Ae^{(3+4i)} + Be^{(3-4i)}x &= Ae^{3x}e^{4ix} + Be^{3x}e^{-4ix} \\ &= e^{3x}(Ae^{4ix} + Be^{-4ix}) \\ &= e^{3x}[A(\cos 4x + i \sin 4x) \\ &\quad + B(\cos[-4x] + i \sin[-4x])] \\ &= e^{3x}[(A+B)\cos 4x \\ &\quad + i(A-B)\sin 4x] \end{aligned}$$

Here we have used the facts that

$$\cos(-x) = \cos x$$
$$\sin(-x) = -\sin x$$

and the Euler formula

$$e^{a+ib} = e^{a}(\cos b + i \sin b)$$

Finally, we consider the case in which the quadratic equation has a multiple root.

Example 13. $y'' + 6y' + 9y = 0$

$$r^2 + 6r + 9 = (r+3)^2 = 0$$

Thus, $r_1 = r_2 = -3$. *The general solution in this case is*

$$Ae^{-3x} + Bxe^{-3x}$$

(Later, we shall indicate how this result comes about.)

If (L) is not homogeneous, i.e., if $g(x)$ is not identically zero, then the procedures for solving the differential equation become more complicated and, in some cases, less effective. First we observe by a straightforward calculation that if $\bar{y}(x)$ is a given solution of (L), then if

$$Ae^{r_1 x} + Be^{r_2 x}$$

is the general solution of the corresponding homogeneous equation [i.e., equation (L) with $g(x) = 0$ for all x] it follows that

$$Ae^{r_1 x} + Be^{r_2 x} + \bar{y}(x)$$

is a solution of (L). In fact, every solution of (L) can be written in this form by a strategic choice of constants A and B.

Example 14. *Straightfoward calculation shows that one solution of*

$$y'' + 4y = x^2 + \cos x$$

is

$$\tfrac{1}{4}x^2 - \tfrac{1}{8} + \tfrac{1}{3}\cos x$$

The corresponding homogeneous equation is

$$y'' + 4y = 0$$

and its general solution is

$$\begin{aligned} Ae^{+2ix} + Be^{-2ix} &= A[\cos 2x + i\ \sin 2x] \\ &\quad + B[\cos 2x - i \sin 2x] \\ &= (A+B)(\cos 2x) + i(A-B)\sin 2x \\ &= A_1 \cos 2x + B_1 \sin 2x \end{aligned}$$

Then any solution of the differential equation may be written as

$$\frac{1}{4}x^2 - \frac{1}{8} + \frac{1}{3}\cos x + A_1 \cos 2x + B_1 \sin 2x$$

Thus the practical problem of solving (L) becomes that of finding the general solution of the corresponding homogeneous equation and just one solution (with no given initial conditions) of equation (L) itself. This conclusion is clearly of practical importance in solving (L), but it is also of considerable general importance because it holds for many other linear equations, for example, partial differential equations, integral equations, and abstract functional equations.

Now we consider the problem of finding one solution of the inhomogeneous equation. We describe first the method of undetermined coefficients which is simple but not always applicable. We suppose that $g(x)$ in equation (L) has a simple form, i.e., that $g(x)$ is a sum of products of polynomials, exponential functions e^{kx}, and trigonometric functions $\cos kx$ and $\sin kx$. Then, looking at $g(x)$, we make a guess at the form of the solution.

Example 15. $y'' + 4y = x^2 + \cos x$. *A reasonable guess at a solution is*

$$Ax^2 + Bx + C + D\cos x + E\sin x$$

and we try to determine the constant coefficients A, B, C, D, E. The derivative and second derivative of the guessed solution are

$$2AX + B - D\sin x + E\cos x$$

$$2A - D\cos x - E\sin x$$

Substituting into the differential equation, we get

$$\begin{aligned}&2A - D\cos x - E\sin x + 4Ax^2 + 4Bx + 4C\\&+ 4D\cos x + 4E\sin x = x^2 + \cos x\end{aligned}$$

or

$$\begin{aligned}&(4A-1)x^2 + 4Bx + 4C + 2A + (3D-1)\cos x\\&+ 3E\sin x = 0\end{aligned}$$

Setting the coefficients equal to zero yields

$$\begin{aligned}A &= \tfrac{1}{4}\\B &= 0\\C &= -\tfrac{1}{2}A = -\tfrac{1}{8}\\D &= \tfrac{1}{3}\\E &= 0\end{aligned}$$

So a solution is

$$\tfrac{1}{4}x^2 - \tfrac{1}{8} + \tfrac{1}{3}\cos x$$

and the general solution is

$$a\cos 2x + b\sin 2x + \tfrac{1}{4}x^2 - \tfrac{1}{8} + \tfrac{1}{3}\cos x$$

where a, b are constants.

Notice that if we consider the equation

$$y'' + 4y = x^2 + \cos 2x$$

where $g(x)$ contains the term $\cos 2x$, which is a solution of the corresponding homogeneous equation

$$y'' + 4y = 0$$

then the procedure used above fails because suppose we make the 'reasonable' guess at a solution as

$$Ax^2 + Bx + C + D\cos 2x + E\sin 2x$$

then we arrive at the equation

$$(4A-1)x^2 + 4Bx + (2A + 4C) - \cos 2x = 0$$

But no matter what values we use for A, B, C, this equation cannot be satisfied for all x because then $\cos 2$ *x would be identically equal to a polynomial, which is certainly not true. For this case, it is necessary to use a more complicated 'guess' obtained by multiplying the terms by an appropriate power of x [5, p. 155].*

We have seen that the method of undetermined coefficients, although simple and straightforward, has serious limitations, especially on the function $g(x)$. There is a far more powerful and general method, called the method of variation of constants or variation of parameters. (It must be acknowledged that the price of this power and generality is a considerable amount of calculation.) The method of variation of parameters has the additional virtue that it can be used in a much more general context. Consequently we will postpone describing it until we consider more general problems (systems of linear first-order equations) in Sec. 5.6.

5.5 LINEAR EQUATIONS WITH NONCONSTANT COEFFICIENTS

It would seem reasonable that if we consider a linear equation with coefficients which are not constant but are simple functions of x (e.g., a function like x^2) that such an equation would not be too difficult to solve. This is, however, not true. The reason is that as long as we deal with equations with constant coefficients, the solutions are composed of sums of terms of the form

$$x^r e^{(a+ib)x} = x^r e^{ax}(\cos bx + i\sin bx)$$

where r is a nonnegative integer and a and b are real. Thus the solutions are finite sums and products of integer powers of x, exponential functions e^{ax}, and trigonometric functions $\sin bx$, $\cos bx$. (This is shown in the general case in Sec. 5.6.) But as soon as we venture into the realm of equations with nonconstant coefficients, the solutions become less familiar (e.g., Bessel functions) or are completely unknown. There is no easy answer to the question of how to proceed if confronted by a linear equation in which the coefficients are functions of x. Part of the reason for this is that certain equations with nonconstant coefficients often arise in physical and engineering applications. Consequently these equations have been studied extensively and for a long time. Among the most important of these equations are the Bessel equation, the Legendre equation, and the Laguerre equation. (For an introduction to these equations and further references, see Ref. 5.) One may have the good fortune or technical insight to recognize that the given equation is or can be transformed into one of these much studied equations. (See, for example, the beautiful discussions in Ref. 5, pp. 277–283, of the Schrödinger equation for the hydrogen atom.) But if the given equation cannot be treated in this way, there is no alternative but to use a numerical method, i.e., put the equation on a computer.

Much of the study of such important equations as the Bessel equation is based on the method of power series solutions. We describe this technique as follows:

Example 16. *To find the solutions of*

$$y'' + y = 0$$

such that

$$y(0) = 0 \qquad y'(0) = 1$$

It is easy to show by the earlier method that the solution of this problem is $y(x) = \sin x$. *However, let us suppose that we do not have this answer and start by assuming that the solution can be represented as an infinite series, i.e.,*

$$y(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n + \cdots$$

Since

$$y'(x) = a_1 + 2a_2 x + \cdots + n a_n x^{n-1} + \cdots$$

and

$$y''(x) = 2a_2 + \cdots + n(n-1)a_n x^{n-2} + \cdots$$

then substituting in the equation we have

$$2a_2 + \cdots + n(n-1)a_n x^{n-2} + \cdots + a_0 + a_1 x + \cdots + a_{n-2}x^{n-2} + \cdots = 0 \qquad (*)$$

Since $(*)$ *must hold for all* x, *then the coefficient of each power of* x *must be zero. Thus we obtain*

$$\begin{aligned} a_2 &= -\frac{a_0}{2} \\ a_3 &= -\frac{a_1}{3 \times 2} \\ &\vdots \\ a_n &= -\frac{a_{n-2}}{n(n-1)} \end{aligned}$$

Since $y(0) = a_0 = 0$, *then if* n *is even*

$$a_n = 0$$

Since $y'(0) = a_1 = 1$, *then*

$$a_3 = -\frac{1}{3 \times 2}$$

and if n *is odd, i.e., if* $n = 2p + 1$, *then*

$$a_n = \frac{(-1)^p}{n!}$$

and the solution is

$$x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

which is a power series (the Maclaurin series) for $y(x) = \sin x$.

Two serious questiosn arise immediately about this procedure. First, if we manage to get a power series in this way, how do we determine whether the series converges or diverges for various values of x? This question is readily answered by application of some classical theory. (See, e.g., Ref. 5, Chap. 3.) But even if we can show that the power series converges, this gives us no practical knowledge of how the function represented by the series behaves. One can try approximating the function by using the first few terms of the power series, but we would not know, in general, how good this approximation is.

On the other hand, it should be emphasized that the power series approach has been and remains very important in the study of the particular equations mentioned earlier: Bessel's equation, the Legendre equation, and others.

5.6 SYSTEMS OF LINEAR FIRST-ORDER DIFFERENTIAL EQUATIONS

So far, we have dealt only with second-order linear equations. It is easy to guess how some of the methods we have described might be carried over to third-order or higher-order equations. However, instead of pursuing this direction we proceed at once to the study of systems of first-order linear equations. There are several reason for choosing this direction. First, a single nth-order equation can be regarded as a system of n linear first-order equations. Thus the investigation of linear first-order systems is a broadening of our study. Secondly, by taking the first-order system viewpoint, we can utilize matrix theory to obtain a coherent and complete description of how to treat the problems. Finally, systems of first-order linear equations are important for many applications.

A system of linear first-order differential equations is a set of equations

$$\begin{aligned}\frac{dx_1}{dt} &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n + b_1 \\ \frac{dx_2}{dt} &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n + b_2 \\ &\vdots \\ \frac{dx_n}{dt} &= a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n + b_n\end{aligned} \tag{S}$$

where a_{ij} $(i = 1, \ldots, n; j = 1, \ldots, n)$ will be assuemd to be constant and $b_1, \ldots, b_n$ are functions of t. If

$$b_i(t) \equiv 0 \qquad i = 1, \ldots, n$$

the system is said to be *homogeneous*. In system (S) the t is the independent variable and solving system (S) means finding functions $x_1(t), \ldots, x_n(t)$ which satisfy system (S). Specifying an initial condition for the solution means specifying initial values $x_1(t_0), \ldots, x_n(t_0)$, all at the same value t_0 of the independent variable.

Each of the equations in (S) is a first-order equation, and we show first how to write a single nth-order linear equation as a system of first-order equations. We consider the equation

$$\begin{aligned}&\frac{d^n x}{dt^n} + a_n\frac{d^{n-1}x}{dt^{n-1}} + a_{n-1}\frac{d^{n-2}x}{dt^{n-2}} + \cdots + a_3\frac{d^2x}{dt^2} \\ &\quad + a_2\frac{dx}{dt} + a_1 x + b(t) = 0\end{aligned} \tag{E}$$

Let

$$\begin{aligned}x_1 &= x \\ x_2 &= \frac{dx}{dt} = \frac{dx_1}{dt} \\ x_3 &= \frac{d^2x}{dt^2} = \frac{d}{dt}\left[\frac{dx}{dt}\right] = \frac{dx_2}{dt} \\ &\vdots \\ x_n &= \frac{d^{n-1}x}{dt^{n-1}} = \frac{d}{dt}\left[\frac{d^{n-2}x}{dt^{n-2}}\right] = \frac{dx_{n-1}}{dt}\end{aligned}$$

Then equation (E) can be written as the system

$$\begin{aligned}\frac{dx_1}{dt} &= x_2 \\ \frac{dx_2}{dt} &= x_3 \\ &\vdots \\ \frac{dx_{n-1}}{dt} &= x_n \\ \frac{d^n x}{dt^n} = \frac{dx_n}{dt} &= -a_1x_1 - a_2x_2 - \cdots - a_{n-1}x_{n-1} \\ &\quad - a_n x_n - b(t)\end{aligned}$$

Although it is by no means obvious at this stage, it turns out that it is better to treat the single nth-order equation as a special case of a system of first-order equations. We shall see why later. But in order to develop the theory for systems of first-order equations, we need a few facts about matrices.

5.6.1 Some Theory of Matrices

A matrix is a rectangular array (m rows, n columns) of numbers

$$\begin{bmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & & \\ a_{m1} & \ldots & a_{mn}\end{bmatrix}$$

Sometimes the matrix is denoted by $[a_{ij}]$.

5.6.1.1 Sums, Determinants, and Products of Matrices

The sum of two matrices is defined for two matrices with the same number of rows and columns as follows:

$$\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \ldots & a_{mn}\end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & \ldots & b_{1n} \\ b_{21} & b_{22} & \ldots & b_{2n} \\ \vdots & & & \\ b_{m1} & b_{m2} & \ldots & b_{mn}\end{bmatrix}$$
$$= \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \ldots & a_{1n}+b_{1n} \\ a_{21}+b_{21} & a_{22}+b_{22} & \ldots & a_{2n}+b_{2n} \\ \vdots & & & \\ a_{m1}+b_{m1} & a_{m2}+b_{m2} & \ldots & a_{mn}+b_{mn}\end{bmatrix}$$

Multiplication of a matrix by a number (or scalar) is defined as follows: if c is a number then the product of c and the matrix $[a_{ij}]$ is

$$c\begin{bmatrix} a_{11} & \dots & a_{1n} \\ a_{21} & & a_{2n} \\ & & \\ a_{m1} & & a_{mn} \end{bmatrix} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ a_{21} & & a_{2n} \\ & & \\ a_{m1} & & a_{mn} \end{bmatrix} c$$

$$= \begin{bmatrix} ca_{11} & \dots & ca_{1n} \\ ca_{21} & & ca_{2n} \\ & & \\ ca_{m1} & & ca_{mn} \end{bmatrix}$$

We will need to consider only square matrices ($m = n$) and single column matrices ($n = 1$). (A single column matrix is a vector.)

If A is a square matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ \vdots & & & \\ a_{m1} & a_{m2} & & a_{mm} \end{bmatrix}$$

the determinant of A, denoted by $\det A$, is defined inductively as follows. If $m = 2$, $\det A = a_{11}a_{22} - a_{12}a_{21}$. If $\det A$ is defined for A, an $(m-1) \times (m-1)$ matrix, let B be an $m \times m$ matrix,

$$B = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & & & \\ b_{m1} & b_{m2} & \dots & b_{mm} \end{bmatrix}$$

Let $B(i, j)$ be the $(m-1) \times (m-1)$ matrix obtained by deleting the ith row and the jth column of matrix B. Then $\det B$ can be defined to be

$$\det B = \sum_{j=1}^{n} (-1)^{i+j} b_{ij} B(i, j)$$

where i is a fixed integer with $1 \leq i \leq m$ or $\det B$ can be defined as

$$\det B = \sum_{i=1}^{m} (-1)^{i+j} b_{ij} B(i, j)$$

where j is fixed with $1 \leq j \leq m$.

All these expressions for $\det B$ yield the same result. The first equation is called expansion by the ith row. The second equation is called expansion by the jth column.

Example 17. *Expansion by the first row:*

$$\begin{bmatrix} 1 & -2 & 7 \\ 3 & 2 & 6 \\ -5 & 0 & 4 \end{bmatrix} = (1)[2(4) - 6(0)] + (-1)(-2)[3(4) - 6(-5)] + (7)[3(0) - (2)(-5)]$$
$$= 8 + 2(42) + 7(10)$$
$$= 8 + 84 + 70$$
$$= 162$$

Expansion by the second column:

$$\begin{bmatrix} 1 & -2 & 7 \\ 3 & 2 & 6 \\ -5 & 0 & 4 \end{bmatrix} = (-1)(-2)[12 + 30] + (2)[(4) - 7(-5)]$$
$$= 2(42) + 2(39)$$
$$= 84 + 78$$
$$= 162$$

If A and B are two square matrices, their products are defined to be

$$AB = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & \\ & & & \\ a_{m1} & \dots & \dots & a_{mm} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & & \\ & & & \\ b_{m1} & b_{m2} & & b_{mm} \end{bmatrix}$$

$$= \begin{bmatrix} c_{11} & c_{12} & \dots & x_{1m} \\ c_{m1} & c_{m2} & & c_{mm} \end{bmatrix}$$

where

$$c_{11} = a_{11}b_{11} + a_{12}b_{21} + \dots + a_{1m}b_{m1}$$
$$= \sum_{j=1}^{m} a_{1j}b_{j1}$$

That is, c_{11} is the dot product of the first row of A with the first column of B. Also

$$c_{12} = a_{11}b_{12} + a_{12}b_{22} + \dots + a_{1m}b_{m2}$$
$$= \sum_{j=1}^{m} a_{1j}b_{j2}$$

That is, c_{12} is the dot product of the first row of A with the second column of B. Generally, if $k = 1, \dots, m$ and $\ell = 1, \dots, m$, then

$$c_{k\ell} = \sum_{j=1}^{m} a_{kj}b_{j\ell}$$

Similarly

$$BA = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & & & \\ b_{m1} & b_{m2} & \dots & b_{mm} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ & \dots & & \\ a_{m1} & a_{m2} & \dots & a_{mm} \end{bmatrix}$$

$$= \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1m} \\ d_{21} & d_{22} & \dots & d_{2m} \\ & \dots & & \\ d_{m1} & d_{m2} & \dots & d_{mm} \end{bmatrix}$$

where

$$d_{k\ell} = \sum_{j=1}^{m} b_{kj} a_{j\ell}$$

WARNING. In general $AB \neq BA$. That is, multiplication of matrices is not commutative, as the following example shows.

Example 18

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \qquad B = \begin{bmatrix} 2 & 3 \\ 1 & 4 \end{bmatrix}$$

Then

$$AB = \begin{bmatrix} 4 & 11 \\ 10 & 25 \end{bmatrix} \qquad BA = \begin{bmatrix} 11 & 16 \\ 13 & 18 \end{bmatrix}$$

The $m \times m$ matrix

$$I = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}$$

in which all the diagonal entries, a_{ii}, are equal to 1 *and every entry off the diagonal, a_{ij} with $i \neq j$, is equal to* 0, *is called the* identity matrix *because if A is any $m \times m$ matrix then*

$$AI = IA = A$$

If A has an inverse A^{-1}, i.e., if there exists a matrix A^{-1} such that $AA^{-1} = I$, the identity matrix, then A and A^{-1} commute:

$$AA^{-1} = A^{-1}A = I$$

It can be proved that A has an inverse if and only if $\det A \neq 0$.

Multiplication of a square matrix and a vector is defined as follows:

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ & \dots & & \\ a_{m1} & a_{m2} & \dots & a_{mm} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{m} a_1 j b_j \\ \sum_{j=1}^{m} a_{2j} b_j \\ \sum_{j=1}^{m} a_{mj} b_j \end{bmatrix}$$

5.6.1.2 Exponent of a Matrix

The *exponent of a matrix* is a very useful concept in systems of linear differential equations. It is defined by using the infinite series for the familiar exponential function, i.e.,

$$e^x = 1 + x + \frac{x^2}{2!} + \dots + \frac{x^n}{n!} + \dots \qquad (*)$$

If A is a square matrix, we define the exponent of A as

$$e^A = I + A + \frac{A^2}{2!} + \dots + \frac{A^n}{n!} + \dots \qquad (**)$$

For this definition to make sense it is necessary to say what we mean by convergence of an infinite series of matrices and to show that the above infinite series of matrices converges. In order to define the convergence, we must introduce a definition of distance between matrices A and B, denoted by $|A - B|$. This distance is defined as

$$|A - B| = \sum_{i,j=1}^{m} |a_{ij} - b_{ij}|$$

The formal definition of convergence is based on this distance and using the convergence of the familiar exponential series $(*)$, it can be proved that $(**)$ converges.

5.6.1.3 Eigenvalues and Eigenvectors of a Matrix

An important concept for matrices is an *eigenvalue* of a matrix. Given a matrix $A = [a_{ij}]$, then an eigenvalue of A is a number λ such that

$$\det \begin{bmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} - \lambda & & a_{2m} \\ & & & \\ a_{m1} & a_{m2} & & a_{mm} - \lambda \end{bmatrix} = 0$$

Example 19. *Find the eigenvalues of*

$$A = \begin{bmatrix} 4 & 2 & 8 \\ 1 & 13/3 & 16/3 \\ -1 & -4/3 & -7/3 \end{bmatrix}$$

The eigenvalues are the solutions of the polynomial equation

$$\begin{aligned} &\det \begin{bmatrix} 4-\lambda & 2 & 8 \\ 1 & 13/3-\lambda & 16/3 \\ -1 & -4/3 & -(7/3)-\lambda \end{bmatrix} \\ &= -\left[(\lambda-4)\left(\lambda-\tfrac{13}{3}\right)\left(\lambda+\tfrac{7}{3}\right)\right] + (4-\lambda)\left[-\left(\tfrac{16}{3}\right)\left(-\tfrac{4}{3}\right)\right] \\ &\quad + 2\left(\lambda-\tfrac{9}{3}\right) + 8\left(-\tfrac{4}{3}+\tfrac{13}{3}-\lambda\right) = 0 \\ &\quad - \lambda^3 + 6\lambda^2 - 11\lambda + 6 = 0 \end{aligned}$$

or

$$\lambda^3 - 6\lambda^2 + 11\lambda - 6 = 0$$

The number 6 must be divisible by any integer solution of this polynomial equation. So the possible integer solutions include 2 and 3. Using synthetic division, we have

$$\begin{array}{cccc|l} 1 & -6 & +11 & -6 & \underline{3} \\ & +3 & -9 & +6 & \\ \hline 1 & -3 & 2 & & \end{array}$$

That is: $\lambda^3 - 6\lambda^2 + 11\lambda - 6 = (\lambda - 3)(\lambda^2 - 3\lambda + 2)$. *So the solutions are* $\lambda = 1, 2, 3$.

The eigenvalues are the solutions of a polynomial equation, that is, if A is an $m \times m$ matrix, the eigenvalues of A are the solutions of a polynomial equation of degree m. Hence by a theorem from algebra, the matrix A has m eigenvalues if each eigenvalue is counted with its multiplicity.

It is straightforward to prove that the product of the eigenvalues is $\det A$. Thus 0 is an eigenvalue of A if and only if $\det A = 0$.

Later we will need the concept of an eigenvector. If λ is an eigenvalue of matrix A, then

$$\det(A - \lambda I) = 0$$

As will be shown later (after Example 25), it follows that there is a nonzero vector x such that

$$(A - \lambda I)x = 0$$

Vector x is called an *eigenvector* of A. Note that if x is an eigenvector and c is a nonzero constant, then cx is also an eigenvector.

5.6.1.4 Canonical Forms of a Matrix

With one more bit of matrix theory, we reach the result which is used to given an explicit formula for the solutions of a system of m first-order homogeneous equations with constant coefficients. If A is an $m \times m$ matrix with entries a_{ij} which are real or complex, then there is a constant matrix P with an inverse P^{-1} such that

$$P^{-1}AP = J$$

where J is a matrix called the *Jordan canonical form* and J has the form

$$\begin{bmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_s \end{bmatrix}$$

where all the entries not written in are zero and J_j ($j = 1, \ldots, s$) has the form

$$J_j = \begin{bmatrix} \lambda_j & 1 & & & \\ & \lambda_j & 1 & & \\ & & \ddots & & \\ & & & & 1 \\ & & & & \lambda_j \end{bmatrix}$$

where λ_j is an eigenvalue of A. (All the entries not written in are zero.) Each eigenvalue of A appears in at least one of the matrices J_j. (The eigenvalues λ_j are, in general, not distinct.)

If all the entries of A are real, then there is a real constant matrix P with an inverse P^{-1} such that

$$P^{-1}AP = \tilde{J}$$

where $\tilde{J}$ is a matrix called the *real canonical form*, all the entries of $\tilde{J}$ are real, and $\tilde{J}$ has the form

$$\tilde{J} = \begin{bmatrix} \tilde{J}_1 & & & \\ & \tilde{J}_2 & & \\ & & \ddots & \\ & & & \tilde{J}_s \end{bmatrix}$$

and $\tilde{J}_j$ is associated with eigenvalue λ_j and has one of the two following forms: if λ_j is real,

$$\tilde{J}_j = \begin{bmatrix} \lambda_j & 1 & & & \\ & \lambda_j & 1 & & \\ & & \ddots & \ddots & \\ & & & & 1 \\ & & & & \lambda_j \end{bmatrix}$$

Each λ_j appears in at least one $\tilde{J}_j$. If λ_j is a complex eigenvalue,

$$\lambda_j = \alpha_j + i\beta_j$$

where α_j and β_j are real and $\beta_j > 0$. (Note that since A is real, then if $\alpha_j + i\beta_j$ is an eigenvalue, so is $\alpha_j - i\beta_j$.) Then

$$\tilde{J}_j = \begin{bmatrix} \alpha_j & \beta_j & 1 & 0 & & & & \\ -\beta_j & \alpha_j & 0 & 1 & & & & \\ & & & \ddots & & & & \\ & & & & \alpha_j & \beta_j & 1 & 0 \\ & & & & -\beta_j & \alpha_j & 0 & 1 \\ & & & & & & \alpha_j & \beta_j \\ & & & & & & -\beta_j & \alpha_j \end{bmatrix}$$

Corresponding to each $\lambda_j = \alpha_j + i\beta_j$, there is at least one $\tilde{J}_j$.

Example 20. *The eigenvalues of*

$$\begin{bmatrix} -19 & 30 & 0 \\ -15 & 23 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

are the solutions of the polynomial equation

$$\det \begin{bmatrix} -19-\lambda & 30 & 0 \\ -15 & 23-\lambda & 0 \\ 0 & 0 & 8-\lambda \end{bmatrix} = 0$$

or, expanding the determinant by the third column,

$$(8-\lambda)[(-19-\lambda)(23-\lambda)+450] = 0$$

$$(8-\lambda)[(-19)23 + (19-23)\lambda + \lambda^2 + 450] = 0$$

$$(8-\lambda)[\lambda^2 - 4\lambda - 437 + 450] = 0$$

$$(8-\lambda)[\lambda^2 - 4\lambda + 13] = 0$$

The eigenvalues are 8 and

$$\frac{4 \pm \sqrt{16-52}}{2}$$

or

$$\frac{4+6i}{2} \qquad \frac{4-6i}{2}$$

or

$$2+3i \qquad 2-3i$$

Then

$$\tilde{J} = \begin{bmatrix} 2 & 3 & 0 \\ -3 & 2 & 0 \\ 0 & 0 & 8 \end{bmatrix}$$

In general, finding the Jordan canomical form or the real canonical form can be a lengthy process even if m is fairly small. (For a more detailed description of the canonical forms, a description of how to calculate them, and an example with $m = 7$, see Ref. 9, Chap. 2.)

5.6.2 Solutions of a Homogeneous System of Linear Differential Equations

With these facts about matrices, we return to the study of systems of linear differential equations. We consider a linear homogeneous system with constant coefficients

$$\begin{aligned} \frac{dx_1}{dt} &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m \\ \frac{dx_2}{dt} &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2m}x_m \\ &\vdots \\ \frac{dx_m}{dt} &= a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mm}x_m \end{aligned} \qquad (\mathcal{L})$$

We may rewrite $(\mathcal{L})$ in matrix form as

$$\begin{bmatrix} dx_1/dt \\ dx_2/dt \\ \vdots \\ dx_m/dt \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ & & \vdots & \\ a_{m1} & a_{m2} & \dots & a_{mm} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

If we introduce the notation

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

$$\frac{dx}{dt} = \begin{bmatrix} dx_1/dt \\ dx_2/dt \\ \vdots \\ dx_m/dt \end{bmatrix}$$

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ & & \vdots & \\ a_{m1} & a_{m2} & & a_{mm} \end{bmatrix}$$

then the linear system ($\mathcal{L}$) can be written in the much briefer form

$$\frac{dx}{dt} = Ax \qquad (\mathcal{L}_\text{m})$$

Before discussing the problem of solving ($\mathcal{L}$) or ($\mathcal{L}_\text{m}$), we state a couple of facts about the set of solutions of ($\mathcal{L}$). These facts are stated in terms of the notion of linear independence. We will then discuss the meaning of linear independence.

The *basic facts* are:

1. System ($\mathcal{L}$) has m linearly independent solutions.
2. Every solution of ($\mathcal{L}$) can be expressed as a linear combination of m linearly independent solutions.

(For proofs of these statements, see Ref. 9, Chap. 2.)

The notion of linear independence is of crucial importance throughout the study of linear equations of all sorts. Unfortuantely, it has an elusive quality that is difficult to deal with, at least at first. Part of this difficulty stems from the fact that the definition of linear independence is given in negative terms.

First we define linear dependence. Functions $f_1(t), \ldots, f_n(t)$ are *linearly dependent* if there exist constants $a_1, \ldots, a_n$, not all zero, such that for all t

$$a_1 f_1(t) + a_2 f_2(t) + \cdots + a_n f_n(t) = 0$$

(Speaking informally, we would say that $f_1, \ldots, f_n$ are linearly dependent if one of them can be written as a linear combination of the others.) The vectors

$$\begin{bmatrix} x_{11}(t) \\ x_{21}(t) \\ \vdots \\ x_{m1}(t) \end{bmatrix}, \begin{bmatrix} x_{12}(t) \\ x_{22}(t) \\ \vdots \\ x_{m2}(t) \end{bmatrix}, \ldots, \begin{bmatrix} x_{1n}(t) \\ x_{2n}(t) \\ \vdots \\ x_{mn}(t) \end{bmatrix}$$

are *linearly dependent* if there exist constants $a_1, \ldots, a_n$, not all zero, such that for all t,

$$a_1 \begin{bmatrix} x_{11}(t) \\ x_{21}(t) \\ \\ x_{m1}(t) \end{bmatrix} + a_2 \begin{bmatrix} x_{12}(t) \\ x_{22}(t) \\ \vdots \\ x_{m2}(t) \end{bmatrix} + \cdots + a_n \begin{bmatrix} x_{1n}(t) \\ x_{2n}(t) \\ \vdots \\ x_{mn}(t) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

or

$$\begin{bmatrix} a_1 x_{11}(t) \\ a_1 x_{21}(t) \\ \vdots \\ a_1 x_{m1}(t) \end{bmatrix} + \begin{bmatrix} a_2 x_{12}(t) \\ a_2 x_{22}(t) \\ \vdots \\ a_2 x_{m2}(t) \end{bmatrix} + \cdots + \begin{bmatrix} a_n x_{1n}(t) \\ a_n x_{2n}(t) \\ \vdots \\ a_n x_{mn}(t) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

(The zero vector on the right is sometimes denoted simply by 0.) This last vector equation is equivalent to the scalar equations

$$\begin{aligned} a_1 x_{11}(t) + a_2 x_{12}(t) + \cdots + a_n x_{1n}(t) &= 0 \\ a_1 x_{21}(t) + a_2 x_{22}(t) + \cdots + a_n x_{2n}(t) &= 0 \\ &\vdots \\ a_1 x_{m1}(t) + a_2 x_{m2}(t) + \cdots + a_n x_{mn}(t) &= 0 \end{aligned}$$

(Note that in these definitions, any or all of the functions $f_j(t)$, $x_{ij}(t)$ may be constant functions.)

Example 21. *The constant vectors*

$$\begin{bmatrix} 3 \\ 2 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 12 \\ 8 \end{bmatrix}$$

are linearly dependent because

$$(-4)\begin{bmatrix} 3 \\ 2 \end{bmatrix} + \begin{bmatrix} 12 \\ 8 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Example 22. *To show that the vectors*

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \begin{bmatrix} 17 \\ 18 \end{bmatrix}$$

are linearly dependent, we show that there are numbers a_1, a_2, such that

$$a_1 \begin{bmatrix} 1 \\ 2 \end{bmatrix} + a_2 \begin{bmatrix} 3 \\ 4 \end{bmatrix} + \begin{bmatrix} 17 \\ 18 \end{bmatrix} = 0$$

or

$$\begin{aligned} a_1 + 3a_2 &= -17 \\ 2a_1 + 4a_2 &= -18 \end{aligned}$$

These simultaneous equations have a unique solution for a_1 and a_2 because

$$\det \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix} = -2 \neq 0$$

(Actually it can be proved that if n is any integer, then a set of $(n+1)$ constant n-vectors must be linearly dependent.)

Functions $f_1(t), \dots, f_n(t)$ are *linearly independent* if they are not linearly dependent. Similarly, vectors are linearly independent if they are not linearly dependent. Thus if we want to show that a set of functions or vectors is linearly independent then we must show that there is no set of numbers $a_1, \dots, a_n$ with the properties described in the definition of linear dependence. It is by no means obvious how to go about this.

Example 23. *The functions*

$$\begin{aligned} f_1(t) &= 1 \\ f_2(t) &= t \\ f_3(t) &= t^2 \end{aligned}$$

are linearly independent because if a_1, a_2, a_3 are fixed numbers, then the equation

$$a_1 + a_2 t + a_3 t^2 = 0$$

holds for at most two values of t (the solutions of the quadratic equation) unless $a_1 = 0$, $a_2 = 0$, and $a_3 = 0$.

Example 24. *Similarly the functions*

$$\begin{aligned} f_1(t) &= 1 \\ f_2(t) &= t^{12} \\ f_3(t_3) &= t^{17} \end{aligned}$$

are linearly independent because the equation

$$a_1 + a_2 t^{12} + a_3 t^{17} = 0$$

holds for, at most, 17 values of t (the solutions of the polynomial equation of degree 17) unless $a_1 = 0$, $a_2 = 0$, and $a_3 = 0$.

Example 25. *The vectors*

$$\begin{bmatrix} 1 \\ -2 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \\ 6 \end{bmatrix}, \begin{bmatrix} 5 \\ 0 \\ 2 \end{bmatrix}$$

are linearly independent because if

$$a_1 \begin{bmatrix} 1 \\ -2 \\ 4 \end{bmatrix} + a_2 \begin{bmatrix} 1 \\ 3 \\ 6 \end{bmatrix} + a_3 \begin{bmatrix} 5 \\ 0 \\ 2 \end{bmatrix} = 0$$

then

$$\begin{aligned} a_1 + a_2 + 5a_3 &= 0 \\ -2a_1 + 3a_2 &= 0 \qquad (*) \\ 4a_1 + 6a_2 + 2a_3 &= 0 \end{aligned}$$

But since the coefficient determinant is

$$\det \begin{bmatrix} 1 & 1 & 5 \\ -2 & 3 & 0 \\ 4 & 6 & 2 \end{bmatrix} = 5(-12 - 12) + 2(3 + 2)$$

$$= -120 + 10 \neq 0$$

then the system of equations $(*)$ has the unique solution $a_1 = a_2 = a_3 = 0$.

The same kind of argument shows that more generally, we have: n constant n-vectors are linearly independent if the corresponding determinant is nonzero. Also if the n-vectors are linearly independent, then the corresponding determinant is nonzero. Finally, it is easy to show that eigenvectors exist: since

$$\det(A - \lambda I) = 0$$

then the columns of $A - \lambda I$ are linearly dependent. That is, there exists a nonzero vector x such that

$$(A - \lambda I)x = 0$$

Example 26. *Suppose we have two 3-vectors*

$$\begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \end{bmatrix}, \begin{bmatrix} x_{12} \\ x_{22} \\ x_{32} \end{bmatrix}$$

then if

$$\det \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \neq 0$$

the vectors are linearly independent. To prove this suppose that

$$a_1 \begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \end{bmatrix} + a_2 \begin{bmatrix} x_{12} \\ x_{22} \\ x_{32} \end{bmatrix} = 0$$

that is

$$\begin{aligned} a_1 x_{11} + a_2 x_{12} &= 0 \\ a_1 x_{21} + a_2 x_{22} &= 0 \\ a_1 x_{31} + a_2 x_{32} &= 0 \end{aligned}$$

Then the equations

$$\begin{aligned} a_1 x_{11} + a_2 x_{12} &= 0 \\ a_1 x_{21} + a_2 x_{22} &= 0 \end{aligned}$$

can be solved for a_1, a_2. *Since*

$$\det\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \neq 0$$

the solution is $a_1 = 0$, $a_2 = 0$.

The basic facts stated earlier essentially tell us that in order to find the solutions of system $(\mathcal{L})$, it is sufficient to find m linearly independent solutions. The first step in such a search is to observe that the matrix equation

$$\frac{dx}{dt} = Ax \qquad (\mathcal{L}_M)$$

which is another form of $(\mathcal{L})$, is a generalization of the simple scalar equation

$$\frac{dx}{dt} = ax$$

and we know that a solution of this equation is e^{ta}. This suggests that the exponential

$$e^{tA} = I + tA + \frac{t^2A^2}{2!} + \cdots + \frac{t^nA^n}{n} + \cdots \qquad \text{(F)}$$

is, in some sense, a solution of $(\mathcal{L}_M)$. In fact it is not difficult to prove that the m columns of the matrix e^{tA} are m linearly independent solutions of $(\mathcal{L}_M)$. [A matrix with this property is called a *fundamental matrix* of $(\mathcal{L}_M)$.] Indeed, for $t = 0$, we have

$$e^{0A} = I$$

and hence the solution $x(t)$ of $(\mathcal{L}_M)$ such that

$$x(0) = \bar{x}$$

where $\bar{x}$ is a given vector, is

$$x(t) = e^{tA}\bar{x}$$

A glance at equation (F) is enough to suggest that these statements have no practical value. If A is, say, a 3 × 3 matrix, how can the powers of A be calculated? That is, how can an explicit form for the infinite series be obtained? Also how can we determine what this infinite series converges to for various values of t? These would seem to be difficult problems, but they are fairly easy to solve if we use the Jordan canonical form or the real canonical form described earlier. Since we are interested mainly in the case in which all the entries in A are real and in finding real solutions we will use the real canonical form. We stated earlier that there exists a real matrix P with inverse P^{-1} such that

$$P^{-1}AP = \tilde{J}$$

where $\tilde{J}$ is the real canonical form described in Sec. 5.6.1.3. Multiplying this equation on the left by P and on the right by P^{-1}, we obtain

$$A = P\tilde{J}P^{-1}$$

and

$$\begin{aligned} A^2 = (P\tilde{J}P^{-1})(P\tilde{J}P^{-1}) &= P\tilde{J}(P^{-1}P)\tilde{J}P^{-1} \\ &= P(\tilde{J})^2P^{-1} \end{aligned}$$

and for any integer n

$$A^n = P(\tilde{J})^nP^{-1}$$

Then by equation (F),

$$\begin{aligned} e^{tA} &= I + tP\tilde{J}P^{-1} + \frac{t^2P(\tilde{J})^2P^{-1}}{2!} + \cdots \\ &\qquad + \frac{t^nP(\tilde{J})^nP^{-1}}{n!} + \cdots \\ &= P\left[1 + t\tilde{J} + \frac{t^2(\tilde{J})^2}{2!} + \cdots + \frac{t^n(\tilde{J})^n}{n!} + \cdots\right]P^{-1} \\ &= Pe^{t\tilde{J}}P^{-1} \end{aligned}$$

Calculating $(\tilde{J})^n$ is not difficult and with some simple, although rather lengthy, computations the expression $e^{t\tilde{J}}$ can be explicitly determined. (For a detailed description of this, see Ref. 9, Chap. 2.) Here we will just sketch the results. First, we have

$$e^{t\tilde{J}} = \begin{bmatrix} e^{t\tilde{J}_1} & & & \\ & e^{t\tilde{J}_2} & & \\ & & \ddots & \\ & & & e^{t\tilde{J}_s} \end{bmatrix}$$

So it is sufficient to exhibit $e^{t\tilde{J}}$ where $\tilde{J}$ has the form, if the eigenvalue λ is complex,

$$\tilde{J} = \begin{bmatrix} \alpha & \beta & 1 & 0 & & & & \\ -\beta & \alpha & 0 & 1 & & & & \\ & & \ddots & & & & & \\ & & & & \alpha & \beta & 1 & 0 \\ & & & & -\beta & \alpha & 0 & 1 \\ & & & & & & \alpha & \beta \\ & & & & & & -\beta & \alpha \end{bmatrix}$$

or, if λ is real,

$$\tilde{J} = \begin{bmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda & 1 \\ & & & & \lambda \end{bmatrix}$$

Simple calculations show that if λ is real and J is such an $m \times m$ matrix, then

$$e^{tJ} = \begin{bmatrix} e^{t\lambda} & t^{t\lambda} & \cdots & \dfrac{t^{m-1}}{(m-1)!}e^{t\lambda} \\ & e^{t\lambda} & te^{t\lambda}\ldots & \\ & & \ddots & \\ & & & e^{t\lambda} \end{bmatrix}$$

If λ is complex, then simple but lengthier calculations show that e^{tJ} has the following form, which we indicate only briefly:

$$e^{tJ} = \begin{bmatrix} e^{\alpha t}\cos\beta t & e^{\alpha t}\sin\beta t & te^{\alpha t}\cos\beta t & te^{\alpha t}\sin\beta t & \\ -e^{\alpha t}\sin\beta t & e\alpha t\cos\beta t & -te^{\alpha t}\sin\beta t & te^{\alpha t}\cos\beta t & \ldots \\ & & e^{\alpha t}\cos\beta t & e^{\alpha t}\sin\beta t & \\ & & -e^{\alpha t}\sin\beta t & e^{\alpha t}\cos\beta t & \ldots \\ & & \vdots & & \end{bmatrix}$$

(These expressions appear to be rather complicated, but it is important to notice that each term is simply an integer power of t multiplied by a real exponential or a sine or a cosine.) Finally, it is not difficult to show that since

$$e^{tA} = Pe^{t\tilde{J}}P^{-1}$$

is a fundamental matrix, so is $Pe^{t\tilde{J}}$. [An $m \times m$ matrix whose m columns are linearly independent solutions of $(\mathcal{L}_M)$ is a *fundamental* matrix of $(\mathcal{L}_M)$.]

Example 27. *Find the solution of*

$$\frac{dx}{dt} = 4x + 2y + 8z$$
$$\frac{dy}{dt} = x + \frac{13}{3}y + \frac{16}{3}z$$
$$\frac{dz}{dt} = -x - \frac{4}{3}y - \frac{7}{3}z$$

which satisfies the initial condition

$$x(0) = 1$$
$$y(0) = -1$$
$$z(0) = 4$$

By Example 19, the eigenvalues of the coefficient matrix

$$A = \begin{bmatrix} 4 & 2 & 8 \\ 1 & 13/3 & 16/3 \\ -1 & -4/3 & -7/3 \end{bmatrix}$$

are the number 1, 2, 3. Hence the Jordan canonical form of A is

$$J = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

It is easy to see that if n is a positive integer, then

$$(tJ)^n = \begin{bmatrix} t^n & 0 & 0 \\ 0 & (2t)^n & 0 \\ 0 & 0 & (3t)^n \end{bmatrix}$$

and it follows that

$$e^{tJ} = \begin{bmatrix} e^t & 0 & 0 \\ 0 & e^{2t} & 0 \\ 0 & 0 & e^{3t} \end{bmatrix}$$

As remarked above, a fundamental matrix of the system of differential equation is Pe^{tJ}.

It remains to determine the matrix P. As is shown in a detailed discussion of canonical forms (e.g., Ref. 9, Chap. 2), the columns of P are eigenvectors associated with the eigenvalues. If $\lambda = 1$, *we solve the following equation in order to find an eigenvector:*

$$\begin{bmatrix} 4-1 & 2 & 8 \\ 1 & (13/3)-1 & 16/3 \\ -1 & -4/3 & -(7/3)-1 \end{bmatrix}\begin{bmatrix} p_{11} \\ p_{21} \\ p_{31} \end{bmatrix} = 0$$

or

$$\begin{bmatrix} 3 & 2 & 8 \\ 1 & 10/3 & 16/3 \\ -1 & -4/3 & -10/3 \end{bmatrix}\begin{bmatrix} p_{11} \\ p_{21} \\ p_{31} \end{bmatrix} = 0$$

or

$$3p_{11} + 2p_{21} + 8p_{31} = 0 \qquad (1)$$
$$p_{11} + \tfrac{10}{3}p_{21} + \tfrac{16}{3}p_{31} = 0 \qquad (2)$$
$$-p_{11} - \tfrac{4}{3}p_{21} - \tfrac{10}{3}p_{31} = 0 \qquad (3)$$

Adding (2) *and* (3) *yields*

$$2p_{21} + 2p_{31} = 0 \qquad \text{or} \qquad p_{21} = -p_{31}$$

Multiplying (3) *by 5/2 and adding to* (2) *yields*

$$-\tfrac{3}{2}p_{11} - \tfrac{9}{3}p_{31} = 0$$
$$-3p_{11} - 6p_{31} = 0$$
$$p_{11} = -2p_{31}$$

So an eigenvector associated with eigenvector $\lambda = 1$ *is (if we take* $p_{31} = 1$*)*

$$\begin{bmatrix} -2 \\ -1 \\ 1 \end{bmatrix}$$

If $\lambda = 2$, *we solve*

$$\begin{bmatrix} 4-2 & 2 & 8 \\ 1 & (13/3)-2 & 16/3 \\ -1 & -4/3 & -(7/3)-2 \end{bmatrix} \begin{bmatrix} p_{12} \\ p_{22} \\ p_{32} \end{bmatrix} = 0$$

$$2p_{12} + 2p_{22} + 8p_{32} = 0 \tag{4}$$

$$p_{12} + \tfrac{7}{3}p_{22} + \tfrac{16}{3}p_{32} = 0 \tag{5}$$

$$-p_{12} - \tfrac{4}{3}p_{22} - \tfrac{13}{3}p_{32} = 0 \tag{6}$$

Adding (5) *and* (6), *we obtain*

$$p_{22} + p_{32} = 0 \qquad \text{or} \qquad p_{22} = -p_{32}$$

Substituting in (4) *yields*

$$2p_{12} + 6p_{32} = 0$$

or

$$p_{12} = -3p_{32}$$

So an eigenvector associated with the eigenvalue $\lambda = 2$ *is (if we take* $p_{32} = 1$*)*

$$\begin{bmatrix} -3 \\ -1 \\ 1 \end{bmatrix}$$

If $\lambda = 3$ *we solve*

$$\begin{bmatrix} 4-3 & 2 & 8 \\ 1 & (13/3)-3 & 16/3 \\ -1 & -4/3 & -(7/3)-3 \end{bmatrix} \begin{bmatrix} p_{13} \\ p_{23} \\ p_{33} \end{bmatrix} = 1$$

$$p_{13} + 2p_{23} + 8p_{33} = 0 \tag{7}$$

$$p_{13} + \tfrac{4}{3}p_{23} + \tfrac{16}{3}p_{33} = 0 \tag{8}$$

$$-p_{13} - \tfrac{4}{3}p_{23} - \tfrac{16}{3}p_{33} = 0 \tag{9}$$

Subtracting (8) *from* (7) *yields*

$$\tfrac{2}{3}p_{23} + \tfrac{8}{3}p_{33} = 0 \qquad \text{or} \qquad p_{23} = -4p_{33}$$

Substituting in (9) *yields*

$$-p_{13} + \tfrac{4}{3}(4p_{33}) - \tfrac{16}{3}p_{33} = 0 \qquad \text{or} \qquad p_{13} = 0$$

So an eigenvalue associated with eigenvalue $\lambda = 3$ *is (if we take* $p_{33} = 1$*)*

$$\begin{bmatrix} 0 \\ -4 \\ 1 \end{bmatrix}$$

Now let matrix P have these three eigenvectors as its columns. That is,

$$P = \begin{bmatrix} -2 & -3 & 0 \\ -1 & -1 & -4 \\ 1 & 1 & 1 \end{bmatrix}$$

Then since the columns of P are eigenvectors, we have

$$AP = \begin{bmatrix} -2 & 2(-3) & 3(0) \\ -1 & 2(-1) & 3(-4) \\ 1 & 2(1) & 3(1) \end{bmatrix}$$

and

$$P^{-1}AP = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2(1) & 0 \\ 0 & 0 & 3(1) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

A fundamental matrix of our system is

$$Pe^{tJ}$$

where

$$J = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

but

$$e^{tJ} = \begin{bmatrix} e^t & 0 & 0 \\ 0 & e^{2t} & 0 \\ 0 & 0 & e^{3t} \end{bmatrix}$$

and hence the fundamental matrix is

$$Pe^{tJ} = \begin{bmatrix} -2 & -3 & 0 \\ -1 & -1 & -4 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} e^t & 0 & 0 \\ 0 & e^{2t} & 0 \\ 0 & 0 & e^{3t} \end{bmatrix}$$

$$= \begin{bmatrix} -2e^t & -3e^{2t} & 0 \\ -e^t & -e^{2t} & -4e^{3t} \\ e^t & e^{2t} & e^{3t} \end{bmatrix}$$

Since Pe^{tJ} *is a fundamental matrix then the desired solution is a linear combination of the columns of* Pe^{tJ}. *That is, there exist constants a, b, c such that at* $t = 0$

$$a\begin{bmatrix} -2e^t \\ -e^t \\ e^t \end{bmatrix} + b\begin{bmatrix} -3e^{2t} \\ -e^{2t} \\ e^{2t} \end{bmatrix} + c\begin{bmatrix} 0 \\ -4e^{3t} \\ e^{3t} \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 4 \end{bmatrix}$$

Taking $t = 0$, *we have*

$$-2a - 3b = 1$$
$$-a - b - 4c = -1$$
$$a + b + c = 4$$

Solving these three equations for a, b, c, we obtain: $c = -1$, $b = -11$, $a = 16$. *Hence the desired solution is*

$$x(t) = 16(-2e^t) - 11(-3e^{2t})$$
$$y(t) = -16e^t + 11e^{2t} + 4e^{3t}$$
$$z(t) = 16e^t - 11e^{2t} - e^{3t}$$

These calculations show that even for a simple, low-dimensional example, the solutions are not quickly obtained. It is clear that the difficulties increase if n is much larger. Also we have not described at all how to deal with the case of eigenvalues of multiplicity greater than 1 [9, Chap. 2]. To do so, we must introduce generalized eigenvectors in order to find the matrix P. There is one cheering note in this welter of complications. If we deal with a system which stems from a single nth-order equation, and if λ is a multiple eigenvalue, then λ appears in just one 'box' J_j in the Jordan canonical form [9, Chap. 2]. Thus we obtain a straightforward algebraic explanation of the form of the solutions of the single nth-order equation. (The form of the solutions for the single second-order equation with a multiple eigenvalue was described in Example 13.)

5.6.3 Solutions of Nonhomogeneous Linear Systems

Finally, we turn to the problem of solving a nonhomogeneous linear system with constant coefficients, i.e., a system

$$\frac{dx_1}{dt} = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m + b_1(t)$$
$$\frac{dx_2}{dt} = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2m}x_m + b_2(t)$$
$$\vdots$$
$$\frac{dx_m}{dt} = a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mm}x_m + b_m(t)$$

Using the matrix notation introduced earlier and letting $b(t)$ denote the vector

$$\begin{bmatrix} b_1(t) \\ b_2(t) \\ \vdots \\ b_m(t) \end{bmatrix}$$

then we may rewrite this system as

$$\frac{dx}{dt} = Ax + b(t) \qquad (1)$$

Let $M(t)$ be a fundamental matrix of the corresponding homogeneous equation

$$\frac{dx}{dt} = Ax \qquad (2)$$

Then we have the following result:

Variation-of-Constants Formula (also called the variation-of-parameters formula). *The solution* $x(t)$ *of* (1) *such that* $x(t)$ *satisfies the initial condition*

$$x(t_0) = x_0$$

is given by the formula

$$x(t) = y(t) + M(t)\int_{t_0}^{t} [M(s)]^{-1} b(s)\, ds$$

where $y(t)$ *is the solution of* (2) *such that*

$$y(t_0) = x_0$$

The proof of this formula is a straightforward calculation [9, Chap. 2]. It is worth noting that this is a generalization of the formulation for the solutions of a single first-order linear equation, i.e., the formula given in Sec. 5.2.2.

We have looked only at linear systems with constant coefficients. If we permit the coefficients a_{ij} to be functions of t, the problems become far more difficult. This is to be expected because if we consider a single-second-order linear homogeneous equation whose coefficients are not constants, then we move from the realm where all the solutions are sums of terms of the familiar form e^{rt} or $t^k e^{rt}$ into another country where the solutions are entirely novel: Bessel functions, Legendre functions, etc. In fact there is very little general theory of linear systems with coefficients a_{ij} which are functions of t. One exception is the Floquet theory for systems in which each $a_{ij}(t)$ has period T [9, Chap. 2]. Although useful, the Floquet theory is limited in scope and yields few explicit results.

5.7 NONLINEAR EQUATIONS

A second direction of generalization is to consider equations or systems which have terms in which the unknown functions appear with exponent different from 1. Such equations are termed *nonlinear* equations. (Examples 2, 4, 6, and 7 are nonlinear equations.) In some rough sense, most equations are nonlinear. More precisely, the class of linear equations is highly specialized. As this suggests, there are many applications in which we must deal with nonlinear equations. Among these applications is celestial mechanics, which was the motivation for the remarkable work of Poincaré and Lyapunov. Later impetus for study came from the intensive study of radio circuits, which began in the 1920s. More recent applications have arisen in chemistry and biology: for example, population models and mathematical descriptions of electrically excitable cells (neurons, cardiac components). These applications have inspired a lot of study of nonlinear differential equations in the last hundred years. But although tremendous strides have been made in our knowledge, the results remain somewhat fragmentary. There is no general theory like that for linear systems, which is embodied in the variation-of-constants formula. That is, we have no convenient general formulas which automatically yield solutions. There are, of course, many special cases which can be dealt with, such as Examples 2, 4, 6, and 7. But all of these examples satisfy some very special hypotheses. (Examples, 2, 4, 6, and 7 were all first-order equations.)

Besides the lack of general formulas for the solutions of nonlinear equations, there is another serious complication which arises in nonlinear equations. We have seen that the solutions of linear equations with constant coefficients are defined for all values of the independent variable, i.e., the domain of the solution includes all the real numbers. (The reason for this is that each component of each solution consists of a sum of terms of the form $ct^k e^{(a+ib)t}$ where a, b, c and k are constants.)

However, even a very simple example reveals the unpleasant fact that there may be strong limitations on the domain of the solution of a nonlinear equation.

Example 28. *Find the solution of*

$$\frac{dx}{dt} = x^2$$

such that $x(0) = 1$. *(Since the dependent variable* x *appears with exponent 2, this equation is certainly nonlinear.)*

Solution: using separation of variables, we have

$$\begin{aligned} \frac{dx}{x^2} &= dt \\ -\frac{1}{x} &= t + C \\ x &= -\frac{1}{t+C} \\ x(0) = 1 &= -\frac{1}{C} \end{aligned}$$

Therefore $C = -1$ and $x(t) = -1/(t-1)$. But then $x(t)$ is defined for $t < 1$; however, as t approaches 1, the solution $x(t)$ increases without bound.

5.7.1 Classical Techniques

By the end of the 1800s, mathematicians had, largely by struggling with nonlinear equations in celestial mechanics, become aware of the deep problems inherent in the study of nonlinear equations. The result was that various techniques were introduced which are used to investigate different aspects or properties of solutions. Our next step is to describe briefly some of these techniques.

5.7.1.1 Perturbation Theory

One of the oldest techniques for dealing with nonlinear equations is the perturbation method. The basic idea of the perturbation method is to represent the problem as a simple problem which is perturbed by the addition of a small but more complicated term. For example, the equation

$$\frac{d^2x}{dt^2} + x + \varepsilon x^3 = 0$$

where ε is a small parameter is a special case of Duffing's equation, which arises in nonlinear mechanics. The unperturbed equation is

$$\frac{d^2x}{dt^2} + x = 0$$

which is by now a familiar equation and easy to solve. The idea is then to look for solutions of the perturbed equation near solutions of the unperturbed equation. This idea is the basis for a large amount of important theory. It was developed extensively by Poincaré in his studies of celestial mechanics and used widely in studies of radio circuitry in the 1920s. It is also widely used in other fields: sound, mechanics, and quantum mechancis. However, a description of the subject is far beyond the reach of this chapter. (For an interesting introductory discussion of perturbation theory, see Ref. 10, Chap. 25; see also Ref. 9, Chap. 7, and Ref. 11.)

5.7.1.2 Poincaré–Bendixson Theory

The geometrical viewpoint was developed largely by Poincaré and has been used and extended ever since. The most complete development is for systems of the form

$$\begin{aligned} \frac{dx}{dt} &= P(x,y) \\ \frac{dy}{dt} &= Q(x,y) \end{aligned} \tag{NL}$$

Let us assume that P and Q have continuous partial derivatives of second order for all (x, y). Each solution of the system is a pair of functions $(x(t), y(t))$ which describes a curve in the xy-plane. From the standard existence and uniqueness theorem is follows that there is one such curve through each point in the xy-plane. Thus no two curves intersect each other and no curve crosses itself. To get an idea of how the curves behave we use a viewpoint introduced earlier in the study of first-order equations (see Example 10). If a solution curve passes through a point (x_0, y_0), then the slope of the curve at (x_0, y_0) is

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{Q(x_0, y_0)}{P(x_0, y_0)}$$

Thus we can associate with each point (x_0, y_0) an arrow which indicates the slope of the curve at (x_0, y_0). As earlier, if enough of these arrows are sketched, then a picture of the solution curves begins to emerge. This picture becomes clearer if we investigate the solutions which are single points. Suppose $(\bar{x}, \bar{y})$ is a point such that

$$P(\bar{x}, \bar{y}) = Q(\bar{x}, \bar{y}) = 0$$

Then the pair of constant functions

$$\left.\begin{array}{l} x(t) = \bar{x} \\ y(t) = \bar{y} \end{array}\right\} \quad \text{for all } t$$

is a solution of (NL). Such a single-point solution is called a *critical point*, a *singular point*, or an *equilibrium point*. The behavior of solutions near the critical point can be determined if the determinant of the matrix

$$M = \begin{bmatrix} \dfrac{\partial P}{\partial x}(\bar{x}, \bar{y}) & \dfrac{\partial P}{\partial y}(\bar{x}, \bar{y}) \\ \dfrac{\partial Q}{\partial x}(\bar{x}, \bar{y}) & \dfrac{\partial Q}{\partial y}(\bar{x}, \bar{y}) \end{bmatrix}$$

is nonzero. The solutions of the linear approximation to (NL), i.e., the system

$$\begin{bmatrix} dx/dt \\ dy/dt \end{bmatrix} = M \begin{bmatrix} x \\ y \end{bmatrix} \tag{LA}$$

can be described in detail and these results can then be used to obtain considerable information about the solutions near $(\bar{x}, \bar{y})$ of the system (NL). Here we will just describe briefly the behavior near $(\bar{x}, \bar{y})$ of the solutions of (LA). If both the eigenvalues of M (i.e., both the solutions of the quadratic equation

$$\det[M - \lambda I] = 0$$

are positive, all the solutions of (LA) near $(\bar{x}, \bar{y})$ move away from $(\bar{x}, \bar{y})$. Typical such behavior is shown in Fig. 2(a). If both the eigenvalues are negative, all the solutions near $(\bar{x}, \bar{y})$ move toward $(\bar{x}, \bar{y})$ as indicated typically in Fig. 2(b). If one eigenvalue is positive and the other negative, the behavior is more complicated. Most of the solutions move away from $(\bar{x}, \bar{y})$, but two of them approach $(\bar{x}, \bar{y})$. Such an equilibrium point is called a *saddle point*; a typical saddle point is shown in Fig. 2(c). If the eigenvalues are complex, then since the entries in matrix M are real the two eigenvalues will be complex conjugates $a + ib$ and $a - ib$. If $a > 0$ $(a < 0)$ the solutions spiral outward from $(\bar{x}, \bar{y})$ [inward toward $(\bar{x}, \bar{y})$] as shown in Fig. 2(d) [(e)]. If $a = 0$, the solutions describe closed curves as shown in Fig. 2(f).

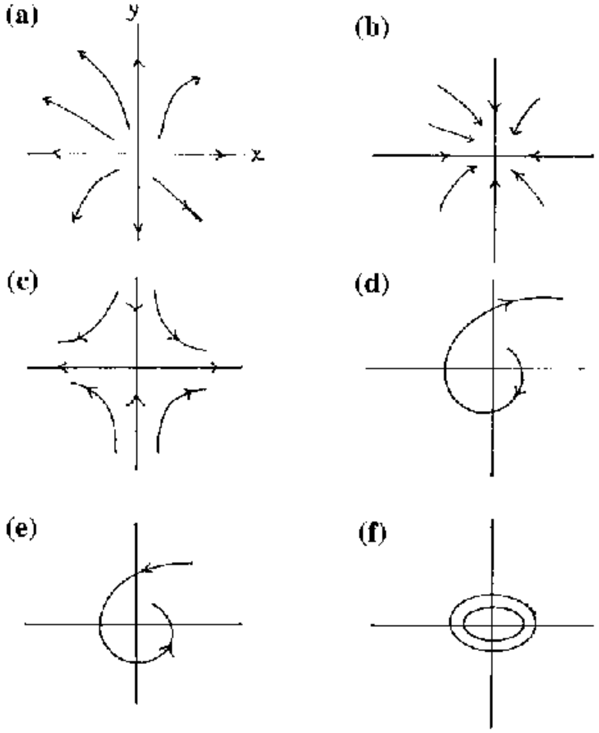


Figure 2

The analysis by which these pictures are obtained is quite simple [9, Chap. 3]. But other seemingly simple questions present more serious problems. Having used the equilibrium points to help get a picture of the solution curves, we turn to a question which arises very often in applications: Does the system (NL) have a periodic solution? In geometrical terms, this question becomes: Is there a solution which describes a closed curve? A solution $(x(t), y(t))$ describes a closed curve if there is a number T such that for all t

$$(x(t + T), y(t + T)) = (x(t), y(t))$$

An answer to this question is given by the Poincaré–Bendixson theorem which can be stated informally as follows. If $(x(t), y(t))$ is a solution of (NL) with the following properties:

1. For t beyond some value, the solution $(x(t), y(t))$ is bounded. (More precisely, there is a value t_0 and a positive number M such that if $t \geq t_0$, then $|x|(t)| + |y(t)| \leq M$.)
2. If $(\bar{x}, \bar{y})$ is an equilibrium point of (NL), then $(x(t), y(t))$ does not approach $(\bar{x}, \bar{y})$ as t increases without bound. [More precisely, there is a disc D in the xy-plane with center $(\bar{x}, \bar{y})$ and a value t_1, such that if $t > t_1$, then $(x(t), y(t))$ is outside D.]

Then either:

1. $(x(t), y(t))$ is a periodic solution, or
2. $(x(t), y(t))$ spirals toward a periodic solution as indicated in Fig. 3.

The Poincaré–Bendixson theorem is intuitively very reasonable. Roughly speaking, it says that if a solution is bounded and does not 'pile up' on an equilibrium point, then either D 'piles up' on a periodic solution or is itself a periodic solution. However, a rigorous proof of the theorem is surprisingly complicated.

Some of the geometrical theory described above for the system (NL) can be extended to systems of n equations where $n > 2$. Generally speaking, studying the solution curves in n-space turns out to be a fruitful approach for many problems. But it has drawbacks. If $n > 3$, we lose the possibility of completely visualizing the solution curve. Also the Poincaré–Bendixson theorem is no longer valid if $n > 2$.

5.7.1.3 Stability Theory

The equilibrium points of (LA) described in the preceding section illustrate the simplest aspects of the concept of *stability*, a subject largely developed by the great Russian mathematician Lyapunov. Speaking informally, we say that a solution of a differential equation is *stable* if every solution which gets close to the solution stays close to the solution. The solution is *asymptotically stable* if every solution which gets close to the solution actually approaches the solution. If neither of these conditions holds, the solution is *unstable*. The equilibrium points in Fig. 2(b) and (e) are asymptotically stable; the equilibrium points in Fig. 2(a), (c), and (d) are unstable. The equilibrium point in Fig. 2(f) is stable. For applications, stability is a very important property because in many cases we can expect that only the solutions which have some stability will predict the actual behavior of the physical system that is modeled by the differential equations. The reason for this is that the physical system is often subject to small disturbances which are not included in the description given by the differential equation. Such disturbances can be interpreted as 'kicking' the system a small distance from the solution that the system is 'on.' If the system were at equilibrium and described by one of the asymptotically stable equilibrium points, then after a 'kick' the system would tend to return to the equilibrium point. But if the equilibrium point were unstable then after even a very small 'kick,' the system would tend to move away from the equilibrium point.

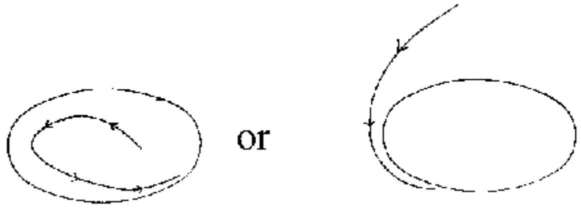


Figure 3

Stability theory has been studied extensively for many years, and a good deal of theory has been developed. Here we will merely point out one fundamental results for linear systems.

Let

$$\frac{dx}{dt} = Ax$$

be a linear homogeneous system with constant coefficients. Then the solution $x = 0$ has the following stability properties: if the eigenvalues of A all have negative real parts, then $x = 0$ is asymptotically stable; if at least one eigenvalue of A has positive real part then $x = 0$ is unstable; if all the eigenvalues of A are pure imaginary, then $x = 0$ is stable but it is not asymptotically stable. These results can be proved by looking at the fundamental matrix

$$e^{tA} = Pe^{tJ}P^{-1}$$

which we discussed earlier.

5.7.2 Modern Techniques

More recent approaches to the study of nonlinear equations include qualitative or topological studies and chaos theory.

The topological studies, which use results such as fixed point theory, had their start in the work of Poincaré and Lyapunov. They provide useful information about the existence and properties of solutions and are a useful complement to the more conventional studies. However, a description of topological techniques is beyond the scope of this chapter.

The basic idea of chaos theory lies in the following observation. We consider a system of nonlinear equations:

$$\frac{dx_1}{dt} = f_1(x_1, \ldots, x_n)$$

$$\vdots$$

$$\frac{dx_n}{dt} = f_n(x_1, \ldots, x_n)$$

with $n \geq 3$, in which all the functions $f_1, \ldots, f_n$ are fairly simple and are 'well behaved' (say, have derivatives of all orders) so that the usual existence and uniqueness theorems are applicable. Then it may happen that the solutions of the system display complicated behavior and also that solutions which have almost the same initial condition may behave very differently. This observation was already formulated by Poincaré before 1900 but it did not receive the attention it deserved until the advent of computers which made possible extensive and detailed numerical studies. The nummerical studies illustrate vividly the complicated and disorderly solution behavior which sometimes occurs. This behavior is called deterministic chaos. A particularly striking example of deterministic chaos is displayed by the solutions of the famous Lorenz system of three nonlinear differential equations. A meterological model, the Lorenz system was introduced and studied in the 1960s.

5.8 NUMERICAL OR COMPUTER ANALYSIS

There are many software programs which can be used to obtain accurate numerical approximations to the solutions of a given differential equation. Such numerical analysis can be used even if one has no idea how to analyze the differential equation by using one of the techniques described earlier. Moreover, even if one of those techniques can be applied there is, sooner or later, numerical analysis that must be carried out. An example would be to find the solutions of an n-dimensional homogeneous linear system

$$\frac{dx}{dt} = Ax$$

where n is not small, e.g., $n = 11$. To find the eigenvalues, one must find the solutions of a polynomial equation of degree 11. Also the eigenvectors must be calculated, and finally one must calculate Pe^{tJ}. If the linear equation is inhomogeneous, then if the variation-of-constants formula is used, we must also calculate an integral, which calculation may itself be a nontrivial chore.

In some important applications, we are wholly dependent on the use of numerical techniques. For example, the orbits of satellites are determined accurately by numerical calculations of the solutions of the appropriate differential equations.

In view of these remarks, the natural question that arises is: Why not just 'put the differential equation on the computer' rather than attempt to use one of the "pencil-and-paper" techniques that we described earlier? This is a valid question and deserves a detailed answer.

Before attempting to answer this question, we discuss a little further the kind of results that can be obtained by using a computer program. We have already referred to the numerical approximations to solutions that can be obtained with the computer. In addition some software programs (e.g., Maple [4]) do symbolic work, i.e., the program can be used to obtain a formula for the general solution. Thus we have a quick way to solve the differential equation provided that the equation can be treated symbolically by the program. On the other hand, the program may not have the resources to deal with a given differential equation even though the differential equation has a closed solution, i.e., a solution which can be represented explicitly or implicitly in terms of known functions. Then the computer program can be used only to obtain a numerical approximation to a solution of the differential equation.

Thus if one is confronted with a differential equation, the first step is to try to apply whatever pencil-and-paper techniques one knows; then, if this fails, to seek a formula for the solution by using the symbol resources of whatever computer program is available. If neither of these directions yields the desired solution, then one must make a decision whether to search further for a pencil-and-paper technique or settle for a numerical approximation to the solution which can be obtained from the computer.

The further search for a pencil-and-paper method can be started by consulting a textbook such as Ref. 5 and then continued in the collections of differential equations [6–8]. This procedure may yield a closed solution. But the danger is that time may be invested in a vain search.

On the other hand, the numerical approximation is not a wholly satisfactory description of the solution.

Also the computer analysis yields only information about single solutions of the differential equation. Little or no information is obtained about the structure of the set of solutions. For example, the stability properties of the solutions remain unknown. An example of this is shown in the study of differential equations which model electrically excitable cells (nerve fibers, neurons, cardiac fibers). These differential equations are messy-looking nonlinear systems, and their study has been largely limited to numerical studies. However, an analytical study of some of these equations reveals that the structure of the set of solutions of the differential equation is quite simple, surprisingly so in view of the appearance of the equations.

REFERENCES

1. Mathematical Tables from Handbook of Chemistry and Physics. 10th ed. Cleveland: Chemical Rubber Publishing Co., 1954.
2. HB Dwight. Tables of Integrals and Other Mathematical Data. 4th ed. New York: Macmillan, 1961.
3. W Grobner, N Hafreiter. Integraltafel: erste Teil, Unbestimmte Integrale; zweiter Teil, Bestimmte Integrale. Wien: Springer-Verlag, 1965.
4. Maple U Release 4, copyright 1981–1996 by Waterloo Maple Inc.
5. CH Edwards Jr, DE Penney. Elementary Differential Equations with Boundary Value Problems. 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1989.
6. E Kamke. Differentialgleichungen Lösungsmethoden und Lösungen. Band 1, Gewöhnliche Differentialgleichungen, 3. Auflage. New York: Chelsea Publishing Company, 1959.
7. GM Murphy. Ordinary Differential Equations and Their Solutions. New York: Van Nostrand, 1960.
8. AD Polyanin, VF Zaitsev. Handbook of Exact Solutions for Ordinary Differential Equations. Boca Raton, FL: CRC Press, 1995.
9. J Cronin. Differential Equations: Introduction and Qualitative Theory, 2nd ed. New York: Marcel Dekker, 1994.
10. MD Greenberg. Foundations of Applied Mathematics. Englewood Cliffs, NJ: Prentice-Hall, 1978.
11. RE O'Malley Jr. Singular Perturbation Methods for Ordinary Differential Equations. New York: Springer-Verlag, 1991.

Chapter 1.6

Boolean Algebra

Ki Hang Kim
Alabama State University, Montgomery, Alabama

6.1 INTRODUCTION

The theory of Boolean algebra is relatively simple but it is endowed with an elegant structure and rich in practical applications. Boolean algebra is named after the British mathematician George Boole (1815–1864). For Boole's pioneering work, see Refs 1 and 2. Boolean algebras have been extensively studied by Schröder [3], Huntington [4], Birkhoff [5], Stone [6], Halmos [7], Sikorski [8], and Hohn [9]. In this chapter, we present a concise summary of Boolean algebra and its applications.

A *Boolean algebra* is a mathematical system $(\beta, \vee, \wedge)$ consisting of a nonempty set $\beta = \{a, b, c, \ldots\}$ and two binary operations $\vee$ (vee) and $\wedge$ (wedge) defined on β such that:

b_1. Both operations are associative; that is

$$(a \vee b) \vee c = a \vee (b \vee c)$$
$$(a \wedge b) \wedge c = a \wedge (b \wedge c)$$

b_2. Both operations are commutative; that is,

$$a \vee b = b \vee a \qquad a \wedge b = b \wedge a$$

b_3. Each operation is distributive with respect to the other; that is,

$$a \vee (b \vee c) = (a \vee b) \wedge (a \vee c)$$
$$a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c)$$

b_4. β contains an identity element 0 with respect to $\vee$ and an identity element 1 with respect to $\wedge$; that is,

$$a \vee 0 = a \qquad a \wedge 1 = a$$

The element 0 is called the *zero element* and the element 1 is called the *unit* (or *universal) element*, respectively, of β.

b_5. For each $a \in \beta$, there exists an element $\bar{a} \in \beta$ such that

$$a \vee \bar{a} = 1 \qquad a \wedge \bar{a} = 0$$

The element $\bar{a}$ is called a *complement* of a.

Notice that there exists a complete symmetry in the postulates b_1–b_5 with respect to the operations $\vee$ and $\wedge$ and also in the identities of b_4. Therefore, we can state the following principle:

Principle of Duality. *If in any statement deduced from the five postulates b_1–b_5 we interchange $\vee$ and $\wedge$, and 0 and 1, then we obtain a valid statement.*

Example 1. *Let $S = \{a, b\}$, $a < b$. We define $a \vee b = \sup\{a, b\}$ and $a \wedge b = \inf\{a, b\}$. (Accordingly, we can also define the operations as $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$.) The tables for these operations are as Tables 1 and 2 .*

Clearly $(S, \vee, \wedge)$ is the simplest and most fundamental nontrivial Boolean algebra.

Example 2. *Let D be the set of positive integral divisors of 6: that is, $D = \{1, 2, 3, 6\}$. For all $a, b \in D$, we*

Table 1
Multiplication for $\vee$

$\vee$	a	b
a	a	b
a	b	b

Table 2
Multiplication for $\wedge$

$\wedge$	a	b
a	a	a
b	a	b

define $a \vee b$ and $a \wedge b$ to be respectively the least common multiple (lcm) and greatest common divisor (gcd) of a and b. In other words, $a \vee b = \text{lcm}\ \{a, b\}$ and $a \wedge b = \text{gcd}\ \{a, b\}$. The tables for these operations are Tables 3 and 4.

Table 3 Lowest Common Multiple Multiplication Table

$\vee$	1	2	3	6
1	1	2	3	6
2	2	2	6	6
3	3	6	3	6
6	6	6	6	6

Table 4 Greatest Common Divisor Multiplication Table

$\wedge$	1	2	3	6
1	1	1	1	1
2	1	2	1	2
3	1	1	3	3
6	1	2	3	6

A quick inspection of these tables yields the fact that the integer 1 plays the role of the zero element and the integer 6 plays the role of the unit element and the tables also yield the various complements as follows: for all $a \in D$, $\bar{a} = 6/a$; $\bar{1} = 6$, $\bar{2} = 3$, $\bar{3} = 2$, $\bar{6} = 1$. Therefore, $(D, \vee, \wedge)$ is a Boolean algebra.

Example 3. *Let $S = \{a, b, c, d\}$ together with the operations defined in Tables 5 and 6. The system $(S, \vee, \wedge)$ is a Boolean algebra. The verification is left as an exercise.*

Table 5 Four-Element $\vee$ Multiplication Table

$\vee$	a	b	c	d
a	a	b	c	d
b	b	b	d	d
c	c	d	c	d
d	d	d	d	d

Table 6 Four-Element $\wedge$ Multiplication Table

$\wedge$	a	b	c	d
a	a	a	a	a
b	a	b	a	b
c	a	a	c	c
d	a	b	c	d

Example 4. *Let $P(X)$ be the power set of a nonempty, finite set X. Then $(P(X), \cup, \cap)$ is a Boolean algebra, taking $\vee$ and $\wedge$ as $\cup$ and $\cap$, respectively. Take "–" as complementation relative to X, $0 = \emptyset$, $1 = X$, respectively. Therefore, the operations $\vee$ and $\wedge$ may be denoted by $\cup$ and $\cap$, respectively.*

Infinite Boolean algebras are not used much in the theory of switching circuits but do occur in other areas of mathematics such as measure theory, logic, probability, and topology. Since we are mainly interested in the applications of Boolean algebras, we will only concentrate on finite Boolean algebras.

The simplest infinite Boolean algebras are infinite Cartesian products of $\{0, 1\}$. These act almost identically to finite Boolean algebras.

Example 5. *All measurable subsets of the real numbers form a Boolean algebra which allows not only finite union and intersection but also countable union and intersection.*

Example 6. *All sets obtained from the open and closed sets of a topology by finite union and intersection form a Boolean algebra.*

Example 7. *Given any set, a Boolean algebra is obtained by taking all its finite subsets and all the complements of finite subsets.*

We now present another characterization of Boolean algebra. For a comprehensive treatment of lattice theory, see Birkhoff [5]. A *lattice* is a mathematical system $(L, \vee, \wedge)$ consisting of a nonempty set $L = \{a, b, c, \ldots\}$ and two binary operations $\vee$ (join) and $\wedge$ (meet) defined on L such that:

- l_1. Associativity: $(a \vee b) \vee c = a \vee (b \vee c)$, $(a \wedge b) \wedge c = a \wedge (b \wedge c)$.
- l_2. Commutativity: $a \vee b = b \vee a, a \wedge b = b \wedge a$.
- l_3. Absorption: $a \vee (a \wedge b) = a$, $a \wedge (a \vee b) = a$.

If in addition the distributive law:

- l_4. $a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c)$, $a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c)$ holds, then L is called a *distributive lattice.*

L is said to be *modular* if and only if

- l_5. $a \wedge [b \vee (a \wedge c)] = (a \wedge b) \vee (a \wedge c)$.

Since l_1–l_3 hold in every Boolean algebra, every Boolean algebra is a lattice.

Example 8. *Let* $X = \{1, 2, 3\}$. *Then* $P(X) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 3\}, X\}$. *The subset lattice of* $P(X)$ *is shown in Fig. 1.*

Example 9. *Let* $P(X)$ *be the same as in Example 4. Then* $(P(X), \vee, \wedge)$ *is a lattice.*

As a consequence of the above example the operations $\vee$ and $\wedge$ are frequently interchanged with set-theoretical "union" (lattice-theoretical "join") and set-theoretical "intersection" (lattice-theoretical "meet"), respectively. However, some authors use "+" (addition and "·" (multiplication) instead of $\vee$ ($\cup$) and $\wedge$ ($\cap$), respectively. For brevity and simplicity, from now on we use "+" instead of $\vee$ and "·" instead of $\wedge$, respectively. Furthermore, we usually suppress the dot "·" of $a \cdot b$ and simply write ab.

We list some important properties of Boolean algebras which can be deduced from the five postulates b_1–b_5.

- p_1. The identities 0 and 1 are unique.
- p_2. Idempotency: for $a \in \beta, a + a = a, aa = a$.
- p_3. Dominant element: for $a \in \beta$, $a + 1 = 1$, $a0 = 0$.
- p_4. Absorption: for a, $b \in \beta$, $a + (ab) = a$, $a(a + b) = a$.
- p_5. Complementation: $\bar{0} = 1, \bar{1} = 0$.
- p_6. Every $a \in \beta$ has a unique complement.

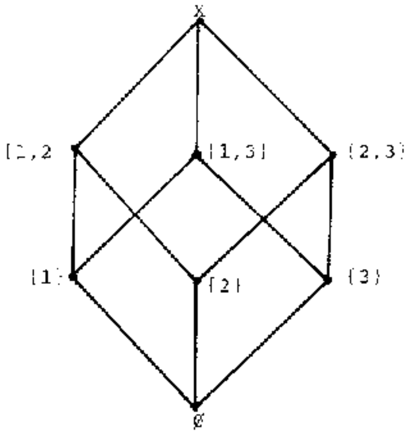


Figure 1 Subset lattice of $P(\{1, 2, 3\})$.

- p_7. Involution: for $a \in \beta, \bar{\bar{a}} = a$.
- p_8. De Morgan's law: for a, $b \in \beta$, $\overline{a + b} = \overline{ab}$, $\overline{ab} = \bar{a} + \bar{b}$.

A ring is called *Boolean* if all of its elements are idempotent under the second operation.

Example 10. *The ring of integers modulo 2,* $(Z_2, \oplus, \otimes)$ *is a Boolean ring, since* $0 \otimes 0 = 0$ *and* $1 \otimes 1 = 1$.

Example 11. *The ring* $(P(X), \Delta, \cap)$ *of subsets of a nonempty set X is a Boolean ring, where* Δ *is the symmetrical difference of sets. The verification is left as an exercise.*

6.2 BOOLEAN FUNCTION

Let $\beta\{x_1, x_2, \ldots, x_n\}$ be a Boolean algebra. A *constant* of β is any symbol, such as "0" and "1", which represents a specified element of β. A *variable* is any symbol which represents an arbitrary element of β. In the expression $x_1 + x_2x_3$ we call x_1 and x_2x_3 *monomials* and the entire expression $x_1 + x_2x_3$ a *Boolean polynomial.* Any expression such as $x_1 + x_2$, x_1x_3, $(x_1 + x_2x_3)$ consisting of a finite number of elements of β will be called a *Boolean function* and we will denote it by $f(x_1, x_2, x_3)$ (for short f). For example, $[(x_1 + x_2)(x_2 + x_3)]x_4$ is a function of four variables x_1, x_2, x_3, and x_4.

A Boolean polynomial is said to be in *disjunctive normal form* (or *canonical form*) if it is a sum of monomials in which each variable or its complement appears exactly once.

Example 12. *This polynomial is in disjunctive normal form*

$$f = \bar{x}_1 x_2 \bar{x}_3 + \bar{x}_1 \bar{x}_2 x_3 + x_1 x_2 x_3$$

In disjunctive normal form of a function f, a given monomial M will appear if and only if the function f is 1 *when we make* $x_i = 1$ *whenever* x_i *appears in M and* $x_i = 0$ *whenever* $\bar{x}_i$ *appears in M.*

Example 13. *In the above example the first monomial is* 1 *provided* $x_1 = 0, x_2 = 1, x_3 = 0$. *For then all of* $\bar{x}_1$, x_2, $\bar{x}_3$ *are* 1 *and so is their product. But in all other cases at least one variable is zero and the product is zero.*

This argument shows that the disjunctive normal form is unique, and moreover that it exists since if we add one monomial to M whenever the function is 1 at some set of x_i values then we get a function equal to that function at each set of x_i values.

Example 14. *Suppose f has the truth table of Table 7. Then* $f = 1$ *at the three triples* $(0, 0, 0)$, $(0, 1, 0)$, *and* $(0, 1, 1)$. *Add the corresponding monomials to get*

$$f = \bar{x}_1 \bar{x}_2 \bar{x}_3 + \bar{x}_1 x_2 \bar{x}_3 + \bar{x}_1 x_2 x_3$$

We can also obtain the disjunctive normal form for general polynomial by expanding products of polynomials and replacing absent variables in monomials by $x_i + \bar{x}_i$.

Example 15. *Suppose* $f = (x_1 + x_2)(\bar{x}_1 + x_3) + x_1(x_2 + x_3)$. *Expand the products to get* $f = x_1\bar{x}_1 + x_2\bar{x}_1 + x_1x_3 + x_2x_3 + x_1x_2 + x_1x_3$. *Now* $x_1\bar{x}_1 = 0$, *and combine terms to get* $f = x_2\bar{x}_1 + x_1x_3 + x_2x_3 + x_1x_2$. *Now since* x_3 *is missing in the first term, and* $x_3 + \bar{x}_3 = 1$, *rewrite it as* $\bar{x}_1 x_2(x_3 + \bar{x}_3)$.

Table 7 Truth Table of Boolean Function *f*

x_1	x_2	x_3	f
0	0	0	1
0	0	1	0
0	1	0	1
1	0	0	0
1	1	0	0
1	0	1	0
0	1	1	1
1	1	1	0

Do the same for other terms:

$$\begin{aligned}
&\bar{x}_1 x_2(x_3 + \bar{x}_3) + x_1(x_2 + \bar{x}_2)x_3 + (x_1 + \bar{x}_1)x_2x_3 \\
&\quad + x_1x_2(x_3 + \bar{x}_3) \\
&= \bar{x}_1x_2x_3 + \bar{x}_1x_2\bar{x}_3 + x_1x_2x_3 + x_1\bar{x}_2x_3 + x_1x_2x_3 \\
&\quad + \bar{x}_1x_2x_3 + x_1x_2x_3 + x_1x_2\bar{x}_3 \\
&= x_1x_2x_3 + \bar{x}_1x_2x_3 + \bar{x}_1x_2\bar{x}_3 + x_1\bar{x}_2x_3 + x_1x_2\bar{x}_3
\end{aligned}$$

With n variables, the number of monomials is 2^n, since for each variable we have two choices: complement it or not, give choices on previous variables. This means a Boolean function is determined by its 2^n coefficients which can each be 0 or 1. *This means that there are* 2^{2^n} *Boolean functions of n variables.*

Example 16. *In two variables, there are* $2^2 = 4$ *monomials:* x_1x_2, $x_1\bar{x}_2$, $\bar{x}_1x_2$, $\bar{x}_1\bar{x}_2$, *and* $2^4 = 16$ *functions.*

The conjuctive normal form is dual to the disjunctive normal form. It is a product of sums S such that each variable or its complement occurs just once as a summand in S.

Example 17. *This is in conjunctive normal form*

$$(x_1 + x_2 + \bar{x}_3)(\bar{x}_1 + x_2 + x_3)$$

In the conjunctive normal form, if x_i *appears in a factor then set* $x_i = 0$, *if* $\bar{x}_i$ *appears set* $x_i = 1$. *Then the factor and hence the whole expansion is zero. This indicates that a given factor appears in the conjunctive normal form for a function f if and only if when we make this substitution the function is zero. By considering those values which make f zero, we can expand any function in conjunctive normal form.*

Suppose f has the same values as in Table 7. Then f is zero at the triples $(0, 0, 1)$, $(1, 0, 0)$, $(1, 0, 1)$, $(1, 1, 0)$, $(1, 1, 1)$. *Therefore f is the product*

$$\begin{aligned}
&(x_1 + x_2 + \bar{x}_3)(\bar{x}_1 + x_2 + x_3)(\bar{x}_1 + x_2 + \bar{x}_3) \\
&(\bar{x}_1 + \bar{x}_2 + x_3)(\bar{x}_1 + \bar{x}_2 + \bar{x}_3)
\end{aligned}$$

We may also expand the function using the dual distributive law and replacing summands 0 by $x_i\bar{x}_i$.

Example 18. *Let* $f = (x_1 + x_2)(\bar{x}_1 + x_3) + x_1(x_2 + x_3)$.

$$\begin{aligned}
f &= [(x_1 + x_2)(\bar{x}_1 + x_3) + x_1][(x_1 + x_2)(\bar{x}_1 + x_3) \\
&\quad + x_2 + x_3] \\
&= (x_1 + x_2 + x_1)(\bar{x}_1 + x_3 + x_1) \\
&\quad (x_1 + x_2 + x_2 + x_3)(\bar{x}_1 + x_3 + x_2 + x_3) \\
&= (x_1 + x_2)(1)(x_1 + x_2 + x_3)(\bar{x}_1 + x_2 + x_3) \\
&= (x_1 + x_2 + x_3\bar{x}_3)(x_1 + x_2 + x_3)(\bar{x}_1 + x_2 + x_3) \\
&= (x_1 + x_2 + x_3)(x_1 + x_2 + \bar{x}_3)(\bar{x}_1 + x_2 + x_3)
\end{aligned}$$

Still another way to obtain the conjunctive normal form of f is to obtain the disjucntive normal form of $\bar{f}$ and take its dual.

Example 19. *Let f be as in the Example 14. The disjunctive normal form of $\bar{f}$ is $\bar{x}_1\bar{x}_2x_3 + x_1\bar{x}_2\bar{x}_3 + x_1\bar{x}_2x_3 + x_1x_2\bar{x}_3 + x_1x_2x_3$.*

Then f is its complement, a product of one factor for each summand in $\bar{f}$. The summand $\bar{x}_1\bar{x}_2x_3$ goes to $\overline{\bar{x}_1\bar{x}_2x_3} = x_1 + x_2 + \bar{x}_3$ and so on. This gives once again the product

$$(x_1 + x_2 + \bar{x}_3)(\bar{x}_1 + x_2 + x_3)(\bar{x}_1 + x_2 + \bar{x}_3)$$
$$(\bar{x}_1 + \bar{x}_2 + x_3)(\bar{x}_1 + \bar{x}_2 + \bar{x}_3)$$

6.3 SWITCHING FUNCTIONS

Shannon [10] was the first one to apply Boolean algebra to digital circuitry. For an excellent account of the switching functions, see Hohn [11] and Dornhoff and Hohn [12].

In two-state circuits, a *switch* or *contact* may be in the *open state* or in the *closed state*. With an *open contact* (or *open path*) in a circuit we assign the symbol "0" and with a *closed contact* (or *closed path*) we assign the symbol "1." That is, we assign the value 1 or 0 to any circuit according as current does or does not flow through it. Therefore, we can construct a *two-element Boolean algebra*, which is also known as *switching algebra* as follows.

Example 20. *Let $\beta_0 = \{0, 1\}$. We define $+$, by Tables 8 and 9. This system has applications in both switching theory and the algebra of propositions, where 0 is "false" and 1 is "true."*

Table 8 Addition Table for β_0

$+$	0	1
0	0	1
1	1	1

Table 9 Multiplication Table for β_0

$\cdot$	0	1
0	0	0
1	0	1

Let $\beta = \{x_1, x_2, \ldots, x_n\}$ where x_i is either 0 or 1 for every i. Then the Boolean function

$$f(x_1, x_2, \ldots, x_n) = \begin{cases} 0 \\ 1 \end{cases}$$

is called a switching function.

Example 21. *Let x and y be the circuit variables which apply to a two-terminal circuit consisting of two contacts x and y. Thus, we have the parallel and series connection, shown in Figs. 2 and 3.*

The two circuits can be represented by switching functions (Table 10).

The disjunctive normal form gives a way to represent any Boolean function by a switching circuit.

Example 22. *Let $f(x, y, z) = \bar{x}y\bar{z} + \bar{x}\bar{y}z + xyz$. This is represented by Figs. 4 and 5 .*

In general, we have one series circuit for each monomial and then put all those circuits in parallel. However, in very many cases, this form will not be the simplest form.

Example 23. *The function $f(x, y, z) = x + yz$ in disjunctive normal form is $xyz + xy\bar{z} + x\bar{y}z + x\bar{y}\bar{z} + \bar{x}yz$,*

Table 10 Values of Switching Circuits

x	y	$x+y$	xy
0	0	0	0
1	0	1	0
0	1	1	0
1	1	1	1

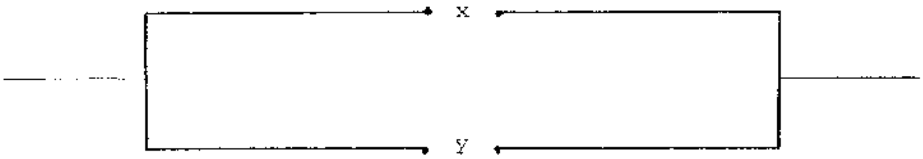


Figure 2 Parallel connection: $x + y$.

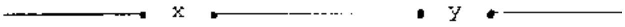


Figure 3 Series connection: xy.

as shown in Fig. 6, but the original function is as shown in Fig. 7, which is simpler.

The terms appearing in the disjunctive normal form are known as *minterms*.

In realizing Boolean functions by switching circuits, it is important to have as few switches as possible in order to save money, space, and time.

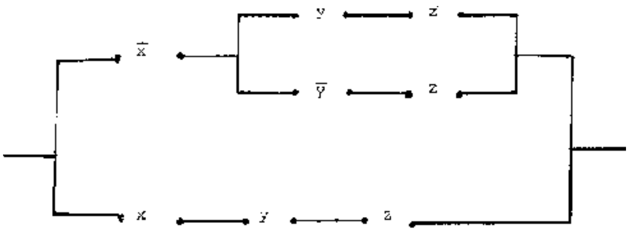


Figure 4 First switching circuit for $f(x, y, z)$.

We may consider minimizing a representation of a Boolean function as a sum of products of variables and its complements. We say that one polynomial P_1 is shorter than another polynomial P_2 if (1) P_1 has no more terms than P_2, and (2) P_1 has fewer total appearances of variables than P_2.

Example 24. *$x_1x_2 + x_1\bar{x}_2 = x_1$ so the former is not minimal.*

Example 25. *$x_1\bar{x}_2 + x_2\bar{x}_3 + x_3\bar{x}_1 = \bar{x}_1x_2 + \bar{x}_2x_3 + \bar{x}_3x_1$ so minimal forms need not be unique.*

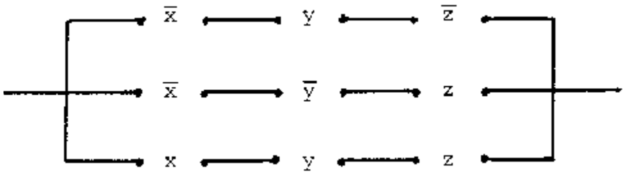


Figure 5 Second switching circuit for $f(x, y, z)$.

The basic way of simplifying an expression is to replace $Mx_i + M\bar{x}_i$ by M where M is a monomial.

A product $P = z_1z_2 \ldots z_k$ of variables and complements which implies a function f is called an *implicant* of f. This is equivalent to $P \leq f$ or $P + f = f$. An implicant is *prime* if no subproduct of P is also an implicant.

Example 26. *Any term of a Boolean polynomial is an implicant.*

Example 27. *x_1x_2 is an implicant of $x_1x_2 + x_1\bar{x}_2$ which is not prime, and x_1 is an implicant which is prime.*

Any minimal expression for a Boolean function f must be a sum of prime implicants, otherwise we can replace any nonprime implicant by a subproduct and get a shorter expression.

Quine's method of finding prime implicants is first to write in a column the terms in the disjunctive normal form of f. Examine each term to see whether it can be combined with a term below, i.e., the two terms differ by complementing a single variable. If so, check it and the term below and write the shortened term in a second column. Then repeat this procedure for the second and later columns.

Example 28. *Consider the Boolean polynomial with disjunctive normal form $x_1x_2x_3 + x_1\bar{x}_2x_3 + x_1x_2\bar{x}_3 + x_1\bar{x}_2\bar{x}_3 + \bar{x}_1x_2x_3$. The first column is*

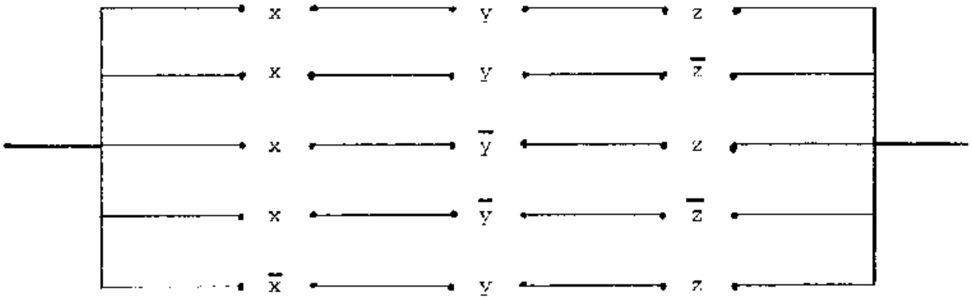


Figure 6 Switching circuit before simplification.

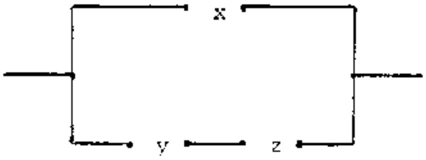


Figure 7 Switching circuit after simplification.

$x_1x_2x_3$

$x_1\bar{x}_2x_3$

$x_1x_2\bar{x}_3$

$x_1\bar{x}_2\bar{x}_3$

$\bar{x}_1x_2x_3$

The first term combines with the second to give x_1x_3, with the third to give x_1x_2, with the fifth to give x_2x_3.

The second combines with the fourth to give $x_1\bar{x}_2$. The third combines with the fourth to give $x_1\bar{x}_3$. So the new table is Table 11. In the second column, x_1x_3 combines with $x_1\bar{x}_3$ to give x_1, x_1x_2 combines with $x_1\bar{x}_2$ to give x_1. This gives Table 12. Then the unchecked terms are the prime implicants of f. Their sum is one formula for f, $x_1 + x_2x_3$, which in this case is minimal.

In general, a shortest formula will use some but not all of the minimal implicants. To determine this we form a table with rows labeled by prime implicants and columns labeled by terms in the disjunctive normal form (Table 13). We put a "√" in a cell when the prime implicant implies the disjunctive term.

Example 29. *Now choose as few rows as possible such that every column has a check. Here we need both x_1 and x_2x_3.*

For n large, these procedures will typically grow exponentially in the amount of time and space required to carry them out and in fact the last part of this method is an example of an NP-complete problem.

Table 11 Computation of Prime Implicants, Second Stage

$x_1x_2x_3\surd$	x_1x_3
$x_1\bar{x}_2x_3\surd$	x_1x_2
$x_1x_2\bar{x}_3\surd$	x_2x_3
$x_1\bar{x}_2\bar{x}_3\surd$	$x_1\bar{x}_2$
$\bar{x}_1x_2x_3\surd$	$x_1\bar{x}_3$

Table 12 Computation of Prime Implicants, Third Stage

$x_1x_2x_3\surd$	$x_1x_3\surd$	x_1
$x_1\bar{x}_2x_3\surd$	$x_1x_2\surd$	
$x_1x_2\bar{x}_3\surd$	x_2x_3	
$x_1\bar{x}_2\bar{x}_3\surd$	$x_1\bar{x}_2\surd$	
$\bar{x}_1x_2x_3\surd$	$x_1\bar{x}_3\surd$	

Next we give an example where not all prime implicants are used.

Example 30. *Let* $f = x_1x_2\bar{x}_3 + x_1\bar{x}_2x_3 + \bar{x}_1x_2x_3 + x_1\bar{x}_2\bar{x}_3 + \bar{x}_1x_2\bar{x}_3 + \bar{x}_1\bar{x}_1x_3$. *Then the monomials are*

$x_1x_2\bar{x}_3$

$x_1\bar{x}_2x_3$

$\bar{x}_1x_2x_3$

$x_1\bar{x}_2\bar{x}_3$

$\bar{x}_1x_2\bar{x}_3$

$\bar{x}_1\bar{x}_2x_3$

The first combines with the fourth to give $x_1\bar{x}_3$, with the fifth to give $x_2\bar{x}_3$. The second combines with the fourth to give $x_1\bar{x}_2$ and with the sixth to give $\bar{x}_2x_3$. The third combines with the fifth to give $\bar{x}_1x_2$ and with the sixth to give $\bar{x}_1x_3$. So we have Table 14 (see pg. 120). All the second column are prime implicants (Table 15, see pg. 120). Any three disjoint rows give a minimal expression for f,

$$x_1\bar{x}_3 + \bar{x}_2x_3 + \bar{x}_1x_2 \quad \text{or} \quad x_2\bar{x}_3 + x_1\bar{x}_2 + \bar{x}_1x_3$$

In practice often some values of Boolean functions are not important, for instance those combinations may never occur. These are called "*don't cares.*" They are dealt with by listing them when finding the prime implicants, but not in Table 15.

Example 31. *Suppose we add $x_1x_2x_3$ as a don't care in Table 14 (Table 16, see pg. 120). Then across the top we list all but the last don't care (Table 17, see pg. 120). So $x_1 + x_2 + x_3$ is the minimal Boolean polynomial.*

Table 13 Computation of Shortest Formula

	$x_1x_2x_3$	$x_1\bar{x}_2x_3$	$x_2x_2\bar{x}_3$	$x_1\bar{x}_2\bar{x}_3$	$\bar{x}_1x_2x_3$
x_1	√	√	√	√	
x_2x_3	√				√

Table 14 Computation of Prime Implicants

$x_1x_2\bar{x}_3$ √	$x_1\bar{x}_3$
$x_1\bar{x}_2x_3$ √	$x_2\bar{x}_3$
$\bar{x}_1x_2x_3$ √	$x_1\bar{x}_2$
$x_1\bar{x}_2\bar{x}_3$ √	$\bar{x}_2x_3$
$\bar{x}_1x_2\bar{x}_3$ √	$\bar{x}_1x_2$
$\bar{x}_1\bar{x}_2x_3$ √	$\bar{x}_1x_3$

6.4 BOOLEAN MATRICES

Let BV_n denote the set of all n-tuples $(v_1, v_2, \ldots, v_n)$ over $\beta_0 = \{0, 1\}$. An element of BV_n is called a *Boolean vector* of *dimension n*. The system $(BV_n, +, \cdot)$ is a Boolean algebra by defining operations as elementwise sums and products. For elementary properties of Boolean vectors, see Kim [13].

Example 32. *Let* $(1, 0, 1, 0),\ (0, 0, 1, 1) \in BV_4$. *Then*
$(1, 0, 1, 0) + (0, 0, 1, 1) = (1, 0, 1, 1)$
$(1, 0, 1, 0)(0, 0, 1, 1) = (0, 0, 1, 0)$.

A *subspace* of BV_n is a subset containing the zero vector $(0, 0, \ldots, 0)$ and closed under addition of vectors. The *span* of a set S of vectors, denoted $\langle S \rangle$, is the intersection of all subspaces containing S.

Example 33

1. *Let* $U = \{(0, 0, 0), (1, 0, 0), (0, 1, 0), (1, 1, 0), (0, 1, 1), (1, 1, 1)\} \in BV_3$. *Then U is a subspace of* BV_3.
2. *Let* $V = \{(0, 0, 0), (1, 0, 0), (0, 0, 1), (1, 1, 0), (0, 1, 1), (1, 1, 1)\} \in BV_3$. *Since* $(1, 0, 0)+ (0, 0, 1) = (1, 0, 1) \notin BV_3$, *V is not a subspace of* BV_3.

Table 15 Determination of Minimal Expression for f

	$x_1x_2\bar{x}_3$	$x_1\bar{x}_2x_3$	$\bar{x}_1x_2x_3$	$x_1\bar{x}_2\bar{x}_3$	$\bar{x}_1x_2\bar{x}_3$	$\bar{x}_1\bar{x}_2x_3$
$x_1\bar{x}_3$	√			√		
$x_2\bar{x}_3$	√				√	
$x_1\bar{x}_2$		√		√		
$\bar{x}_2x_3$		√				√
$\bar{x}_1x_2$			√		√	
$\bar{x}_1x_3$			√			√

Table 16 Computation of Prime Implicants with "Don't Care"

$x_1x_2\bar{x}_3$ √	$x_1\bar{x}_3$ √	x_1
$x_1\bar{x}_2x_3$ √	$x_2\bar{x}_3$ √	x_2
$\bar{x}_1x_2x_3$ √	x_1x_2 √	x_3
$x_1\bar{x}_2\bar{x}_3$ √	$x_1\bar{x}_1$ √	
$\bar{x}_1x_2\bar{x}_3$ √	$\bar{x}_2x_3$ √	
$\bar{x}_1\bar{x}_2x_3$ √	x_1x_3 √	
$x_1x_2x_3$ √	$\bar{x}_1x_2$ √	
	$\bar{x}_1x_3$ √	
	x_2x_3 √	

By a *Boolean matrix* we mean a matrix over β_0. One can also define a Boolean matrix over an arbitrary Boolean algebra. The Boolean matrices behave quite differently from matrices over a field. Let B_{mn} denote the set of all $m \times n$ Boolean matrices. If $m = n$, we just write B_n. For a comprehensive treatment of Boolean matrices and their applications, see Kim [13].

Boolean matrix addition and multiplication are the same as in the case of matrices over a field except the concerned sums and products are over β_0. There are two other operations on Boolean matrices.

1. Logical product: $A \odot B = (a_{ij}b_{ij})$.
2. Complement: $A^C = (\bar{a}_{ij})$.

Example 34. *Let*

$$A = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix} \qquad B = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \in B_4$$

Table 17 Omission of "Don't Care" to Find Shortest Formula

	$x_1x_2\bar{x}_3$	$x_1\bar{x}_2x_3$	$\bar{x}_1x_2x_3$	$x_1\bar{x}_2\bar{x}_3$	$\bar{x}_1x_2\bar{x}_3$	$\bar{x}_1\bar{x}_2x_3$
x_1	√	√		√		
x_2	√		√		√	
x_3		√	√			√

Then

$$A+B=\begin{bmatrix}1&0&1&1\\0&1&0&0\\0&0&1&1\\1&1&0&0\end{bmatrix}+\begin{bmatrix}0&0&1&0\\1&0&0&1\\0&1&1&0\\0&0&1&1\end{bmatrix}$$

$$=\begin{bmatrix}1&0&1&1\\1&1&0&1\\0&1&1&1\\1&1&1&1\end{bmatrix}$$

$$AB=\begin{bmatrix}1&0&1&1\\0&1&0&0\\0&0&1&1\\1&1&0&0\end{bmatrix}\begin{bmatrix}0&0&1&0\\1&0&0&1\\0&1&1&0\\0&0&1&1\end{bmatrix}$$

$$=\begin{bmatrix}0&1&1&1\\1&0&0&1\\0&1&1&1\\1&0&1&1\end{bmatrix}$$

$$A\odot B=\begin{bmatrix}1&0&1&1\\0&1&0&0\\0&0&1&1\\1&1&0&0\end{bmatrix}\odot\begin{bmatrix}0&0&1&0\\1&0&0&1\\0&1&1&0\\0&0&1&1\end{bmatrix}$$

$$=\begin{bmatrix}0&0&1&0\\0&0&0&0\\0&0&1&0\\0&0&0&0\end{bmatrix}$$

$$A^C=\begin{bmatrix}1&0&1&1\\0&1&0&0\\0&0&1&1\\1&1&0&0\end{bmatrix}^C=\begin{bmatrix}0&1&0&0\\1&0&1&1\\1&1&0&0\\0&0&1&1\end{bmatrix}$$

$$A^T=\begin{bmatrix}1&0&1&1\\0&1&0&0\\0&0&1&1\\1&1&0&0\end{bmatrix}^T=\begin{bmatrix}1&0&0&1\\0&1&0&1\\1&0&1&0\\1&0&1&0\end{bmatrix}$$

$$[A^T=(a_{ji})]$$

The *inequality* of Boolean matrices is defined by $A \leq B$ if and only if $a_{ij} \leq b_{ij}$ for all i, j. This is a partial order relation. Similarly, one can define a strict partial order relation $A < B$ if and only if $A \leq B$ and $a_{ij} < b_{ij}$ for some i and j.

Example 35. *Let*

$$\begin{bmatrix}1&0\\0&1\end{bmatrix},\begin{bmatrix}1&1\\0&1\end{bmatrix},\begin{bmatrix}1&1\\1&1\end{bmatrix}\in B_2$$

Then

$$\begin{bmatrix}1&1\\0&1\end{bmatrix}\leq\begin{bmatrix}1&1\\1&1\end{bmatrix}$$

$$\begin{bmatrix}1&0\\0&1\end{bmatrix}<\begin{bmatrix}1&1\\0&1\end{bmatrix}$$

The *row space* of $A \in B_{mn}$ is the span of the set of all rows of A. Similarly, the *column space* of A is the span of the set of all columns of A. Let $R(A)$ $[C(A)]$ denote the row (column) space of A.

Example 36. *Let*

$$A=\begin{bmatrix}1&0&0\\1&1&0\\1&0&1\end{bmatrix}\in B_3$$

Then $R(A)=\{0,0,0),(1,0,0),(1,1,0),(1,0,1),(1,1,1)\}$, *and* $C(A)=\{(0,0,0)^T,(0,1,0)^T,(0,0,1)^T,\ (0,1,1)^T,(1,1,1)^T\}$.

Let $A \in B_{mn}$. Then both $R(A)$ and $C(A)$ are lattices, respectively. The join of two elements is their sum and the meet of two elements is the sum of the elements of $R(A)$ which are less than or equal to both elements. Certainly, 0 is the universal lower bound while the sum of all the elements of $R(A)$ is the universal upper bound. Here 0 denotes the zero vector.

Example 37. *Let A be the same as in the above example (Figs. 8, 9 see pg. 122).*

For a binary relation from $X=\{x_1,\ x_2,\ldots,x_m\}$ to $Y=\{y_1,y_2,\ldots,y_n\}$, we can associate a $m\times n$ Boolean matrix $A=(a_{ij})$ to each binary relation R by the following rule:

$$a_{ij}=\begin{cases}1, & (x_i,y_j)\in R\\ 0, & (x_i,y_j)\notin R\end{cases}$$

This gives a one-to-one correspondence between binary relations from X to Y and $m\times n$ Boolean matrices. Under this correspondence unions of binary relations become Boolean matrix sums and compositions of binary relations become Boolean matrix products.

Example 38. *Let* $X=\{1,2,3,4\}$ *and* $Y=\{1,2,3\}$. *Let* $R=\{(1,2),(2,3),(3,1),(3,2),(4,3)\}$. *Then the Boolean matrix corresponding to R is*

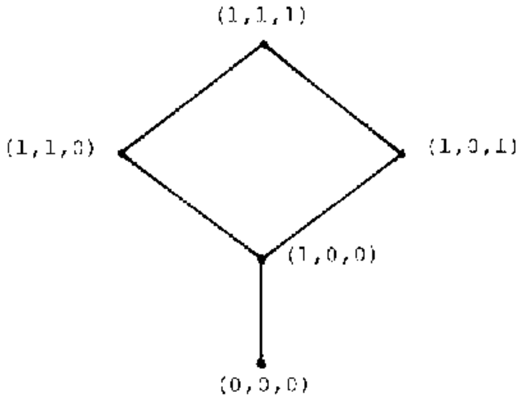


Figure 8 Lattice of $R(A)$.

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

A *directed graph* (for short *digraph*) consists of points called *vertices*, and line segments with arrows from certain vertices to certain others. The line segments are called *edges*. A digraph represents a binary relation R from a set X to a set Y if its vertices are labeled to correspond to the elements of X and Y and an edge is drawn from x to y if and only if $(x, y) \in R$. For an excellent treatment of digraph theory, see Harary et al. [14].

Example 39. *Let* $X = \{1, 2, 3, 4, 5\}$ *and* $Y = \{1, 2, 3, 4\}$. *Let* $R = \{(1, 1), (1, 4), (2, 2), (2, 3), (3, 1), (3, 4), (4, 3), (4, 4), (5, 1), (5, 4)\}$. *Then the digraph corresponding to R is shown in Fig. 10.*

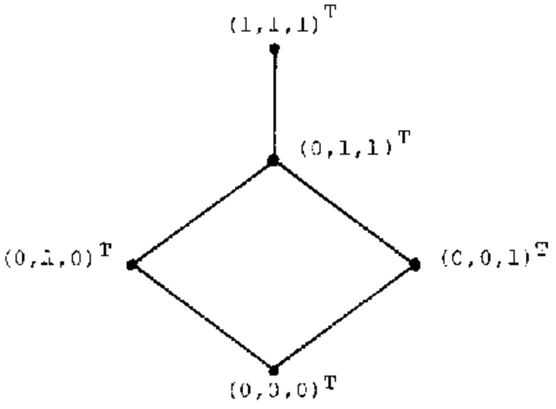


Figure 9 Lattice of $C(A)$.

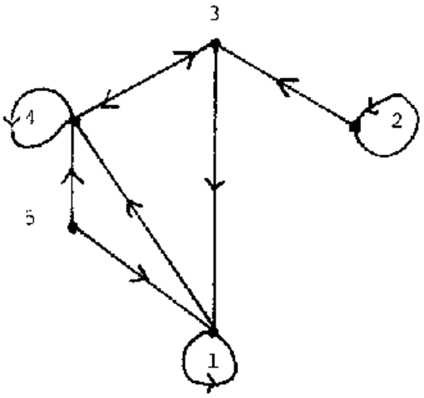


Figure 10 Digraph of a binary relation.

We give an application of Boolean matrices to switching circuits due to Hohn and Schissler [9]. Switching circuits can be drawn by a variety of graphs which cannot always be analyzed into parallel and series components. The question of analysis is: Given a circuit in graphical form, when will current flow between any pair of points?

Example 40. *Consider Fig. 11. Here x, y, u, and v are switches. We are interested in current flows among pairs of vertices, 1, 2, 3, and 4.*

The *primitive connection matrix* of a graph with vertex set V is the $|V| \times |V|$ matrix whose (i, j)-entry is labeled with the product of all switches going between vertex i and vertex j (assumed to be in series). Its (i, j)-entries are 1. Here $|V|$ denotes the cardinality of V.

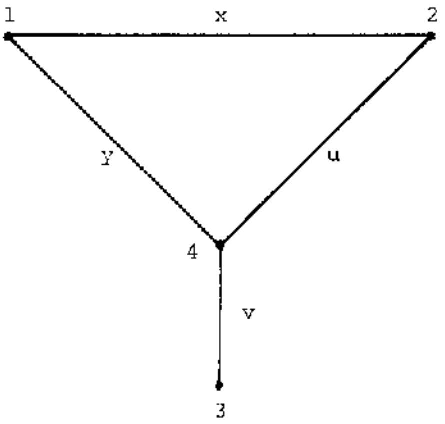


Figure 11 A general switching circuit.

Example 41. *The primitive connection matrix in Example 40 is*

$$\begin{bmatrix} 1 & x & 0 & y \\ x & 1 & 0 & u \\ 0 & 0 & 1 & v \\ y & u & v & 1 \end{bmatrix}$$

Our goal is to produce a matrix representing whether current will flow between any set of vertices in a subset S of V.

An *output matrix* of a circuit has (i, j)-entry some Boolean polynomial P such that for each setting of all switches P is 1 if and only if current flows between those vertices.

To obtain an output matrix, we may first remove all vertices outside the set S. To remove a vertex v_r, delete its row and column and then to every remaining entry (i, j) add the product of the (i, r)- and (r, j)-entries in the original matrix.

Example 42. *If we remove vertex 4 from the above example we obtain*

$$M = \begin{bmatrix} 1+yy & x+yu & vy \\ x+yu & 1+uu & uv \\ vy & uv & 1+vv \end{bmatrix}$$

Note that the matrix M so obtained is symmetrical and reflexive ($M \geq I$) *where I is an identity matrix. We may by Boolean algebra simplify this to*

$$M = \begin{bmatrix} 1 & x+yu & vy \\ x+yu & 1 & uv \\ vy & uv & 1 \end{bmatrix}$$

Now take the least power of this matrix which is idempotent ($M^2 = M$). *The reason is that nth powers of matrices have entries representing all products of entries of length n walks in their graphs. For a reflexive Boolean matrix,* $M^n \geq M^{n-1}$. *So this eventually gives all possible routes for the current to take between any two given points.*

For a reflexive n-square Boolean matrix its $(n-1)$ power will equal all subsequent powers, this corresponds to any matrix being reachable from another by a path of length at most $(n-1)$ if it is reachable at all. Powers of large Boolean matrices can be computed rapidly by repeatedly squaring them.

Example 43. *Here it is enough to take the second power of the* 3×3 *matrix in the last example. That is,*

$$\begin{bmatrix} 1 & x+yu+uvy & vy+uvx+yuv \\ x+yu+uvy & 1 & xvy+yuvy+uv \\ vy+xuv+yuv & xvy+yuvy+uv & 1 \end{bmatrix}$$
$$= \begin{bmatrix} 1 & x+yu & vy+uvx \\ x+yu & 1 & uv+xvy \\ vy+uvx & uv+xvy & 1 \end{bmatrix}$$

This is the required output function.

Given a required output, circuits may be synthesized by reversing this sequence of steps. However, there are a very great number of ways to synthesize a given circuit and it is not obvious which synthesis will have the fewest circuit components.

One could synthesize any circuit by adding a new vertex for every term but one, in every off-main-diagonal 1 entry. For example, synthesize

$$M = \begin{bmatrix} 1 & xy+uv \\ xy+uv & 1 \end{bmatrix}$$

We add a new vertex 3 so that $m_{13}m_{12}$ is the second term uv of m_{12} (see Fig. 12).

In the following we give an application of Boolean matrices to automata [15]. A *finite-state machine* consists of a finite set S (*internal states*), an initial state $\sigma \in S$, a finite set X (*input symbols*), a finite set Y (*output symbols*), and functions

$$f : S \times X \rightarrow S \qquad (\textit{transition function})$$
$$g : S \times X \rightarrow Y \qquad (\textit{output function})$$

A *finite-state nondeterministic machine* differs in that f and g are binary relations instead of functions. That is, the next state and output are not uniquely determined by the previous state.

Nondeterministic machines are of interest even though their behavior is not completely predictable. Sometimes deterministic machines can be modeled by nondeterministic machines with fewer states.

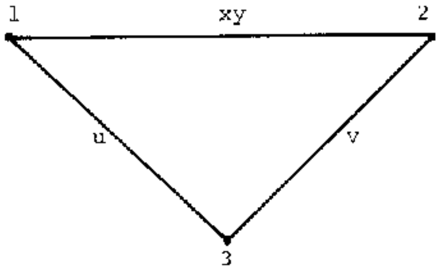


Figure 12 Equivalent connection graph.

Example 44. *We can produce a deterministic finite-state machine which will add two numbers of any number of digits where the input set $X = Y$ is $\{(a, b) : a, b \in \{0, 1, 2, \ldots, 9\}\}$. The set of internal states is $\{0, 1\}$, the possible carries. The transition function calculates the next carry from the previous carry c and the two input states:*

$$f(a, b, c) = \begin{cases} 0 & \text{if } a + b + c < 10 \\ 1 & \text{if } a + b + c \geq 10 \end{cases}$$

The output function produces the next digit of the output. Any deterministic machine is also a nondeterministic machine.

A *semigroup* is a set G together with an associative binary operation defined on G. The *semigroup of a deterministic finite automaton* is the set of all functions from S to S which are finite compositions of the functions $f_x(s) = f(s, x)$. Its operation is composition of functions. The *semigroup of a nondeterministic finite automaton* is the set of all binary relations from S to S which are finite compositions of the binary relations $f_x(s) = f(s, x)$ from S to S, under the operation composition of binary relations.

Example 45. *The semigroup of the machine in Example 44 consists of the three monotone functions f_0 ,f_1,f_2; $\{0, 1\}$ to itself: $f_0(s) = 0, f_1(s) = 1, f_2(s) = s$. For instance, if $x = (0, 0)$ we never have a carry, giving f_0. If $x = (9, 9)$ we always have a carry, giving f_1. If $x = (0, 9)$ then we have a carry if and only if the previous carry is 1, giving f_2.*

Semigroups of machines can be used to tell something about the number of states used to produce a given output. For example, we can show there is no machine analogous to the adder above which can do arbitrary multiplication. Suppose it has n states. Consider the product of the two numbers $10\ldots01$, $10\ldots01$ with n zeros.

The output will be $10\ldots020\ldots01$ with two sequences of n zeros. It will be output in reverse order, and the final $0\ldots01$ occur when both inputs are 0. Given the state s_{n+2} after the last 1's are input, the function f_{00} must have the property that $f_{00}(s_{n+2})$ lies in a state yielding output 0 for $j = 1, 2, \ldots, n$ and in a state yielding output 1 for $j = n + 1$.

However, there does not exist any such transformation on a set of n states because $f_{00}^j(s_n)$ for $j = 1, 2, \ldots, n$ will realize all states that could possibly occur for higher powers.

We can more precisely describe the asymptotic behavior of any function f on a set of n elements as follows: there is an integer k called the *index*, which is at most $(n - 1)$, and an integer d called the *period*, such that for any $j \geq k, f^{k+d} = f^k$. The integer d must divide $n!$. There is also an index and a period for the set $f^k(s)$ for each $s \in S$, and its period is at most n.

Example 46. *On $\{0, 1, \ldots, 9\}$ consider the function such that $f(x) = x - 1$ if $x \geq 1$ and $f(0) = 2$. This has index 7, since f^7 is the least power mapping the whole set into $\{0, 1, 2\}$ where it cycles, and the period is 3, since $f(2) = 1, f(1) = 0, f(0) = 2$.*

Boolean matrices A also have an index k and a period d such that $A^{j+d} = A^j$ if $j \geq d$. For n-square Boolean matrices, the index is at most $(n - 1)^2 + 1$ and the period is the period of some permutation on n elements, a divisor of $n!$.

Example 47. *Consider the Boolean matrix*

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

Its powers are the semigroup of a machine defined as follows: $S = \{1, 2, 3, 4\}$, $X = \{0\}$, $Y = \{1, 2, 3, 4\}$, $f(i, x)$ contains the element j if and only if $a_{ij} = 1$ and $g(i, x) = i$.

We have

$$A^2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} \qquad A^4 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix}$$

$$A^8 = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \qquad A^9 = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

$$A^{10} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Then A^{10} equals all larger powers and the index is $10 = (4 - 1)^2 + 1$.

Example 48. *The* 5×5 *Boolean matrix*

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

is the direct sum of a 2-cycle and a 3-cycle. It can be made into the semigroup of a nondeterministic machine in the same way. Its period is the least common multiple of 2, 3, or 6.

6.5 GENERALIZATION OF BOOLEAN ALGEBRA

The two-element Boolean algebra $\beta_0 = \{0, 1\}$ only expresses two extreme or opposite relationships such as "negative" and "positive," "no" and "yes," "off" and "on," and "false" and "true." Thus, in order to express degrees of relationships, we introduce a new algebra known as an *incline algebra* to expand β_0 to a closed unit interval [0, 1]. For a detailed account of incline algebra, see Cao et al. [16].

A *semiring* is a nonempty set S provided with two binary operations "$+$" and "$\cdot$" such that S is closed and associative under both operations, and commutative under $+$, satisfying also the distributive law.

Example 49. *Any Boolean algebra is a semiring under Boolean operations.*

In the following two examples, "$+$" stands for ordinary addition and "$\cdot$" stands for ordinary multiplication.

Example 50. *Let* Z^+ *denote the set of all positive integers. Then* $(Z^+, +, \cdot)$ *is a semiring.*

Example 51. *Let* $M_n(Z^+)$ *denote the set of all n-square matrices over* Z^+*. Then* $[M_n(Z^+), +, \cdot]$ *is a semiring.*

An *incline* is a semiring which also satisfies

i_1. Idempotence under $+$: $a + a = a$.
i_2. Absorption law: $a + ab = a$,
$a + ba = a$.

The absorption law implies that any product $ab \le a$ or $ab \le b$. Therefore, these operations tend to make quantities "slide downhill." Accordingly, we decided to call it incline. The first letter of the Korean alphabet is "7" (pronounced "gee-yeok"), which looks like a slide downhill and so 7 denotes an arbitrary incline.

Example 52. *The two-element Boolean algebra* $\beta_0 = \{0, 1\}$ *is an incline under Boolean operations.*

Example 53. *Let* Q^+ (R^+) *denote the set of all positive rationals (reals). Let* $Z^-(Q^-, R^-)$ *denote the set of all negative integers (rationals, reals). Then (1)* $\mathrm{Z}^+ (\mathrm{Q}^+, \mathrm{R}^+)$ *is not an incline under ordinary addition and ordinary multiplication. Similarly, (2)* Z^- (Q^-, R^-) *is not an incline under ordinary addition and multiplication. However,* Z^- (Q^-, R^-) *is an incline under the operations* $a + b = \sup\{a, b\}$ *and* $ab = a + b$ *(ordinary addition) for* $a, b \in Z^-$ (Q^-, R^-)*.*

Let IV_n for an incline 7 denote the Cartesian product incline $7 \times 7 \times \cdots \times 7$ (n factors). An element of IV_n is called an *incline vector of dimension n*. The system $(\mathrm{IV}_n, +, \cdot)$ is an incline under the operations $u + v = u_i + v_i$ and $u \cdot v = u_i v_i$ where $u = (u_1, \ldots, u_n)$, $v = (v_1, \ldots, v_n) \in IV_\mathrm{n}$.

Example 54. *Let 7 have operations* $\sup\{a, b\}$ *and* ab*. Let* $(0.01, 0.9, 0.2)$, $(0.5, 0.8, 0.12) \in IV_3$*. Then* $(0.01, 0.9, 0.2) + (0.5, 0.8, 0.12) = (0.5, 0.9, 0.2)$ *and* $(0.01, 0.9, 0.2) \cdot (0.5, 0.8, 0.12) = (0.005, 0.72, 0.024)$.

The matrices over an incline 7 are called *incline matrices*. Let $M_{mn}(7)$ $[M_n(7)]$ denote the set of all $m \times n$ $(n \times n)$ incline matrices. The system $[M_{mn}(7), +, \cdot]$ is an incline under the operations $A + B = \sup\{a_{ij}, b_{ij}\}$ and $A \cdot B = A \odot B = (a_{ij} b_{ij})$ (elementwise product) for all $A = (a_{ij})$, $B = (b_{ij}) \in M_{mn}(7)$.

Example 55. *Let*

$$A = \begin{bmatrix} 0.1 & 0.5 \\ 0.6 & 0.3 \end{bmatrix} \qquad B = \begin{bmatrix} 0.2 & 0.3 \\ 0.7 & 0.8 \end{bmatrix} \in M_2(7).$$

Then

$$A + B = \begin{bmatrix} 0.1 & 0.5 \\ 0.6 & 0.3 \end{bmatrix} + \begin{bmatrix} 0.2 & 0.3 \\ 0.7 & 0.8 \end{bmatrix} = \begin{bmatrix} 0.2 & 0.5 \\ 0.7 & 0.8 \end{bmatrix}$$

$$AB = \begin{bmatrix} 0.1 & 0.5 \\ 0.6 & 0.3 \end{bmatrix} \begin{bmatrix} 0.2 & 0.3 \\ 0.7 & 0.8 \end{bmatrix} = \begin{bmatrix} 0.35 & 0.4 \\ 0.21 & 0.24 \end{bmatrix}$$

$$A \odot B = \begin{bmatrix} 0.1 & 0.5 \\ 0.6 & 0.3 \end{bmatrix} \odot \begin{bmatrix} 0.2 & 0.3 \\ 0.7 & 0.8 \end{bmatrix} = \begin{bmatrix} 0.02 & 0.15 \\ 0.42 & 0.24 \end{bmatrix}$$

The *inequality* of incline matrices is defined by $A \le B$ if and only if $a_{ij} \le b_{ij}$ for all i, j. *Transpose* is defined by $A^T = (a_{ji})$.

Example 56

$$\begin{bmatrix} 0.2 & 0.5 \\ 0.1 & 0.7 \end{bmatrix} \leq \begin{bmatrix} 0.2 & 0.9 \\ 0.3 & 0.8 \end{bmatrix}$$

$$\begin{bmatrix} 0.2 & 0.5 \\ 0.1 & 0.7 \end{bmatrix}^T = \begin{bmatrix} 0.2 & 0.1 \\ 0.5 & 0.7 \end{bmatrix}$$

We introduce three basic and practical inclines. We expand $\beta_0 = \{0, 1\}$ to a closed unit interval $F = [0, 1]$ and the operations are defined by $a + b = \sup\{a, b\}$ and $a \cdot b = \inf\{a, b\}$ for all a, $b \in F$. This is called a *fuzzy algebra* and vectors over F are called *fuzzy sets*, respectively. In fact, fuzzy sets were first invented by Menger [17] in 1951 who called them *hazy sets*. However, they were independently rediscovered by Zadeh [18] who explored and popularized the subject. For basic facts and applications of fuzzy sets, see Dubois and Prade [19].

Example 57. *The fuzzy algebra is an incline under the operations maximum and minimum.*

Let F be the first basic incline, the fuzzy algebra, and denote it by 7_1. The second basic incline 7_2 is defined on F by $a + b = \inf\{a, b\}$ and $a \cdot b = \inf\{1, a + b\}$. The third basic incline 7_3 is denoted on F by $a + b = \sup\{a, b\}$ and $a \cdot b = ab$ (ordinary multiplication).

Since basic inclines essentially differ in their second operation, these determine the suitability of an incline to a particular application. (1) 7_1 with idempotent operation is suited to an application involving order properties only or a totally controlled system. (2) In 7_2, for a $\neq 0$,

$$\underbrace{a \cdot a \ldots a}_{n\,\text{factors}} = 1$$

for some n. Therefore, it is suited to applications having a strong convergence to a finite state. (3) 7_3 is suitable to applications in which constant proportionality between cause and effect holds. In the following we give an example dealing with application of 7_2 [20].

Fuzzy algebras and inclines have been used by many workers to model control systems. Here we model a simple industrial system which responds to disturbances by recovering to a satisfactory equilibrium.

The use of maximum is natural in a control system because the controller wants to maximize his or her utility. The incline 7_2 is used here because (1) the situation is linear and (2) there is a cutoff point which should not be exceeded.

A *linear system* is a dynamical system over time with

$$x\langle t+1\rangle = (x\langle t\rangle A) + B$$

where $x\langle t\rangle = (x_1, x_2, \ldots, x_n)$ is the state at time t, $A = (a_{ij})$ is a matrix describing transitions and $B = (b_1, b_2, \ldots, b_n)$ is a vector describing some external factor.

We assume a_{ij} is the desired balance between factor x_i and factor x_j. In this situation, the automatic controller is causing the transitions. Therefore, $a_{ij} = c$ means that factor x_j should not exceed factor x_i by more than c. In order for the process to be safe, we must impose $(b_1, b_2, \ldots, b_n)$ as upper limits on these n quantities.

If $x_i > a_{ij} + x_j$ then the prevalence of factor x_i over factor x_j results in an inferior product. Subject to these conditions the quantity of production is maximized when all x_i are as large as possible. Then $x_i = \inf\{a_{ij} + x_j, b_i\}$ which is precisely an equilibrium in our system.

Example 58. *Consider an automated process of manufacturing in which $x_1\langle t\rangle$ is pressure, $x_2\langle t\rangle$ is temperature, and $x_3\langle t\rangle$ acidity.*

Let $x\langle 0\rangle = (0, 0.1, 0.3)$, $B = (0.5, 0.5, 0.5)$,

$$A = \begin{bmatrix} 1 & 0.2 & 0.1 \\ 0.2 & 1 & 0.6 \\ 0.5 & 0.3 & 1 \end{bmatrix}$$

Then

$$\begin{aligned}
x\langle 1\rangle &= (x\langle 0\rangle A) + B \\
&= \left\{ (0, 0.1, 0.3) \begin{bmatrix} 1 & 0.2 & 0.1 \\ 0.2 & 1 & 0.6 \\ 0.5 & 0.3 & 1 \end{bmatrix} \right\} \\
&\quad + (0.5, 0.5, 0.5) \\
&= (0.3, 0.2, 0.1) + (0.5, 0.5, 0.5) \\
&= (0.3, 0.2, 0.1) \\
x\langle 2\rangle &= (0.4, 0.4, 0.4) \\
x\langle 3\rangle &= (0.5, 0.5, 0.5) \\
x\langle 4\rangle &= (0.5, 0.5, 0.5) \\
x\langle 5\rangle &= (0.5, 0.5, 0.5) \\
x\langle 6\rangle &= (0.5, 0.5, 0.5) \\
&\ \vdots \\
x\langle k\rangle &= (0.5, 0.5, 0.5) \qquad k \in Z^+
\end{aligned}$$

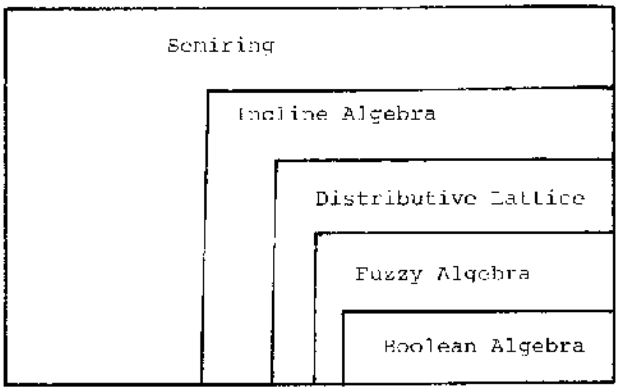


Figure 13 Inclines of various semirings.

Therefore, (0.5, 0.5, 0.5) *is the equilibrium and so it is an ideal state.*

We conclude by illustrating the relationship between the various algebras and the distinct characteristics of the various algebras mentioned in the chapter (Fig. 13, Table 18).

Table 18 Properties of Types of Semirings

	Algebras		
Properties	β_0	F	7_{i}
Ordered structure	×	×	×
Degree of intensity		×	×
Real number operations			×
Parameter of proportionality			×

ACKNOWLEDGMENTS

The author is very grateful to Fred Roush for his very constructive criticism and suggestions on the various stages of rough drafts. Also, the author would like to thank Kay Roush for very carefully proofreading the manuscript.

REFERENCES

1. G Boole. The Mathematical Analysis of Logic, Being an Essay Toward a Calculus of Deductive Reasoning. Cambridge, England: Mcmillan, 1847.
2. G Boole. An Investigation of the Laws of Thought, on Which are Founded the Mathematical Theories of Logic and Probabilities. Cambridge, England: Macmillan, 1854.
3. E Schröder. Algebra der Logik. Leipzig, 1880.
4. EV Huntington. Sets of independent postulates for the algebra of logic. Trans Math Soc 5: 288–309, 1904.
5. G Birkhoff. Lattice Theory, vol XXV. Providence, RI: Amer Math Soc Col Pub, 1967.
6. MH Stone. The theory of representations for Boolean algebras. Trans Am Math Soc 40: 37–111, 1936.
7. PR Halmos. Lecture on Boolean Algebra. Princeton, NJ: Van Nostrand, 1963.
8. R Sikorski. Boolean Algebra. Berlin: Springer-Verlag, 1964.
9. FE Hohn, LR Schissler. Boolean matrices and the design of combinatorial relay circuits. Bell Syst Tech J 34: 177–202, 1955.
10. CE Shannon. A symbolic analysis of relay and switching circuits. Trans Am Inst Elect Eng 57: 713–723, 1938.
11. FE Hohn. Applied Boolean Algebra. Toronto: Macmillan, 1969.
12. LL Dornhoff, FE Hohn. Applied Modern Algebra. New York: Macmillan, 1978.
13. KH Kim. Boolean Matrix Theory and Applications. New York: Dekker, 1982.
14. F Harary, RZ Norman, D Cartwright. Structural Models: An Introduction to the Theory of Directed Graph. New York: Wiley, 1965.
15. KH Kim, FW Roush. Automata on One Symbol. Studies in Pure Mathematics. Boston: Birkhauser, 1983, pp 423–425.
16. ZQ Cao, KH Kim, FW Roush. Incline Algebra and Applications. Chichester, England: Horwood; New York: Wiley, 1984.
17. K Menger. Selected Papers in Logic and Foundations, Didactics, Economics. New York: Reidel, 1979.
18. LA Zadeh. Fuzzy sets. Inform Control 8: 338–353, 1965.
19. D Dubois, H Prade. Fuzzy Sets and Systems Theory and Applications. New York: Academic Press, 1980.
20. KH Kim, FW Roush. Applications of inclines. Proceedings of International Conference on Information and Knowledge Engineering, Dalian, China, 1995, pp 190–196.

Chapter 1.7

Algebraic Structures and Applications

J. B. Srivastava
Indian Institute of Technology, Delhi, New Delhi, India

7.1 GROUPS

In this section we study one of the most important and useful algebraic structures, that of a *group*. The recent developments in computer applications have made it possible to automate a wide range of systems and operations in the industry by building ever more intelligent systems to cope with the growing demands. Group theory and group theoretical methods have played a vital role in certain crucial areas in these developments. Group theory is a vast subject with wide applications both within mathematics and in the real-world problems. We present here some of the most basic concepts with special emphasis on permutation groups and matrix groups.

Let S be a nonempty set. A binary operation on S is any function from $S \times S$ to S. We shall denote an arbitrary operation on S by "$\star$." Thus $\star : S \times S \to S$ sending $(x, y) \to x \star y$ assigns to each ordered pair (x, y) of elements of S an element $x \star y$ in S. Binary operation $\star$ on S is said to be associative if $(x \star y) \star z = x \star (y \star z)$ for all $x, y, z \in S$. $(S, \star)$ is called a semigroup if S is a nonempty set and $\star$ defines an associative binary operation on S.

$(G, \star)$ is said to be a *group* if G is nonempty set and $\star$ is a binary operation on G satisfying the following properties (axioms):

1. $(x \star y) \star z = x \star (y \star z)$ for all $x, y, z \in G$.
2. There exists an element $e \in G$ such that $e \star x = x$ for every $x \in G$.
3. For every $x \in G$, there exists an element $y \in G$ with $y \star x = e$.

If $(G, \star)$ is a group, as above, then it can be proved that $x \star e = x$ for every $x \in G$ and in axiom 3, $y \star x = e$ if and only if $x \star y = e$. Further, e in axiom 2 and y in axiom 3 are unique. In the group $(G, \star)$, the element e is called the identity of the group G and y is called the inverse of x in G. Axiom 1 says that every group is a semigroup. Thus a group is a semigroup in which there exists a unique identity element e with $e \star x = x = x \star e$ and for every x there exists a unique y such that $y \star x = e = x \star y$.

7.1.1 Examples of Groups

7.1.1.1 Abelian Groups

A group $(G, \star)$ is said to be Abelian or commutative if $x \star y = y \star x$ for all $x, y \in G$. The following examples are well known:

1. $(\mathcal{Z}, +) =$ the group of all integers (positive, negative and 0) under addition.
2. $(\mathbb{R}, +) =$ the group of all real numbers under addition.
3. $(\mathbb{C}, +) =$ the group of all complex numbers under addition.
4. $(\mathbb{M}_n(\mathbb{R}), +) =$ the group of all $n \times n$ real matrices under addition.
5. $(\mathbb{R}^\star, \cdot) =$ all nonzero real numbers under multiplication.

6. $(\mathbb{C}^{\star}, \cdot)$ = all nonzero complex numbers under multiplication.
7. $G = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \right\}$ under multiplication.

7.1.1.2 Permutation Groups

Let X be a nonempty set. Any one-to-one, onto function from X to X is called a permutation on X. The set of all permutations on X form a group under multiplication which is by definition a composition of functions. Let S_X denote the set of all permutations on X and let $\sigma, \rho \in S_X$. Then σ and ρ are one-to-one, onto functions from X to X. Their product $\sigma\rho = \sigma \circ \rho$ is defined by $\sigma\rho(x) = (\sigma \circ \rho)(x) = \sigma(\rho(x))$ for every $x \in X$. Clearly $\sigma\rho$ is one-to-one, onto and hence $\sigma\rho \in S_X$. Composition of functions is associative. The function $e : X \rightarrow X$ defined by $e(x) = x$ for every $x \in X$ is the identity. Also every one-to-one, onto function has an inverse which is also one-to-one, onto. Thus S_X under multiplication (which is composition) is a group.

When $X = \{x_1, x_2, \ldots, x_n\}$ is a finite set, a function from X to X is one-to-one if and only if it is onto. In this case the group of all permutations S_X on X is denoted by S_n. In particular, if $X = \{1, 2, 3, \ldots, n\}$, the group of all permutations on X under multiplication is denoted by S_n and is called the symmetrical group of degree n. It may be noted that the symmetrical group S_n is non-Abelian for all $n \geq 3$.

7.1.1.3 Matrix Groups

1. $GL_n(\mathbb{R}) = GL(n; \mathbb{R})$ = the group of all $n \times n$ nonsingular real matrices under matrix multiplication. Here nonsingular means invertible or equivalently having nonzero determinant. $GL_n(\mathbb{R})$ is called the general linear group of degree n over the reals $\mathbb{R}$.
2. $GL_n(\mathbb{C}) = GL(n; \mathbb{C})$ = the general linear group of degree n over the field of all complex numbers.
3. $SL_n(\mathbb{R}) = SL(n; \mathbb{R}) = \{A \in GL_n(\mathbb{R}) \mid \det(A) = 1\}$ = the special linear group of degree n over $\mathbb{R}$.
4. $SL_n(\mathbb{C})$ = the special linear group of degree n over $\mathbb{C}$.
5. $U(n) = U_n(\mathbb{C})$ = the n-dimensional unitary group is the multiplicative group of all $n \times n$ complex unitary matrices. $A \in GL_n(\mathbb{C})$ is unitary if $A^{-1} = (\overline{A})^t$ = the conjugate transpose of A.
6. $O(n) = O(n; \mathbb{R})$ = the n-dimensional real orthogonal group = the group of all $n \times n$ real orthogonal matrices under multiplication. $A \in O(n) \Leftrightarrow A^tA = I = AA^t \Leftrightarrow A^{-1} = A^t$.
7. $SU(n)$ = the special unitary group of degree $n = \{A \in U(n) \mid \det(A) = 1\}$ under matrix multiplication.
8. $SO(n)$ = the special orthogonal group of degree $n = \{A \in O(n) \mid \det(A) = 1\}$ under matrix multiplication.

7.1.1.4 Miscellaneous

1. *Dihedral groups*:

$$D_{2n} = \langle x, y \mid x^n = 1, y^2 = 1, y^{-1}xy = x^{-1} \rangle$$

   This is the group generated by two elements x and y with prdouct rules given by $x^n = 1$, $y^2 = 1$, $yx = x^{-1}y$, $x^{-1} = x^{n-1}$, $y^{-1} = y$ having 1 as the identity.
2. *Quaternion groups*: $Q_8 = \{\pm 1, \pm i, \pm j, \pm k\}$ with multiplication rules $i^2 = j^2 = k^2 = -1$, $ij = k = -ji$, $jk = i = -kj$, $ki = j = -ik$ and identity 1.
3. $\{1, -1, i, -i\}$ under multiplication where $i = \sqrt{-1}$ is the complex number.
4. $G_n = \{z \in \mathbb{C} \mid z^n = 1\}$ = the multiplicative group of all nth roots of unity belonging to $\mathbb{C}$. Here n is fixed but arbitrary.

7.1.2 Basic Concepts

Let $(G, \star)$ be a group. A nonempty subset H of G is a subgroup of G if H is a group under the binary operation of G. This means (1) $e \in H$, (2) $x, y \in H \Rightarrow x \star y \in H$, (3) $x \in H \Rightarrow x^{-1} \in H$. It is easy to see that H is a subgroup of G if and only if $x \star y^{-1} \in H$ for all $x, y \in H$. A subgroup H of G is said to be a normal subgroup of G if $g \star H = H \star g$ for each $g \in G$. Thus H is a normal subgroup of G if and only if $g^{-1} \star H \star g = H = g \star H \star g^{-1}$ for all $g \in G$.

If $x, y \in G$, then y is said to be a conjugate of x if $y = g \star x \star g^{-1}$ or $y = g^{-1} \star x \star g$ for some $g \in G$. If H is a subgroup of G, then a conjugate of H is $g \star H \star g^{-1}$ or $g^{-1} \star H \star g$ for any $g \in G$.

If H is a subgroup of G, we define $g \star H = \{g \star h \mid h \in H\}$ = the left coset of H by $g \in G$ and $H \star g = \{h \star g \mid h \in H\}$ = the right coset of H by g. Two cosets are either identical or disjoint. Further, $H \star g = H \Leftrightarrow g \in H \Leftrightarrow g \star H = H$. Using this, we get $H \star g_1 = H \star g_2 \Leftrightarrow g_1 \star g_2^{-1} \in H$. Similarly, $g_1 \star H = g_2 \star H \Leftrightarrow g_1^{-1} \star g_2 \in H$.

For a finite group G, we denote by $|G|$ the order of the group G. Here $|G|$ denotes the number of distinct elements in G. If H is a subgroup of G, then define $|G : H| =$ the index of H in $G =$ the number of distinct left cosets of H in $G =$ the number of distinct right cosets of H in G.

If H is a normal subgroup of G, then the quotient group or factor group G/H is defined by $G/H = \{H \star g = g \star H \mid g \in G\} =$ the set of all right (left) cosets of H in G and

$$(H \star g_1) \star (H \star g_2) = H \star (g_1 \star g_2)$$

$(H \star g)^{-1} = H \star g^{-1}$ and identity G/H is $H = H \star e$.

Let $(G_1, \star_1)$ and $(G_2, \star_2)$ be groups which may be identical. Then a function $f : G_1 \to G_2$ is called a group homomorphism if $f(x \star_1 y) = f(x) \star_2 f(y)$ for all $x, y \in G_1$. If $f : G_1 \to G_2$ is a homomorphism, then $K = \mathrm{Ker} f =$ the kernel of f is defined by

$$\mathrm{Ker} f = \{x \in G_1 \mid f(x) = e_2 = \text{the identity of } G_2\}$$

It is not difficult to see that $K = \mathrm{Ker} f$ is a normal subgroup of G_1. Also $\mathrm{Im} f = \{f(x) \mid x \in G_1\}$ is a subgroup of G_2.

An isomorphism of groups is a one-to-one, onto group homomorphism. We write $G_1 \cong G_2$ for G_1 is isomorphic to G_2 if there exists an isomorphism from G_1 to G_2.

If S is a subset of G, then $\langle S \rangle$ denotes the smallest subgroup of G containing the subset S. If $G = \langle S \rangle$, then we say that G is generated by the subset S. If $S = \{x\}$ is a singleton subset and $G = \langle x \rangle$, then G is called a cyclic group. Thus G is a cyclic group if and only if it is generated by a single element. G is a finite cyclic group if $G = \langle x \rangle$ for some $x \in G$ and $x^n = e$ for some positive integer n. The group G is called infinite cyclic if $G = \langle x \rangle$ for some $x \in G$ and $x^n \neq e$ for any positive integer n. If x is an element of a group G, then x has infinite order if x^n *nee* for any positive integer n. An element x of a group has finite order if $x^n = e$ for some positive integer n. The least positive integer n such that $x^n = e$ is called the order of x and is denoted by $o(x) = n$ the order of x. In fact, $o(x) = |\langle x \rangle| =$ the order of the cyclic subgroup generated by x.

7.1.3 Main Theorems

Now onward, for the sake of convenience, we shall use xy for $x \star y$ whenever there is no confusion. In this section, we state without proof some theorems in group theory and explain their importance. We start with the basic results on finite groups.

Lagrange's Theorem. *Let H be a subgroup of a finite group G. Then*

$$|G| = |H||G : H|$$

Remarks:

1. From Lagrange's theorem it is clear that the order and the index of any subgroup divide the order of the whole group.
2. The converse of Lagrange's theorem is false in the sense that if $|G| = n$ and m divides n, then G need not have a subgroup of order m.

A pivotal role is played by the following most famous theorems on finite groups.

Sylow's Theorem. *Let G be a finite group having $|G| = p^n m$, p a prime, p and m relatively prime. Then:*

1. *G has at least one subgroup of order p^n.*
2. *Any two subgroups of G having order p^n are conjugate in G.*
3. *The number n_p of Sylow p-subgroups (subgroups of order p^n) of G is of the form $n_p = 1 + kp$, $k = 0, 1, 2, \ldots$ and n_p divides $|G|$.*

Remarks:

1. Subgroups of G of order p^n in the above theorem are called Sylow p-subgroups of G.
2. Theorems 1, 2, and 3 above are called Sylow's first, second, and third theorem respectively.
3. Sylow's theorems and Lagrange's theorem are the most basic theorems in the study of finite groups.
4. It is known that if $|G| = p$, p a prime, then G is cyclic and if $|G| = p^2$, then G is Abelian.

For arbitrary groups, finite or infinite, the following theorem is quite useful.

Fundamental Theorem of Homomorphism. *Let G_1 and G_2 be two groups and let $f : G_1 \to G_2$ be a group homomorphism. Then*

$$G_1/\mathrm{Ker} f \cong \mathrm{Im} f$$

Thus, if f is onto, then $G_1/\mathrm{Ker} f \cong G_2$.

Remarks:

1. Define $f : GL_n(\mathbb{R}) \to \mathbb{R}^\star = \mathbb{R}\backslash\{0\}$ by $f(A) = \det(A)$. Then $f(AB) = \det(AB) = \det(A)\det(B) = f(A)f(B)$. Thus f is a group homomorphism. Clearly, Ker $f = \{A \in GL_n(\mathbb{R}) \mid f(a) = \det(A)$

$= a\} = SL_n(\mathbb{R})$. By the above theorem $GL_n(\mathbb{R})/SL_n(\mathbb{R}) \cong (R^\star, \cdot)$.

2. Define $f : (\mathbb{R}, +) \to (\mathbb{R}^\star_+, \cdot)$ by $f(x) = e^x$. Then $f(x+y) = e^{x+y} = e^x e^y = f(x)f(y)$. Thus f is a homomorphism, Ker $f(= (0)$. In fact f is an isomorphism. Here $(\mathbb{R}^\star_+, \cdot)$ is the multiplicative group of all positive real numbers.

7.1.4 Permutation Groups

Permutations arise naturally in several concrete problems. The symmetrical group S_n of degree n consists of all permutations on $\{1, 2, 3, \dots, n\}$ and the binary operation is product which is composition. We discuss in detail the permutations because of their importance in several practical problems.

Let $\sigma \in S_n$ be any permutation. Then σ can be represented by displaying its values:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & \dots & n \\ \sigma(1) & \sigma(2) & \sigma(3) & \dots & \sigma(n) \end{pmatrix}$$

The order of the symmetrical group S_n is $n!$. $\sigma \in S_n$ fixes k if $\sigma(k) = k$. The most simple kind of permutations are transpositions. A permutation $\tau \in S_n$ is called a transposition if $\tau(i) = j$, $\tau(j) = i$ and $\tau(k) = k$ for all $k \neq i, j$ where $1 \leq i, j, k \leq n$ with $i \neq j$. We write this transposition as $\tau = (i, j)$. Thus $\tau = (r, s)$ means $r \to s$, $s \to r, k \to k$ for $k \ \neq r, s, r \neq s$, i.e., $\tau(r) = s$, $\tau(s) = r$, $r \neq s$, and $\tau(k) = k$ for $k \neq r, s$. Clearly, every transposition τ has order 2, i.e., $\tau^2 = e$, where e is the identity permutation and hence $\tau^{-1} = \tau$.

A permutation $\rho \in S_n$ is called a cycle of length r if there exist distinct integers $i_1, i_2, \dots, i_r$ between 1 and n such that

$$\begin{array}{ll} \rho(i_1) = i_2 & \rho(i_2) = i_3, \dots, \rho(i_{r-1}) = \rho(i_r) \\ \rho(i_r) = i_1 & \rho(k) = k \end{array}$$

for all other integers between 1 and n. It is denoted by $\rho = (i_1, i_2, \dots, i_r)$. Thus S_3 is explicitly given by

$$S_3 = \left\{ e = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \tau_1 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \right.$$

$$\tau_2 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}, \tau_3 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix},$$

$$\left. \sigma = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \rho = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} \right\}$$

$$= \{e, (1, 2), (1, 3), (2, 3), (1, 2, 3), (1, 3, 2)\}$$

Now $\tau_1\tau_2 = (1\ 2)(1\ 3) = (1\ 3\ 2) = \rho$ but $\tau_2\tau_1 = (1\ 3)(1\ 2) = (1\ 2\ 3) = \sigma$. Thus $\tau_1\tau_2 \neq \tau_2\tau_1$ and hence S_3 is non-Abelian. In fact S_n is non-Abelian for $n \geq 3$.

Two cycles, $\sigma = (i_1 i_2 \dots i_r)$ and $\rho = (j_1, j_2 \dots j_s)$, in S_n are said to be disjoint if $\{i_1, i_2, \dots, i_r)$ and $\{j_1, j_2, \dots, j_s\}$ are disjoint as sets, i.e., they do not have common indices. Disjoint cycles commute. It is a routine computation to verify that any nonidentity permutation can be written as a product of disjoint cycles and writing this way is unique except for the order in which the cycles are written. Since 1-cycles are identity permutations, we ignore them. This is best illustrated by an example:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 \\ 7 & 10 & 4 & 6 & 11 & 9 & 3 & 2 & 1 & 8 & 5 & 14 & 13 & 12 \end{pmatrix}$$
$$= (1\ 7\ 3\ 4\ 6\ 9)(2\ 10\ 8)(5\ 11)(12\ 14)$$

Here $\sigma \in S_{14}$ has been expressed as a product of disjoint cycles which commute.

Remarks:

1. A 2-cycle is simply a transposition.
2. An r-cycle $(i_1 i_2 \dots i_r)$ can be expressed as a product of $(r-1)$ transpositions as

   $$(i_1 i_2 \dots i_r) = (i_1 i_r)(i_1 i_{r-1}) \dots (i_1 i_2)$$

   which are not disjoint.
3. Any permutation in $S_n, n \geq 2$, can be written as a product of transpositions. It is a well-known result that while writing a permutation as a product of transpositions, the number of transpositions is either always even or always odd. It is never possible to write the same permutation as a product of even number of transpositions as well as a product of odd number of transpositions.
4. $\sigma \in S_n$ is called an *even* permutation if σ is a product of even number of transpositions. σ is called an odd permutation if σ can be written as a product of odd number of transpositions.
5. Let $A_n = \{\sigma \in S_n | \sigma \text{ is even}\}$. Then A_n is a subgroup of S_n and it is called the alternating group of degree n. The mapping $f : S_n \to \{1, -1\}$ sending

   $$f(\sigma) = \begin{cases} 1 & \text{if } \sigma \text{ is even} \\ -1 & \text{if } \sigma \text{ is odd} \end{cases}$$

   is an onto group homomorphism with Ker$f = A_n$. Thus $S_n/A_n \cong \{1, -1\}$. Here $\{1, -1\}$ is a group under multiplication. Thus order $|A_n| = n!/2$ and index $|S_n : A_n| = 2$. Also A_n is a normal subgroup of S_n.

6. Any subgroup of the symmetrical group S_n is called a permutation group.

7.1.5 Group Actions

Let G be a group and let X be a nonempty set. A group action of G on X is a function from $G \times X \to X$ sending (g, x) to $g \cdot x \in X$ satisfying the following:

1. $(g_1 \cdot g_2) \cdot x = g_1 \cdot (g_2 \cdot x)$ for all $g_1, g_2 \in G$ and all $x \in X$.
2. $e \cdot x = x$ for all $x \in X$, where e is the identity of the group G.

Given a group action, there exists a group homomorphism $\rho : G \to S_X$ defined by $g \to \rho_g$ and $\rho_g : X \to X$ maps x to $g \cdot x$. Conversely, every group homomorphism $\rho : G \to S_X$ defines a unique group action where $g \cdot x = \rho_g(x)$ for $g \in G$ and $x \in X$.

Given a group action of the group G on a set X, we have the following concepts:

1. The orbits of the group action are defined in a natural way. If $x \in X$, then $O_x = O(x) = \{g \cdot x \mid g \in G\}$ = the orbit of x under the given G-action.
2. The stabilizer subgroup of $x \in X$ is $G_x = \{g \in G \mid g \cdot x = x\}$ = the stabilizer subgroup of x.
3. Points on the same orbit have conjugate stabilizers. Suppose x and y belong to the same orbit. Then there exists a $g \in G$ such that $y = g \cdot x$. Then it can be shown that the stabilizers of x and y are conjugate as $G_y = gG_xg^{-1}$.

In this context, we have the following well-known theorems.

Orbit-Stabilizer Theorem. *Let G be a finite group acting on a finite nonempty set X. Then the size of the orbit $|O_x| = |G : G_x|$ = the index of the stabilizer subgroup for each $x \in X$. Further, the size of each orbit divides the order $|G|$ of the group.*

The next theorem gives the count.

The Counting Orbit Theorem. *If G and X are as in the previous theorem then the number of distinct orbits is given by*

$$\frac{1}{|G|}\sum_{g \in G} |X^g|$$

where $|X^g|$ denotes the number of distinct elements of X left fixed by g in the sense $g \cdot x = x$.

7.1.6 Group of Rigid Motions

We briefly describe the group of all rigid motions (the Euclidean group). This is the group of all distance-preserving functions from $\mathbb{R}^n \to \mathbb{R}^n$. When $n = 2$ or 3, these groups are of great practical importance and contain translations, rotations, and reflections. Here $\mathbb{R}^n = \{(x_1, x_2, \ldots, x_n) \mid x_i \in \mathbb{R}, 1 \leq i \leq n\}$ consists of all ordered n-tuples of real numbers, and the distance between two points $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ is defined by $d(x, y)$ where $d(x, y) = \|x - y\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$. A function $T : \mathbb{R}^n \to \mathbb{R}^n$ is distance preserving if $d(Tx, Ty) = \|Tx - Ty\| = \|x - y\| = d(x, y)$ for all $x, y \in \mathbb{R}^n$. All distance-preserving functions from $\mathbb{R}^n$ to $\mathbb{R}^n$ form a group under composition of functions. Such functions are called rigid motions of $\mathbb{R}^n$. When $n = 2$, $\mathbb{R}^2$ defines a plane and when $n = 3$, $\mathbb{R}^3$ defines the usual three-dimensional space.

If $w \in \mathbb{R}^n$, $T_w : \mathbb{R}^n \to \mathbb{R}^n$ given by $T_w(x) = x + w$ for $x \in \mathbb{R}^n$ defines a translation. $SO(2)$ = the group of all 2×2 orthogonal matrices having determinant 1 define the group of all rotations of the plane and $SO(3)$ similarly defines the group of all rotations of the three-dimensional Euclidean space. Reflections can be represented by orthogonal matrices having determinant -1. Every rigid motion is the composite of a translation and an orthogonal transformation which may be a rotation or a reflection in the above sense. Geometrical applications help in understanding the physical phenomena.

7.2 RINGS AND FIELDS

In this section we deal with algebraic structures having two binary operations satisfying certain axioms.

A ring R is a nonempty set together with two binary operations "+" and "·" called addition and multiplication, which satisfy the following axioms:

1. $(R, +)$ is an Abelian group with identity denoted by 0.
2. $(R, \cdot)$ is a semigroup with identity denoted by 1.
3. Distributive laws hold: for all $a, b, c \in R$,

$$(a + b) \cdot c = a \cdot c + b \cdot c$$
$$a \cdot (b + c) = a \cdot b + a \cdot c$$

Note: in axiom 2 some authors do not assume the identity.

R is said to be a commutative ring if $a \cdot b = b \cdot a$ for all $a, b \in R$. A commutative ring with identity 1 is called an integral domain if for $a, b \in R, a \neq 0, b \neq 0$ always implies $a \cdot b \neq 0$. A field is a commutative ring with identity $1 \neq 0$ in which every nonzero element has a multiplicative inverse, i.e., $(R^{\star}, \cdot)$ is also an Abelian group where $R^{\star} = R \backslash \{0\}$ = all nonzero elements of R.

7.2.1 Examples

7.2.1.1 Commutative Rings

1. $(\mathcal{Z}, +, \cdot) = \mathcal{Z}$ = the ring of all integers under usual addition and multiplication.
2. $\mathbb{R}[X]$ = the ring of all real polynomials under usual addition and multiplication.
3. $R = \mathcal{C}([0, 1]) = \{f : [0, 1] \rightarrow R | f$ is continous$\}$, the ring of all continuous real-valued functions defined on the closed interval [0, 1] with addition and multiplication of functions defined pointwise

$$(f + g)(x) = f(x) + g(x)$$
$$(f \cdot g)(x) = f(x)g(x) \qquad \text{for all } x \in [0, 1]$$

4. $\mathcal{Z}[i] = \{a + bi \in \mathbb{C} | a, b \in \mathcal{Z}\}$, the ring of gaussian integers consisting of all complex numbers having their real and imaginary parts integers, and addition and multiplication inherited from complex numbers.
5. $\mathcal{Z}_n = \{0, 1, 2, \ldots, (n-1)\}$ with addition and multiplication defined modulo n. Here $r \equiv s$ mod(n) means $r - s$ is divisible by n. $n \equiv 0$ mod (n).

7.2.1.2 Examples of Fields

1. a. $\mathbb{R}$ = the field of all real numbers with usual addition and multiplication.
   b. $\mathbb{C}$ = the field of all complex numbers with usual addition and multiplication.
   c. $\mathbb{Q}$ = the field of all rational numbers.
2. $\mathcal{Z}_p$ = the field of all integers modulo p, p a prime. Here $\mathcal{Z}_p = \{0, 1, 2, \ldots, (p-1)\}$, p a prime and addition and multiplication is defined modulo p.
3. $K(X)$ = the field of all rational functions with coefficients in a given field K

$$= \left\{ \frac{f(X)}{g(X)} \mid f(X), g(X) \in K[X] \right.$$
$$\left. \text{polynomials with } g(X) \neq 0 \right\}$$

By taking $K = \mathbb{R}, \mathbb{C}, \mathbb{Q}$, or $\mathcal{Z}_p$ we get special examples of fields of rational functions.

7.2.1.3 Noncommutative Rings

1. $M_n(K)$ = the ring of all $n \times n$ matrices with entries from the field K with the usual addition and multiplication of matrices, $n \geq 2$. $M_n(\mathbb{R})$, $M_n(\mathbb{C})$ etc. give concrete examples of practical importance.
2. $\mathcal{H} = \{a + bi + cj + dk \mid a, b, c, d \in \mathbb{R}\}$ the ring of all real quaternions with addition $(a + bi + cj + dk) + (a' + b'i + c'j + d'k) = (a + a') + (b + b')i + (c + c')j + (d + d')k$ and multiplication adefined distributively using the rules: $i^2 = j^2 = k^2 = -1$, $ij = k, ji = -k; jk = i, kj = -i; ki = j, ik = -j$.

7.2.2 Basic Concepts

Since most of the rings from the application point of view are generally commutative, the emphasis will be on such rings.

Let R be a ring. S is called a subring of R if $a, b \in S \Rightarrow a \pm b, a \cdot b, 1 \in S$. Further, a nonempty set I is called an ideal of R ifs $a, b \in I \Rightarrow a \pm b \in I$ and $a \in I, r \in R \Rightarrow a \cdot r, r \cdot a \in I$. In this situation $(I, +)$ is a subgroup of $(R, +)$.

If I is an ideal of a ring R, then the quotient ring or factor ring is given by

$R/I = \{I + a \mid a \in R\}$ with addition and multiplication

defined by

$$(I + a) + (I + b) = I + (a + b)$$
$$(I + a) \cdot (I + b) = I + (a \cdot b)$$

Let R_1 and R_2 be two rings which may be the same. A function $f : R_1 \rightarrow R_2$ is called a ring homomorphism if it preserves addition and multiplication, i.e.,

$$f(a + b) = f(a) + f(b) \qquad f(ab) = f(a)\, f(b)$$

for all $a, b \in R_1$. Here "+" and "$\cdot$" denote + and $\cdot$ of R_1 and R_2 depending upon where the elements belong to.

If $f : R_1 \rightarrow R_2$ is a ring homomorphism, then the kernel of f is defined as

$$\text{Ker} f = \{a \in R_1 \mid f(a) = 0 \text{ in } R_2\}$$

It is easy to see that Ker$f$ is an ideal of R_1 and Im$f = \{f(a) \mid a \in R_1\}$ is a subring of R_2.

An isomorphism is a one-to-one, onto ring homomorphism. If $f : R_1 \rightarrow R_2$ is an isomorphism, then we saythat R_1 is isomorphic to R_2 via f. In general, we write $R_1 \cong R_2$ and read R_1 is isomorphic to R_2 if there exists an isomorphism from R_1 to R_2.

Note: in homomorphism, we always assume that $f(1) = 1$, i.e., the homomorphism maps 1 of R_1 to 1 of R_2.

Fundamental Homomorphism Theorem for Rings. *Let* $f : R_1 \rightarrow R_2$ *be a ring homomorphism with kernel* $\mathrm{Ker} f$. *Then*

$$R_1/\mathrm{Ker} f \cong \mathrm{Im} f$$

and if f is onto, $R_1/\mathrm{Ker} f \cong R_2$.

7.2.3 Polynomial Rings

In this section, we shall study polynomial rings over the field $\mathbb{R}$ of all real numbers and the field $\mathbb{C}$ of all complex numbers. The study of these rings leads to the understanding of algebraic and projective geometry which has a wide range of applications in such diverse areas as robotics and computer vision. Much of this material is very well presented in Cox et al. [1] and uses the computation of Grobner bases as a basic tool.

$\mathbb{R}[X]$ and $\mathbb{C}[X]$ denote the polynomial rings in single variable X. Also $\mathbb{R}[X_1, X_2, \ldots, X_n]$ and $\mathbb{C}[X_1, X_2, \ldots, X_n]$ for $n \geq 2$ denote the polynomial rings in n commuting variables $X_1, X_2, \ldots, X_n$. Over the field $\mathbb{C}$ of complex numbers, the following theorem plays the dominant role.

The Fundamental Theorem of Algebra. *Every nonconstant polynomial in* $\mathbb{C}\,[X]$ *has a root in* $\mathbb{C}$.

Remark. If $f(X) \in \mathbb{C}\,[X]$ is nonconstant, then repeated application of the above theorem gives $f(x) = cx \prod_{i=1}^{n}(X - \alpha_i)$; $c, \alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbb{C}$ and $n =$ degree of $f(X)$.

The behavior of polynomials in several variables is much more difficult. Understanding of these polynomials requires the knowledge of commutative rings.

REFERENCES

1. D Cox, J Little, D O'Shea. Ideals, Varieties and Algorithms. Springer-Verlag, UTM, 1992.
2. M Artin. Algebra. Englewood Cliffs, NJ: Prentice Hall, 1991.
3. IN Herstein. Topics in Algebra, 2nd ed. New York: Wiley, 1975.

Chapter 2.1

Measurement and Control Instrumentation Error-Modeled Performance

Patrick H. Garrett
University of Cincinnati, Cincinnati, Ohio

1.1 INTRODUCTION

Modern technology leans heavily on the science of measurement. The control of industrial processes and automated systems would be very difficult without accurate sensor measurements. Signal-processing functions increasingly are being integrated within sensors, and digital sensor networks directly compatible with computer inputs are emerging. Nevertheless, measurement is an inexact science requiring the use of reference standards and an understanding of the energy translations involved more directly as the need for accuracy increases. Seven descriptive parameters follow:

Accuracy: the closeness with which a measurement approaches the true value of a measurand, usually expressed as a percent of full scale.

Error: the deviation of a measurement from the true value of a measurand, usually expressed as a precent of full scale.

Tolerance: allowable error deviation about a reference of interest.

Precision: an expression of a measurement over some span described by the number of significant figures available.

Resolution: an expression of the smallest quantity to which a quantity can be represented.

Span: an expression of the extent of a measurement between any two limits.

A general convention is to provide sensor measurements in terms of signal amplitudes as a percent of full scale, or %FS, where minimum–maximum values correspond to 0 to 100%FS. This range may correspond to analog signal levels between 0 and 10 V (unipolar) with full scale denoted as $10\,V_{FS}$. Alternatively, a signal range may correspond to ±50%FS with signal levels between ±5 V (bipolar) and full scale denoted at $\pm 5V_{FS}$.

1.2 INSTRUMENTATION AMPLIFIERS AND ERROR BUDGETS

The acquisition of accurate measurement signals, especially low-level signals in the presence of interference, requires amplifier performance beyond the typical capabilities of operational amplifiers. An instrumentation amplifier is usually the first electronic device encountered by a sensor in a signal-acquisition channel, and in large part it is responsible for the data accuracy attainable. Present instrumentation amplifiers possess sufficient linearity, stability, and low noise for total error in the microvolt range even when subjected to temperature variations, and is on the order of the nominal thermocouple effects exhibited by input lead connections. High common-mode rejection ratio (CMRR) is essential for achieving the amplifier performance of interest with regard to interference rejection, and for establishing a signal ground reference at the amplifier that can accommodate the presence of ground–return potential differences. High amplifier input impedance is also necessary to preclude input signal loading and voltage divider effects from finite source impedances, and to accommodate source-impedance imbalances without degrading CMRR. The precision gain values possible with instrumentation amplifiers, such as 1000.000, are equally important to obtain accurate scaling and registration of measurement signals.

The instrumentation amplifier of Fig. 1 has evolved from earlier circuits to offer substantially improved performance over subtractor instrumentation amplifiers. Very high input impedance to $10^9\,\Omega$ is typical with no resistors or their associated temperature coefficients involved in the input signal path. For example, this permits a $1\,\mathrm{k}\Omega$ source impedance imbalance without degrading CMRR. CMRR values to 10^6 are achieved with $A_{\mathrm{V_{diff}}}$ values of 10^3 with precision internal resistance trimming.

When conditions exist for large potentials between circuits in a system an isolation amplifier should be considered. Isolation amplifiers permit a fully floating sensor loop because these devices provide their own input bias current, and the accommodation of very high input-to-input voltages between a sensor input and the amplifier output ground reference. Off-ground V_{cm} values to ±10 V, such as induced by interference coupled to signal leads, can be effectively rejected by the CMRR of conventional operational and instrumentation amplifiers. However, the safe and linear accommodation of large potentials requires an isolation mechanism as illustrated by the transformer circuit of Fig. 2. Light-emitting diode (LED)-phototransistor optical coupling is an alternate isolation method which sacrifices performance somewhat to economy. Isolation amplifiers are especially advantageous in very noisy and high voltage environments and for breaking ground loops. In addition, they provide galvanic isolation typically on the order of $2\,\mu\mathrm{A}$ input-to-output leakage.

The front end of an isolation amplifier is similar in performance to the instrumentation amplifier of Fig. 1 and is operated from an internal dc–dc isolated power converter to insure isolation integrity and for sensor excitation purposes. Most designs also include a $100\,\mathrm{k}\Omega$ series input resistor R to limit the consequences of catastrophic input fault conditions. The typical amplifier isolation barrier has an equivalent circuit of $10^{11}\,\Omega$ shunted by 10 pF representing R_{iso} and C_{iso}, respectively. An input-to-output V_{iso} rating of ±2500 V peak is common, and is accompanied by an isolation-mode rejection ratio (IMRR) with reference to the output. Values of CMRR to 10^4 with reference to the input common, and IMRR values of 10^8 with reference

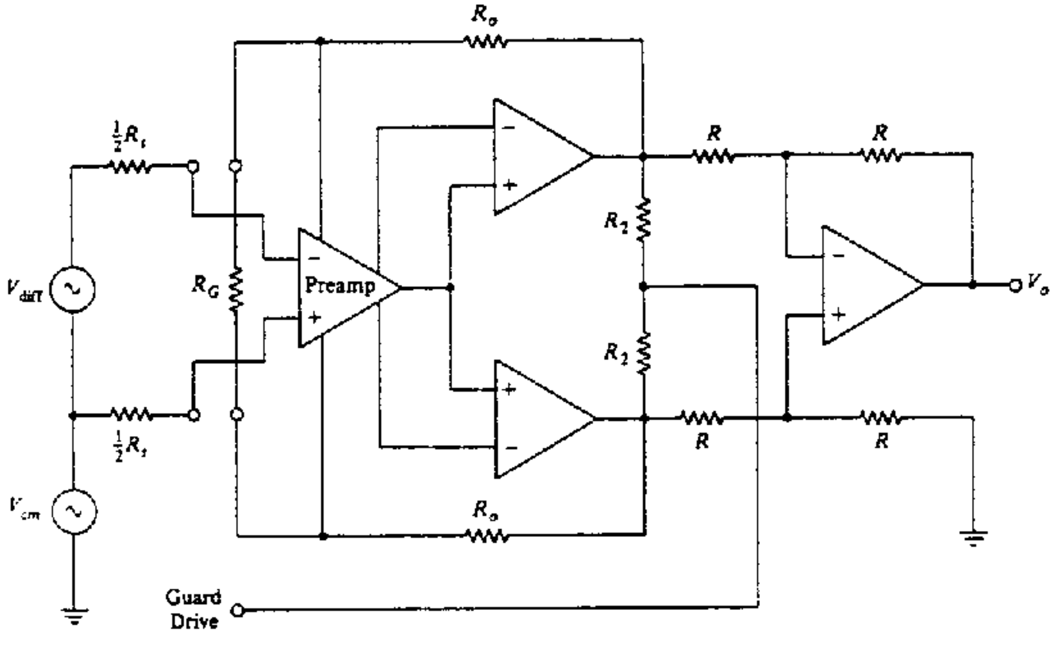


Figure 1 High-performance instrumentation amplifier.

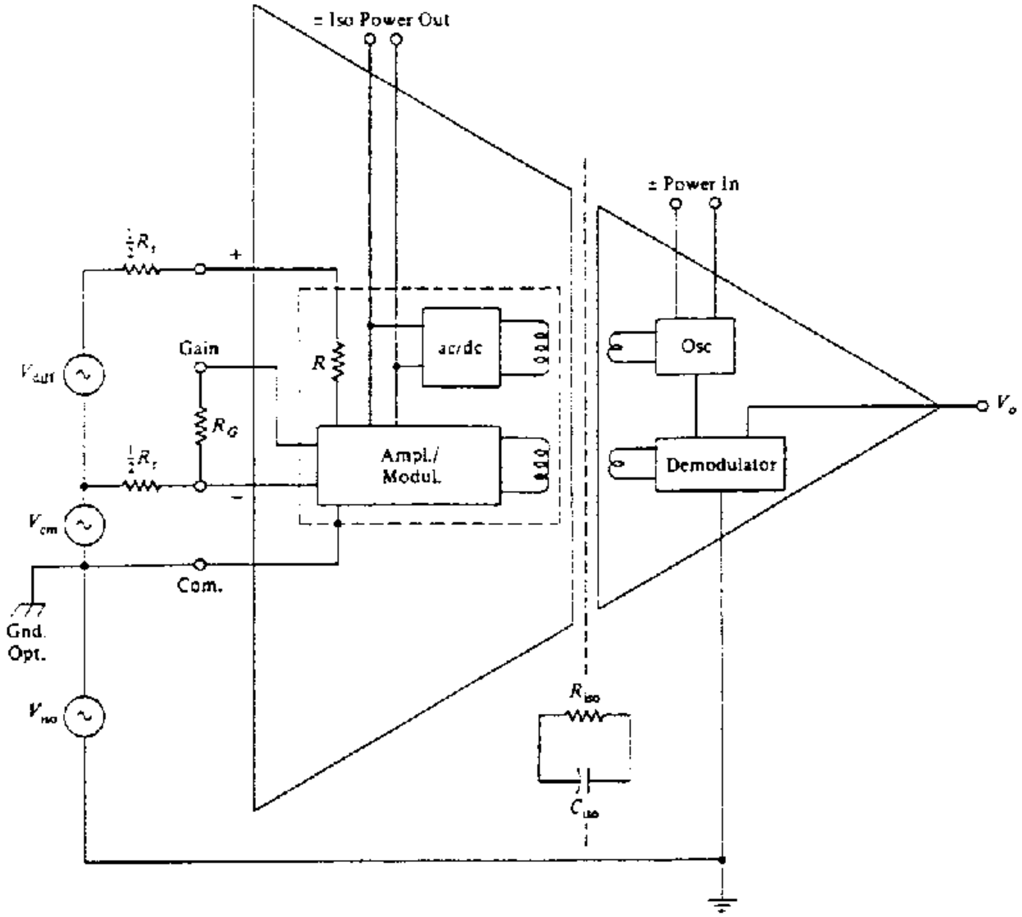


Figure 2 Isolation instrumentation amplifier.

to the output are available at 60 Hz. This dual rejection capability makes possible the accommodation of two sources of interference, V_{cm} and V_{iso}, frequently encountered in sensor applications. The performance of this connection is predicted by Eq. (1), where nonisolated instrumentation amplifiers are absent the V_{iso}/IMRR term:

$$V_0 = A_{v_{diff}} V_{diff}\left(1 + \frac{1}{\text{CMRR}}\frac{V_{cm}}{V_{diff}}\right) + \frac{V_{iso}}{\text{IMRR}}$$

where (1)

$$A_{v_{diff}} = 1 + \frac{2R_0}{R_G}$$

The majority of instrumentation-amplifier applications are at low frequencies because of the limited response of the physical processes from which measurements are typically sought. The selection of an instrumentation amplifier involves the evaluation of amplifier parameters that will minimize errors associated with specific applications under anticipated operating conditions. It is therefore useful to perform an error evaluation in order to identify significant error sources and their contributions in specific applications. Table 1 presents parameter specifications for example amplifiers described in five categories representative of available contemporary devices. These parameters consist of input voltage and current errors, interference rejection and noise specifications, and gain nonlinearity. Table 2 provides a glossary of amplifier parameter definitions.

The instrumentation amplifier error budget tabulation of Table 3 employs the parameters of Table 1 to obtain representative amplifier error, expressed both as an input-amplitude-threshold uncertainty in volts and as a percent of the full-scale output signal. These error totals are combined from the individual device parameter errors by

Table 1 Example Amplifier Parameters

	Subtractor amplifier OP-07	Three-amplifier AD624	Isolation amplifier BB3456	Low-bias amplifier OPA 103	CAZ DC amplifier ICL 7605
V_{os}	$60\,\mu\mathrm{V}$	$25\,\mu\mathrm{V}$	$0.25\,\mu\mathrm{V}$	$100\,\mu\mathrm{V}$	$2\,\mu\mathrm{V}$
dV_{os}/dT	$0.2\,\mu\mathrm{V}/{}^{\circ}\mathrm{C}$	$0.2\,\mu\mathrm{V}/{}^{\circ}\mathrm{C}$	$1\,\mu\mathrm{V}/{}^{\circ}\mathrm{C}$	$1\,\mu\mathrm{V}/{}^{\circ}\mathrm{C}$	$0.05\,\mu\mathrm{V}/{}^{\circ}\mathrm{C}$
I_{os}	$0.8\,\mathrm{nA}$	$10\,\mathrm{nA}$	$10\,\mu\mathrm{A}$	$0.2\,\mathrm{pA}$	$150\,\mathrm{pA}$
dI_{os}/dT	$5\,\mathrm{pA}/{}^{\circ}\mathrm{C}$	$20\,\mathrm{pA}/{}^{\circ}\mathrm{C}$	$0.3\,\mathrm{nA}/{}^{\circ}\mathrm{C}$	$7\%I_{\mathrm{os}}/{}^{\circ}\mathrm{C}$	$1\,\mathrm{pA}/{}^{\circ}\mathrm{C}$
S_{r}	$0.17\,\mathrm{V}/\mu\mathrm{s}$	$5\,\mathrm{V}/\mu\mathrm{s}$	$0.5\,\mu\mathrm{V}/\mu\mathrm{s}$	$1.3\,\mathrm{V}/\mu\mathrm{s}$	$0.5\,\mathrm{V}/\mu\mathrm{s}$
$f_{\mathrm{hi}}@A_{\mathrm{v_{diff}}} = 10^3$	$600\,\mathrm{Hz}$	$25\,\mathrm{kHz}$	$1\,\mathrm{kHz}$	$1\,\mathrm{kHz}$	$10\,\mathrm{Hz}$
CMRR (IMRR)	10^5	10^6	$10^4(10^6)$	10^4	10^5
V_{n} rms	$10\,\mathrm{nV}/\sqrt{\mathrm{Hz}}$	$4\,\mathrm{nV}/\sqrt{\mathrm{Hz}}$	$7\,\mathrm{nV}/\sqrt{\mathrm{Hz}}$	$30\,\mathrm{nV}/\sqrt{\mathrm{Hz}}$	$200\,\mathrm{nV}/\sqrt{Hz}$
$f(A_{\mathrm{V}})$	0.01%	0.001%	0.01%	0.01%	0.01%
dA_{V}/dT	R_{tempco}	$5\,\mathrm{ppm}/{}^{\circ}\mathrm{C}$	$10\,\mathrm{ppm}/{}^{\circ}\mathrm{C}$	R_{tempco}	$15\,\mathrm{ppm}/{}^{\circ}\mathrm{C}$
$R_{i_{\mathrm{cm}}}$	$1.2 \times 10^{11}\,\Omega$	$10^9\,\Omega$	$5 \times 10^9\,\Omega$	$10^{14}\,\Omega$	$10^{12}\,\Omega$
$R_{i_{\mathrm{diff}}}$	$3 \times 10^7\,\Omega$	$10^9\,\Omega$	$10^7\,\Omega$	$10^{13}\,\Omega$	$10^{12}\,\Omega$

$$\varepsilon_{\mathrm{amp1RTI}} = \overline{V_{\mathrm{os}}} + \overline{I_{\mathrm{os}}R_{\mathrm{s}}} + \overline{f(A_{\mathrm{v}})\frac{V_{\mathrm{FS}}}{A_{\mathrm{v_{diff}}}}} + \left[\left(\frac{dV_{\mathrm{os}}}{dT}\,dT\right)^2 + \left(\frac{V_{\mathrm{cm}}}{\mathrm{CMRR\,(IMRR)}}\right)^2 + (6.6\,V_{\mathrm{n}}\sqrt{f_{\mathrm{hi}}})^2\left(\frac{dA_{\mathrm{v}}}{dT}\,dT\,\frac{V_{\mathrm{FS}}}{A_{\mathrm{v_{diff}}}}\right)^2\right]^{1/2} \tag{2}$$

$$\varepsilon_{\mathrm{amp1\%FS}} = \varepsilon_{\mathrm{amp1RTI}}\frac{A_{\mathrm{v_{diff}}}}{V_{\mathrm{FS}}} \times 100\% \tag{3}$$

The barred parameters denote mean values, and the unbarred parameters drift and random values that are combined as the root-sum-square (RSS). Examination of these amplifier error terms discloses that input offset voltage drift with temperature is a consistent error, and the residual V_{cm} error following upgrading by amplifier CMRR is primarily significant with the subtractor instrumentation amplifier. Amplifier referred-to-input internal rms noise V_{n} is converted to peak–peak at a 3.3σ confidence (0.1% error) with multiplication by 6.6 to relate it to the other dc errors in accounting for its crest factor. The effects of both gain nonlinearity and drift with temperature are also referenced to the amplifier input, where the gain nonlinearity represents an average amplitude error over the dynamic range of input signals.

Table 2 Amplifier Parameter Glossary

V_{os}	Input offset voltage
dV_{os}/dT	Input-offset-voltage temperature drift
I_{os}	Input offset current
dI_{os}/dT	Input-offset-current temperature drift
$R_{i_{\mathrm{diff}}}$	Differential input impedance
$R_{i_{\mathrm{cm}}}$	Common-mode input impedance
S_{r}	Slew rate
V_{n}	Input-referred noise voltage
I_{n}	Input-referred noise current
$A_{v_{\mathrm{o}}}$	Open-loop gain
$A_{\mathrm{v_{cm}}}$	Common-mode gain
$A_{\mathrm{v_{diff}}}$	Closed-loop differential gain
$f(A_{\mathrm{v}})$	Gain nonlinearity
dA_{v}/dT	Gain temperature drift
f_{hi}	$-3\,\mathrm{dB}$ bandwidth
CMRR (IMRR)	Common-mode (isolation-mode) numerical rejection ratio

The error budgets for the five instrumentation amplifiers shown in Table 3 include typical input conditions and consistent operating situations so that their performance may be compared. The total errors obtained for all of the amplifiers are similar in magnitude and represent typical in-circuit expectations. Significant to the subtractor amplifier is that V_{cm} must be limited to about 1 V in order to maintain a reasonable total error, whereas the three-amplifier instrumentation amplifier can accommodate V_{cm} values to 10 V at the same or reduced total error.

1.3 INSTRUMENTATION FILTERS

Lowpass filters are frequently required to bandlimit measurement signals in instrumentation applications to achieve frequency-selective operation. The application of an arbitrary signal set to a lowpass filter can result in a significant attenuation of higher frequency components, thereby defining a stopband whose boundary is influenced by the choice of filter cutoff

Table 3 Amplifier Error Budgets ($A_{v_{\text{diff}}} = 10^3$, $V_{\text{FS}} = 10\,\text{V}$, $\Delta T = 20°\text{C}$, $R_{\text{tol}} = 1\%$, $R_{\text{tempco}} = 50\,\text{ppm/°C}$)

		Subtractor	Three-	Isolation	Low-bias	CAZ DC
		amplifier	amplifier	amplifier	amplifier	amplifier
	Amplifier parameters	OP-07	AD624	BB3456	OPA103	ICL7605
Input conditions	V_{cm}	±1 V	± 10 V	±1000 V	±100 mV	±100 mV
	R_{s}	1 kΩ	1 kΩ	1 kΩ	10 MΩ	1 kΩ
Offset group	V_{os}	Nulled	Nulled	Nulled	Nulled	$\bar{2}\,\mu\text{V}$
	$\dfrac{dV_{\text{os}}}{dT}\Delta T$	4 μV	5 μV	20 μV	20 μV	1 μV
	$I_{\text{os}}R_{\text{s}}$	$\overline{0.8}\,\mu\text{V}$	$\overline{10}\,\mu\text{V}$	$\overline{10}\,\mu\text{V}$	$\bar{2}\,\mu\text{V}$	$\overline{0.15}\,\mu\text{V}$
Interference group	$\dfrac{V_{\text{cm}}}{\text{CMRR (IMRR)}_{\text{inckt}}}$	30 μV	10 μV	(10 μV)	12 μV	1 μV
	$6.6V_{\text{n}}\sqrt{f_{\text{h1}}}$	1.6 μV	4.1 μV	1.5 μV	6.2 μV	4.1 μV
Linearity group	$f(A_{\text{v}})\dfrac{V_{\text{FS}}}{A_{v_{\text{diff}}}}$	$\bar{1}\,\mu\text{V}$	$\overline{0.1}\,\mu\text{V}$	$\bar{1}\,\mu\text{V}$	$\bar{1}\,\mu\text{V}$	$\bar{1}\,\mu\text{V}$
	$\dfrac{dA_{\text{v}}}{DT}\Delta T\dfrac{V_{\text{FS}}}{A_{v_{\text{diff}}}}$	10 μV	1 μV	2 μV	10 μV	3 μV
Combined error	$\epsilon_{\text{ampi RTI}}$	34 μV	22 μV	33 μV	29 μV	8 μV
	$\epsilon_{\text{ampi\%FS}}$	0.34%	0.22%	0.33%	0.29%	0.08%

frequency, with the unattenuated frequency components defining the filter passband. For instrumentation purposes, approximating the lowpass filter amplitude responses described in Fig. 3 is beneficial in order to achieve signal bandlimiting with minimum alteration or addition of errors to a passband signal of interest. In fact, preserving the accuracy of measurement signals is of sufficient importance that consideration of filter charcterizations that correspond to well-behaved functions such as Butterworth and Bessel polynomials are especially useful. However, an ideal filter is physically unrealizable because practical filters are represented by ratios of polynomials that cannot possess the discontinuities required for sharply defined filter boundaries.

Figure 3 describes the Butterworth and Bessel lowpass amplitude response where n denotes the filter order or number of poles. Butterworth filters are characterized by a maximally flat amplitude response in the vicinity of dc, which extends toward its $-3\,\text{dB}$ cutoff frequency f_{c} as n increases. Butterworth attenuation is rapid beyond f_{c} as filter order increases with a slightly nonlinear phase response that provides a good approximation to an ideal lowpass filter. Butterworth filters are therefore preferred for bandlimiting measurement signals.

Table 4 provides the capacitor values in farads for unity-gain networks tabulated according to the number of filter poles. Higher-order filters are formed by a cascade of the second- and third-order networks shown. Figure 4 illustrates the design procedure with a 1 kHz-cutoff two-pole Butterworth lowpass filter including frequency and impedance scaling steps. The choice of resistor and capacitor tolerance determines the accuracy of the filter implementation such as its cutoff frequency and passband flatness. Filter response is typically displaced inversely to passive-component tolerance, such as lowering of cutoff frequency for component values on the high side of their tolerance.

Table 5 presents a tabulation of the example filters evaluated for their amplitude errors, by

$$\varepsilon_{\text{filter\%FS}} = \frac{0.1}{f/f_{\text{c}}} \sum_{0}^{f/0.1f_{\text{c}}} (1.0 - A(f)) \times 100\% \qquad (4)$$

over the specified filter passband intervals. One-pole RC and three-pole Bessel filters exhibit comparable errors of 0.3%FS and 0.2%FS, respectively, for signal bandwidths that do not exceed 10% of the filter cutoff frequency. However, most applications are better

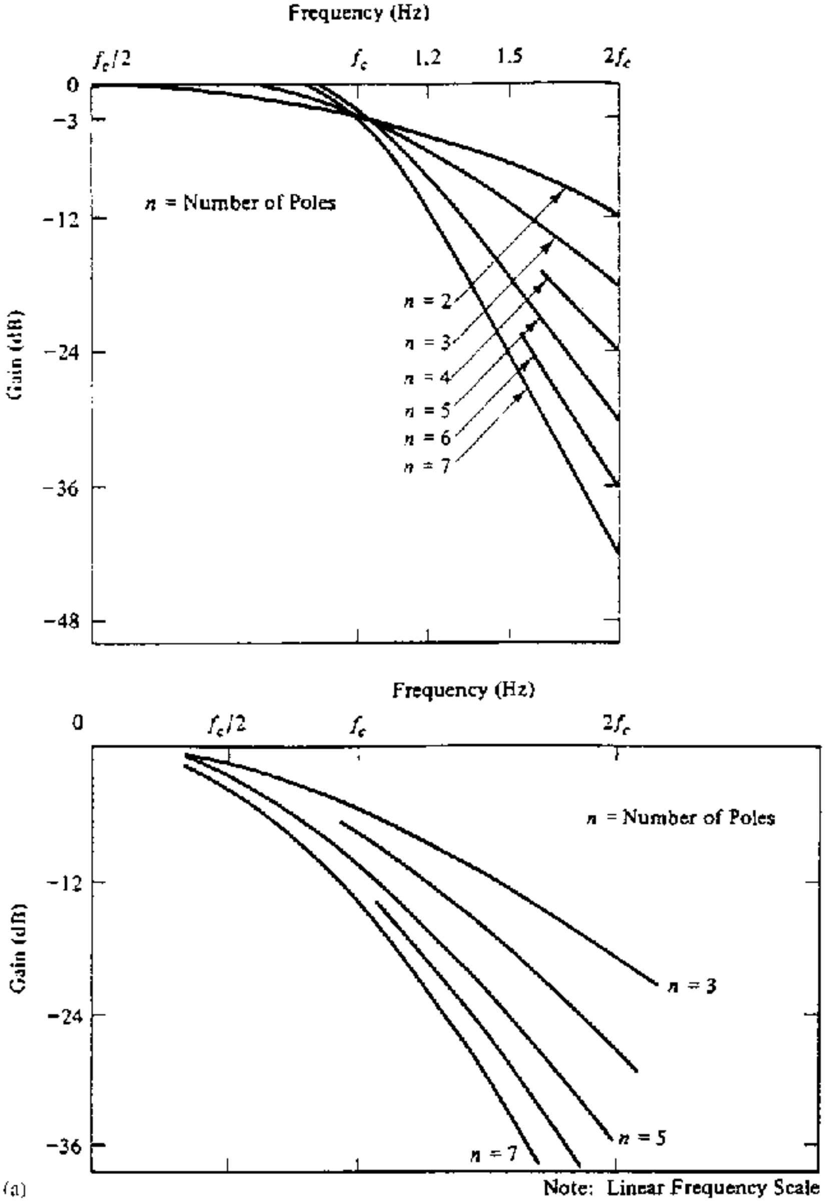
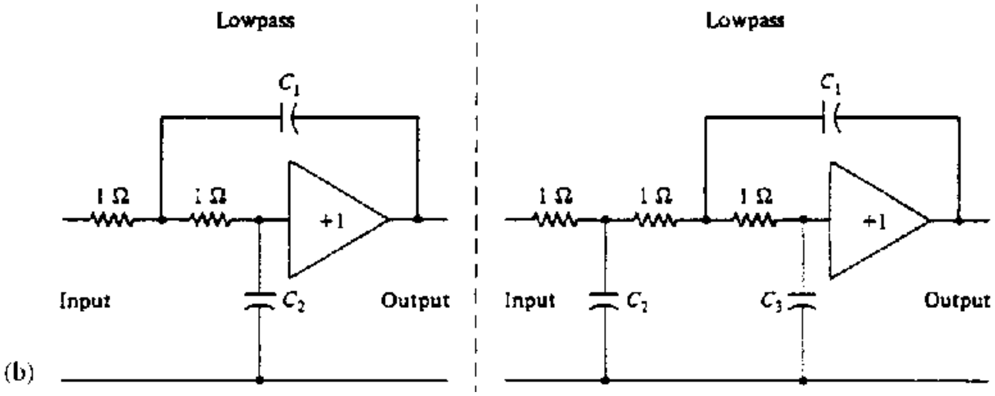


Figure 3 (a) Butterworth and (b) Bessel lowpass filters.

served by the three-pole Butterworth filter which offers an average amplitude error of 0.2%FS for signal passband occupancy up to 50% of the filter cutoff, plus good stopband attenuation. While it may appear inefficient not to utilize a filter passband up to its cutoff frequency, the total bandwidth sacrificed is usually small. Higher filter orders may also be evaluated when greater stopband attenuation is of interest with substitution of their amplitude response $A(f)$ in Eq. (4).

1.4 MEASUREMENT SIGNAL CONDITIONING

Signal conditioning is concerned with upgrading the quality of a signal of interest coincident with measurement acquisition, amplitude scaling, and signal bandlimiting. The unique design requirements of a typical analog data channel, plus economic constraints of achieving necessary performance without incurring the costs of overdesign, benefit from the instrumenta-



tion error analysis presented. Figure 5 describes a basic signal-conditioning structure whose performance is described by the following equations for coherent and random interference:

$$\varepsilon_{\text{coherent}} = \frac{V_{\text{cm}}}{V_{\text{diff}}}\left[\frac{R_{\text{diff}}}{R_{\text{cm}}}\right]^{1/2}\frac{A_{\text{v}_{\text{cm}}}}{A_{\text{v}_{\text{diff}}}}\left[1+\left(\frac{f_{\text{coh}}}{f_{\text{c}}}\right)^{2n}\right]^{-1/2} \times 100\% \qquad (5)$$

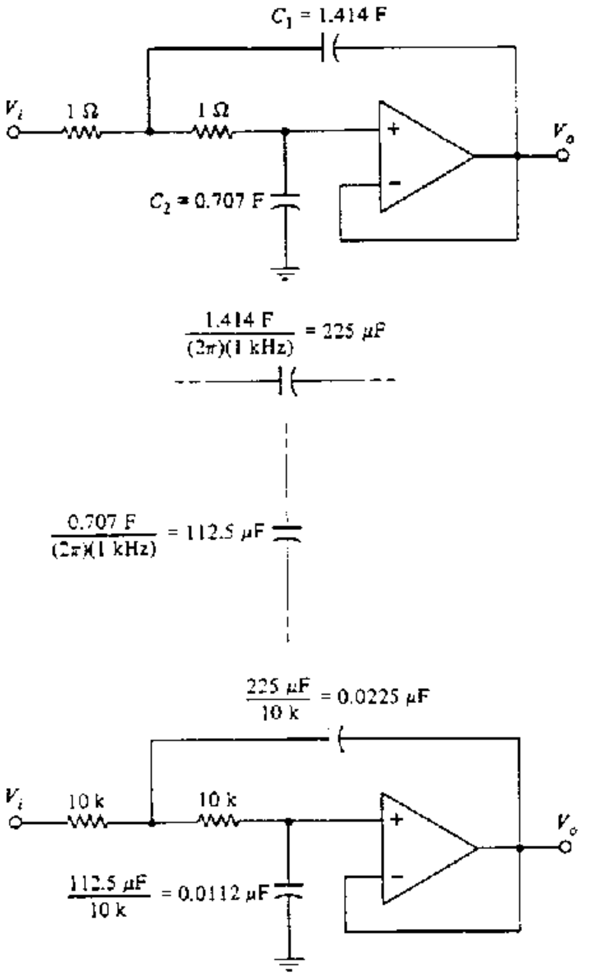
$$\varepsilon_{\text{random}} = \frac{V_{\text{cm}}}{V_{\text{diff}}}\left[\frac{R_{\text{diff}}}{R_{\text{cm}}}\right]^{1/2}\frac{A_{\text{v}_{\text{cm}}}}{A_{\text{v}_{\text{diff}}}}\sqrt{2}\left[\frac{f_{\text{c}}}{f_{\text{hi}}}\right]^{1/2} \times 100\% \qquad (6)$$

$$\varepsilon_{\text{measurement}} = \left[\varepsilon^2_{\text{sensor}} + \varepsilon^2_{\text{amplifier}} + \varepsilon^2_{\text{filter}} + \varepsilon^2_{\text{random}} + \varepsilon^2_{\text{coherent}}\right]^{1/2} \cdot n^{-1/2} \qquad (7)$$

Input signals V_{diff} corrupted by either coherent or random interference V_{cm} can be sufficiently enhanced by the signal-conditioning functions of Eqs. (5) and (6), based upon the selection of amplifier and filter parameters, such that measurement error is principally determined by the hardware device residual errors derived in previous sections. As an option, averaged measurements offer the merit of sensor fusion whereby total measurement error may be further reduced by the

Table 4 Unity-Gain Filter Network Capacitor Values (Farads)

	Butterworth			Bessel		
Poles	C_1	C_2	C_3	C_1	C_2	C_3
2	1.414	0.707		0.907	0.680	
3	3.546	1.392	0.202	1.423	0.988	0.254
4	1.082	0.924		0.735	0.675	
	2.613	0.383		1.012	0.390	
5	1.753	1.354	0.421	1.009	0.871	0.309
	3.235	0.309		1.041	0.310	
6	1.035	0.966		0.635	0.610	
	1.414	0.707		0.723	0.484	
	3.863	0.259		1.073	0.256	
7	1.531	1.336	0.488	0.853	0.779	0.303
	1.604	0.624		0.725	0.415	
	4.493	0.223		1.098	0.216	
8	1.091	0.981		0.567	0.554	
	1.202	0.831		0.609	0.486	
	1.800	0.556		0.726	0.359	
	5.125	0.195		1.116	0.186	


C_1 = 1.414 F
V_i
1 Ω
1 Ω
+
−
V_o
C_2 = 0.707 F
$\frac{1.414\text{ F}}{(2\pi)(1\text{ kHz})} = 225\ \mu F$
$\frac{0.707\text{ F}}{(2\pi)(1\text{ kHz})} = 112.5\ \mu F$
$\frac{225\ \mu F}{10\text{ k}} = 0.0225\ \mu F$
V_i
10 k
10 k
+
−
V_o
$\frac{112.5\ \mu F}{10\text{ k}} = 0.0112\ \mu F$

Component values from Table 4 are normalized to 1 rad/s with resistors taken as 1 Ω and capacitors in farads.

The filter is then frequency-scaled by dividing the capacitor values from the table by the cutoff frequency in radians (2π × 1 kHz).

The filter is finally impedance-scaled by multiplying the resistor values by a convenient value (10 k) and dividing the capacitor values by the same value.

Figure 4 Butterworth lowpass filter design example.

Table 5 Filter Passband Errors

Frequency	Amplitude response $A(f)$			Average filter error $\varepsilon_{\text{filter\%FS}}$		
$\frac{f}{f_c}$	1-pole RC	3-pole Bessel	3-pole Butterworth	1-pole RC	3-pole Bessel	3-pole Butterworth
0.0	1.000	1.000	1.000	0%	0%	0%
0.1	0.997	0.998	1.000	0.3	0.2	0
0.2	0.985	0.988	1.000	0.9	0.7	0
0.3	0.958	0.972	1.000	1.9	1.4	0
0.4	0.928	0.951	0.998	3.3	2.3	0
0.5	0.894	0.924	0.992	4.7	3.3	0.2
0.6	0.857	0.891	0.977	6.3	4.6	0.7
0.7	0.819	0.852	0.946	8.0	6.0	1.4
0.8	0.781	0.808	0.890	9.7	7.7	2.6
0.9	0.743	0.760	0.808	11.5	9.5	4.4
1.0	0.707	0.707	0.707	13.3	11.1	6.9

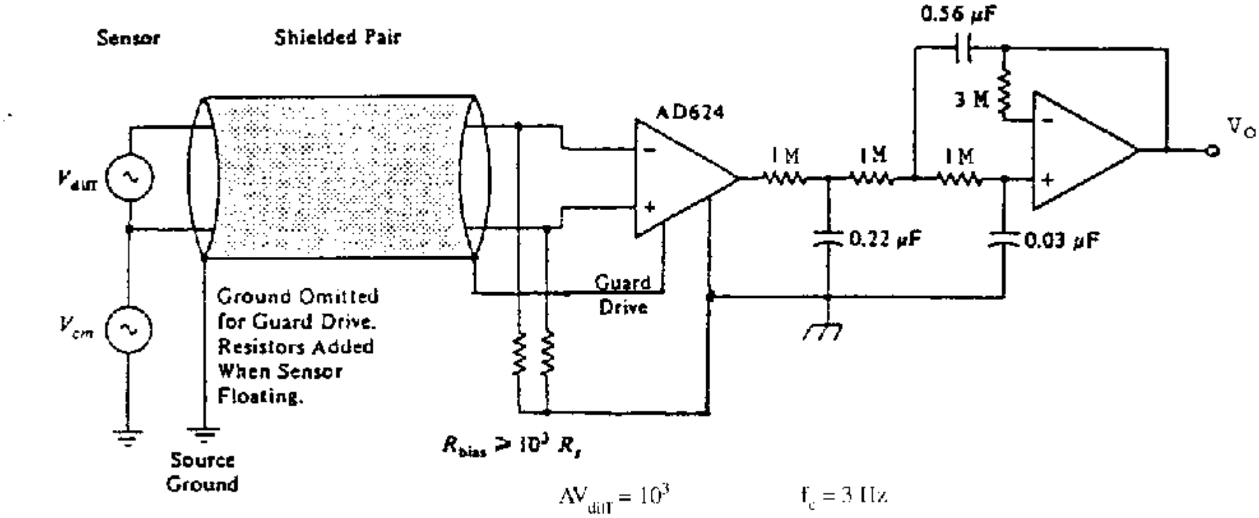


Figure 5 Signal-conditioning channel.

factor $n^{-1/2}$ for n identical signal conditioning channels combined. Note that V_{diff} and V_{cm} may be present in any combination of dc or rms voltage magnitudes.

External interference entering low-level instrumentation circuits frequently is substantial, especially in industrial environments, and techniques for its attenuation or elimination are essential. Noise coupled to signal cables and input power buses, the primary channels of external interference, has as its cause local electric and magnetic field sources. For example, unshielded signal cables will couple 1 mV of interference per kilowatt of 60 Hz load for each lineal foot of cable run on a 1 ft spacing from adjacent power cables. Most interference results from near-field sources, primarily electric fields, whereby the effective attenuation mechanism is reflection by a nonmagnetic material such as copper or aluminum shielding. Both copper-foil and braided-shield twinax signal cables offer attenuation on the order of 90 voltage dB to 60 Hz interference. However, this attenuation decreases by 20 dB per decade of increasing frequency.

For magnetic fields, absorption is the effective attenuation mechanism, and steel or mu-metal shielding is required. Magnetic-field interference is more difficult to shield against than electric-field interference, and shielding effectiveness for a given thickness diminishes with decreasing frequency. For example, steel at 60 Hz provides interference attenuation on the order of 30 voltage dB per 100 mils of thickness. Magnetic shielding of applications is usually implemented by the installation of signal cables in steel conduit of the necessary wall thickness. Additional magnetic-field cancellation can be achieved by periodic transposition of a twisted-pair cable, provided that the signal return current is on one conductor of the pair and not on the shield. Mutual coupling between circuits of a computer input system, resulting from finite signal-path and power-supply impedances, is an additional source of interference. This coupling is minimized by separating analog signal grounds from noisier digital and chassis grounds using separate ground returns, all terminated at a single star-point chassis ground.

Single-point grounds are required below 1 MHz to prevent circulating currents induced by coupling effects. A sensor and its signal cable shield are usually grounded at a single point, either at the sensor or the source of greatest intereference, where provision of the lowest impedance ground is most beneficial. This also provides the input bias current required by all instrumentation amplifiers except isolation types, which furnish their own bias current. For applications where the sensor is floating, a bias-restoration path must be provided for conventional amplifiers. This is achieved with balanced differential R_{bias} resistors of at least 10^3 times the source resistance R_s to minimize sensor loading. Resistors of 50 MΩ, 0.1% tolerance, may be connected between the amplifier input and the single-point ground as shown in Fig. 5.

Consider the following application example. Resistance-thermometer devices (RTDs) offer commercial repeatability to 0.1°C as provided by a 100 Ω platinum RTD. For a 0–100°C measurement range the resistance of this device changes from 100.0 Ω to

138.5 Ω with a nonlinearity of 0.0028°C/°C. A constant-current excitation of 0.26 mA converts this resistance to a voltage signal which may be differentially sensed as V_{diff} from 0 to 10 mV, following a 26 mV amplifier offset adjustment whose output is scaled 0–10 V by an AD624 instrumentation amplifier differential gain of 1000. A three-pole Butterworth lowpass bandlimiting filter is also provided having a 3 Hz cutoff frequency. This signal-conditioning channel is evaluated for RSS measurement error considering an input V_{cm} of up to 10 V rms random and 60 Hz coherent interference. The following results are obtained:

$$\begin{aligned}
\varepsilon_{\text{RTD}} &= \frac{\text{tolerance} + \text{nonlinearity} \times \text{FS}}{\text{FS}} \times 100\% \\
&= \frac{0.1^\circ\text{C} + 0.0028\dfrac{^\circ\text{C}}{^\circ\text{C}} \times 100^\circ\text{C}}{100^\circ\text{C}} \times 100\% \\
&= 0.38\%\text{FS} \\
\varepsilon_{\text{ampl}} &= 0.22\%\text{FS} \qquad (\text{Table 3}) \\
\varepsilon_{\text{filter}} &= 0.20\%\text{FS} \qquad (\text{Table 5}) \\
\varepsilon_{\text{coherent}} &= \frac{10\,\text{V}}{10\,\text{mV}} \left[\frac{10^9\,\Omega}{10^9\,\Omega}\right]^{1/2} \times 10^{-6} \\
&\quad \times \left[1 + \left(\frac{60\,\text{Hz}}{3\,\text{Hz}}\right)^6\right]^{-1/2} \times 100\% \\
&= 1.25 \times 10^{-5}\%\text{FS} \\
\varepsilon_{\text{random}} &= \frac{10\,\text{V}}{10\,\text{mV}} \left[\frac{10^9\,\Omega}{10^9\,\Omega}\right]^{1/2} \times 10^{-6} \\
&\quad \times \sqrt{2}\left[\frac{3\,\text{Hz}}{25\,\text{kHz}}\right]^{1/2} \times 100\% \\
&= 1.41 \times 10^{-3}\%\text{FS} \\
\varepsilon_{\text{measurement}} &= \left[\varepsilon^2_{\text{RTD}} + \varepsilon^2_{\text{ampl}} + \varepsilon^2_{\text{filter}} + \varepsilon_{\text{coherent}} \right. \\
&\quad \left. + \varepsilon^2_{\text{random}}\right]^{1/2} \\
&= 0.48\%\text{FS}
\end{aligned}$$

An RTD sensor error of 0.38%FS is determined for this measurement range. Also considered is a 1.5 Hz signal bandwidth that does not exceed one-half of the filter passband, providing an average filter error contribution of 0.2%FS from Table 5. The representative error of 0.22%FS from Table 3 for the AD624 instrumentation amplifier is employed for this evaluation, and the output signal quality for coherent and random input interference from Eqs. (5) and (6), respectively, is 1.25×10^{-5}%FS and 1.41×10^{-3}%FS. The acquisition of low-level analog signals in the presence of appreciable intereference is a frequent requirement in data acquisition systems. Measurement error of 0.5% or less is shown to be readily available under these circumstances.

1.5 DIGITAL-TO-ANALOG CONVERTERS

Digital-to-analog (D/A) converters, or DACs, provide reconstruction of discrete-time digital signals into continuous-time analog signals for computer interfacing output data recovery purposes such as actuators, displays, and signal synthesizers. These converters are considered prior to analog-to-digital (A/D) converters because some A/D circuits require DACs in their implementation. A D/A converter may be considered a digitally controlled potentiometer that provides an output voltage or current normalized to a full-scale reference value. A descriptive way of indicating the relationship between analog and digital conversion quantities is a graphical representation. Figure 6 describes a 3-bit D/A converter transfer relationship having eight analog output levels ranging between zero and seven-eighths of full scale. Notice that a DAC full-scale digital input code produces an analog output equivalent to FS − 1 LSB. The basic structure of a conventional D/A converter incudes a network of switched current sources having MSB to LSB values according to the resolution to be represented. Each switch closure adds a binary-weighted current increment to the output bus. These current contributions are then summed by a current-to-voltage converter

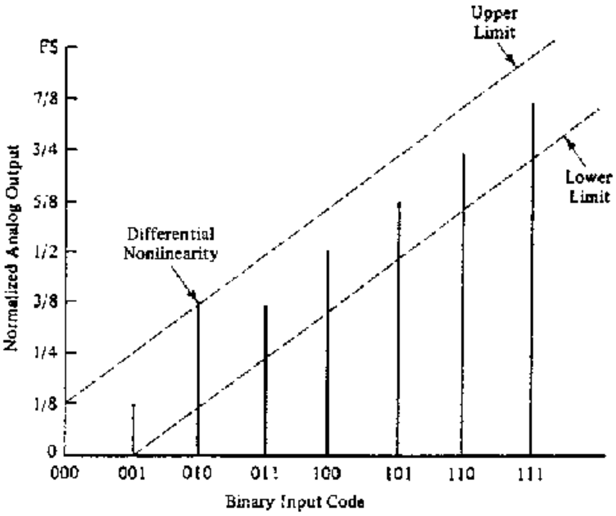


Figure 6 Three-bit D/A converter relationships.

amplifier in a manner appropriate to scale the output signal. Figure 7 illustrates such a structure for a 3-bit DAC with unipolar straight binary coding corresponding to the representation of Fig. 6.

In practice, the realization of the transfer characteristic of a D/A converter is nonideal. With reference to Fig. 6, the zero output may be nonzero because of amplifier offset errors, the total output range from zero to FS – 1 LSB may have an overall increasing or decreasing departure from the true encoded values resulting from gain error, and differences in the height of the output bars may exhibit a curvature owing to converter nonlinearity. Gain and offset errors may be compensated for leaving the residual temperature-drift variations shown in Table 6, where gain temperature coefficient represents the converter voltage reference error. A voltage reference is necessary to establish a basis for the DAC absolute output voltage. The majority of voltage references utilize the bandgap principle, whereby the $V_{\rm be}$ of a silicon transistor has a negative temperature coefficient of $-2.5\,{\rm mV}/^\circ{\rm C}$ that can be extrapolated to approximately 1.2 V at absolute zero (the bandgap voltage of silicon).

Converter nonlinearity is minimized through precision components, because it is essentially distributed throughout the converter network and cannot be eliminated by adjustment as with gain and offset error. Differential nonlinearity and its variation with temperature are prominent in data converters in that they describe the difference between the true and actual outputs for each of the 1-LSB code changes. A DAC with a 2-LSB output change for a 1-LSB input code change exhibits 1 LSB of differential nonlinearity as shown. Nonlinearities greater than 1 LSB make the converter output no longer single valued, in which case it is said to be nonmonotonic and to have missing codes.

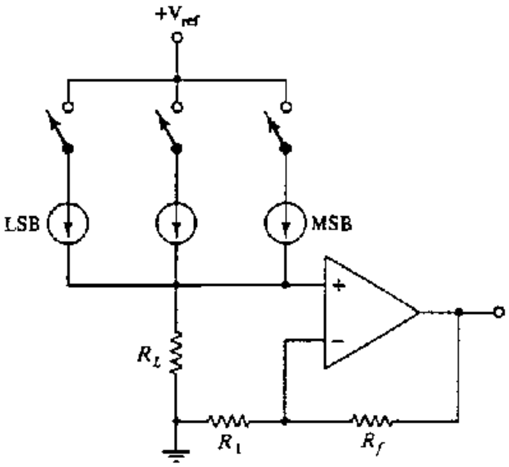


Figure 7 Three-bit D/A converter circuit.

Table 6 Representative 12-Bit D/A Errors

Differential nonlinearity (1/2 LSB)	0.012%
Linearity temp. coeff. (2 ppm/°C)(20°C)	0.004
Gain temp. coeff. (20 ppm/°C)(20°C)	0.040
Offset temp. coeff. (5 ppm/°C)(20°C)	0.010
$\epsilon_{\rm D/A}$	0.05%FS

1.6 ANALOG-TO-DIGITAL CONVERTERS

The conversion of continuous-time analog signals to discrete-time digital signals is fundamental to obtaining a representative set of numbers which can be used by a digital computer. The three functions of sampling, quantizing, and encoding are involved in this process and implemented by all A/D converters as illustrated by Fig. 8. We are concerned here with A/D converter devices and their functional operations as we were with the previously described complementary D/A converter devices. In practice one conversion is performed each period T, the inverse of sample rate $f_{\rm s}$, whereby a numerical value derived from the converter quantizing levels is translated to an appropriate output code. The graph of Fig. 9 describes A/D converter input–output relationships and quantization error for prevailing uniform quantization, where each of the levels q is of spacing $2^{-n}(1 - {\rm LSB})$ for a converter having an n-bit binary output wordlength. Note that the maximum output code does not correspond to a full-scale input value, but instead to $(1 - 2^{-n})$FS because there exist only $(2^n - 1)$ coding points as shown in Fig. 9.

Quantization of a sampled analog waveform involves the assignment of a finite number of amplitude levels corresponding to discrete values of input signal $V_{\rm i}$ between 0 and $V_{\rm FS}$. The uniformly spaced quantization intervals 2^{-n} represent the resolution limit for an n-bit converter, which may also be expressed as the quantizing interval q equal to $V_{\rm FS}/(2^n - 1)$V. These relationships are described by Table 7. It is useful to match A/D converter wordlength in bits to a required analog input signal span to be represented digitally. For example, a 10 mV-to-10 V span (0.1%–100%) requires a minimum converter wordlength n of 10 bits. It will be shown that additional considerations are involved in the conversion

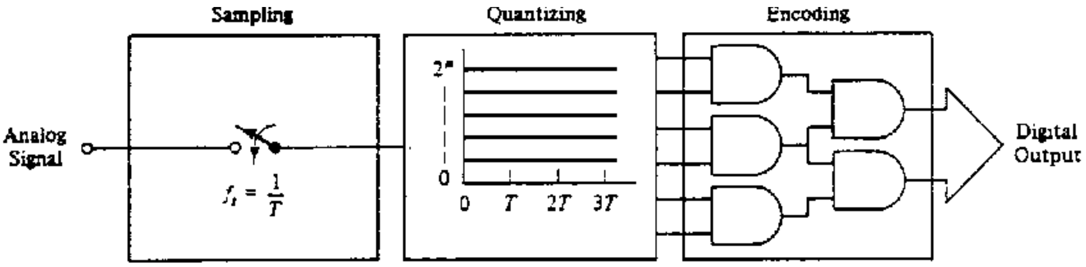


Figure 8 Analog-to-digital converter functions.

of an input signal to an n-bit accuracy other than the choice of A/D converter wordlength, where the dynamic range of a digitized signal may be represented by an n-bit wordlength without achieving n-bit data accuracy. However, the choice of a long wordlength A/D converter will beneficially minimize both quantization noise and A/D device error and provide increased converter linearity.

The mechanization of all A/D converters is by either the integrating method or the voltage-comparison method. The successive-approximation voltage-comparison technique is the most widely utilized A/D converter for computer interfacing primarily because its constant conversion period T is independent of input signal amplitude, making its timing requirements conveniently uniform. This feedback converter operates by comparing the output of an internal D/A converter with the input signal at a comparator, where each bit of the converter wordlength n is sequentially tested during n equal time subperiods to develop an output code representative of the input signal amplitude. The conversion period T and sample/hold (S/H) acquisition time t_{acq} determine the maximum data conversion throughput rate $f_s \leq (T + t_{\mathrm{acq}})^{-1}$ shown in Fig. 10. Figure 11 describes the operation of a successive-approximation converter. The internal elements are represented in the 12-bit converter errors of Table 8, where differential nonlinearity and gain temperature coefficient are derived from the internal D/A converter and its reference, and quantizing noise as the 1/2 LSB uncertainty in the conversion process. Linearity temperature coefficient and offset terms are attributable to the comparator, and long-term change is due to shifts occurring from component aging. This evaluation reveals a two-binary-bit derating in realizable accuracy below the converter wordlength. High-speed, successive-approximation A/D converters require high-gain fast comparators, particularly for accurate conversion at extended wordlengths. The comparator is therefore critical to converter accuracy, where its performance is ultimately limited by the influence of internal and external noise effects on its decision threshold.

Integrating converters provide noise rejection for the input signal at an attenuation rate of $-20\,\mathrm{dB}/$decade of frequency. Notice that this noise improvement capability requires integration of the signal plus noise during the conversion period, and therefore is not provided when a sample-hold device precedes the converter. A conversion period of 16 2/3 ms will provide a useful null to the conversion of 60 Hz interference, for example. Only voltage-comparison converters actually need a S/H to satisfy the A/D-conversion process requirement for a constant input signal.

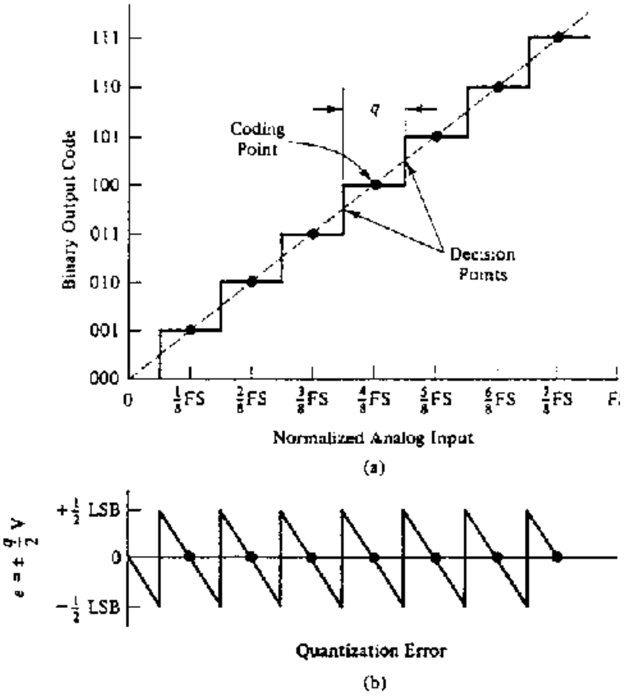


Figure 9 Three-bit A/D converter relationships.

Table 7 Decimal Equivalents of Binary Quantities

Bits, n	Levels, 2^n	LSB weight, 2^{-n}	$\varepsilon_{\%\mathrm{FS}}$(1 LSB)
1	2	0.5	50.0
2	4	0.25	25.0
3	8	0.125	12.5
4	16	0.0625	6.25
5	32	0.03125	3.12
6	64	0.015625	1.56
7	128	0.0078125	0.78
8	256	0.00390625	0.39
9	512	0.001953125	0.19
10	1,024	0.0009763625	0.097
11	2,048	0.00048828125	0.049
12	4,096	0.000244140625	0.024
13	8,192	0.0001220703125	0.012
14	16,384	0.00006103515625	0.006
15	32,768	0.000030517578125	0.003
16	65,536	0.0000152587890625	0.0015
17	131,072	0.00000762939453125	0.0008
18	262,144	0.000003814697265625	0.0004
19	524,288	0.0000019073486328125	0.0002
20	1,048,576	0.00000095367431640625	0.0001

Dual-slope integrating converters perform A/D conversion by the indirect method of converting an input signal to a representative time period that is totaled by a counter. Features of this conversion technique include self-calibration that makes it immune to component temperature drift, use of inexpensive components in its mechanization, and the capability for multiphasic integration yielding improved resolution of the zero endpoint as shown in Fig. 12. Operation occurs in three phases. The first is the autozero phase that stores the converter analog offsets on the integrator with the input grounded. During the second phase, the input signal is integrated for a constant time T_1. In the final phase, the input is connected to a reference of opposite polarity. Integration then proceeds to zero during a variable time T_2 while clock pulses are totaled to represent the amplitude of the input signal. The representative errors of Table 8 show slightly better performance for dual-slope compared with successive-approximation converters, but their speed differences belie this advantage. The self-calibration, variable conversion time, and lower cost features of dual-slope converters make them especially attractive for instrumentation applications.

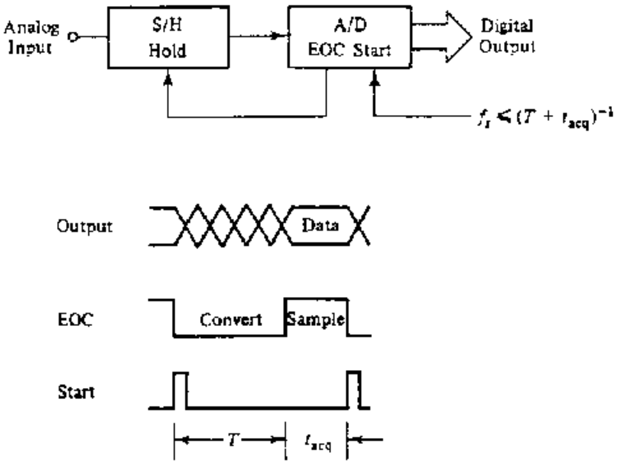


Figure 10 Timing relationships for S/H–A/D conversion.

Sample/hold component errors consist of contributions from acquisition time, capacitor charge droop and dielectric absorption, offset voltage drift, and hold-mode feedthrough. A representative S/H error budget is shown in Table 9. Hold-capacitor voltage droop dV/dt is determined primarily by the output amplifier bias-current requirements. Capacitor values in the 0.01– 0.001 μF range typically provide a balance for reasonable droop and acquisition errors. Capacitor dielectric absorption error is evident as voltage creep following repetitive changes in capacitor charging

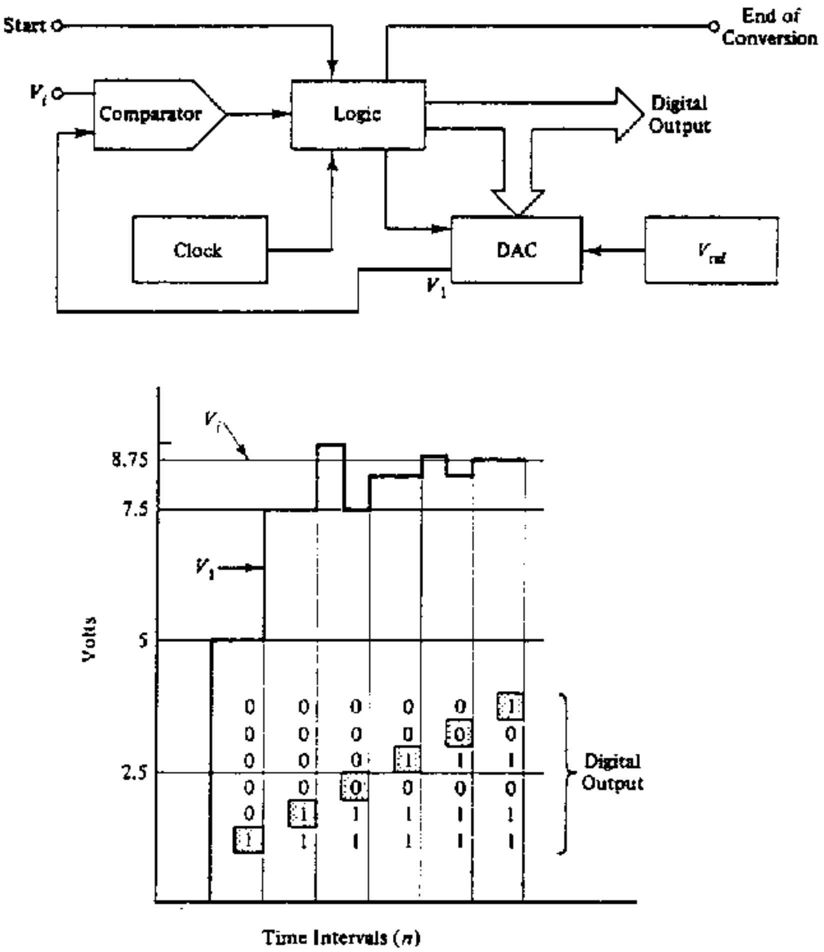


Figure 11 Successive-approximation A/D conversion.

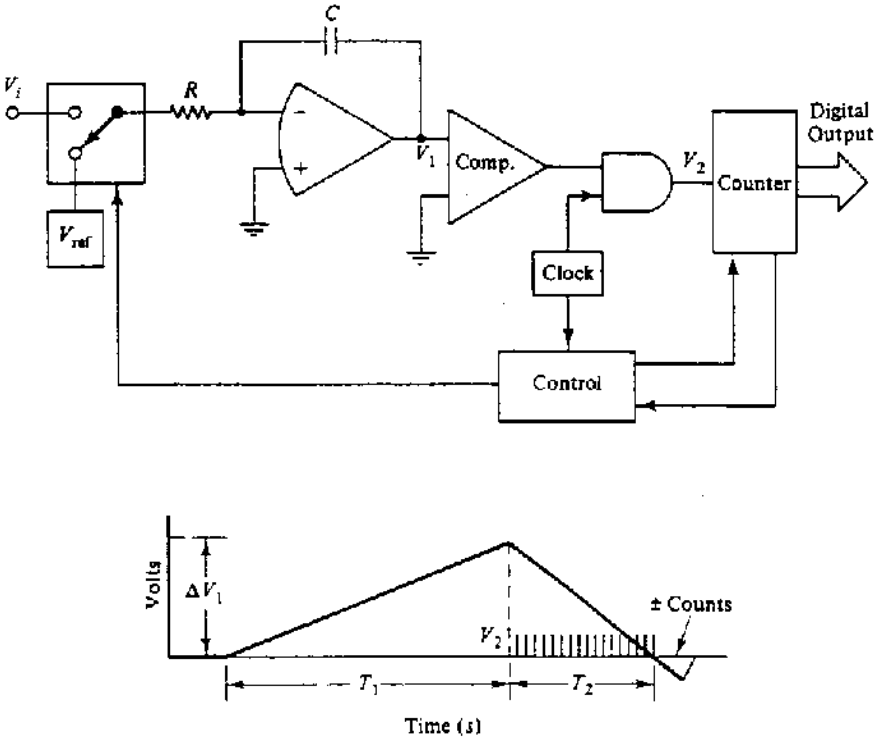
Table 8 Representative 12-Bit A/D Errors

12-bit successive approximation	
Differential nonlinearity (1/2 LSB)	$\overline{0.012\%}$
Quantizing uncertainty (1/2 LSB)	0.012
Linearity temp. coeff. (2 ppm/°C)(20°C)	0.004
Gain temp. coeff. (20 ppm/°C)(20°C)	0.040
Offset (5 ppm/°C)(20°C)	0.010
Long-term change	0.050
$\epsilon_{A/D}$	0.080%FS
12-bit dual slope	
Differential nonlinearity (1/2 LSB)	$\overline{0.012\%}$
Quantizing uncertainty (1/2 LSB)	0.012
Gain temp. coeff. (25 ppm/°C)(20°C)	0.050
Offset temp.coeff. (2 ppm/°C)(20°C)	0.004
$\epsilon_{A/D}$	0.063%FS

resulting from incomplete dielectric repolarization. Polycarbonate capacitors exhibit 50 ppm dielectric absorption, polystyrene 20 ppm, and Teflon 10 ppm. Hold-jump error is attributable to that fraction of the logic signal transferred by the capacitance of the switch at turnoff. Feedthrough is specified for the hold mode as the percentage of an input sinusoidal signal that appears at the output.

1.7 SIGNAL SAMPLING AND RECONSTRUCTION

The provisions of discrete-time systems include the existence of a minimum sample rate for which theoretically exact signal reconstruction is possible from a sampled sequence. This provision is significant in that signal sampling and recovery are considered



$$\Delta V_1 = \frac{1}{RC} \cdot V_i \cdot T_{1_{\text{constant}}}$$

$$= \frac{1}{RC} \cdot V_{\text{ref}} \cdot T_{2_{\text{variable}}}$$

$$T_2 = \frac{V_i \cdot T_1}{V_{\text{ref}}}$$

Figure 12 Dual-slope A/D conversion.

simultaneously, correctly implying that the design of real-time data conversion and recovery systems should also be considered jointly. The following interpolation formula analytically describes this approximation $\hat{x}(t)$ of a continuous time signal $x(t)$ with a finite number of samples from the sequence $x(nT)$ as illustrated by Fig. 13:

$$
\begin{aligned}
\hat{x}(t) &= F^{-1}\{f[x(nT)] * H(f)\} \qquad (8)\\
&= \sum_{n=-x}^{x} \left(T \int_{-\text{BW}}^{\text{BW}} x(nT)\, e^{-j2\pi f nT} \right) e^{j2\pi ft}\, df\\
&= T \sum_{n=-x}^{x} x(nT) \frac{e^{j2\pi \text{BW}(t-nT)} - e^{-j2\pi \text{BW}(t-nT)}}{j2\pi(t-nT)}\\
&= 2T\text{BW} \sum_{n=-x}^{x} x(nT) \frac{\sin 2\pi \text{BW}(t-nT)}{2\pi \text{BW}(t-nT)}
\end{aligned}
$$

$\hat{x}(t)$ is obtained from the inverse Fourier transform of the input sequence and a frequency-domain convolution with an ideal interpolation function $H(f)$, result-

Table 9 Representative Sample/Hold Errors

Acquisition error	0.01%
Droop (25 μV/μs)(2 μs hold) in $10V_{\text{FS}}$	0.0005
Dielectric absorption	0.005
Offset (50 μV/°C)(20°C) in $10V_{\text{FS}}$	0.014
Hold-jump error	0.001
Feedthrough	0.005
$\epsilon_{\text{S/H}}$	0.02%FS

Table 10 Signal Interpolation Functions

Interpolator	$A(f)$	Intersample error (%FS)
D/A	$\mathrm{sinc}(f/f_s)$	$\left[\dfrac{V_{\mathrm{FS}}^2}{1.644V_{\mathrm{S}}^2\left[\mathrm{sinc}^2\left(1-\frac{\mathrm{BW}}{f_s}\right)+\mathrm{sinc}^2\left(1+\frac{\mathrm{BW}}{f_s}\right)\right]}\right]^{-1/2} \times 100\%$
D/A + 1-pole RC	$[1+(f/f_c)^2]^{-1/2}$	$\left[\dfrac{V_{\mathrm{FS}}^2}{V_S^2\left\{\left[\mathrm{sinc}^2\left(1-\frac{\mathrm{BW}}{f_s}\right)\right]\left[1+\left(\frac{f_s-\mathrm{BW}}{f_c}\right)^{2n}\right]^{-1}+\left[\mathrm{sinc}^2\left(1+\frac{\mathrm{BW}}{f_s}\right)\right]\left[1+\left(\frac{f_s+\mathrm{BW}}{f_c}\right)^{2n}\right]^{-1}\right.}\right]^{-1/2} \times 100\%$
D/A + Butterworth n-pole lowpass	$[1+(f/f_c)^{2n}]^{-1/2}$	($f_s \pm$ BW substituted for f in $A(f)$

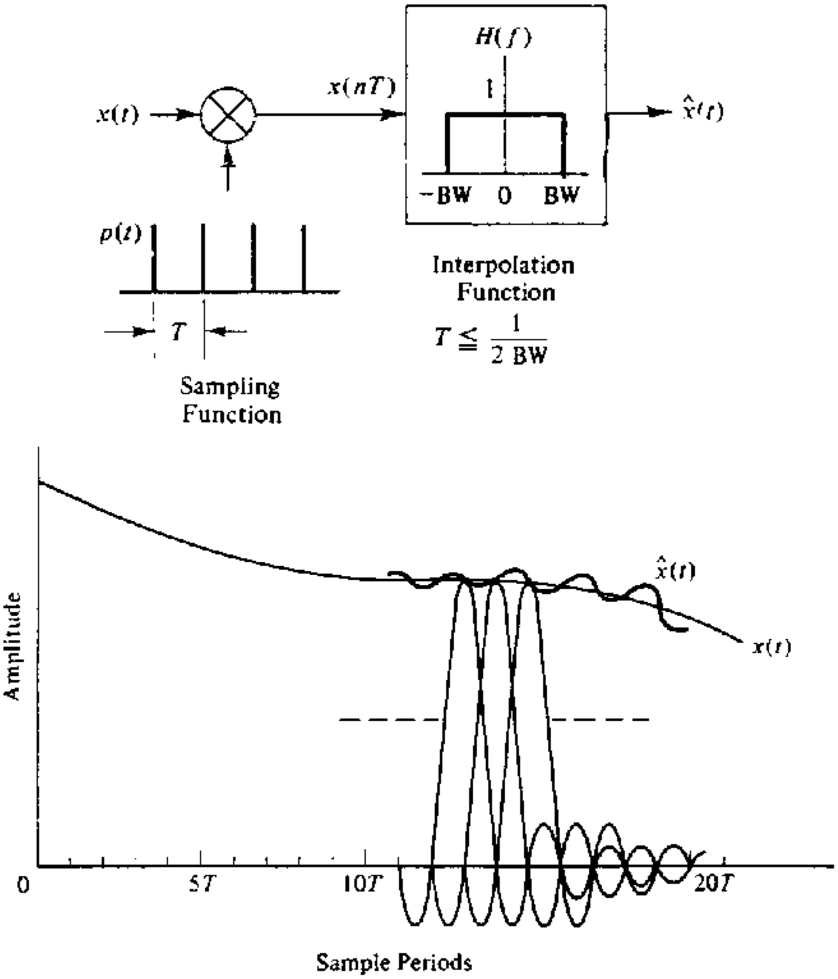


Figure 13 Ideal signal sampling and recovery.

ing in a time-domain sinc amplitude response owing to the rectangular characteristic of $H(f)$. Due to the orthogonal behavior of Eq. (8), however, only one nonzero term is provided at each sampling instant by a summation of weighted samples. Contributions of samples other than the ones in the immediate neighborhood of a specific sample, therefore, diminish rapidly because the amplitude response of $H(f)$ tends to decrease. Consequently, the interpolation formula provides a useful relationship for describing recovered bandlimited sampled-data signals of bandwidth BW with the sampling period T chosen sufficiently small to prevent signal aliasing where sampling frequency $f_s = 1/T$.

It is important to note that an ideal interpolation function $H(f)$ utilizes both phase and amplitude information in reconstructing the recovered signal $\hat{x}(t)$, and is therefore more efficient than conventional bandlimiting functions. However, this ideal interpolation function cannot be physically realized because its impulse response is noncausal, requiring an output that anticipates its input. As a result, practical interpolators for signal recovery utilize amplitude information that can be made efficient, although not optimum, by achieving appropriate weighting of the reconstructed signal.

Of key interest is to what accuracy can an original continuous signal be reconstructed from its sampled values.

It can be appreciated that the determination of sample rate in discrete-time systems and the accuracy with which digitized signals may be recovered requires the simultaneous consideration of data conversion and reconstruction parameters to achieve an efficient allocation of system resources. Signal to mean-squared-error relationships accordingly represent sampled and recovered data intersample error for practical interpolar functions in Table 10. Consequently, an intersample error of interest may be achieved by substitution of a selected interpolator function and solving for the sampling frequency f_s by iteration, where asymptotic convergence to the performance provided by ideal interpolation is obtained with higher-order practical interpolators.

The recovery of a continuous analog signal from a discrete signal is required in many applications. Providing output signals for actuators in digital control systems, signal recovery for sensor acquisition systems, and reconstructing data in imaging systems are but a few examples. Signal recovery may be viewed from either time-domain or frequency-domain perspectives. In time-domain terms, recovery is similar to interpolation procedures in numerical analysis with the criterion being the generation of a locus that reconstructs the true signal by some method of connecting the discrete data samples. In the frequency domain, signal recovery involves bandlimiting by a linear filter to attenuate the repetitive sampled-data spectra above baseband in achieving an accurate replica of the true signal.

A common signal recovery technique is to follow a D/A converter by an active lowpass filter to achieve an output signal quality of interest, accountable by the convergence of the sampled data and its true signal representation. Many signal power spectra have long time-average properties such that linear filters are especially effective in minimizing intersample error. Sampled-data signals may also be applied to control actuator elements whose intrinsic bandlimited amplitude response assist with signal reconstruction. These terminating elements often may be characterized by a single-pole RC response as illustrated in the following section.

An independent consideration associated with the sampling operation is the attenuation impressed upon the signal spectrum owing to the duration of the sampled-signal representation $x(nT)$. A useful criterion is to consider the average baseband amplitude error between dc and the full signal bandwidth BW expressed as a percentage of departure from full-scale response. This average sinc amplitude error is expressed by

$$\varepsilon_{\text{sinc\%FS}} = \frac{1}{2}\left(1 - \frac{\sin(\pi \text{BW} T)}{\pi \text{BW} T}\right) \times 100\% \tag{9}$$

and can be reduced in a specific application when it is excessive by increasing the sampling rate f_s. This is frequently referred to as oversampling.

A data-conversion system example is provided by a simplified three-digit digital dc voltmeter (Fig. 14). A dual-slope A/D conversion period T of 16 2/3 ms provides a null to potential 60 Hz interference, which is essential for industrial and field use, owing to sinc nulls occurring at multiples of the integration period T. A 12-bit converter is employed to achieve a nominal data converter error, while only 10 bits are required for display excitation considering 3.33 binary bits per decimal digit. The sampled-signal error evaluation considers an input-signal rate of change up to an equivalent bandwidth of 0.01 Hz, corresponding to an f_s/BW of 6000, and an intersample error determined by zero-order-hold (ZOH) data, where V_s equals V_{FS}:

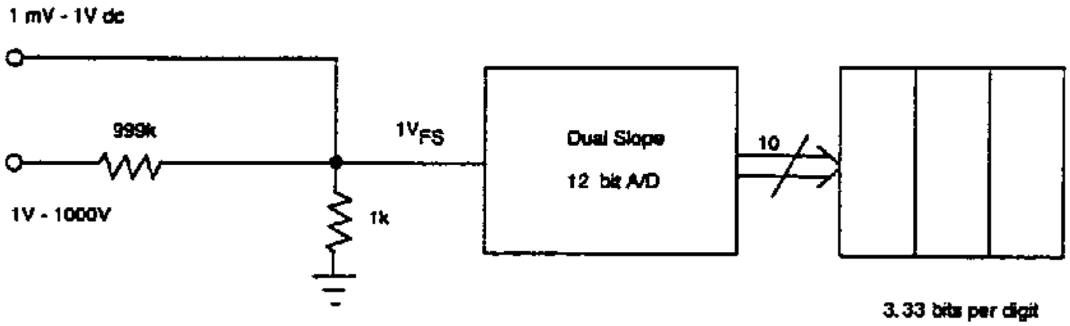


Figure 14 Three-digit digital voltmeter example.

$$\varepsilon_{\text{intersample}} = \left[\frac{V_{\text{FS}}^2}{1.644 \times V_{\text{S}}^2\left[\text{sinc}^2\left(1-\frac{\text{BW}}{f_s}\right)+\text{sinc}^2\left(\frac{1+\text{BW}}{f_s}\right)\right]}\right]^{-1/2} \times 100\%$$

$$= \left[\frac{1}{1.644\left\{\left[\frac{\sin\left(\pi\left(1-\frac{1}{6000}\right)\right)}{\pi\left(1-\frac{1}{6000}\right)}\right]^2+\left[\frac{\sin\left(\pi\left(1+\frac{1}{6000}\right)\right)}{\pi\left(1+\frac{1}{6000}\right)}\right]^2\right\}}\right]^{-1/2}$$

$$\times\ 100\%$$

$$= 0.033\%\text{FS}$$

$$\varepsilon_{\text{sinc}} = \frac{1}{2}\left(1-\frac{\sin(\pi(\text{BW})/f_s)}{\pi(\text{BW})/f_s}\right) \times 100\%$$

$$= \frac{1}{2}\left(1-\frac{\sin(\pi/6000)}{\pi/6000}\right) \times 100\%$$

$$= 0.000001\%\text{FS}$$

$$\varepsilon_{\text{A/D}} = 0.063\%\text{FS} \qquad \text{(Table 8)}$$

$$\varepsilon_{\substack{\text{sampled}\\ \text{signal}}} = \left[\varepsilon_{\text{intersample}}^2 + \varepsilon_{\text{sinc}}^2 + \varepsilon_{\text{A/D}}\right]^{1/2}$$

$$= 0.07/\%\text{FS}$$

The RSS error of 0.07/% exceeds 10 bits required for a three-digit display with reference to Table 7.

1.8 DIGITAL CONTROL SYSTEM ERROR

The design of discrete-time control loops can benefit from an understanding of the interaction of sample rate and intersample error and their effect on system performance. The choice of sample rate influences stability through positioning of the closed-loop transfer function pole locations in the z-domain with respect to the origin. Separately, the decrease in intersample error from output interpolation provided by the closed-loop bandwidth of the control system reduces the uncertainty of the controlled variable. Since the choice of sample rate also influences intersample error, an analysis of a digital control loop is instructive to illustrate these interrelationships.

Figure 15 describes an elementary discrete-time control loop with a first-order process and unity feedback. All of the process, controller, and actuator gains are represented by the single constant K with the compensator presently that of proportional control. The D/A converter represents the influence of the sampling period T, which is z-transformed in the closed-loop transfer function of the following equations:

$$C(Z) = \frac{k(1-e^{-T})}{Z-e^{-T}(1+K)+K}R(z)$$

$$\text{(transfer function)} \tag{10}$$

$$= \frac{k(1-e^{-0.1})}{Z-e^{-0.1}(2)+1}\,\frac{Z}{Z-1}$$

$$\text{(unit step input } K=1,\ T=0.1\text{ sec)} \tag{11}$$

$$= \frac{-0.5Z}{Z-0.8}+\frac{0.5Z}{Z-1}$$

$$\text{(by partial fractions)} \tag{12}$$

$$C(n) = [(-0.5)(0.8)^n + (0.5)(1)^n]U(n)$$

$$\text{(inverse transforming)} \tag{13}$$

$$= 0.50 \text{ final value } (n \text{ large})$$

The denominator of the transfer function defines the influence of the gain K and sampling period T on the pole positions, and hence stability. Values are substituted to determine the boundary between stable and unstable regions for control loop performance evaluated at the z-plane unit circle stability boundary of $z = 1$. This relationship is plotted in Fig. 15.

Calculation of the −3dB closed-loop bandwidth BW for both first- and second-order processes is necessary for the determination of interpolated intersample error of the controlled-variable C. For first-order processes, the closed-loop BW is obtained in terms of the rise time t_r between the 10% and 90% points of the controlled-variable amplitude response to a step input

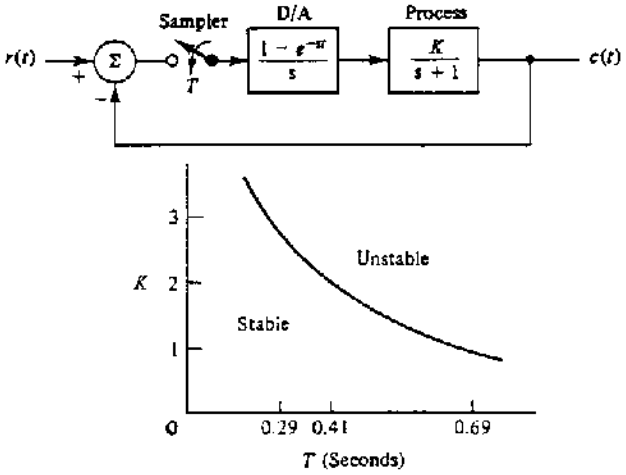


Figure 15 Elementary digital control loop.

as defined in Table 11. The constant 0.35 defines the ratio of 2.2 time constants, required for the response to rise between 10% and 90% of the final value, to 2π radians for normalization to frequency in Hertz. Validity for digital control loops is achieved by acquiring t_r from a discrete-time plot of the controlled-variable amplitude response. Table 11 also defines the bandwidth for a second-order process which is calculated directly with knowledge of the natural frequency, sampling period, and damping ratio.

In the interest of minimizing sensor-to-actuator variability in control systems the error of a controlled variable of interest is divisible into an analog measurement function and digital conversion and interpolation functions. Instrumentation error models provide a unified basis for combining contributions from individual devices. The previous temperature measurement signal conditioning associated with Fig. 5 is included in this temperature control loop, shown by Fig. 16, with the averaging of two identical 0.48%FS error measurement channels to effectively reduce that error by $n^{-1/2}$ or $2^{-1/2}$, from Eq. (7), yielding 0.34%FS. This provides repeatable temperature measurements to within an uncertainty of 0.34°C, and a resolution of 0.024°C provided by the 12-bit digital data bus wordlength.

The closed-loop bandwidth is evaluated at conservative gain and sampling period values of $K = 1$ and $T = 0.1$ sec ($f_s = 10\,\text{Hz}$), respectively, for unit-step excitation at $r(t)$. The rise time of the controlled variable is evaluated from a discrete-time plot of $C(n)$ to be 1.1 sec. Accordingly, the closed-loop bandwidth is found from Table 11 to be 0.318 Hz. The intersample error of the controlled variable is then determined to be 0.143%FS with substitution of this bandwidth value and the sampling period $T(T = 1/f_s)$ into the one-pole process-equivalent interpolation function obtained from Table 10. These functions include provisions for scaling signal amplitudes of less than full scale, but are taken as V_S equalling V_{FS} for this example. Intersample error is therefore found to be directly proportional to process closed-loop bandwidth and inversely proportional to sampling rate.

The calculations are as follows:

$$\varepsilon_{\text{measurement}} = 0.48\%\S \qquad \text{(Fig. 5)}$$
$$\varepsilon_{\text{S/H}} = 0.02\%\S \qquad \text{(Table 9)}$$
$$\varepsilon_{\text{A/D}} = 0.08\%\S \qquad \text{(Table 8)}$$
$$\varepsilon_{\text{D/A}} = 0.05\%\S \qquad \text{(Table 6)}$$
$$\varepsilon_{\text{sinc}} = \frac{1}{2}\left(\frac{1-\sin\pi 0.318\,\text{Hz}/10\,\text{Hz}}{\pi(0.318''/10'')}\right)\times 100\% = 0.08\%\text{FS}$$

$$\varepsilon_{\text{intersample}} = \left[\frac{1}{\left[\dfrac{\sin\left(\pi\left(1-\dfrac{0.318\,\text{Hz}}{10\,\text{Hz}}\right)\right)}{\pi\left(1-\dfrac{0.318\,\text{Hz}}{10\,\text{Hz}}\right)}\right]^2\left[1+\left(\dfrac{10\,\text{Hz}-0.318\,\text{Hz}}{0.318\,\text{Hz}}\right)^2\right]^{-1} + \left[\dfrac{\sin\left(\pi\left(1+0.318\,\dfrac{\text{Hz}}{10\,\text{Hz}}\right)\right)}{\pi\left(1+\dfrac{0.318\,\text{Hz}}{10\,\text{Hz}}\right)}\right]^2\left[1+\left(\dfrac{10\,\text{Hz}+0.318\,\text{Hz}}{0.318\,\text{Hz}}\right)^2\right]^{-1}}\right]^{-1/2}$$
$$\times 100\% = 0.143\%\text{FS}$$

$$\varepsilon_{\substack{\text{controlled}\\\text{variable}}} = \left[\begin{array}{c}(\varepsilon_{\text{measurement}}\times 2^{-1.2})^2 + \varepsilon^2_{\text{S/H}} + \varepsilon^2_{\text{A/D}}\\ +\varepsilon^2_{\text{D/A}} + \varepsilon^2_{\text{sinc}} + \varepsilon^2_{\text{intersample}}\end{array}\right]^{1/2} = 0.39\%\text{FS}$$

Table 11 Process Closed-Loop Bandwidth

Process	−3dB BW of controlled variable
First order	$\text{BW} = \dfrac{0.35}{1.1t_r}\,\text{Hz}$ (t_r from $C(n)$)
Second order	$\text{BW} = \frac{1}{2\pi}\left(-a+\frac{1}{2}\sqrt{a^2+4\omega_n^4}\right)^{1/2}$ Hz where $a = 4\sigma^2\omega_n^2 + 4\sigma\omega_n^3 T - 2\omega_n^2 - \omega_n^4 T^2$ (natural frequency ω_n, sample period T sec, damping ratio σ)

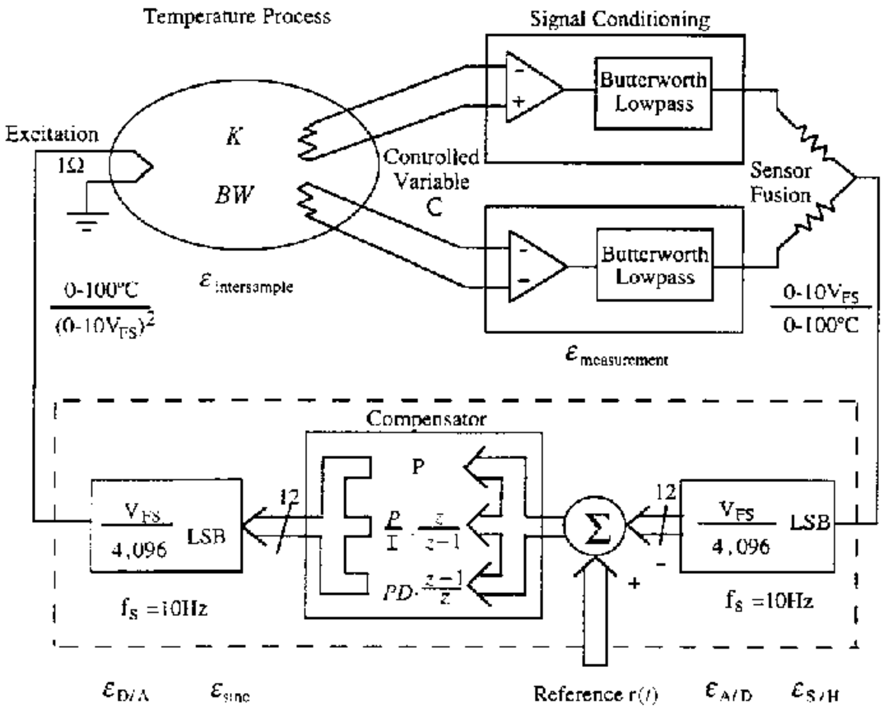


Figure 16 Process controlled-variable defined error.

The addition of interpolation, sinc, and device errors results in a total rss controlled-variable error of 0.39%FS, corresponding to 8-bit binary accuracy. This 0.39%FS defined error describes the baseline variability of the control loop and hence the process quality capability. It is notable that control-loop tracking cannot achieve less process disorder than this defined-error value regardless of the performance enabled by process identification and tuning of the PID compensator.

BIBLIOGRAPHY

1. JW Gardner. Microsensors. New York: John Wiley, 1994.
2. G Tobey, J Graeme, L Huelsman. Operational Amplifiers: Design and Applications. New York: McGraw-Hill, 1971.
3. J Graeme. Applications of Operational Amplifiers: Third-Generation Techniques. New York: McGraw-Hill, 1973.
4. PH Garrett. Computer Interface Engineering for Real-Time Systems. Englewood Cliffs, NJ: Prentice-Hall, 1987.
5. PR Geffe. Toward high stability in active filters. IEEE Spect 7(May): 1970.
6. P Allen, L Huelsman. Theory and Design of Active Filters. New York: McGraw-Hill, 1980.
7. PH Garrett. Optimize transducer/computer interfaces. Electron Des (May 24): 1977.
8. S Laube. Comparative analysis of total average filter component error. Senior Design Project, Electrical Engineering Technology, University of Cincinnati, 1983.
9. M Budai. Optimization of the signal conditioning channel. Senior Design Project, Electrical Engineering Technology, University of Cincinnati, 1978.
10. LW Gardenshire. Selecting sample rates. ISA J April: 1964.
11. AJ Terri. The Shannon sampling theorem – its various extensions and applications: a tutorial review. Proc IEE 65 (11): 1977.
12. N Weiner, Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications. Cambridge, MA: MIT Press, 1949.
13. E Zuch. Data Acquisition and Conversion Handbook. Mansfield, MA: Datel-Intersil, 1977.
14. ER Hnatek. A User's Handbook of D/A and A/D Converters. New York: John Wiley, 1976.
15. PH Garrett. Advanced Instrumentation and Computer I/O Design. New York: IEEE Press, 1994.

Chapter 2.2

Fundamentals of Digital Motion Control

Ernest L. Hall, Krishnamohan Kola, and Ming Cao
University of Cincinnati, Cincinnati, Ohio

2.1 INTRODUCTION

Control theory is a foundation for many fields, including industrial automation. The concept of control theory is so broad that it can be used in studying the economy, human behavior, and spacecraft design as well as the design of industrial robots and automated guided vehicles. Motion control systems often play a vital part of product manufacturing, assembly, and distribution. Implementing a new system or upgrading an existing motion control system may require mechanical, electrical, computer, and industrial engineering skills and expertise. Multiple skills are required to understand the tradeoffs for a systems approach to the problem, including needs analysis, specifications, component source selection, and subsystems integration. Once a specific technology is selected, the supplier's application engineers may act as members of the design team to help ensure a successful implementation that satisfies the production and cost requirements, quality control, and safety.

Motion control is defined [1] by the American Institute of Motion Engineers as: "The broad application of various technologies to apply a controlled force to achieve useful motion in fluid or solid electromechanical systems."

The field of motion control can also be considered as mechatronics [1]: "Mechatronics is the synergistic combination of mechanical and electrical engineering, computer science, and information technology, which includes control systems as well as numerical methods used to design products with built-in intelligence."

Motion control applications include the industrial robot [2] and automated guided vehicles [3–6]. Because of the introductory nature of this chapter, we will focus on digital position control; force control will not be discussed.

2.2 MOTION CONTROL ARCHITECTURES

Motion control systems may operate in an open loop, closed-loop nonservo, or closed-loop servo, as shown in Fig. 1, or a hybrid design. The open-loop approach, shown in Fig. 1(a), has input and output but no measurement of the output for comparison with the desired response. A nonservo, on–off, or bang–bang control approach is shown in Fig. 1(b). In this system, the input signal turns the system on, and when the output reaches a certain level, it closes a switch that turns the system off. A proportion, or servo, control approach is shown in Fig. 1(c). In this case, a measurement is made of the actual output signal, which is fed back and compared to the desired response. The closed-loop servo control system will be studied in this chapter.

The components of a typical servo-controlled motion control system may include an operator interface, motion control computer, control compensator, electronic drive amplifiers, actuator, sensors and transducers, and the necessary interconnections. The actua-

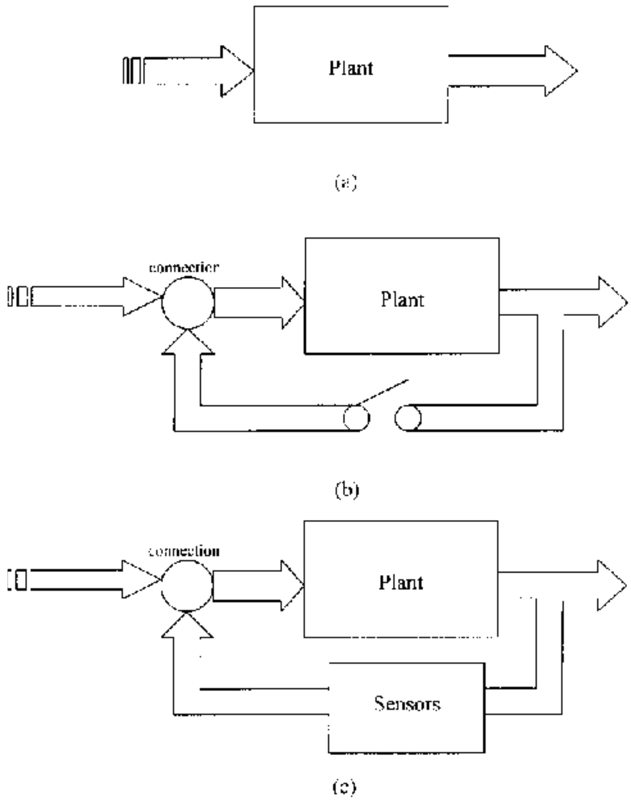


Figure 1 Motion control systems may operate in several ways such as (a) open loop, (b) closed-loop nonservo, or (c) closed-loop servo.

tors may be powered by electromechanical, hydraulic, or pneumatic power sources, or a combination.

The operator interface may include a combination of switches, indicators, and displays, including a computer keyboard and a monitor or display. The motion control computer generates command signals from a stored program for a real-time operation. The control compensator is a special prgram in the motion control computer. Selecting the compensator parameters is often a critical element in the success of the overall system. The drive amplifiers and electronics must convert the low power signals from the computer to the higher power signals required to drive the actuators. The sensors and transducers record the measurements of position or velocity that are used for feedback to the controller. The actuators are the main drive devices that supply the force or torque required to move the load. All of these subsystems must be properly interconnected in order to function properly.

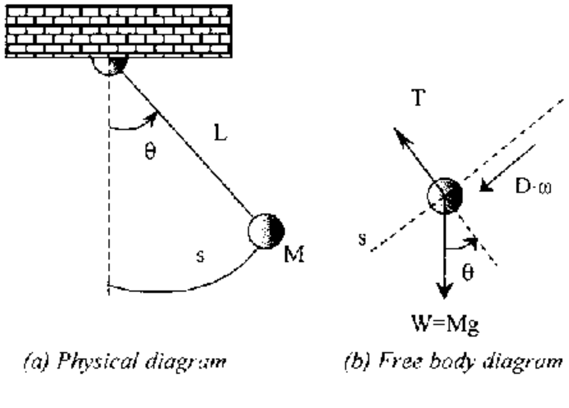
2.3 MOTION CONTROL EXAMPLE

Consider the simple pendulum shown in Fig. 2 that has been studied for more than 2000 years. Aristotle first observed that a bob swinging on a string would come to rest, seeking a lower state of energy. Later, Galileo Galilei made a number of incredible, intuitive inferences from observing the pendulum. Galileo's conclusions are even more impressive considering that he made his discoveries before the invention of calculus.

2.3.1 Flexible-Link Pendulum

The pendulum may be described as a bob with mass, M, and weight given by $W = Mg$, where g is the acceleration of gravity, attached to the end of a flexible cord of length, L as shown in Fig. 2. When the bob is displaced by an angle θ, the vertical weight component causes a restoring force to act on it. Assuming that viscous damping, from resistance in the medium, with a damping factor, D, causes a retarding force proportional to its angular velocity, ω, equal to $D\omega$. Since this is a homogeneous, unforced system, the starting motion is set by the initial conditions. Let the angle at time $\theta(t = 0)$ be 45°. For definiteness let the weight, $W = 40\,\text{lb}$, the length, $L = 3\,\text{ft}$, $D = 0.1\,\text{lb sec}$ and $g = 32.2\,\text{ft/s}^2$.

The analysis is begun by drawing a free-body diagram of the forces acting on the mass. We will use the tangent and normal components to describe the forces acting on the mass. The free-body diagram shown in Fig. 2(b) and Newton's second law are then used to derive a differential equation describing the dynamic response of the system. Forces may be balanced in any direction; however, a particularly simple form of the



(a) Physical diagram

(b) Free body diagram

Figure 2 Pendulum as studied by Galileo Galilei.

equation for pendulum motion can be developed by balancing the forces in the tangential direction:

$$\sum F_t = Ma_t \tag{1}$$

This gives the following equation:

$$-Mg\sin\theta - D\frac{d\theta}{dt} = Ma_t \tag{2}$$

The tangential acceleration is given in terms of the rate of change of velocity or arc length by the equation

$$a_t = \frac{dv}{dt} = \frac{d^2s}{dt^2} \tag{3}$$

Since the arc length, s, is given by

$$s = L\theta \tag{4}$$

Substituting s into the differential in Eq. (3) yields

$$a_t = L\frac{d^2\theta}{dt^2} \tag{5}$$

Thus, combining Eqs. (2) and (5) yields

$$-Mg\sin\theta - D\frac{d\theta}{dt} = Ma_t = ML\frac{d^2\theta}{dt^2} \tag{6}$$

Note that the unit of each term is force. In imperial units, W is in lb_f, g is in ft/sec^2, D is in lb sec, L is in feet, θ is in radians, $d\theta/dt$ is in rad/sec and $d^2\theta/dt^2$ is in rad/sec^2. In SI units, M is in kg, g is in m/sec^2, D is in kg m/sec, L is in meters, θ is in radians, $d\theta/dt$ is in rad/sec, and $d^2\theta/dt^2$ is in rad/sec^2.

This may be rewritten as

$$\frac{d^2\theta}{dt^2} + \frac{D}{ML}\frac{d\theta}{dt} + \frac{g}{L}\sin\theta = 0 \tag{7}$$

This equation may be said to describe a *system*. While there are many types of systems, systems with no output are difficult to observe, and systems with no input are difficult to control. To emphasize the importance of position, we can describe a *kinematic* system, such as $y = T(x)$. To emphasize time, we can describe a *dynamic* system, such as $g = h(f(t))$. Equation (7) describes a dynamic response. The differential equation is nonlinear because of the $\sin\theta$ term.

For a linear system, $y = T(x)$, two conditions must be satisfied:

1. If a constant, a, is multiplied by the input, x, such that ax is applied as the input, then the output must be multiplied by the same constant:

$$T(ax) = ay \tag{8}$$

2. If the sum of two inputs is applied, the output must be the sum of the individual outputs and the principal of superposition must hold as demonstrated by the following equations:

$$T(x_1 + x_2) = y_1 + y_2 \tag{9}$$

where

$$T(x_1) = y_1 \tag{10}$$

and

$$T(x_2) = y_2 \tag{11}$$

Equation (7) is nonlinear because the sine of the sum of two angles is not equal to the sum of the sines of the two angles. For example, $\sin 45^\circ = 0.707$, while $\sin 90^\circ = 1$.

Invariance is an important concept for systems. In an optical system, such as reading glasses, position invariance is desired, whereas, for a dynamic system time invariance is very important.

Since an arbitrary input function, $f(t)$ may be expressed as a weighted sum of impulse functions using the Dirac delta function, $\delta(t-\tau)$. This sum can be expressed as

$$f(t) = \int_{-\infty}^{\infty} f(\tau)\,\delta(t-\tau)\,d\tau \tag{12}$$

(Note that t is the time the output is observed and τ is the time the input is applied.)

The response of the linear system to this arbitrary input may be computed by

$$g(t) = h\left[\int_{-\infty}^{\infty} f(\tau)\,\delta(t-\tau)\,d\tau\right] \tag{13}$$

Thus by the property of linearity we obtain

$$g(t) = \int_{-\infty}^{\infty} f(\tau)\,h[\delta(t-\tau)]\,d\tau \tag{14}$$

Therefore, the response of the linear system is characterized by the response to an impulse function. This leads to the definition of the impulse response, $h(t,\tau)$, as

$$h(t,\tau) = h[\delta(t-\tau)] \tag{15}$$

Since the system response may vary with the time the input is applied, the general computational form for the output of a linear system is the superposition integral called the Fredholm integral equation [7,8]:

$$g(t) = \int_{\alpha}^{\beta} f(\tau)\, h(t, \tau)\, d\tau \tag{16}$$

The limits of integration are important in determining the form of the computation. Without any assumptions about the input or system, the computation must extend over an infinite interval.

$$g(t) = \int_{-\infty}^{+\infty} f(\tau)\, h(t, \tau)\, d\tau \tag{17}$$

An important condition of realizability for a continuous system is that the response be nonanticipatory, or casual, such that no output is produced before an input is applied:

$$h(t, \tau) = 0 \qquad \text{for } t - \tau < 0 \tag{18}$$

The causality condition leads to the computation:

$$g(t) = \int_{-\infty}^{t} f(\tau)\, h(t, \tau)\, d\tau \tag{19}$$

With the condition that $f(t) = 0$ for $t < 0$, the computation reduces to

$$g(t) = \int_{0}^{t} f(\tau)\, h(t, \tau)\, d\tau \tag{20}$$

If the system is time invariant, then

$$h(t, \tau) = h(t - \tau) \tag{21}$$

This leads to the familiar convolution equation:

$$g(t) = \int_{0}^{t} f(\tau) h(t - \tau)\, d\tau \tag{22}$$

The reason that linear systems are so important is that they are widely applicable and that a systematic method of solution has been developed for them. The relationship between the input and output of a linear, time-invariant system is known to be a convolution relation. Furthermore, transformational techniques, such as the Laplace transform, can be used to convert the convolution into an equivalent product in the transform domain. The Laplace transform $F(s)$ of $f(t)$ is

$$F(s) = \int_{0}^{\infty} f(t)\, e^{-st}\, dt \tag{23}$$

The convolution theorem states that

$$G(s) = H(s)\, F(s) \tag{24}$$

where

$$G(s) = \int_{0}^{\infty} g(t)\, e^{-st}\, dt \tag{25}$$

and

$$H(s) = \int_{0}^{\infty} h(t)\, e^{-st}\, dt \tag{26}$$

(Note that this theorem shows how to compute the convolution with only multiplication and transform operations.) The transform, $H(s)$, of the system function, $h(t)$, is called the system transfer function. For any input, $f(t)$, its transform, $F(s)$, can be computed. Then multiplying by $H(s)$ yields the transform $G(s)$. The inverse Laplace transform of $G(s)$ gives the output time response, $g(t)$.

This transform relationship may also be used to develop block diagram representations and algebra for linear systems, which is very useful to simplify the study of complicated systems.

2.3.1.1 Linear-Approach Modeling

Returning to the pendulum example, the solution to this nonlinear equation with $D \neq 0$ involves the elliptical function. (The solutions of this nonlinear system will be investigated later using Simulink.[1]) Using the approximation $\sin\theta = \theta$ in Eq. (7) gives the linear approximation

$$\frac{d^2\theta}{dt^2} + \frac{D}{ML}\frac{d\theta}{dt} + \frac{g}{L}\theta = 0 \tag{27}$$

When $D = 0$, Eq. (27) simplifies to the linear differential equation for simple harmonic motion:

$$\frac{d^2\theta}{dt^2} + \frac{g}{L}\theta = 0 \tag{28}$$

A Matlab[1] m-file may be used to determine the time response to the linear differential equation. To use Laplace transforms in Matlab, we must use the linear form of the system and provide initial conditions, since no forcing function is applied.

Remembering that the Laplace transform of the derivative is

$$L\left\{\frac{d\theta}{dt}\right\} = s\Theta(s) - \theta(0^-) \tag{29}$$

and

[1] Matlab and Simulink are registered trademarks of the Math Works, Inc.

$$L\left\{\frac{d^2\theta}{dt^2}\right\} = s^2\Theta(s) - s\theta(0^-) - \frac{d\theta(0^-)}{dt} \tag{30}$$

Taking the Laplace transform of the linear differential Eq. (27) gives

$$s^2\Theta(s) - s\theta(0^-) - \frac{d\theta(0^-)}{dt} + \frac{D}{ML}[s\Theta(s) - \theta(0^-)] + \frac{g}{L}\Theta(s) = 0 \tag{31}$$

This may be simplified to

$$\Theta(s) = \frac{s\theta(0^-) - \frac{D}{ML}\theta(0^-) + \frac{d\theta(0^-)}{dt}}{s^2 + \frac{D}{ML}s + \frac{g}{L}} \tag{32}$$

(Note that the initial conditions act as a forcing function for the system to start it moving.) It is more common to apply a step function to start a system. The unit step function is defined as

$$u(t) = \begin{cases} 1 & \text{for } t \geqslant 0 \\ 0 & \text{for } t < 0 \end{cases} \tag{33}$$

(Note that the unit step function is the integral of the delta function.) It may also be shown that the Laplace transform of the delta function is 1, and that the Laplace transform of the unit step function is $1/s$.

To use Matlab to solve the transfer function for $\theta(t)$, we must tell Matlab that this is the output of some system. Since $G(s) = H(s)F(s)$, we can let $H(s) = 1$ and $F(s) = \Theta(s)$. Then the output will be $G(s) = \Theta(s)$, and the impulse function can be used directly. If Matlab does not have an impulse response but it does have a step response, then a slight manipulation is required. [Note that the impulse response of system $G(s)$ is the same as the step response of system $s\ (G(s))$.]

The transform function with numerical values substituted is

$$\Theta(s) = \frac{45(s - 0.0268)}{s^2 + 0.0268s + 10.73} \tag{34}$$

Note that $\theta(0) = 45^\circ$ and $d\theta(0)/dt = 0$. We can define $T0 = \theta(0)$ for ease of typing, and express the numerator and denominator polynomials by their coefficients as shown by the num and den vectors below.

To develop a Matlab m-file script using the step function, define the parameters from the problem statement:

```
T0=45
D=0.1
M=40/32.2
L=3
G=32.3
num=[T0,D*T0/(M*L),0];
den=[1,D/(M*L),G/L];
t=0:0.1:10;
step(num,den,t);
grid on
title ('Time response of the pendulum
linear approximation')
```

This m-file or script may be run using Matlab and should produce an oscillatory output. The angle starts at 45° at time 0 and goes in the negative direction first, then oscillates to some positive angle and dampens out. The period,

$$T = 2\pi\sqrt{\frac{L}{g}} \tag{35}$$

in seconds (or frequency, $f = 1/T$ in cycles/second or hertz) of the response can be compared to the theoretical solution for an undamped pendulum given in Eq. (35) [9]. This is shown in Fig. 3.

2.3.1.2 Nonlinear-Approach Modeling

To solve the nonlinear system, we can use Simulink to develop a graphical model of the system and plot the time response. This requires developing a block diagram solution for the differential equation and then constructing the system using the graphical building

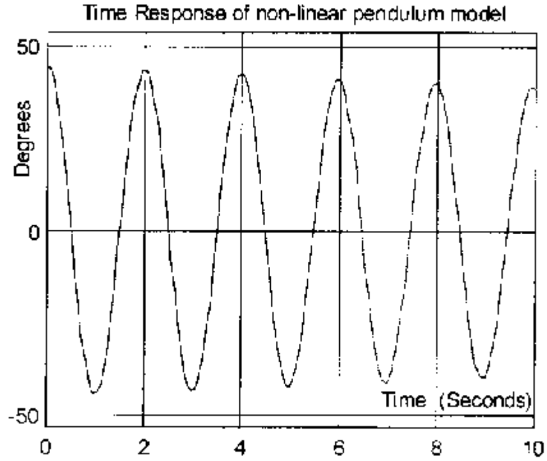


Figure 3 Pendulum response with linear approximation, $\theta(0+) = 45^\circ$.

blocks of Simulink. From this block diagram, a simulation can be run to determine a solution.

To develop a block diagram, write the differential equation in the following form:

$$\frac{d^2\theta(t)}{dt^2} = \frac{-D}{ML}\frac{d\theta}{dt} - \frac{g}{L}\sin\theta(t) \tag{36}$$

Note that this can be drawn as a summing junction with two inputs and one output. Then note that θ can be derived from $d^2\theta/dt^2$ by integrating twice. The output of the first integrator gives $d\theta/dt$. An initial velocity condition could be put at this integration. A pick-off point could also be put here to be used for velocity feedback. The output of the second integrator gives θ. The initial position condition can be applied here. This output position may also be fed back for the position feedback term. The constants can be implemented using gain terms on amplifiers since an amplifier multiplies its input by a gain term. The sine function can be represented using a nonlinear function.

The motion is started by the initial condition, $\theta(0+) = 45^\circ$, which was entered as the integration constant on the integrator which changes $d\theta/dt$ to θ. Note that the sine function expects an angle in radians, not degrees. Therefore, the angle must be converted before computing the sine. In addition, the output of the sine function must be converted back to degrees. A block diagram of this nonlinear model is shown in Fig. 4. The mathematical model to analyze such a nonlinear system is complicated. However, a solution is easily obtained with the sophisticated software of Simulink. The response of this nonlinear system is shown in Fig. 5. Note that it is very similar to the response of the linear system with an amplitude swinging between $+45^\circ$ and -45°, and a period slightly less than 2 sec, indicating that the linear system approximation is not bad. Upon close inspection, one would see that the frequency of the nonlinear solution is not, in fact, constant.

2.3.2 Rigid-Link Pendulum

Consider a related problem, the dynamic response for the mechanical system model of the human leg shown in Fig. 6. The transfer function relates the output angular position about the hip joint to the input torque supplied by the leg muscle. The model assumes an input torque, $T(t)$, viscous damping, D at the hip joint, and inertia, J, around the hip joint. Also, a component of the weight of the leg, $W = Mg$, where M is the mass of the leg and g is the acceleration of gravity, creates a nonlinear torque. Assume that the leg is of uniform density so that the weight can be applied at the centroid at $L/2$ where L is the length of the leg. For definiteness let $D = 0.01\,\text{lb sec}$, $J = 4.27\,\text{ft lb sec}^2$, $W = Mg = 40\,\text{lb}$, $L = 3\,\text{ft}$. We will use a torque amplitude of $T(t) = 75\,\text{ft lb}$.

The pendulum gives us a good model for a robot arm with a single degree of freedom. With a rigid link, it is natural to drive the rotation by a torque applied to the pinned end and to represent the mass at the center of mass of the link. Other physical variations lead to different robot designs. For example, if we mount the rigid link horizontally and then articulate it, we reduce

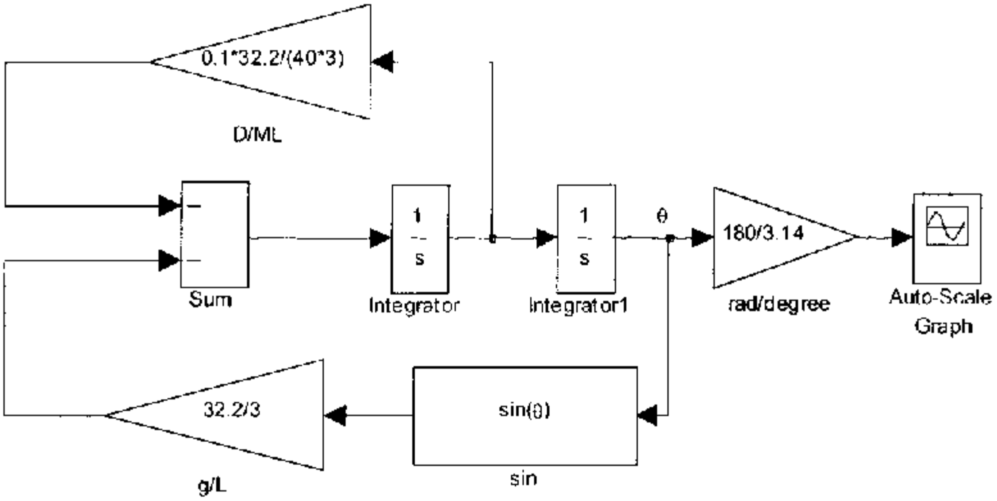


Figure 4 Block diagram entered into Simulink to solve the nonlinear system.

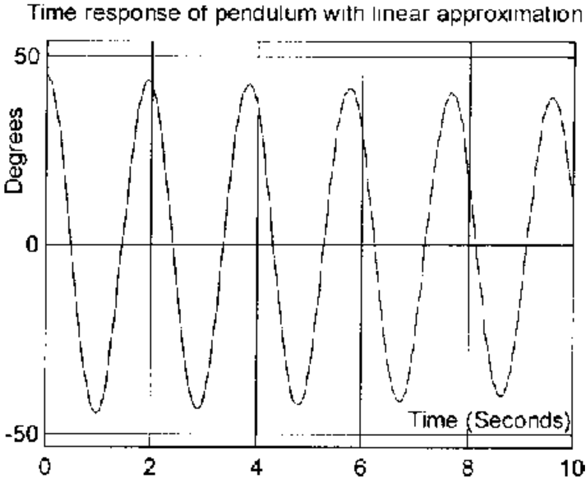


Figure 5 Nonlinear pendulum response simulation with Simulink. Note the location of the peaks.

the effect of gravity on the motion. This option is used in the Cartesian, cylindrical, and SCARA (Selective Compliant Articulated Robot for Assembly) robot designs. The spherical and vertically articulated industrial robots have the rigid link pointed upward. A gantry or an overhead-mounted, vertically articulated industrial robot has the link pointing down as in the simple pendulum.

We can draw a free-body diagram, as shown in Fig. 6, for the application of the rotational form of Newton's second law about the pinned end to balance the torque. The angle of motion is shown with positive direction counterclockwise. The motion is resisted by three torques: the component of weight is $(MgL/2)\sin\theta$; the damping torque is $D(d\theta/dt)$; and the inertial torque is $J(d^2\theta/dt^2)$. For a bob mass at the end of a link, the inertia is $J = ML^2$. However, for a distributed link the inertia is only $ML^2/12$.

We can write the differential equation that describes the system dynamic response and obtain both nonlinear and linear equations. A solution can be developed by using the rotary system torque balance.

$$J\frac{d^2\theta}{dt^2} + D\frac{d\theta}{dt} + \frac{MgL}{2}\sin\theta = T(t) \tag{37}$$

Using the small-angle approximation, $\sin\theta = \theta$ gives

$$J\frac{d^2\theta}{dt^2} + D\frac{d\theta}{dt} + \frac{MgL}{2}\theta = T(t) \tag{38}$$

Equation (38) is a linear form. Since it is linear, we can take the Laplace transform to obtain the transfer function between the output and input:

$$\frac{\Theta(s)}{T(s)} = \frac{1}{Js^2 + Ds + \dfrac{MgL}{2}} = \frac{1/J}{s^2 + \dfrac{D}{J}s + \dfrac{MgL}{2J}} \tag{39}$$

It is also interesting to show how the equations simplify for a pendulum mounted in a horizontal plane rather than a vertical plane. For a horizontally articulated pendulum or robot, the weight is perpendicular to the motion and does no work so the equation simplifies as

$$T(t) = J\frac{d^2\theta}{dt^2} + D\frac{d\theta}{dt} \tag{40}$$

This linear system is easier to control and shows why the SCARA robot design has become popular. The SCARA robot assembles items accurately and efficiently as well.

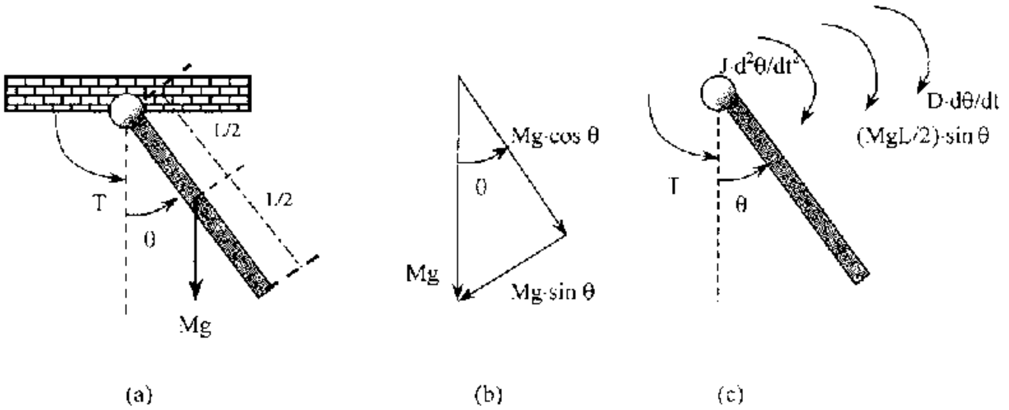


Figure 6 Rigid-link pendulum structure diagram: (a) physical diagram; (b) components of weight vector; (c) free-body torque diagram.

We can also develop a Matlab m-file solution to this linear differential equation:

```
J=4.27;
D=0.1;
M=40/32.2;
g=32.2;
L=3;
num=[0,180/3.14159];%18/3.14159 is to
translate radians into degrees
den=[J,D,M*g*L/2];
t=0:0.1:10;
impulse(num,den,t);%find impulse response
grid on;
xlabel=('Degrees');
ylabel=('Time(seconds)');
title('Unit impulse response of the rigid
link pendulum');
```

When one runs this program using Matlab, it produces the result shown in Fig. 7.

One can also use Simulink to develop a graphical model and solve the nonlinear system. To develop the block diagram recall that $T(t)$ is the input and θ is the output. We can manipulate the differential equation and develop the block diagram. Various forms of the block diagram may be developed depending on how one solves the equation. One form is shown in Fig. 8. When the torque step input is $T(0+) = 75$, the time response is as shown in Fig. 9. Rather than oscillating, the angle output appears to be going to infinity. This corresponds to the rigid link rotating continuously about its axis.

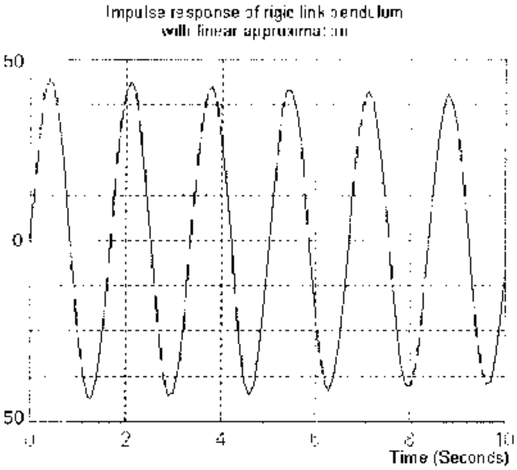


Figure 7 Solution to nonlinear system computed with Simulink.

2.3.2.1 Representation with State Variables

One can also determine the differential equation for the rigid-link pendulum by applying a torque balance around the pinned end for a vertically articulated robot pointed upward using a *state variable* representation [10].

State variables are a basic approach to modern control theory. Mathematically, it is a method for solving an nth-order differential equation using an equivalent set of n, simultaneous, first-order differential equations. Numerically, it is easier to compute solutions to first-order differential equations than for higher-order differential equations. Practically, it is a way to use digital computers and algorithms based on matrix equations to solve linear or nonlinear systems. A system is described in terms of its *state variables*, which are the smallest set of linearly independent variables that describe the system, its *dynamic state variable*, the derivative of the state variable, its *input*, and its *output*. Since state variables are not unique, many different forms may be chosen for solving a particular problem. One particular set which is useful in the solution of nth-order single variable differential equations is the set of phase variables. These are defined in terms of the variable and its derivatives of the variable of the nth-order equation. For example, in the second-order differential equation in θ which we are working with, we can define a vector state variable with components, $x_1 = \theta(t)$ and $x_2 = d\theta(t)/dt$. Two state variables are required because we have a second-order differential equation. We would need N for an Nth-order differential equation. The state vector may be written as the transpose of the row vector: $[x_1, x_2]^T$. We normally use column vectors, not row vectors, for points. The state equations for a linear system always consist of two equations that are usually written as

$$\frac{dx}{dt} = Ax + Bu$$
$$y = Cx + Du \tag{41}$$

where x is the state vector, dx/dt is the dynamic state vector, u is the vector input and y is the vector output. Suppose the state vector x has a dimension of n. For a single input, single output (SISO) system: A is an $n \times n$ constant coefficient matrix called the system matrix; B is a $n \times 1$ constant matrix called the control matrix; C is a $1 \times n$ constant matrix called the output matrix; and D is called the direct feedforward matrix. For the SISO system, D is a 1×1 matrix containing a scalar constant.

Using the phase variables as state variables,

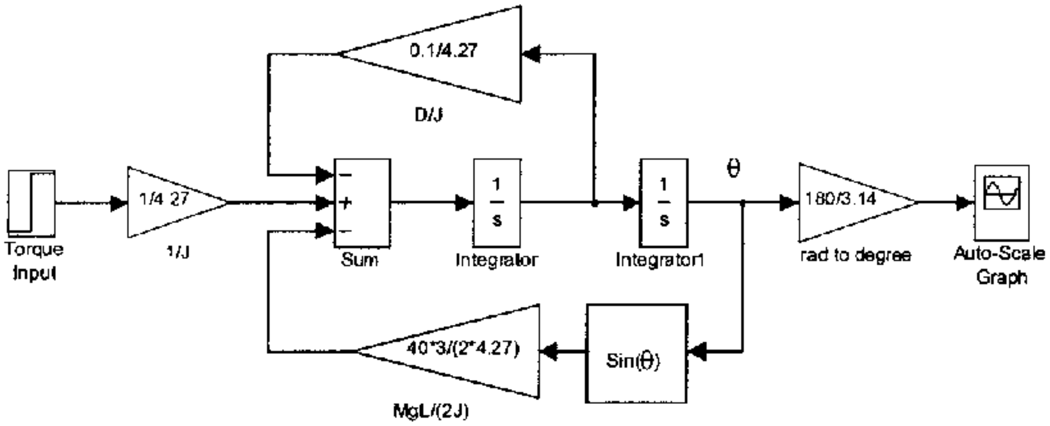


Figure 8 Block diagram of the nonlinear rigid link system with torque input.

$$x_1 = \theta$$
$$x_2 = \frac{d\theta}{dt} \tag{42}$$

Then the dynamic state equations may be developed using the definition and the system equation

$$\frac{dx_1}{dt} = x_2$$
$$\frac{dx_2}{dt} = \frac{-MgL}{J}\sin x_1 - \frac{D}{J}x_2 + \frac{T(t)}{J} \tag{43}$$

$$y = x_1 \tag{44}$$

This is the state variable form for the dynamic system. The output y is simply x_1.

The system can be linearized about the equilibrium point , $x_1 = 0$, and $x_2 = 0$. This again amounts to using the approximation

$$\sin x_1 = x_1 \tag{45}$$

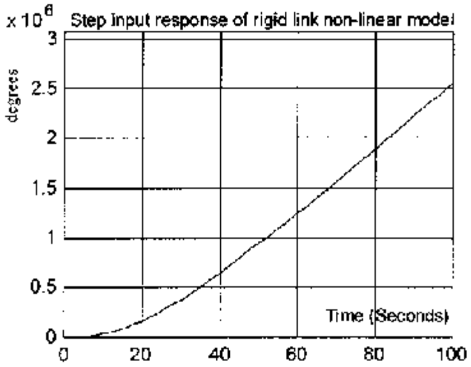


Figure 9 Unit step response of the rigid link model.

and may be used to obtain the linear state equations:

$$x_1 = 0 + \delta x_1$$
$$x_2 = 0 + \delta x_2 \tag{46}$$

The complete state equations are

$$\dot{x} = Ax + Bu$$
$$y = Cx + Du \tag{47}$$

These may be written in the linear state equation's matrix form:

$$\frac{dx_1}{dt} = x_2$$
$$\frac{dx_2}{dt} = -\frac{D}{J}x_2 - \frac{MgL}{J}x_1 + \frac{T(t)}{J} \tag{48}$$
$$y = x_1$$

in which the output is

$$A = \begin{bmatrix} 0 & 1 \\ -MgL/J & -D/J \end{bmatrix} \qquad B = \begin{bmatrix} 0 \\ 1/J \end{bmatrix}$$
$$C = [1 \quad 0] \qquad D = [0] \tag{49}$$

To illustrate the use of state variables, let us determine the equilibrium states of the system. The equilibrium is defined in terms of the unforced system as the points at which the dynamic state vector is equal to zero:

$$\frac{dx_1}{dt} = 0 \tag{50}$$

$$\frac{dx_2}{dt} = 0 \tag{51}$$

which gives

$$x_2 = 0 \tag{52}$$

and

$$\frac{MgL}{2J}\sin x_1 \frac{Dx_2}{J} = 0 \tag{53}$$

So the solutions are

$$x_1 = n\pi \qquad n = 0, 1, 2, \ldots$$
$$x_2 = 0$$

It is possible to use either the state space or the transfer function representation of a system. For example, the transfer function of the linearized rigid link pendulum is developed as described in the next few pages.

Taking the Laplace transform assuming zero initial conditions gives

$$\begin{aligned} T(t) &= J\frac{d^2\theta}{dt^2} + D\frac{d\theta}{dt} + \frac{Mgl}{2}\theta \\ \frac{\Theta(s)}{T(s)} &= \frac{1/J}{s^2 + \frac{Ds}{J} + \frac{MgL}{2J}} \end{aligned} \tag{54}$$

The nonlinear differential equation of the rigid link pendulum can also be put in the "rigid robot" form that is often used to study the dynamics of robots.

$$M(\ddot{q})\ddot{q} + V(\dot{q}, q) + G(q) = T(t) \tag{55}$$

where M is an inertia matrix, q is a generalized coordinate vector, V represents the velocity dependent torque, G represents the gravity dependent torque and T represents the input control torque vector.

$$\begin{aligned} M(\theta) &= J \\ V(\theta) &= D \\ G(\theta) &= \frac{MgL}{2}\theta \\ \theta &= q \\ \tau &= T(t) \end{aligned} \tag{56}$$

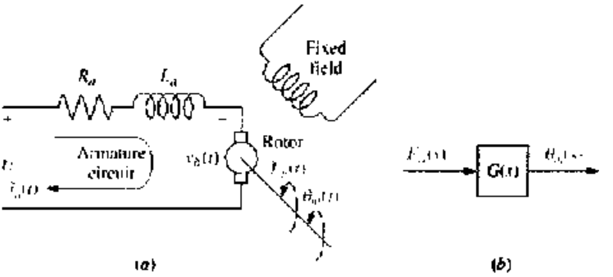


Figure 10 Fixed field DC motor: (a) circuit diagram; (b) block diagram (from Nise, 1995).

2.3.3 Motorized Robot Arm

As previously mentioned, a rigid-link model is in fact the basic structure of a robot arm with a single degree of freedom. Now let us add a motor to such a robot arm.

A DC motor with armature control and a fixed field is assumed. The electrical model of such a DC motor is shown in Fig. 10. The armature voltage, $e_a(t)$ is the voltage supplied by an amplifier to control the motor. The motor has a resistance R_a, inductance L_a, and back electromotive force (emf) constant, K_b. The back emf voltage, $v_b(t)$ is induced by the rotation of the armature windings in the fixed magnetic field. The counter emf is proportional to the speed of the motor with the field strength fixed. That is,

$$v_b(t) = K_b\frac{d\theta}{dt} \tag{57}$$

Taking the Laplace transform gives

$$V_b(s) = sK_b\Theta(s) \tag{58}$$

The circuit equation for the electrical portion of the motor is

$$E_a(s) = R_aI_a(s) + L_asI_a(s) + V_b(s) \tag{59}$$

This may also be written as

$$I_a(s) = \frac{E_a(s) - K_bs\Theta(s)}{L_as + R_a} \tag{60}$$

The torque developed by the motor is proportional to the armature current:

$$T_m(s) = K_tI_a(s) \tag{61}$$

This torque moves the armature and load.

Balancing the torques at the motor shaft gives the torque relation to the angle that may be expressed as follows:

$$T(t) = J\frac{d^2\theta_m}{dt^2} + D\frac{d\theta_m}{dt} \tag{62}$$

where θ_m is the motor shaft angle position, J represents all inertia connected to the motor shaft, and D all friction (air friction, bearing friction, etc.) connected to the motor shaft.

Taking the Laplace transform gives

$$T_m(s) = Js^2\Theta_m(s) + Ds\Theta_m(s) \tag{63}$$

Solving Eq. (63) for the shaft angle, we get

$$\theta_m(s) = \frac{T_m(s)}{Js^2 + Ds} \tag{64}$$

If there is a gear train between the motor and load, then the angle moved by the load is different from the angle moved by the motor. The angles are related by the gear ratio relationship, which may be derived by noting that an equal arc length, S, is traveled by two meshing gears. This can also be described by the following equation:

$$S = R_{\mathrm{m}}\theta_{\mathrm{m}} = R_{\mathrm{L}}\theta_{\mathrm{L}} \tag{65}$$

The gear circumference of the motor's gear is $2\pi R_m$, which has N_m teeth, and the gear circumference of the load's gear is $2\pi R_L$, which has N_L teeth, so the ratio of circumferences is equal to the ratio of radii and the ratio of number of teeth so that

$$N_L\theta_L = N_m\theta_m \tag{66}$$

or

$$\frac{\theta_L}{\theta_m} = \frac{N_m}{N_L} = n \tag{67}$$

The gear ratio may also be used to reflect quantities on the load side of a gear train back to the motor side so that a torque balance can be done at the motor side. Assuming a lossless gear train, it can be shown by equating mechanical, $T\omega_1$, and electrical, EI, power that the quantities such as inertia, J, viscous damping D, and torsional springs with constants K may be reflected back to the motor side of a gear by dividing by the gear ratio squared. This can also be described with the equations below:

$$J_{mL} = \frac{J_L}{n^2} \tag{68}$$

$$D_{mL} = \frac{D_L}{n^2} \tag{69}$$

$$K_{mL} = \frac{K_L}{n^2} \tag{70}$$

Using these relationships, the equivalent load quantities for J and D may be used in the previous block diagram. From Eqs. (59), (60), (61), (64), and (67) we can get the block diagram of the armature-controlled DC motor as shown in Fig. 11.

By simplifying the block diagram shown in Fig. 11, we can get the armature-controlled motor transfer function as

$$G(s) = \frac{\Theta_L(s)}{E(s)} = \frac{K_t n}{s[(Js+D)(L_a s + R_a) + K_b K_t]}$$

$$G(s) = \frac{K_t n}{s[(JL_a s^2 + (JR_a + DL_a)s + DR_a + K_b K_t]} \tag{71}$$

As we can see, this model is of the third order. However, in the servomotor case, the inductance of the armature L_a could usually be ignored. Thus this model could be reduced to a second-order system.

Now, apply this model to a simple example. Suppose a DC motor is used to drive a robot arm horizontally as shown in Fig. 12. The link has a mass, $M = 5\,\mathrm{kg}$, length $L = 1\,\mathrm{m}$, and viscous damping factor $D = 0.1$. Assume the system input is a voltage signal with a range of 0–10 V. This signal is used to provide the control voltage and current to the motor. The motor parameters are given below. The goal is to design a compensation strategy so that a voltage of 0 to 10 V corresponds linearly of an angle of 0° to an angle of 90°. The required response should have an overshoot below 10%, a settling time below 0.2 sec and a steady state error of zero. The motor parameters are given below:

$$J_a = 0.001\,\mathrm{kg\,m^2/s^2}$$

$$D_a = 0.01\,\mathrm{N\,m\,s/rad}$$

$$R_a = 1\,\Omega$$

$$L_a = 0\,\mathrm{H}$$

$$K_b = 1\,\mathrm{V\,s/rad}$$

$$K_t = 1\,\mathrm{N\,m/A}$$

First, consider a system without gears or a gear ratio of 1. The inertia of the rigid link as defined before is

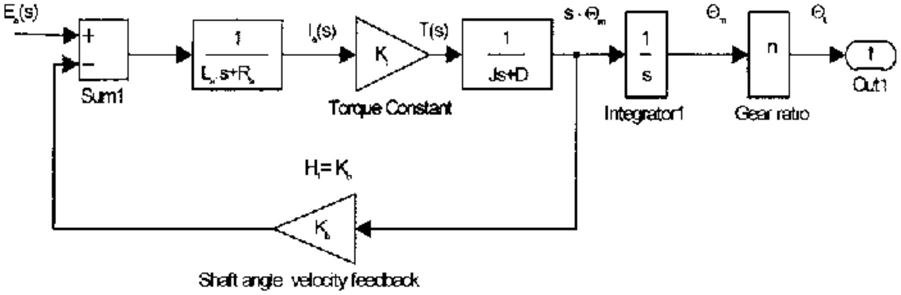


Figure 11 Armature-controlled DC motor block diagram.

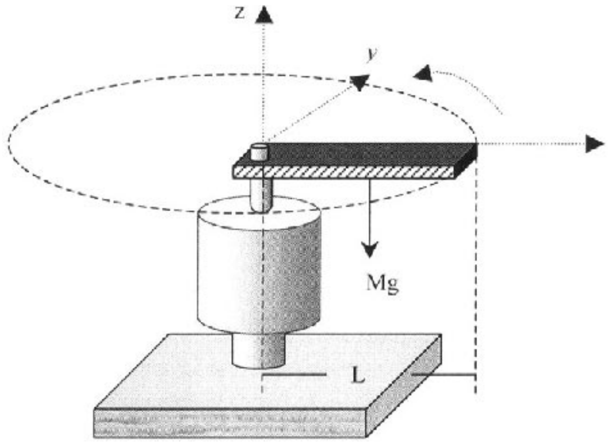


Figure 12 A single joint robot arm driven by an armature-controlled DC motor.

$$J_L = \frac{ML^2}{12} = \frac{5 \times 1^2}{12} = 0.4167\,\mathrm{kg\,m^2} \tag{72}$$

According to the impedance reflection model established before, the total inertia J and total damping factor D are

$$J = J_a + J_L = 0.001 + 0.4167 = 0.4177\,\mathrm{kg\,m^2}$$
$$D = D_a + D_L = 0.01 + 0.1 = 0.11$$

Substituting the known values into Eq. (65) we can get

$$G(s) = \frac{\Theta_L(s)}{E(s)} = \frac{1}{s[(0.4177s + 0.11)(1 + 0 \times s) + 1 \times 1]}$$
$$G(s) = \frac{1}{s(0.4177s + 1.11)} \tag{73}$$

The above process could be calculated with Matlab scripts as follows:

```
J=0.001+0.4167;
D=0.1+0.01;
La=0;
Ra=1;
Kt=1;
Kb=1;
n=1;
Num=Kt*n;
Den=[J*La J*Ra+La*D D*Ra+Kt*Kb 0];
step(Num, Den);
title(‘Step Response of the Motorized Robot
Arm’);
End;
```

The step-response of this system is shown in Fig. 13. As we can see the system does not go to a steady-state value, but to an increasing angle at constant value. This means the armature rotates at a constant speed, which is achieved by its built-in velocity feedback factor K_b.

However, we want the motor to move the robot arm to a proper angular position corresponding to the input. This can be achieved by a positional servomechanism, i.e., we feed back the actual robot arm position, θ_L, convert the position information into voltage via a constant K_p and then negatively feed back this signal back into the system. Accordingly, the feedback signal in voltage is $E_f = \theta_L K_p$, where K_p (V/deg) is the proportional constant depending on the input and desired position output.

A simple position sensor is a potentiometer, a variable resistor. The resistor can measure rotational position. A DC voltage V is applied to both ends of the resistor. R_1 is the resistance from the pointer to one fixed end, while R is the total resistance. R_1 is proportional to the position of the pointer, since the resistor is linearly distributed. Another position sensor is an encoder, it can also feed back digital position information. The mechanism of the encoder will be described later.

The revised system block diagram is shown in Fig. 14. Suppose the voltage between both ends of the potentiometer is 10 V, and that K_p is the ratio of the voltage change to the corresponding angle change. In this case, $K_p = (10 - 0)/(90 - 0) = 0.1111$ V/deg; the gear ratio is 1.

The new transfer function is

$$G'(s) = \frac{G(s)}{1 + G(s)\,K_p}$$
$$G'(s) = \frac{1}{0.4177s^2 + 1.11s + 0.0011} \tag{74}$$

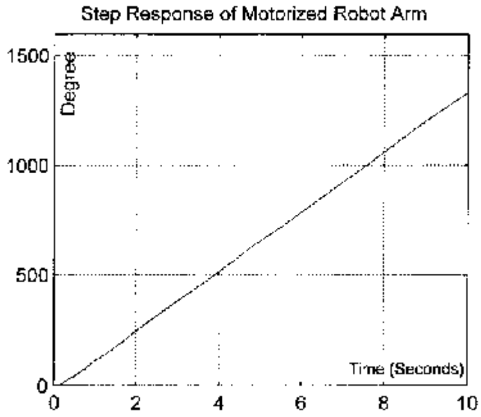


Figure 13 Step response of motorized robot arm.

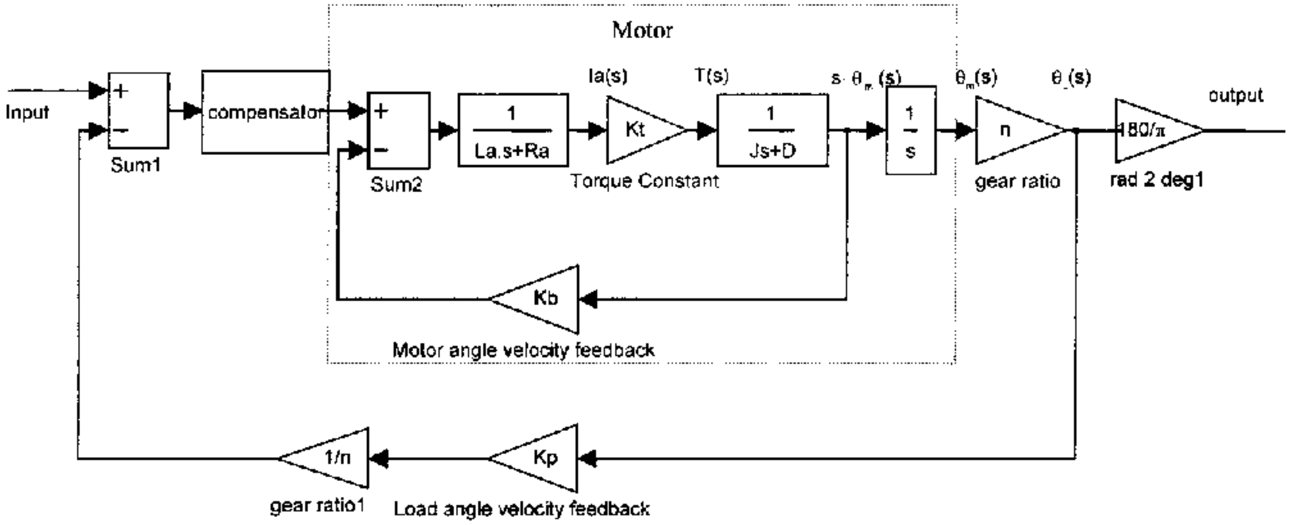


Figure 14 Position and velocity feedback model of the motorized rigid link.

The step response can be determined with the following program:

```
V=10;
Angle=90;
Kp=V/Angle; %feedback voltage/angle
constant
G=tf([1],[0.4177 1.11 0]);
% the transfer function of the velocity
loop
sysclose=feedback (G,Kp);
%the closed loop function of position
feedback
step(sysclose);
end
```

After position feedback, the steady response tends to be stable as shown in Fig. 15. However, the system response is too slow; to make it have faster response speed, further compensation is needed. The following example outlines the building of a compensator for feedback control system.

2.3.4 Digital Motion Control

2.3.4.1 Digital Controller

With the many computer applications in control systems, digital control systems have become more important. A digital system usually employs a computerized controller to control continuous components of a closed-loop system. The block diagram of the digital system is shown in Fig. 16. The digital system first samples the continuous difference data ε, and then, with an A/D converter, changes the sample impulses into digital signals and transfers them into the computer controller. The computer will process these digitral signals with predefined control rules. At last, through the digital-to-analog (D/A) converter, the computing results are converted into an analog signal, $m(t)$, to control those continuous components. The sampling switch closes every T_0 sec. Each time it closes for a time span of h with $h < T_0$. The sampling frequency, f_s, is the reciprocal of $T_0, f_s = 1/T_0$, and $\omega_s = 2\pi/T_0$ is called the sampling angular frequency. The digital controller provides the system with great flexibility. It can

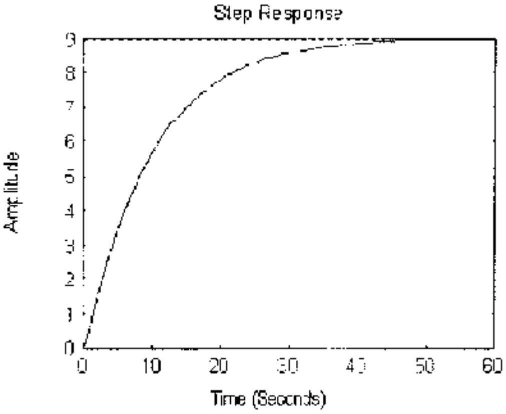


Figure 15 Step response of the motorized robot arm.

achieve compensation values that are hard for analog controllers to obtain.

2.3.4.2 Digital-Controlled Servo System

A servo system is a motor control system that combines various components to automatically control a machine's operation. It is generally composed of a servo drive, a motor, and feedback device. The controller gets a feedback signal and outside control signal and controls the servo drive and motor together to precisely control torque, velocity, or position of the motor shaft. The feedback continuously reports the real-time status, which is compared to the command value. Differences between the command position and feedback signals are automatically adjusted by the closed-loop servo system. This closed loop provides the servo system with accurate, high-performance control of a machine.

A servo motor is characterized by a long, thin configuration. This design permits the motor to provide torque while minimizing the growth in rotor inertia. It results in an improved mechanical time constant and faster time response.

The controller of the servo system is, in most cases, programmable. It allows one piece of equipment to do many related jobs or functions by selecting a different program.

A typical computer-controlled servo system is shown in Fig. 17. It has three main elements: (1) a digital controller, (2) an amplifier, and (3) a motor and an encoder. The controller is composed of a digital filter, a zero-order-hold (ZOH), and a digital-to-analog converter (DAC). The purpose of the controller is to compensate the system to make an unstable system become stable. This is usually achieved by adjusting the parameters of the filter. The controller accepts both encoder feedback and commands from the computer and finds the error between them. The error signal passes through the digital filter, ZOH, and DAC to generate control signals to control the amplifier (AMP). The amplifier amplifies the control signals from the controller to drive the motor. The encoder is usually mounted on the motor shaft. When the shaft moves, it generates electrical impulses, which are processed into digital position information. This position information is then feedback directly into the controller. The mathematical model of the above components could be varied among different products. However, they tend to be the same in most aspects.

Typical digital control systems modeling and design can be illustrated by the following design example.

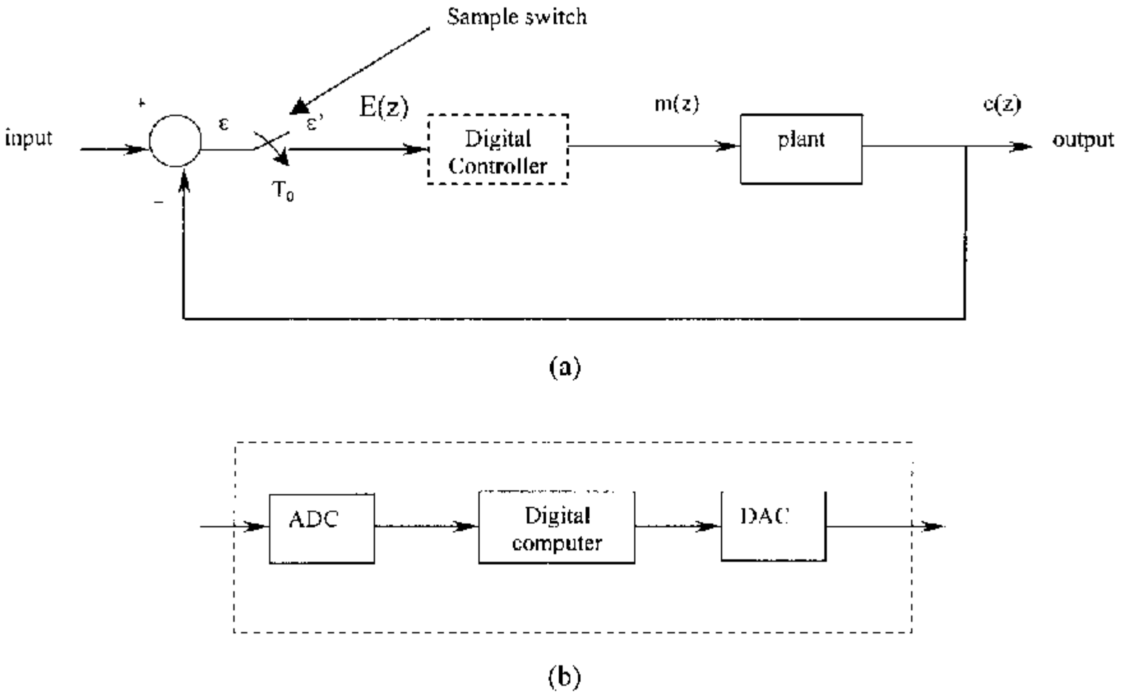


Figure 16 Computer controlled servo system: (a) system; (b) controller.

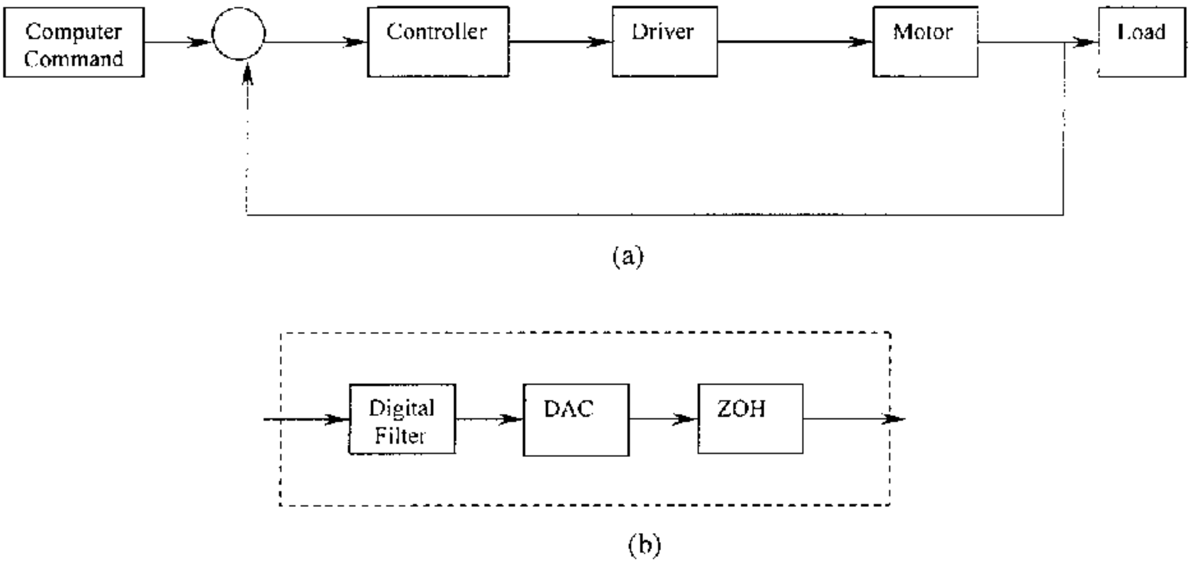


Figure 17 Two representations of digital control systems: (a) digital control system; (b) digital controller.

2.3.5 Digital Motion Control System Design Example

Selecting the right parameters for the position, integral, derivative (PID) controller is the most difficult step for any motion control system. The motion control system of the automatic guided vehicle (AGV) helps maneuver it to negotiate curves and drive around obstacles on the course. Designing a PID controller for the drive motor feedback system of Bearcat II robot, the autonomous unmanned vehicle, was therefore considered one important step for its success.

The wheels of the vehicle are driven independently by two Electrocraft brush-type DC servomotors. Encoders provide position feedback for the system. The two drive motor systems are operated in current loops in parallel using Galil MSA 12-80 amplifiers. The main controller card is the Galil DMC 1030 motion control board and is controlled through a computer.

2.3.5.1 System Modeling

The position-controlled system comprises a position servo motor (Electrocraft brush-type DC motor) with an encoder, a PID controller (Galil DMC 1030 motion control board), and an amplifier (Galil MSA 12-80).

The amplifier model can be configured in three modes, namely, voltage loop, current loop, and velocity loop. The transfer function relating the input voltage V to the motor position P depends upon the configuration mode of the system.

Voltage Loop. In this mode, the amplifier acts as a voltage source to the motor. The gain of the amplifier will be K_v, and the transfer function of the motor with respect to the voltage will be

$$\frac{P}{V} = \frac{K_v}{[K_t s(s\tau_m + 1)(s\tau_e + 1)]} \tag{75}$$

where

$$\tau_m = \frac{RJ}{K_t^2}(s) \qquad \text{and} \qquad \tau_e = \frac{L}{R}(s)$$

The motor parameters and the units are:

K_t: torque constant (N m/A),
R: armature resistance (ohms),
J: combined inertia of the motor and load (kg m^2),
L: armature inductance (Henries).

Current Loop. In this mode the amplifier acts as a current source for the motor. The corresponding transfer function will be as follows:

$$\frac{P}{V} = \frac{K_a K_t}{Js^2} \tag{76}$$

where K_a is the amplifier gain, and K_t and J are as defined earlier.

Velocity Loop. In the velocity mode, a tachometer feedback to the amplifier is incorporated. The transfer function is now the ratio of the Laplace transform of the angular velocity to the voltage input. This is given by

$$\frac{\omega}{V}=\frac{\dfrac{k_aK_t}{J_s}}{1+\dfrac{K_aK_tK_gs}{J_s}}=\frac{1}{K_g(s\tau_1+1)} \tag{77}$$

where

$$\tau_1=\frac{J}{K_aK_tK_g}$$

and therefore

$$\frac{P}{V}=\frac{1}{K_gs(s\tau_1+1)}$$

The Encoder. The encoder is an integral part of the servomotor and has two signals A and B, which are in quadrature and 90° out of phase. Due to the quadrature relationship, the resolution of the encoder is increased to $4N$ quadrature counts/rev, where N is the number of pulses generated by the encoder per revolution.

The model of the encoder can be represented by a gain of

$$K_f=\frac{4N}{2\pi}\ \text{counts/rad} \tag{78}$$

The Controller. The controller in the Galil DMC 1030 board has three elements, namely the digital-to-analog converter (DAC), the digital filter and the zero-order hold (ZOH).

Digital-to-analog converter. The DAC converts a 14-bit number to an analog voltage. The input range of numbers is 16,384 and the output voltage is $\pm 10\,\text{V}$. For the DMC 1030, the DAC gain is given by $K_\text{d}=0.0012\,\text{V/count}$.

Digital filter. This has a discrete system transfer function given by

$$D(z)=\frac{K(z-A)}{z+\dfrac{Cz}{z-1}} \tag{79}$$

The filter parameters are K, A, and C. These are selected by commands KP, KI, and KD, where KP, KI, and KD are respectively the proportional, integral and derivative gains of the PID controller.

The two sets of parameters for the DMC 1030 are related according to the equations

$$\begin{aligned}K&=K_p+K_d\\ A&=\frac{K_d}{(K_p+K_d)}\\ C&=\frac{K_i}{8}\end{aligned} \tag{80}$$

Zero-order hold. The ZOH represents the effect of the sampling process, where the motor command is updated once per sampling period. The effect of the ZOH can be modeled by the transfer function

$$H(s)=\frac{1}{\left(1+s\dfrac{T}{2}\right)} \tag{81}$$

In most applications, $H(s)$ can be approximated as 1.

Having modeled the system, we now have to obtain the transfer functions with the actual system parameters. This is done for the system as follows.

2.3.5.2 System Analysis

The system transfer functions are determined by computing transfer functions of the various components.

Motor and amplifier: The system is operated in a current loop and hence the transfer function of the motor–amplifier is given by

$$\frac{P}{V}=\frac{K_aK_t}{Js^2} \tag{82}$$

Encoder: The encoder on the DC motor has a resolution of 500 lines per revolution. Since this is in quadrature, the position resolution is given by 4 ×500 = 2000 counts per revolution. The encoder can be represented by a gain of

$$K_f=\frac{4\times N}{2\pi}=\frac{2000}{2\pi}=318$$

DAC: from the Galil manual, the gain of the DAC on the DMC 1030 is represented as $K_d=0.0012\,\text{V/count}$.

ZOH: the ZOH transfer function is given by

$$H(s)=\frac{1}{1+s\dfrac{T}{2}}$$

where T is the sampling time. The sampling time in this case is 0.001 sec. Hence the transfer function of the ZOH is

$$H(s)=\frac{2000}{s+2000} \tag{83}$$

2.3.5.3 System Compensation Objective

The analytical system design is aimed at closing the loop at a crossover frequency ω. This crossover frequency is required to be greater than 200 rad/sec. An existing system is taken as a reference and the crossover frequency of that system is used, since the two are similar Ref [11].

The following are the parameters of the system:

1. Time constant of the motor, $K_t = 2.98\,\text{lb in,/A}$ ($0.3375\ \text{N m/A}$).
2. Moment of inertia of the system, $J = 220\,\text{lb in.}^2$ (approx.) [$2.54 \times 10^4\,\text{kg m}^2$ (approx.)].
3. Motor resistance, $R = 0.42\,\Omega$.
4. Amplifier gain in current loop, $K_a = 1.2\,\text{A/V}$.
5. Encoder gain, $K_f = 318\,\text{counts/rev}$.

The design objective is set at obtaining a phase margin of 45°.

The block diagram of the system is shown in Fig. 18.

Motor:

$$M(s) = \frac{K}{Js^2} = \frac{0.3375}{2.54 \times 10^{-4}} = \frac{1330}{s^2} \tag{84}$$

Amplifier:

$$K_a = 1.2 \tag{85}$$

DAC:

$$K_d = \frac{10}{8192} = 0.0012 \tag{86}$$

Encoder:

$$K_f = 318 \tag{87}$$

ZOH:

$$H(s) = \frac{2000}{s + 2000} \tag{88}$$

Compensation filter:

$$G(s) = P + sD \tag{89}$$

$$L(s) = M(s)\,K_a K_f K_d H(s) = \frac{1.21 \times 10^6}{s^2(s + 2000)} \tag{90}$$

The feed-forward transfer function of the system is given by

$$A(s) = L(s)\,G(s) \tag{91}$$

and the open-loop transfer function will be

$$L(j200) = \frac{1.21 \times 10^6}{(j200)^2(j200 + 2000)} \tag{92}$$

The magnitude of $L(s)$ at the crossover frequency of 200 rad/sec is

$$|L(j200)| = 0.015 \tag{93}$$

and the phase of the open-loop transfer function is given by

$$\operatorname{Arg} L(j200) = -180 - \tan^{-1}\left(\frac{200}{2000}\right) = -185^\circ \tag{94}$$

$G(s)$ is selected such that $A(s)$ has a crossover frequency of 200 rad/sec and a phase margin of 45°. This requires that

$$|A(s)| = 1 \tag{95}$$

and

$$\operatorname{Arg}[A(j200)] = -135^\circ \tag{96}$$

But we have $A(s) = L(s)\,G(s)$, therefore we must have

$$|G(j200)| = \frac{|A(j200)|}{|L(j200)|} \approx 66 \tag{97}$$

and

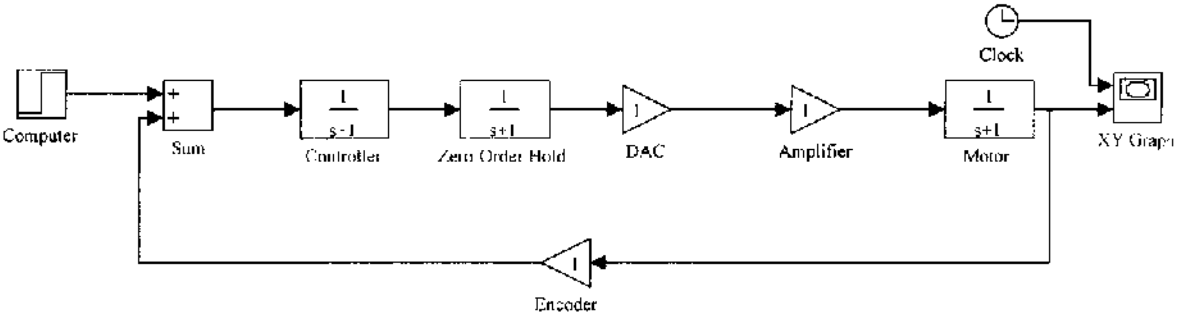


Figure 18 Block diagram of the position controlled servo system.

$$\begin{aligned}\operatorname{Arg}[G(j200)] &= \operatorname{Arg}[A(j200)] = \operatorname{Arg}[L(j200)] \\ &= -135^\circ + 185^\circ = 50^\circ\end{aligned} \tag{98}$$

Hence select the filter function of the form

$$G(s) = P + sD \tag{99}$$

such that at crossover frequency of 200, it would have a magnitude of 66 and a phase of 50°.

$$|G(j200)| = |P + (j200D)| = 66 \tag{100}$$

and

$$\operatorname{Arg}[G(j200)] = \tan^{-1}\left[\frac{200D}{P}\right] = 50^\circ \tag{101}$$

Solving these equations, we get

$$P = 42$$
$$D = 0.25$$

The filter transfer function is given by $G(s) = 0.25s + 42$.

The step response of the compensated system is shown in Fig. 19.

2.3.5.4 System Analysis with Compensator

Now with the filter parameters known, the open-loop and closed-loop transfer functions are computed as follows:

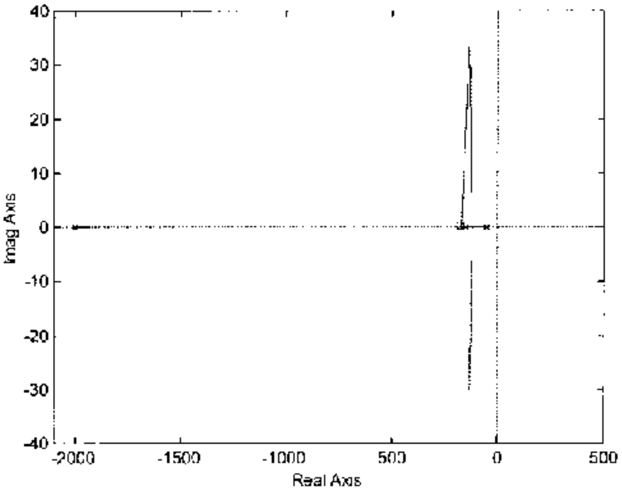


Figure 20 Root locus plot of the compensated system.

$$\text{OLTF} = \frac{9.62s^3 + 2572s^2 + 1.885 \times 10^5 s + 4.104 \times 10^6}{s^5 + 400s^4 + 47{,}500s^3 + 1.5 \times 10^6 s^2} \tag{102}$$

The root locus and Bode plot for the system are shown in Figs. 20 and 21, and it is clear that the system is not stable in the closed loop because it has two poles at the origin. This has to be further compensated by a controller in order to stabilize the closed loop.

A controller with zeros that can cancel the poles at the origin is used. Poles are added at $s = -50$ and $s = -150$ in order to stabilize the closed-loop step response.

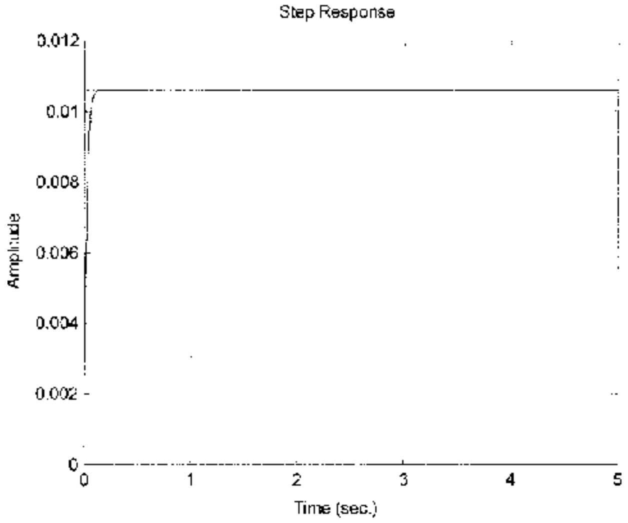


Figure 19 Step response of the compensated system.

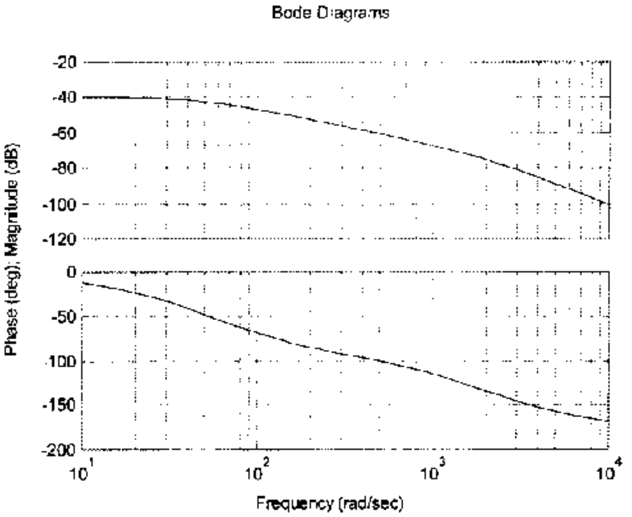
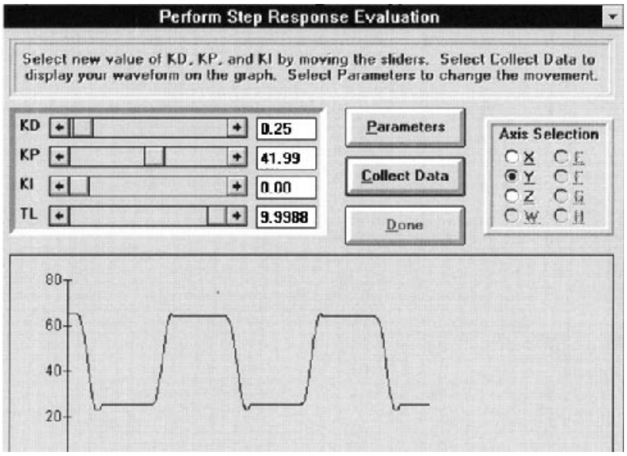


Figure 21 Bode plot of the compensated system.



Perform Step Response Evaluation

Select new value of KD, KP, and KI by moving the sliders. Select Collect Data to display your waveform on the graph. Select Parameters to change the movement.

KD 0.25

KP 41.99

KI 0.00

TL 9.9988

Axis Selection: ○ X ● Y ○ Z ○ W ○ E ○ F ○ G ○ H

Figure 22 Experimental step response.

The controller transfer function is given by

$$G(s) = \frac{s^2}{(s+50)(s+150)} \tag{103}$$

With the controller, the open- and closed-loop transfer functions are given by

$$\mathrm{OLTF} = \frac{30.45 \times 10^3 s + 51.15 \times 10^6}{s^3 + 2200s^2 + 407{,}500s + 15 \times 10^6} \tag{104}$$

and

$$\mathrm{CLTF} = \frac{957.6s + 160{,}876}{s^3 + 2200s^2 + 408{,}457s + 15.16 \times 10^6} \tag{105}$$

The experimental step response plots of the system are shown in Fig. 22.

The analytical values of K_p, K_i, and K_d which are the proportional, integral, and derivative gains, respectively, of the PID controller, are tested for stability in the real system with the help of Galil Motion Control Servo Design Kit Version 4.04.

2.4 CONCLUSIONS

A simple mechanism has been used to illustrate many of the concepts of system theory encountered in controlling motion with a computer. Natural constraints often described by a differential equation are encountered in nature. The parameters such as length and mass of the pendulum have a large impact on its control. Stability and other system concepts must be understood to design a safe and useful system. Analog or continuous system theory must be merged with digital concepts to effect a computer control. The result could be a new, useful, and nonobvious solution to an important practical problem.

REFERENCES

1. D Shetty, RA Kolk. Mechatronics System Design. Boston, MA: PWS Publishing, 1997.
2. H Terasaki, T Hasegawa. Motion planning of intelligent manipulation by a parallel two-fingered gripper equipped with a simple rotating mechanism. IEEE Trans Robot Autom 14(2): 207–218, 1998.
3. K Tchon, R Muszynski. Singular inverse kinematic problem for robotic manipulators: a normal form approach. IEEE Trans Robot and Autom 14(1): 93–103, 1998.
4. G Campion, G Bastin, B D'Andrea-Novel. Structural properties and classification of kinematic and dynamic models of wheeled mobile robots. IEEE Trans Robot Autom 12(1): 47–61, 1996.
5. B Thuilot, B D'Andrea-Novel, A Micaeelli. Modeling and feedback control of mobile robots equipped with several steering wheels. IEEE Trans Robot Autom 12(3): 375–390, 1998.
6. CF Bartel Jr. Fundamentals of Motion Control. Assembly, April 1997, pp 42–46.
7. BW Rust, WR Burris. Mathematical Programming and the Numerical Solution of Linear Equations. New York: Elsevier, 1972.
8. EL Hall. Computer Image Processing and Recognition. New York: Academic Press, 1979, pp 555–567.
9. FP Beer, ER Johnson Jr. Vector Mechanics for Engineers. New York: McGraw-Hill, 1988, pp 946–948.
10. NS Nise. Control Systems Engineering. Redwood City, CA: Benjamin/Cummings, 1995, pp 117–150.
11. J Tal. Motion Control by Microprocessors. Palo Alto, CA: Galil Motion Control, 1989, pp 63, 64.

Chapter 2.3

In-Process Measurement

William E. Barkman
Lockheed Martin Energy Systems, Inc., Oak Ridge, Tennessee

3.1 INTRODUCTION

Manufacturing operations are driven by cost requirements that relate to the value of a particular product to the marketplace. Given this selling price, the system works backward to determine what resources can be allocated to the manufacturing portion of the cost equation. Then, production personnel set up the necessary resources and provide the workpieces that are consumed by the market. Everyone is happy until something changes. Unfortunately, the time constant associated with change in the manufacturing world is usually very short. Requirements often change even before a system begins producing parts and even after production is underway there are typically many sources of variability that impact the cost/quality of the operation. Variability associated with scheduling changes must be accommodated by designing flexibility into the basic manufacturing systems. However, the variability that is related to changing process conditions must be handled by altering system performance at a more basic level.

Error conditions often occur where one or more critical process parameters deviates significantly from the expected value and the process quality is degraded. The sensitivity of the process to these variations in operating conditions depends on the point in the overall manufacturing cycle at which they occur as well as the specific characteristics of a particular process disturbance. Amplitude, a frequency of occurrence, and a direction typically characterize these process errors. In a machining operation, the typical result is a lack of synchronization between the tool and part locations so that erroneous dimensions are produced.

Over time, the amplitudes of process errors are typically limited to a specific range either by their inherent nature or by operator actions. For example, shop temperature profiles tend to follow a specific pattern from day to day, component deflections are directly related to cutting forces, and cutting tools are replaced as they wear out. As multiple process error sources interact, the result is typically a seemingly random distribution of performance characteristics with a given "normal range" that defines the routine tolerances that are achievable with a given set of operations. On the other hand, trends such as increasing operating temperatures due to a heavy workload, coolant degradation, component wear, etc. have a nonrandom component that continues over time until an adjustment is made or a component is replaced.

One solution to the problem of process variation is to build a system that is insensitive to all disturbances; unfortunately, this is rarely practical. A more realistic approach is to use a manufacturing model that defines the appropriate response to a particular process parameter change. This technique can be very successful if the necessary monitoring systems are in place to measure what is really happening within the various manufacturing operations. This approach works because manufacturing processes are deterministic in nature: a cause-and-effect relationship exists between the output of the process and the process parameters. Events occur due to specific causes, not random chance, even though an observer may not recognize the driving force behind a particular action. If the key process characteristics are maintained at a steady-state level then the process output will also remain relatively constant. Conversely, when the process parameters change significantly, the end product is also affected in a noticeable manner.

Recognizing the deterministic nature of manufacturing operations leads to improvements in product quality and lowers production costs. This is accomplished by measuring the important process parameters in real time and performing appropriate adjustments in the system commands. Moving beyond intelligent alterations in control parameters, parts can also be "flagged" or the process halted, as appropriate, when excessive shifts occur in the key process variables. In addition, when an accurate system model is available, this real-time information can also lead to automatic process certification coupled with "sample" certification of process output and the full integration of machining and inspection.

The system elements necessary to accomplish this are an operational strategy or model that establishes acceptable limits of variability and the appropriate response when these conditions are exceeded, a means of measuring change within the process, plus a mechanism for inputting the necessary corrective response. This chapter discusses the selection of the key process measurements, the monitoring of the appropriate process information, and the use of this measurement data to improve process performance.

3.2 PROCESS VARIATION

An important goal in manufacturing is to reduce the process variability and bias to as small a level as is economically justifiable. Process bias is the difference between a parameter's average value and the desired value. Bias errors are a steady-state deviation from an intended target and while they do cause unacceptable product, they can be dealt with through calibration procedures. On the other hand, process variability is a continuously changing phenomenon that is caused by alterations in one or more manufacturing process parameters. It is inherently unpredictable and therefore more difficult to accommodate. Fortunately, real-time process parameter measurements can provide the information needed to deal with unexpected excursions in manufacturing system output. This extension of conventional closed-loop process control is not a complex concept; however, the collection of the necessary process data can be a challenge.

Process variability hinders the efforts of system operators to control the quality and cost of manufacturing operations. This basic manufacturing characteristic is caused by the inability of a manufacturing system to do the same thing at all times, under all conditions. Examples of variability are easily recognized in activities such as flipping a coin and attempting to always get a "heads" or attempting to always select the same card from a complete deck of cards. Machining operations typically exhibit a much higher degree of process control. However, variability is still present in relatively simple operations such as attempting to control a feature diameter and surface finish without maintaining a constant depth of cut, coolant condition/temperature, tooling quality, etc.

Inspecting parts and monitoring the value of various process parameters under different operating conditions collects process variability data. The answers to the following questions provide a starting point in beginning to deal with process variability: What parameters can and should be measured, how much variation is acceptable, is bias a problem (it is usually a calibration issue), what supporting inspection data is required, and does the process model accurately predict the system operation?

Error budgets [1] are an excellent tool for answering many of these questions. It is rarely possible or cost effective to eliminate all the sources of variability in a manufacturing process. However, an error budget provides a structured approach to characterizing system errors, understanding the impact of altering the magnitudes of the various errors, and selecting a viable approach for meeting the desired performance goals. The error budgeting process is based on the assumption that the total process error is composed of a number of individual error components that combine in a predictable manner to create the total system error. The identification and characterization of these error elements and the understanding of their impact on the overall process quality leads to a system model that supports rational decisions on where process improvement efforts should be concentrated.

The procedure for obtaining a viable error budget begins with the identification and characterization of the system errors, the selection of a combinatorial rule for combining the individual errors into a total process error, and the validation of this model through experimental testing. The system model is obtained by conducting a series of experiments in which a relationship is established between individual process parameters and the quality of the workpiece. In a machining operation this involves fabricating a series of parts while keeping all parameters but one at a constant condition. For instance, tool wear can be measured by making a series of identical cuts without changing the cutting tool. Wear measurements made between machining passes provide a wear hsitory that is useful in predicing tool performance. In a similar fashion, a series of diameters can be machined over time (using a tool–workpiece combination that does not exhibit significant wear) without attempting to control the temperature of the coolant. This will produce temperature sensitivity data that can be used to define the degree of temperature control required to achieve a particular workpiece tolerance.

After all the process error sources have been characterized, it is necessary to combine them in some intelligent fashion and determine if this provides an accurate prediction of part quality. Since all errors are not present at the same time, and because some errors will counteract each other it is overly conservative to estimate process performance by simply adding together all the maximum values of the individual error sources. Lawrence Livermore National Laboratory (LLNL) has been quite successful in predicting the performance of precision machine tools using a root-mean-square method for combining the individual error elements into an overall performance predictor [2]. An excellent example of the application of the error budget technique is the LLNL large optics diamond turning machine shown in Fig. 1.

Once the system error model has been validated, a reliable assessment can be made of the impact of reducing, eliminating, or applying a suitable compensation technique to the different error components. Following a cost estimate of the resources required to achieve the elimination (or appropriate reduction) of the various error sources, a suitable course of action can be planned. In general, it is desirable to attempt to reduce the amplitudes of those error sources that can be made relatively small (10% of the remaining dominant error) with only a modest effort. For example, if a single easily corrected error source (or group of error sources) causes 75% of a product feature's error then it is a straightforward decision on how to proceed. Conversely, if this error source is very expensive to eliminate then it may be inappropriate to attempt to achieve the desired tolerances with the proposed equipment. In this case, it is necessary to reevaluate the desired objectives and processing methods and consider alternative approaches. Obviously, a critical element in

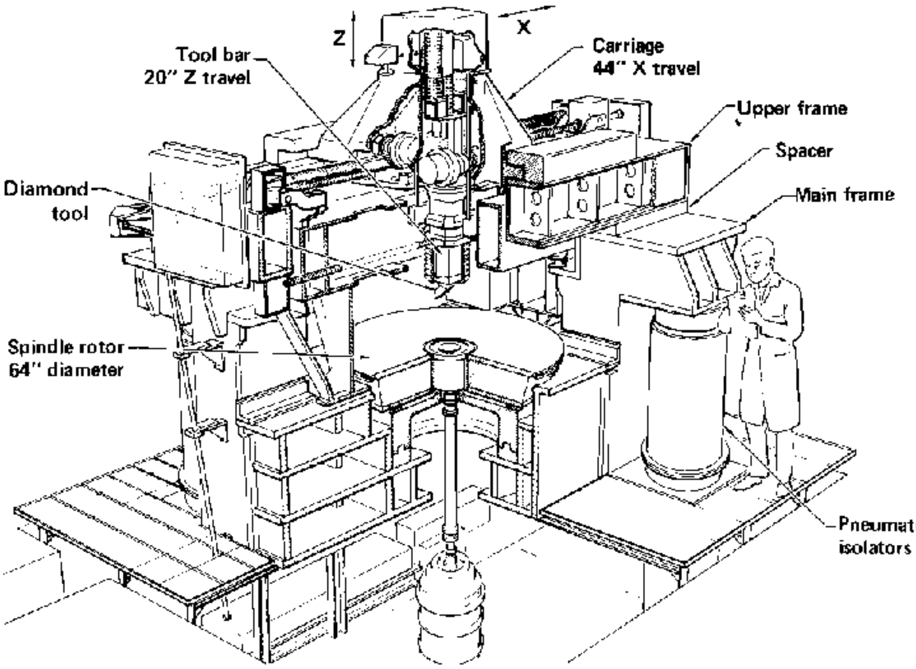


Figure 1 Artist's concept of initial large optics diamond turning machine design (courtesy of Lawrence Livermore National Laboratory).

the above process is the ability to isolate and measure individual process errors.

3.3 IN-PROCESS MEASUREMENTS FOR PROCESS CONTROL

As mentioned above, process parameter information can be used to monitor the condition of a manufacturing operation as well as provide a process control signal to a feedback algorithm. For example, the accuracy of a shaft diameter feature can be enhanced by correcting for cutting tool wear. If errors due to component deflection, machine geometry, etc. are relatively constant, then tool offsets based on the condition of the cutting tool can improve the system performance. At the same time, tool offset variability is introduced by the system operator's inability to determine the amount of compensation needed. If adjustments are made based on historical data, then the system is vulnerable to unexpected changes in factors such as tool performance, material characteristics, operator-induced changes in feeds and speeds, etc. Offsets that are based on product certification results are a little better, since there is a closer tie to the "current process," but the delay between production and inspection can still cause difficulties. In-process measurements offer the best alternative as long as the time required to collect the data is not an unacceptable cost to the production operations. In order to be useful, the in-process measurement data must be easily obtained, an accurate predictor of system performance, and useful to the process operator. Measurement processes that do not meet these criteria provide little, if any, value and only harm the relationship between the shop and the organization that has supported this alteration to the previous manufacturing process.

Figure 2 is an example of a machine tool that uses in-process measurement data to improve the quality of turned workpieces. This machine uses the tool set cycle depicted in Fig. 3 to establish the relationship between the cutting tool and the spindle face and centerline. This avoids the necessity of "touching up" on the part whenever tools are changed and also automatically compensates for tool wear that occurs in the direction of the machine axes. Of course, tool wear occurs at all contact points between the tool and workpiece, and this tool setting algorithm does not compensate for wear or size errors that occur between the tool set locations. This can result in appreciable part errors when using a round-nose tool to machine a tapered section like the one shown in Fig. 3. This occurs

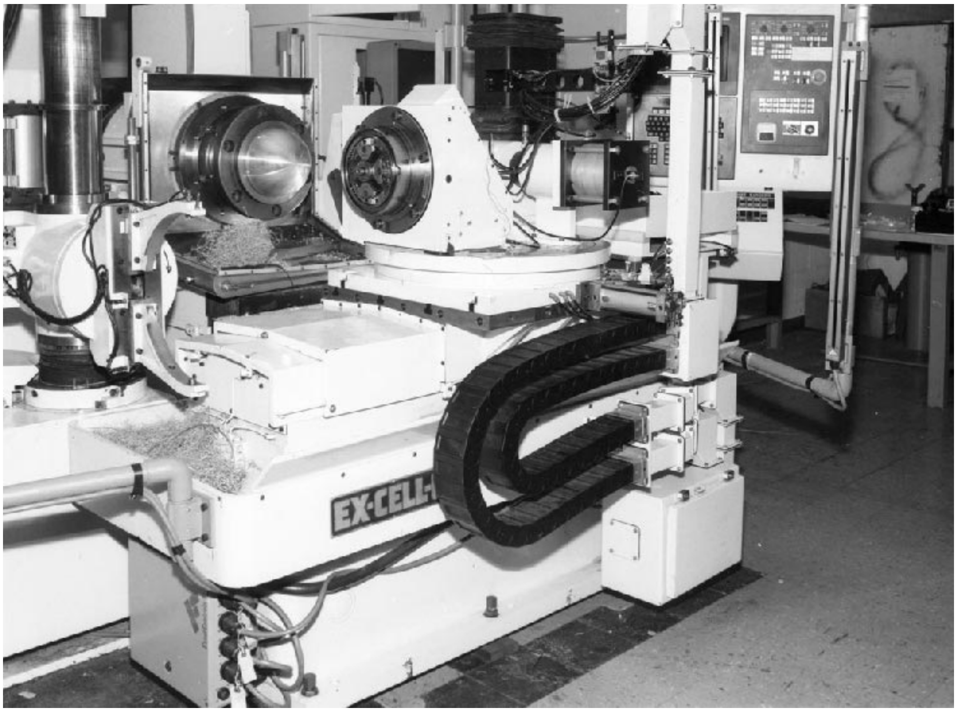


Figure 2 Advanced turning machine with tool and part measurement capability.

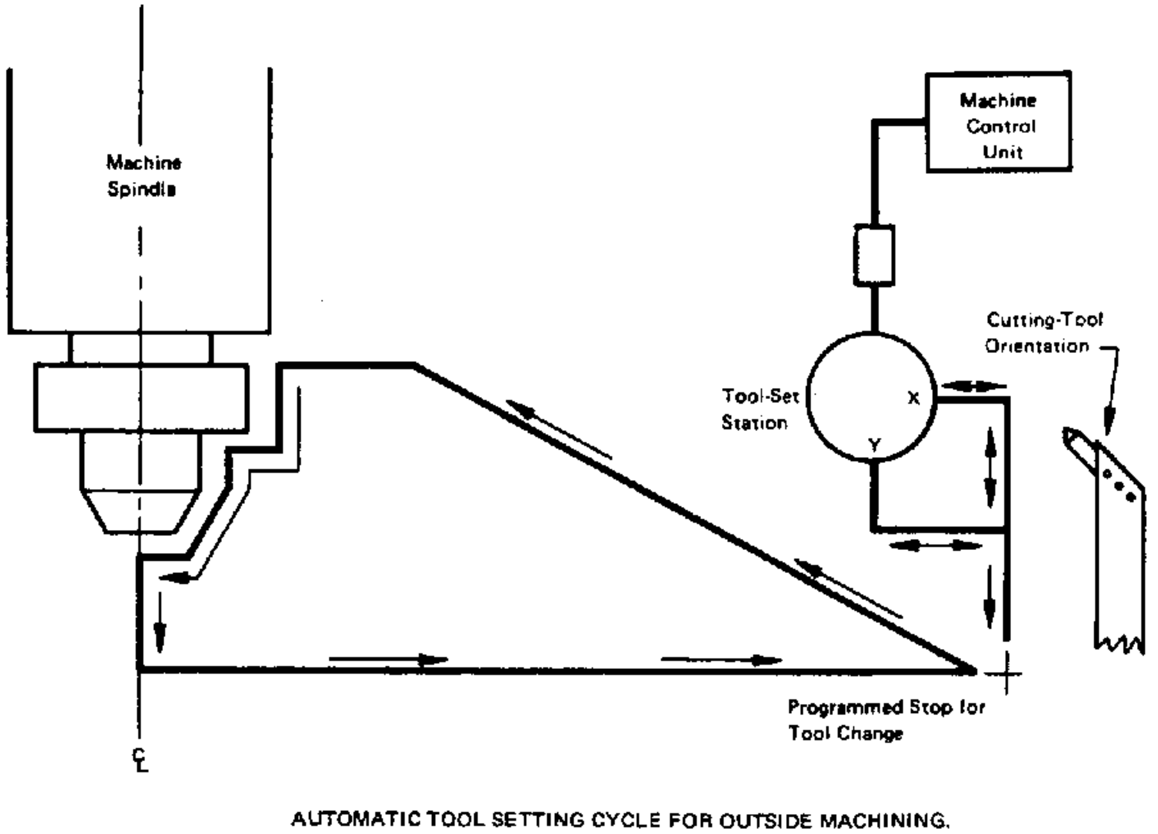

Machine Spindle
Machine Control Unit
Cutting-Tool Orientation
Tool-Set Station
X
Y
Programmed Stop for Tool Change
℄
AUTOMATIC TOOL SETTING CYCLE FOR OUTSIDE MACHINING.

Figure 3 Tool set cycle for establishing relative tool position.

because the location of the physical tool edge may not match the theoretical location associated with a particular radius tool. (For a 45° taper, the cutting tool point of contact would be midway between the two tool set points and the associated error is approximately 40% of the difference in the theoretical and actual tool radii.)

If only a single taper is being machined then an additional tool set cycle could be implemented to determine the tool condition for that particular point of contact. However, a more general solution is possible using an on-machine tool size and shape measurement system. The advanced turning machine shown earlier in Fig. 2 uses an on-machine camera to inspect the size, shape, and location of a cutting tool after the cutter is mounted on the machine's boring bar. The camera measures the location of the tool edge in 1° increments around the nose of the tool and calculates a tool set location as well as an effective tool shape. (The effective tool shape is the actual tool shape adjusted for the projected point of contact as the tool moves around the part.) This is necessary for profiling operations because the cutter's "high points" contact a greater segment of a workpiece than what is "seen by low regions."

The effective tool shape data, obtained in the on-machine tool inspection operation, is used to automatically adjust the theoretical tool path that was produced using the theoretical tool shape. Preliminary machining tests were conducted using worn cutters that exhibited shape errors similar to the data shown in Fig. 4. The circular profile is a relatively challenging shape for a Cartesian-coordinate machine and the results demonstrated that the tool inspection and compensation process could be used to compensate for significant cutter errors.

Additional tests were conducted to evaluate the robustness of the system. In this case, the machine was programmed to produce a spherical contour using a 0.021 in. radius cutter. However, the test was conducted using a 0.032 in. cutter instead of the correctly sized tool. This results in a theoretical error of approximately 0.0045 in. at the point midway between the pole and equator of the test part. Figure 5 shows

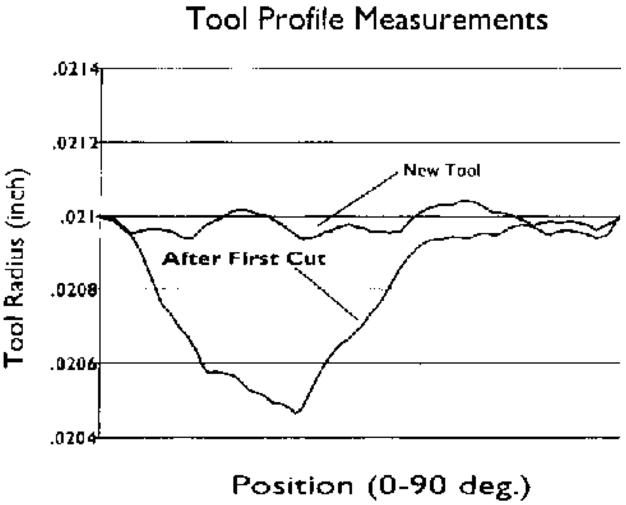


Figure 4 Worn tool shape errors.

the results obtained in this machining test. The profile errors were as expected when no tool path compensation was used. A very significant contour improvement was obtained when the compensation was implemented.

The above example demonstrates many of the concepts discussed throughout this chapter. The machine tool performance was initially tested using an aluminum workpiece, a single-point diamond tool and a coolant temperature control system. The early tests focused on the sensitivity of the system to excursions in the machine coolant. An experiment was conducted in which the coolant temperature was driven through a series of step changes over a 12 hr period. During this time, the machine was moved through a series of simulated tool paths, but no machining was done, so that the part dimensions were only affected by the coolant temperature. Figure 6 shows the temperature response of various machine components plotted along with the coolant temperature. Figure 7 shows the part dimensional response to the temperature changes. This verifies the need to maintain good control of the coolant temperature.

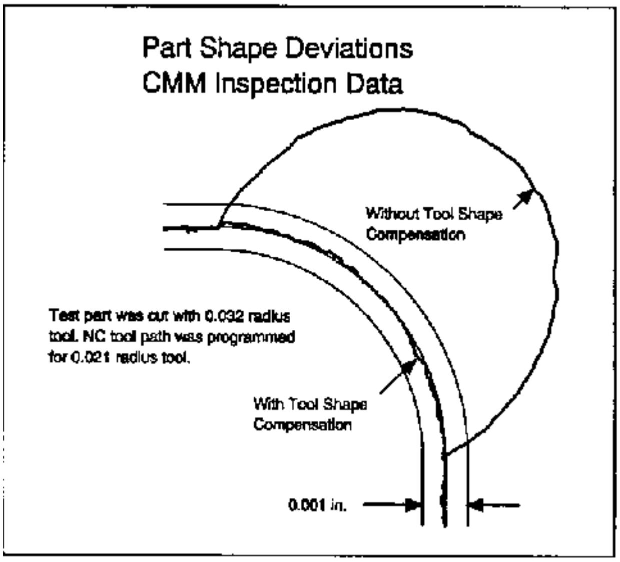

Part Shape Deviations
CMM Inspection Data
Without Tool Shape Compensation
Test part was cut with 0.032 radius tool. NC tool path was programmed for 0.021 radius tool.
With Tool Shape Compensation
0.001 in.

Figure 5 Workpiece inspection results for test using incorrect cutter size.

Additional tests were performed with the coolant temperature control system activated. It was demonstrated that under the relatively ideal cutting conditions, the machine was capable of producing a shape accuracy of approximately 0.0002 in. on a spherical contour. When the workpiece and cutting-tool materials were changed to stainless steel and tungsten carbide respectively, the machined contour was degraded to about 0.002 in. This demonstrated that the most significant error with respect to workpiece contour was the cutting tool wear. Fortunately, it was also noted that the majority of the tool wear occurred on the first pass and the tool was relatively stable for a number of additional machining passes.

Figure 8 shows the tool form errors associated with two machining passes on two different tools. In both cases the wear pattern is essentially unchanged by the second machining operation. This lead to the concept of inspecting the tool form after an initial "wear-in" pass, adjusting the tool path for the effective shape of the worn tool, and then performing the finish-machining operation with the compensated tool path.

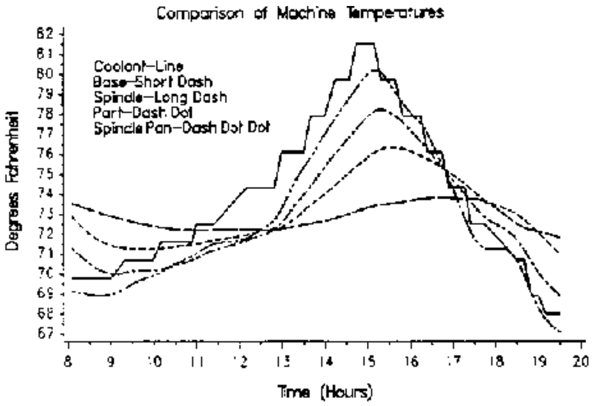


Figure 6 Machine temperature measurements.

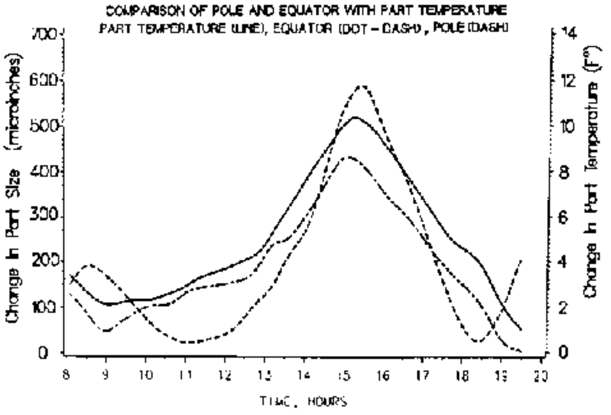


Figure 7 Machine response to temperature excursion.

After significantly reducing the errors associated with the cutting tool, additional tests were conducted to characterize the remaining system error components. It was noted that while the shape of the contour was signficantly improved, there were still occasional size errors that were larger than desired. These size errors were traceable to the cumulative effects of the drift of the tool set station, component deflection, and a relatively small amount of additional tool wear. The solution to this problem was to use an in-process probe to measure the part diameter and perform the appropriate tool offset prior to the finish-machining pass.

3.4 IN-PROCESS MEASUREMENTS FOR PROCESS QUALIFICATION

In addition to improving the accuracy and consistency of manufacturing operations, in-process measurements of critical parameters can be used to provide real-time assurance that the workpiece quality is being maintained at the desired level. Aside from the obvious step of measuring one or more critical dimensions on a finished workpiece, additional process data can be collected that qualifies the process before the part is removed from the machine tool. In the case of the advanced turning machine described above, the machining process was shown to be very repeatable as long as certain key elements were maintained at a constant level. Process consistency was accomplished by focusing on the machine stability and the condition of the cutting tool. The deflection/size errors associated with thermal gradients were avoided by controlling the temperature of the cutting fluid. Tool wear errors were minimized by avoiding the use of a new tool on the finish-machining pass; and in-process inspection cycles were added to correct for errors in initial tool position as well as tool form.

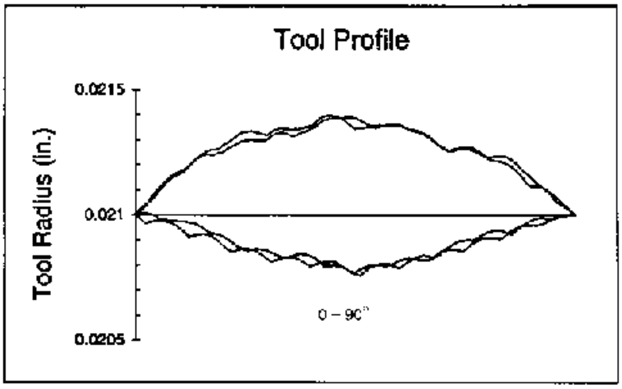


Figure 8 Tool wear comparison data for two different tools.

Each of these operations contributes to the overall accuracy of the system and is also a potential source of disruption to production operations. If the temperature compensation system malfunctions and introduces temperature gradients instead of eliminating them then the machine's tool path accuracy will be degraded. Similarly, if the tool measurement and compensation system or the part probing operation is erratic or awkward to use then it will not add value to the overall process.

Instead of waiting for a post-process inspection step to detect a potential system malfunction, each of these subsystems can be monitored in real time to provide assurance that the correct compensation actions are implemented. The temperature control system can be checked easily by tracking the amplitude of the "temperature-following error." If the difference between the desired coolant temperature and the actual coolant temperature becomes excessive then there is probably an error condition and the system should be halted and checked at an appropariate time.

Monitoring the gaging system is also straightforward. In both tool- and part-gaging operations, an artifact can be checked each time a gaging operation is initiated and the result compared with historical values to estimate the quality of the current measurements. A 1 in. diameter ball is a useful monitor part for the probing system because the ball diameter measurements are sensitive to errors in both probe height and probe repeatability. In a similar fashion, a small gage wire that is permanently mounted in the field of view of the tool measurement camera can provide a viable reference measurement. Artifacts such as these maintain constant dimensions over time and offer a good means of verifying system repeatability and validating the quality of the current measurement process.

Further process performance data can also be gained by comparing the in-process measurement values with post-process certification data. Eventually, sufficient data can be collected to establish a statistical basis for reducing the amount of post-process inspection operations in favor of process certification. Of course, it is generally not appropriate to transfer the inspection burden from the downstream gages to the on-machine systems. This merely creates a pinch point farther upstream in the process. Instead, it is necessary to monitor those critical process parameters that can be used as quality predictors without negatively impacting process throughput.

Additional process information is available by comparing parameters that are common between many part families. The differences between actual and intended dimensions which are common features to multiple part families is a useful technique for tracking process quality in an environment in which the part mix is constantly changing. Deviations in certain part characteristics such as length errors (or diameter errors) can be compared as a measure of system performance and the suitability of cutter offsets. Even though the part sizes may vary widely between different workpieces, the ability of the system to control common features such as a diameter or length is an important system attribute and can be tracked using control charts.

Eventually a model can be constructed that defines the appropriate machining conditions for producing a high-quality product. This model might include the typical amount of tool wear and offsets required for a particular operation as well as the limits that define when external corrective action is required to restore process viability. During the machining cycle, process characteristics such as the size of the cutter offset, the size of key features at an intermediate processing stage, the amount of tool wear on a given pass, etc. can be used to verify that the operation is performing as expected. If all of the critical process attributes fall within the model limits then the process output can be expected to be similar to what has been achieved in the past. However, if one or more of the important system parameters is out of the control limits defined by the process model, then external actions are probably required to restore system performance.

The advanced turning machine mentioned above is an example of how this technique can be applied. This machine can produce complex profiles that require sophisticated inspection machines for product certification yet process performance can be accurately predicted by monitoring a few key parameters. Barring a mechanical or electrical breakdown, the machine's geometry accuracy is quite good as long as there are no temperature gradients in the structure. Monitoring the coolant temperature control sytem gives an accurate prediction of the machine tool path accuracy. Using on-machine probing to compare the size of a small number of features to historical performance records validates the suitability of tool offsets, and changes in tool form define the amount of uncompensated tool wear that can degrade the part quality.

REFERENCES

1. WE Barkman. In-Process Quality Control for Manufacturing. New York: Marcel Dekker, 1989, pp 89–92.
2. RR Donaldson. Large optics diamond turning machine, vol I, final report. Lawrence Livermore National Laboratory, UCRL-52812, Livermore, CA, 1979.

Chapter 3.1

Distributed Control Systems

Dobrivoje Popovic
University of Bremen, Bremen, Germany

1.1 INTRODUCTION

The evolution of plant automation systems, from very primitive forms up to the contemporary complex architectures, has closely followed the progress in instrumentation and computer technology that, in turn, has given the impetus to the vendor to update the system concepts in order to meet the user's growing requirements. This has directly encouraged users to enlarge the automation objectives in the field and to embed them into the broad objectives of the process, production, and enterprise level. The *integrated automation concept* [1] has been created to encompass all the automation functions of the company. This was viewed as an opportunity to optimally solve some interrelated problems such as the efficient utilization of resources, production profitability, product quality, human safety, and environmental demands.

Contemporary industrial plants are inherently complex, large-scale systems requiring complex, mutually conflicting automation objectives to be simultaneously met. Effective control of such systems can only be made feasible using adequately organized, complex, large-scale automation systems like the distributed computer control systems [2] (Fig. 1). This has for a long time been recognized in steel production plants, where 10 million tons per annum are produced, based on the operation of numerous work zones and the associated subsystems like:

Iron zone with coke oven, palletizing and sintering plant, and blast furnace
Steel zone with basic oxygen and electric arc furnace, direct reduction, and continuous casting plant, etc.
Mill zone with hot and cold strip mills, plate bore, and wire and wire rod mill.

To this, the laboratory services and the plant care control level should be added, where all the required calculations and administrative data processing are carried out, statistical reviews prepared, and market prognostics data generated. Typical laboratory services are the:

Test field
Quality control
Analysis laboratory
Energy management center
Maintenance and repair department
Control and computer center

and typical utilities:

Gas and liquid fuel distribution
Oxygen generation and distribution
Chilled water and compressed air distribution
Water treatment
Steam boiler and steam distribution
Power generation and dispatch.

The difficulty of control and management of complex plants is further complicated by permanent necessity of

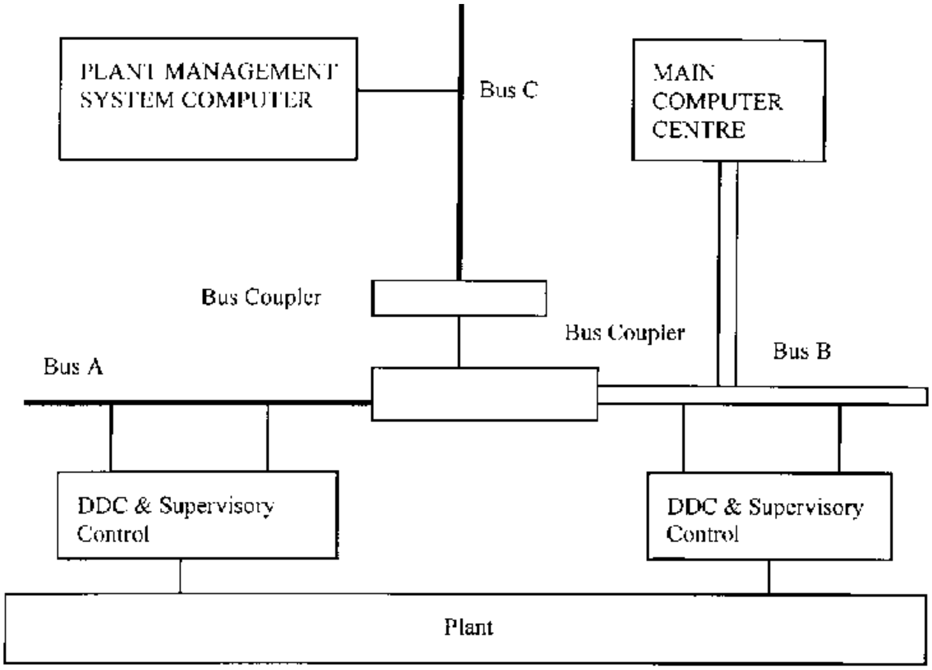

PLANT MANAGEMENT SYSTEM COMPUTER
Bus C
MAIN COMPUTER CENTRE
Bus Coupler
Bus Coupler
Bus A
Bus B
DDC & Supervisory Control
DDC & Supervisory Control
Plant

Figure 1 Distributed computer control system.

steady adaptation to the changing demands, particularly due to the quality variations in the raw materials and the fact that, although the individual subsystems are specific batch-processing plants, they are firmly incorporated into the downstream and upstream processes of the main plant. This implies that the integrated plant automation system has to control, coordinate, and schedule the total plant production process.

On the other hand, the complexity of the hierarchical structure of the plant automation is further expanding because the majority of individual subplants involved are themselves hierarchically organized, like the ore yard, coke oven, sintering plant, BOF/LD (Basic Oxygen Furnace LD-Converter) converter, electric arc furnace, continuous casting, etc.

Onshore and offshore oil and gas fields represent another typical example of distributed, hierarchically organized plants requiring similar automation concepts. For instance, a typical onshore oil and gas production plant consists of a number of oil and gas gathering and separation centers, serving a number of remote degassing stations, where the crude oil and industrial gas is produced to be distributed via long-distance pipelines. The gas production includes gas compression, dehydration, and purification of liquid components.

The remote degassing stations, usually unmanned and completely autonomous, have to be equipped with both multiloop controllers and remote terminal units that should periodically transfer the data, status, and alarm reports to the central computer. These stations should be able to continue to operate also when the communication link to the central computer fails. This is also the case with the gathering and separation centers that have to be equipped with independent microcomputer-based controllers [3] that, when the communication link breaks down, have to automatically start running a preprogrammed, failsafe routine. An offshore oil and gas production installation usually consists of a number of bridge-linked platforms for drilling and production, each platform being able to produce 100,000 or more barrels of crude oil per day and an adequate quantity of compressed and preprocessed gas. Attached to the platforms, beside the drilling modules, are also the water treatment and mud handling modules, power generation facilities, and other utilities.

In order to acquire, preprocess, and transfer the sensing data to the central computer and to obtain control commands from there, a communication link is required and at the platform a supervisory control data acquisition system (SCADA). An additional link is requried for interconnection of platforms for exchange of coordination data.

Finally, a very illustrative example of a distributed, hierarchically organized system is the power system in which the power-generating and power-distributing subsystems are integrated. Here, in the power plant itself, different subsystems are recognizable, like air, gas, combustion, water, steam, cooling, turbine, and generator subsystems. The subsystems are hierarchically organized and functionally grouped into:

Drive-level subsystem
Subgroup-level subsystem
Group-level subsystem
Unit-level subsystem.

1.2 CLASSICAL APPROACH TO PLANT AUTOMATION

Industrial plant automation has in the past undergone three main development phases:

Manual control
Controller-based control
Computer-based control.

The transitions between the individual automation phases have been so vague that even modern automation systems still integrate all three types of control.

At the dawn of industrial revolution and for a long time after, the only kind of automation available was the mechanization of some operations on the production line. Plants were mainly supervised and controlled manually. Using primitive indicating instruments, installed in the field, the plant operator was able to adequately manipulate the likely primitive actuators, in order to conduct the production process and avoid critical situations.

The application of real automatic control instrumentation was, in fact, not possible until the 1930s and 40s, with the availability of pneumatic, hydraulic, and electrical process instrumentation elements such as sensors for a variety of process variables, actuators, and the basic PID controllers. At this initial stage of development it was possible to close the control loop for flow, level, speed, pressure, or temperature control in the field (Fig. 2). In this way, the plants steadily became more and more equipped with field control instrumentation, widely distributed through the plant, able to indicate, record, and/or control individual process variables. In such a constellation, the duty of the plant operator was to monitor periodically the indicated measured values and to preselect and set the controlling set-point values.

Yet, the real breakthrough in this role of the plant operator in industrial automation was achieved in the 1950s by introducing electrical sensors, transducers,

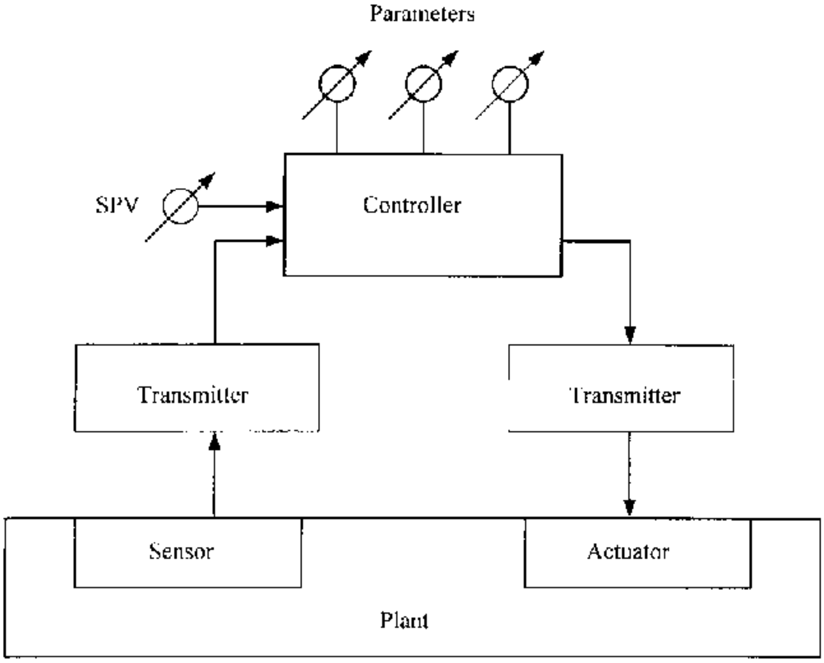


Figure 2 Closed-loop control.

actuators, and, above all, by placing the plant instrumentation in the *central control room* of the plant. In this way, the possibility was given to supervise and control the plant from one single location using some monitoring and command facilities. In fact, the introduction of automatic controllers has mainly shifted the responsibility of the plant operator from manipulating the *actuating* values to the adjustment of controllers' *set-point values*. In this way the operator became a supervisory controller.

In the field of plant instrumentation, the particular evolutionary periods have been marked by the respective state-of-the art of the available instrumentation technology, so that here an instrumentation period is identifiable that is:

Pneumatic and hydraulic
Electrical and electronic
Computer based.

The period of pneumatic and hydraulic plant instrumentation was, no doubt, technologically rather primitive because the instrumentation elements used were of low computational precision. They, nevertheless, have still been highly reliable and—above all—*explosion proof*, so that they are presently still in use, at least in the appropriate control zones of the plant.

Essential progress in industrial plant control has been made by introducing electrical and electronic instrumentation, which has enabled the implementation of advanced control algorithms (besides PID, also cascaded, ratio, nonlinear, etc. control), and considerably facilitated automatic tuning of control parameters. This has been made possible particularly through the computer-based implementation of individual control loops (Fig. 3).

The idea of centralization of plant monitoring and control facilities was implemented by introducing the concept of a central control room in the plant, in which the majority of plant control instrumentation, with the exception of sensors and actuators, is placed. For connecting the field instrumentation elements to the central control room pneumatic and electrical data transmission lines have been installed within the plant. The operation of the plant from the central control room is based on indicating, recording, and alarm elements, situated there, as well as—for better local orientation—on the use of plant mimic diagrams. The use of plant mimic diagrams has proven to be so useful that they are presently still in use. Microcomputers, usually programmed to solve some data acquisition and/or control problems in the field, have been connected, along with other instrumentation elements, to the facilities of the central control room, where the plant operators are in charge of centralized plant monitoring and process control.

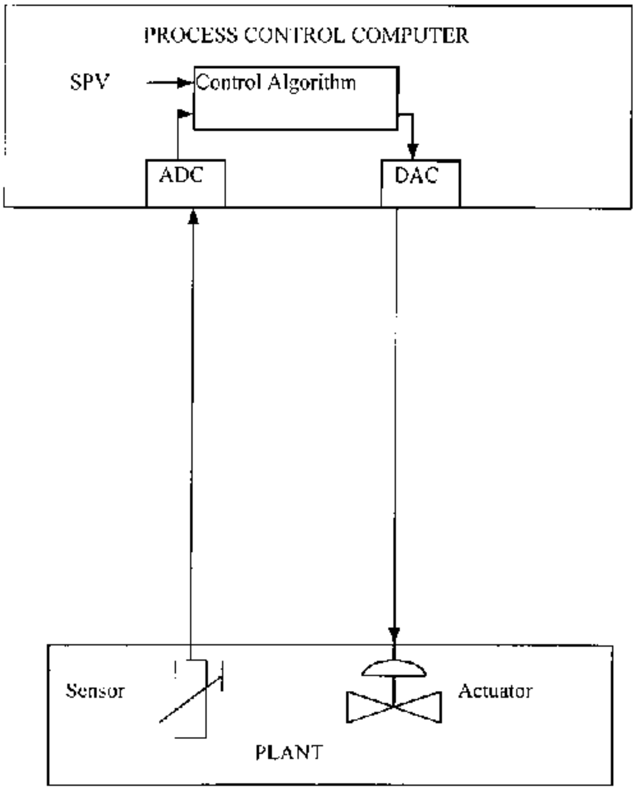


Figure 3 Computer-based control loop.

Closed-loop control is essential for keeping the values of process variables, in spite of internal and external disturbing influences, at prescribed, set-point values, particularly when the control parameters are optimally tuned to the process parameters. In industrial practice, the most favored approach for control parameter tuning is the Ziegler–Nichols method, the application of which is based on some simplified relations and some recommended tables as a guide for determination of the optimal step transition of the loop while keeping its stability margin within some given limits. The method is basically applicable to the stationary, time-invariant processes for which the values of relevant process parameters are known; the control parameters of the loop can be tuned offline. This cannot always hold, so the control parameters have to be optimally tuned using a kind of trial-and-error approach, called the Ziegler–Nichols test. It is an open-loop test through which the pure delay of the loop and its ‘‘reaction rate’’ can be determined, based on which the optimal controller tuning can be undertaken.

1.3 COMPUTER-BASED PLANT AUTOMATION CONCEPTS

Industrial automation has generally been understood as an engineering approach to the control of systems such as power, chemical, petrochemical, cement, steel, water and wastewater treatment, and manufacturing plants [4,5].

The initial automation objectives were relatively simple, reduced to automatic control of a few process variables or a few plant parameters. Over the years, there has been an increasing trend toward simultaneous control of more and more (or of all) process variables in larger and more complex industrial plants. In addition, the automation technology has had to provide a better view of the plant and process state, required for better monitoring and operation of the plant, and for improvement of plant performance and product quality. The close cooperation between the plant designer and the control engineer has, again, directly contributed to the development of better instrumentation, and opened perspectives to implement larger and more complex production units and to run them at full capacity, by guaranteeing high product quality. Moreover, the automation technology is presently used as a valuable tool for solving crucial enterprise problems, and interrelating simultaneous solution of process and production control problems along with the accompanying financial and organizational problems.

Generally speaking, the principal objectives of plant automation are to monitor information flow and to manipulate the material and energy flow within the plant in the sense of optimal balance between the product quality and the economic factors. This means meeting a number of contradictory requirements such as [3]:

Maximal use of production capacity at highest possible production speed in order to achieve maximal production yield of the plant

Maximal reduction of production costs by

  Energy and raw material saving

  Saving of labor costs by reducing the required staff and staff qualification

  Reduction of required storage and inventory space and of transport facilities

  Using low-price raw materials while achieving the same product quality

Maximal improvement of product quality to meet the highest international standards while keeping the quality constant over the production time

Maximal increase of reliability, availability, and safety of plant operation by extensive plant monitoring, back-up measures, and explosion-proofing provisions

Exact meeting of governmental regulations concerning environmental pollution, the ignorance of which incurs financial penalties and might provoke social protest

Market-oriented production and customer-oriented production planning and scheduling in the sense of just-in-time production and the shortest response to customer inquiries.

Severe international competition in the marketplace and steadily rising labor, energy, and raw material costs force enterprise management to introduce advanced plant automation, that simultaneously includes the office automation, required for computer-aided market monitoring, customer services, production supervision and delivery terms checking, accelerated order processing, extensive financial balancing, etc. This is known as integrated enterprise automation and represents the highest automation level [1].

The use of *dedicated comptuers* to solve locally restricted automation problems was the initial computer-based approach to plant automation, introduced in the late 1950s and largely used in the 1960s. At that time the computer was viewed—mainly due to its low reliability and relatively high costs—not so much as a control instrument but rather as a powerful tool to solve some special, clearly defined problems of data acquisition and data processing, process monitoring, production recording, material and energy balancing, production reporting, alarm supervision, etc. This versatile capability of computers has also opened the possibility of their application to laboratory and test field automation.

As a rule, dedicated computers have individually been applied to partial plant automation, i.e., for automation of particular operational units or subsystems of the plant. Later on, one single large mainframe computer was placed in the central control room for centralized, computer-based plant automation. Using such computers, the majority of indicating, recording, and alarm-indicating elements, including the plant mimic diagrams, have been replaced by corresponding application software.

The advent of larger, faster, more reliable, and less expensive process control computers in the mid 1960s even encouraged vendors to place the majority of plant and production automation functions into the single central computer; this was possible due to the enormous progress in computer hardware and software, process and man–machine interface, etc.

However, in order to increase the reliability of the central computer system, some backup provisions have been necessary, such as backup controllers and logic circuits for automatic switching from the computer to the backup controller mode (Fig. 4) so that in the case of computer failure the controllers take over the last set-point values available in the computer and freeze them in the latches available for this purpose. The values can later on be manipulated by the plant operator in a similar way to conventional process control.

In addition, computer producers have been working on some more reliable computer system structures, usually in form of *twin* and *triple computer systems*. In this way, the required availability of a central control computer system of at least 99.95% of production time per year has enormously been increased. To this comes that the troubleshooting and repair time has dramatically been reduced through online diagnostic software, preventive maintenance, and twin-computer modularity of computer hardware, so that the number of really needed backup controllers has been reduced down to a small number of most critical ones.

The situation has suddenly been changed after the microcomputers have increasingly been exploited to solve the control problems. The 8-bit microcomputers, such as Intel's 8080 and Motorola's MC 6800, designed for bytewise data processing, have proved to be appropriate candidates for implementation of *programmable controllers* [6]. Moreover, the 16- and 32-bit microcomputer generation, to which Intel's 8088 and 8086, Motorola's 68000, Zilog's Z 8000 and many others belong, has even gained a relatively high respect within the automation community. They have worldwide been seen as an efficient instrumentation tool, extremely suitable to solve a variety of automation problems in a rather simple way. Their high reliability has placed them at the core of digital, single-loop and multiloop controllers, and has finally introduced the future trend in building automation systems by transferring more and more programmed control loops from the central computer into microcomputers, distributed in the field. Consequently, the duties left to the central computer have been less and less in the area of process control, but rather in the areas of higher-level functions of plant automation such as plant mon-

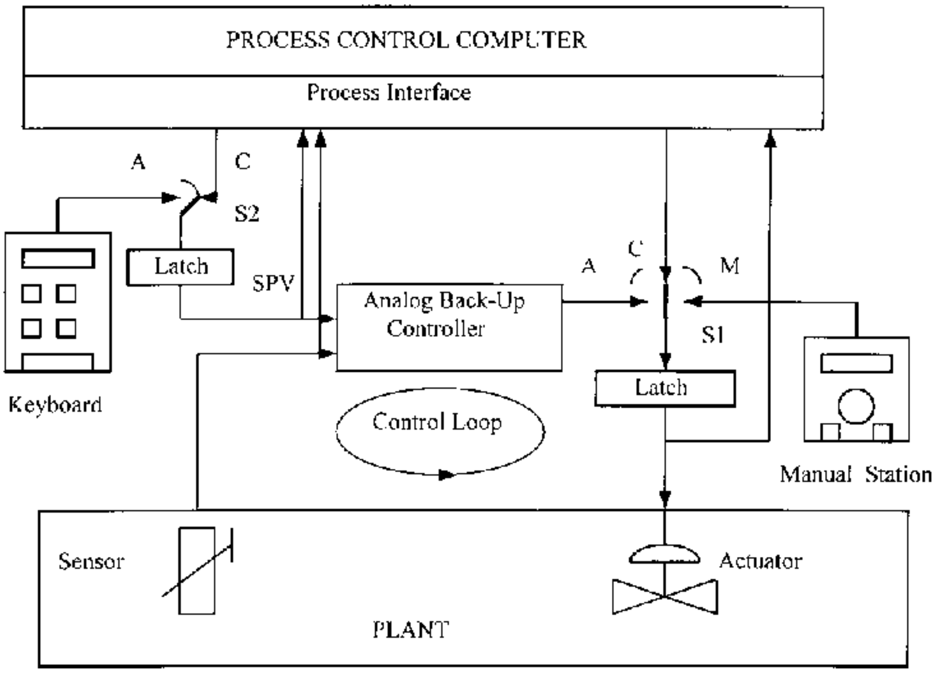


Figure 4 Backup controller mode.

itoring and supervision. This was the first step towards splitting up the functional architecture of a computer-based automation system into at least two hierarchical levels (Fig. 5):

Direct digital control
Plant monitoring and supervision.

The strong tendency to see the process and production control as a unit, typical in the 1970s, soon accelerated further architecture extension of computer-based automation systems by introducing an additional level on top of the process supervisory level: the *production scheduling and control* level. Later on, the need was identified for building the *centralized data files* of the enterprise, to better exploit the available production and storage resources within the production plant. Finally, it has been identified that direct access to the production and inventory files helps optimal production planning, customer order dispatching, and inventory control.

In order to integrate all these strongly interrelated requirements into one computer system, computer users and producers have come to the agreement that the structure of a computer system for integrated plant and production automation should be hierarchical, comprising at least the following hierarchical levels:

Process control
Plant supervision and control
Production planning and plant management.

This structure has also been professionally implemented by computer producers, who have launched an abundant spectrum of distributed computers control systems, e.g.:

ASEA MASTER (ASEA)
CENTUM (Yokogawa)
CONTRONIC P (Harman and Braun)
DCI 4000 (Fisher and Porter)
HIACS 3000 (Hitachi)
LOGISTAT CP 80 (AEG-Telefunken)
MOD 300 (Taylor Instruments)
PLS (Eckardt)
PMS (Ferranti)
PROCONTROL I (BBC)
PROVOX (Fisher Controls)
SPECTRUM (Foxboro)
TDC 3000 (Honeywell)
TeLEPERM M (Siemens)
TOSDIC (Toshiba).

1.4 AUTOMATION TECHNOLOGY

Development of distributed computer control systems evidently depends on the development of their essential parts: hardware, software, and communication links. Thus, to better conceive the real capabilities of modern automation systems it is necessary to review the technological level and the potential application possibilities of the individual parts as constituent subsystems.

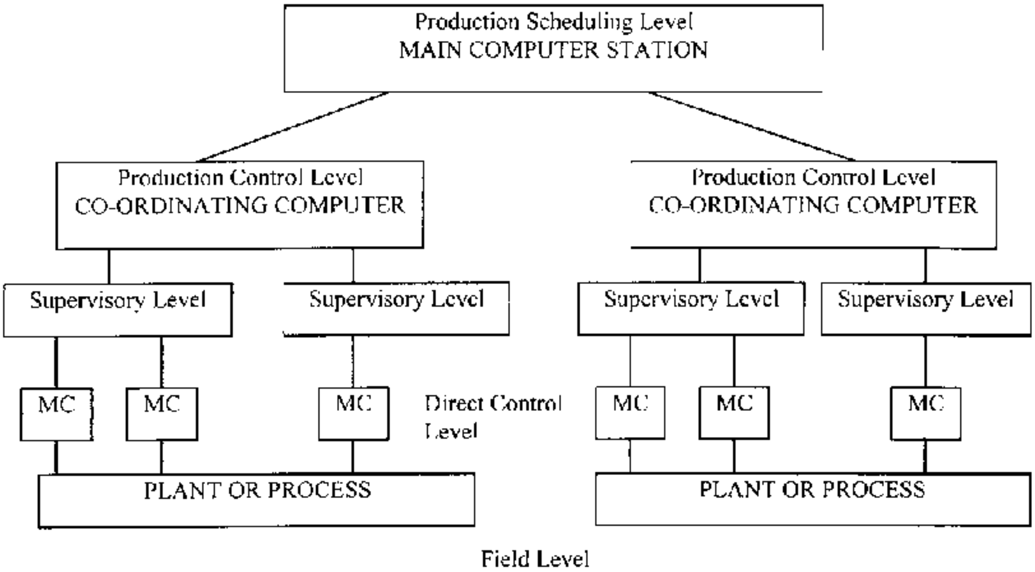


Figure 5 Hierarchical systems level diagram.

1.4.1 Computer Technology

For more than 10 years, the internal, bus-oriented Intel 80×86 and Motorola 680×0 microcomputer architectures have been the driving agents for development of a series of powerful microprocessors. However, the real computational power of processors came along with the innovative design of RISC (reduced instruction set computers) processors. Consequently, the RISC-based microcomputer concept has soon outperformed the *mainstream architecture*. Today, most frequently used RISC processors are the SPARC (Sun), Alpha (DEC), R4X00 (MIPS), and PA-RISC (Hewlett Packard).

Nevertheless, although being powerful, the RISC processor chips have not found a firm domicile within the mainstream PCs, but rather have become the core part of workstations and of similar computational facilities. Their relatively high price has decreased their market share, compared to microprocessor chips. Yet, the situation has recently been improved by introducing *emulation* possibilities that enable compatibility among different processors, so that RISC-based software can also run on conventional PCs. In addition, new microprocessor chips with the RISC architecture for new PCs, such as Power PC 601 and the like, also promote the use of RISCs in automation systems. Besides, the appearance of portable operating systems and the rapid growth the workstation market contributes to the steady decrease of price-to-performance ratio and thus to the acceptance of RISC processors for real-time computational systems.

For process control applications, of considerable importance was the Intel initiative to repeatedly modify its 80×86 architecture, which underwent an evolution in five successive phases, represented through the 8086 (a 5 MIPS, 29,000-transistor processor), 80286 (a 2 MIPS, 134,000-transistor processor), 80386 (an 8 MIPS, 175,000-transistor processor), 80486 (a 37 MIPS 1.2-million-transistor processor), up to the Pentium (a 112 and more MIPs, 3.1-million-transistor processor). Currently, even an over 300 MIPS version of the Pentium is commercially available.

Breaking the 100 MIPS barrier, up to then monopolized by the RISC processors, the Pentium has secured a threat-free future in the widest field of applications, relying on existing systems software, such as Unix, DOS, Windows, etc. This is a considerably lower requirement than writing new software to fit the RISC architecture. Besides, the availability of very advanced system software, such as operating systems like Windows NT, and of real-time and object-oriented languages, has essentially enlarged the application possibilities of PCs in direct process control, for which there is a wide choice of various software tools, kits, and tool boxes, powerfully supporting the computer-aided control systems design on the PCs. Real-time application programs developed in this way can also run on the same PCs, so that the PCs have finally become a constitutional part of modern distributed computer systems [7].

For distributed, hierarchically organized plant automation systems, of vital importance are the computer-based process-monitoring stations, the human–machine interfaces representing human windows into the process plant. The interfaces, mainly implemented as CRT-based color monitors with some connected *keyboard, joystick, mouse, lightpen*, and the like, are associated with individual plant automation levels to function as:

Plant operator interfaces, required for plant monitoring, alarm handling, failure diagnostics, and control interventions.

Production dispatch and production-monitoring interfaces, required for plant production management

Central monitoring interfaces, required for sales, administrative, and financial management of the enterprise.

Computer-based human–machine interfaces have functionally improved the features of the conventional plant monitoring and command facilities installed in the central control room of the plant, and completely replaced them there. The underlying philosophy of new plant-monitoring interfaces (that only those plant instrumentation details and only the process variables selected by the operator are presented on the screen) releases the operator from the visual saturation present in the conventional plant-monitoring rooms where a great number of indicating instruments, recorders, and mimic diagrams is permanently present and has to be continuously monitored. In this way the plant operator can concentrate on monitoring only those process variables requiring immediate intervention.

There is still another essential aspect of process monitoring and control that justifies abandoning the conventional concept of a central control room, where the indicating and recording elements are arranged according to the location of the corresponding sensors and/or control loops in the plant. This hampers the operator in a multialarm case in intervening accordingly because in this case the plant operator has to simultaneously monitor and operationally interrelate the alarmed, indicated, and required command values situated at a relative large mutual distance. Using the screen-oriented displays the plant operator can, upon request, simultaneously display a large number of process and control variables in any constellation. This kind of presentation can even—guided by the situation in the field—be automatically triggered by the computer.

It should be emphasized that the concept of modern human interfaces has been shaped, in cooperation between the vendor designers and the users, for years. During this time, the interfaces have evolved into flexible, versatile, intelligent, user-friendly workplaces, widely accepted in all industrial sectors throughout the world. The interfaces provide the user with a wide spectrum of beneficial features, such as:

Transparent and easily understandable display of alarm messages in chronological sequence that blink, flash, and/or change color to indicate the current alarm status

Display *scrolling* by advent of new alarm messages, while handling the previous ones

Mimic diagram displays showing different details of different parts of the plant by *paging*, *rolling*, *zooming*, etc.

Plant control using mimic diagrams

Short-time and long-time trend displays

Real-time and historical trend reports

Vertical multicolor bars, representing values of process and control variables, alarm limit values, operating restriction values, etc.

Menu-oriented operator guidance with multipurpose help and support tools.

1.4.2 Control Technology

The first computer control application was implemented as direct digital control (DDC) in which the computer was used as a multiloop controller to simultaneously implement tens and hundreds of control loops. In such a computer system conventional PID controllers have been replaced by respective PID control algorithms implemented in programmed digital form in the following way.

The controller output $y(t)$, based on the difference $e(t)$ between the control input $u(t)$ and the set-point value SPV is defined as

$$y(t) = K_p \left[e(t) + \frac{1}{T_R} \int_0^t e(\tau)\, d\tau + T_D \frac{de(t)}{dt} \right]$$

where K_p is the proportional gain, T_R the reset time, and T_D the rate time of the controller.

In the computer, the digital PID control algorithm is based on some discrete values of measured process variables at some equidistant time instants $t_0, t_1, \ldots, t_n$ of sampling, so that one has mathematically to deal with the differences and the sums instead of with derivatives and integrals. Therefore, the discrete version of the above algorithm has to be developed by first differentiating the above equation, getting

$$\dot{y}(t) = K_p \left[\dot{e}(t) + \frac{1}{T_R} e(t) + T_D\, \ddot{e}(t) \right]$$

where $\dot{e}(t)$ and $\ddot{e}(t)$ are the first and the second derivative of $e(t)$, and $\dot{y}(t)$ the first derivative of $y(t)$. The derivatives can be approximated at each sampling point by

$$\dot{y}(k) = (y(k) - y(k-1))/\Delta t$$
$$\dot{e}(k) = (e(k) - e(k-1))/\Delta t$$

and

$$\ddot{e}(k) = (\dot{e}(k) - \dot{e}(k-1))/\Delta t$$

to result in

$$(y(k) - u(k-1))/\Delta t = K_p \left[\frac{e(k) - e(k-1)}{\Delta t} + \frac{1}{T_R} e(k) + T_D \frac{e(k) - 2e(k-1) + e(k-2)}{\Delta t^2} \right]$$

or in

$$\begin{aligned} y(k) = y(k-1) &+ K_p \left(1 + \frac{\Delta t}{T_R} + \frac{T_D}{\Delta t} \right) e(k) \\ &+ K_p(-1 - 2T_D/\Delta t)\, e(k-1) \\ &+ K_p \left(\frac{T_D}{\Delta t} \right) e(k-2) \end{aligned}$$

This is known as the *positional* PDI algorithm that delivers the new output value $y(k)$, based on its previous value $y(k-1)$ and on some additional calculations in which the values of $e(t)$ at three successive samplings are involved. The corresponding velocity version is

$$\Delta y(k) = y(k) - y(k-1)$$

Better resutls can be achieved using the "smoothed" derivative

$$\dot{e}(k) = \frac{1}{n} \sum_{i=0}^{n-1} \frac{e_{k-i} - e_{k-i-1}}{\Delta t}$$

or the "weighted" derivative

$$\dot{e}(k) = \frac{\sum_{i=0}^{n-1} W_i[e(k-i) - e(k-i-1)]}{\Delta t \sum_{i=0}^{n-1} W_i}$$

in which the *weighting factors* are selected, so that

$$W_i = \lambda^i W_0$$

and

$$\sum_{i=0}^{n-1} W_i = 1$$

In this case the final digital form of the PID algorithm is given by

$$\begin{aligned} y(k) = {} & y(k-1) + b_0 e(k) + b_1 e(k-1) + b_2 e(k-2) \\ & + b_3 e(k-3) + b_4 e(k-4) \end{aligned}$$

with

$$b_0 = K_p\left(\frac{1}{6} + \frac{\Delta t}{T_R} + \frac{T_D}{6\Delta t}\right)$$
$$b_1 = K_p\left(\frac{1}{2} + \frac{T_D}{3\Delta t}\right)$$
$$b_2 = K_p\left(-\frac{1}{2} - \frac{T_D}{\Delta t}\right)$$
$$b_3 = K_p\left(-\frac{1}{2} + \frac{T_D}{3\Delta t}\right)$$
$$b_4 = K_p\left(\frac{T_D}{6\Delta t}\right)$$

Another form of discrete PID algorithm, used in the first DDC implementations, was

$$y(k) = K_p\left[e(k) + \frac{1}{T_R}\sum_{i=0}^{k} e(i)\Delta t + T_D \frac{e(k) - e(k-1)}{\Delta t}\right]$$

Due to the sampling, the exact values of measured process variables are known only at sampling instances. Information about the signal values between the sampling instances is lost. In addition, the requirement to hold the sampled value between two sampling instants constantly delays the value by half of the sampling period, so that the choice of a large sampling period is equivalent to the introduction of a relatively long delay into the process dynamics. Consequently, the control loop will respond very slowly to the changes in that set-point value, which makes it difficult to properly manage urgent situations.

The best sampling time Δt to be selected for a given control loop depends on the control algorithm applied and on the process dynamics. Moreover, the shorter the sampling time, the better the approximation of the continuous closed-loop system by its digital equivalent, although this does not generally hold. For instance, the choice of sampling time has a direct influence on pole displacement of the original (continuous) system, whose discrete version can in this way become unstable, unobservable, or uncontrollable.

For systems having only real poles and which are controlled by a sampled-version algorithm, it is recommended to choose the sampling time between 1/6 and 1/3 of the smallest time constant of the system. Some practical recommendations plead for sampling times of 1 to 1.5 sec for liquid flow control, 3 to 5 sec of pressure control, and 20 sec for temperature control.

Input *signal quantization*, which is due to the limited accuracy of the *analog-to-digital converters*, is an essential factor influencing the quality of a digital control loop. The quantization level can here produce a *limit cycle* within the frame of the *quantization error* made.

The use of analog-to-digital converters with a resolution higher than the accuracy of measuring instruments makes this influence component less relevant. The same holds for the quantization of the output signal, where the resolution of the *digital-to-analog* converter is far higher than the resolution of positioning elements (actuators) used. In addition, due to the low-pass behavior of the system to be controlled, the quantization errors of output values of the controller have no remarkable influence on the control quality. Also, the problem of influence of the measurement noise on the accuracy of a digital controllers can be solved by analog or digital *prefiltering* of signals, before introducing it into the *control algorithm*.

Although the majority of distributed control systems is achieving a higher level of sophistication by placing more emphasis on the strategy in the control loops, some major vendors of such systems are already using artificial intelligence technology [8] to implement knowledge-based controllers [9], able to learn online from control actions and their effects [10,11]. Here, particularly the rule-based expert controllers and fuzzy-logic-based controllers have been successfully used in various industrial branches. The controllers enable using the knowledge base around the PID algorithm to make the control loop perform better and to cope with process and system irregularities including the system faults [12]. For example, Foxboro has developed the self-tuning controller EXACT based on a pattern recognition approach [4]. The controller uses a direct performance feedback by monitoring the controlled process variable to determine the action required. It is rule-based *expert controller*, the rules of which allow a faster startup of the plant, and adapt the controller's parameters to the dynamic deviations of plant's parameters, changing set-point values, variations of output load, etc.

Allen–Bradley's programmable controller configuration system (PCCS) provides expert solutions to the programmable controller application problems in some specific plant installations. Also introduced by the same vendor is a programmable vision system (PVS) that performs factory line recognition inspection.

Accol II, of Bristol Babcock, the language of its distributed process controller (DPC), is a tool for building of rule-based control systems. A DPC can be programmed, using heuristic knowledge, to behave in the same way as a human plant operator or a control engineer in the field. The incorporated inference engine can be viewed as a logical progression in the enhancement of an advanced, high-level process control language.

PICON, of LMI, is a real-time expert system for process control, designed to assist plant operators in dealing with multiple alarms. The system can manage up to 20,000 sensing and alarm points and can store and treat thousands of inference rules for control and diagnostic purposes. The knowledge acquisition interface of the system allows building of relatively complex rules and procedures without requiring artificial intelligence programming expertise. In cooperation with LMI, several vendors of distributed computer systems have incorporated PICON into their systems, such as Honeywell, Foxboro, Leeds & Northrup, Taylor Instruments, ASEA–Brown Bovery, etc. For instance, Leeds & Northrup has incorporated PICON into a distributed computer system for control of a pulp and paper mill.

Fuzzy logic controllers [13] are in fact simplified versions of real-time expert controllers, mainly based on a collection of IF-THEN rules and on some *declarative* fuzzy values of input, output, and control variables (classified as LOW, VERY LOW, SMALL, VERY SMALL, HIGH, VERY HIGH, etc.) are able to deal with the uncertainties and to use *fuzzy reasoning* in solving engineering control problems [14,15]. Thus, they can easily replace any manual operator's control action by compiling the decision rules and by heuristic reasoning on compiled database in the field.

Originally, fuzzy controllers were predominantly used as stand-alone, single-loop controllers, particularly appropriate for solving control problems in the situations where the dynamic process behavior and the character of external disturbances is now known, or where the mathematical process model is rather complex. With the progress of time, the fuzzy control software (the *fuzzyfier*, *rule base*, *rule interpreter*, and the *defuzzifier*) has been incorporated into the library of control functions, enabling online configuration of fuzzy control loops within a distributed control system.

In the 1990s, efforts have been concentrated on the use of *neurosoftware* to solve the process control problems in the plant by learning from field data [16]. Initially, neural networks have been used to solve *cognition problems*, such as *feature extraction* and *pattern recognition*. Later on, neurosoftware-based control schemes have been implemented. Networks have even been seen as an alternative technology for solving more complex cognition and control problems based on their massive parallelism and the connectionist learning capability. Although the neurocontrollers have mainly been applied as dedicated controllers in processing plants, manufacturing, and robotics [17], it is nevertheless to be expected that with the advent of low-price neural network hardware the controllers can in many complex situations replace the current programmable controllers. This will introduce the possibility to easily implement intelligent control schemes [18], such as:

Supervised controllers, in which the neural network learns the sensor inputs mapping to corresponding actions by learning a set of training examples, possibly positive and negative

Direct inverse controllers, in which the network learns the inverse system dynamics, enabling the system to follow a planned trajectory, particularly in robot control

Neural adaptive control, in which the network learns the model-reference adaptive behavior on examples

Back-propagation of utility, in which the network adapts an adaptive controller based on the results of related optimality calculations

Adapative critical methods, in which the experiment is implemented to simulate the human brain capabilities.

Very recently also hybrid, *neurofuzzy* approaches have been proposed, that have proven to be very efficient in the area of state estimation, real-time target tracking, and vehicle and robot control.

1.5 SYSTEMS ARCHITECTURE

In what follows, the overall structure of multicomputer systems for plant automation will be described, along with their internal structural details, including data file organization.

1.5.1 Hierarchical Distributed System Structure

The accelerated development of automation technology over many decades is a direct consequence of outstanding industrial progress, innumerable technical innovations, and a steadily increasing demand for high-quality products in the marketplace. Process and production industry, in order to meet the market requirements, was directly dependent on methods and tools of plant automation.

On the other hand, the need for higher and higher automation technology has given a decisive impetus and a true motivation to instrumentation, control, computer, and communication engineers to continually improve methods and tools that help solve the contemporary field problems. A variety of new methods has been proposed, classified into new disciplines, such as signal and system analysis, signal processing, state-space approach of system theory, model building, systems identification and parameter estimation, systems simulation, optimal and adaptive control, intelligent, fuzzy, and neurocontrol, etc. In addition, a large arsenal of hardware and software tools has been developed comprising mainframe and microcomputers, personal computers and workstations, parallel and massively parallel computers (neural networks), intelligent instrumentation, modular and object-oriented software experts, fuzzy and neurosoftware, and the like. All this has contributed to the development of modern automation systems, usually distributed, hierarchically organized multicomputer systems, in which the most advanced hardware, software, and communication links are operationally integrated.

Modern automation systems require distributed structure because of the distributed nature of industrial plants in which the control instrumentation is widely spread throughout the plant. Collection and preprocessing of sensors data requires distributed intelligence and an appropriate field communication system [19]. On the other hand, the variety of plant automation functions to be executed and of decisions to be made at different automation levels require a system architecture that—due to the hierarchical nature of the functions involved—has also to be hierarchical.

In the meantime, a layered, multilevel architecture of plant automation systems has widely been accepted by the international automation community that mainly includes (Fig. 6):

Direct process control level, with process data collection and preprocessing, plant monitoring and data logging, open-loop and closed-loop control of process variables

Plant supervisory control level, at which the plant performance monitoring, and optimal, adaptive, and coordinated control is placed

Production scheduling and control level, production dispatching, supervision, rescheduling and reporting for inventory control, etc.

Plant management level, that tops all the activities within the enterprise, such as market and customer demand analysis, sales statistics, order dispatching, monitoring and processing, production planning and supervision, etc.

Although the manufacturers of distributed computer control systems design their systems for a wide application, they still cannot provide the user with all facilities and all functions required at all hierarchical levels. As a rule, the user is required to plan the distribution system to be ordered. In order for the planning process to be successful, the user has above all to clearly formulate the premises under with the system has to be built and the requirements-oriented functions to be implemented. This should be taken as a selection guide for system elements to be integrated into the future plant automation system, so that the planned system [20]:

Covers all functions of direct control of all process variables, monitors their values, and enables the plant engineers optimal interaction with the plant via sophisticated man–machine interfaces

Offers a transport view into the plant performance and the state-of-the-art of the production schedule

Provides the plant management with the extensive up-to-date reports including the statistical and historical reviews of production and business data

Improves plant performance by minimizing the learning cycle and startup and setup trials

Permits faster adaptation to the market demand tides

Implements the basic objectives of plant automation—production and quality increase, cost

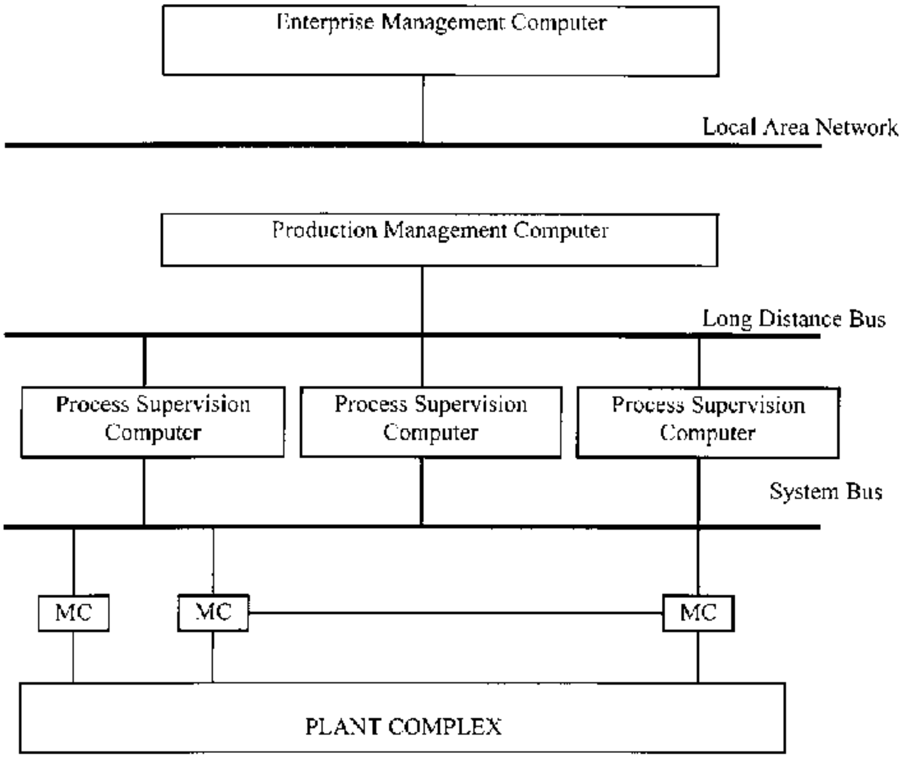

Enterprise Management Computer
Local Area Network
Production Management Computer
Long Distance Bus
Process Supervision Computer
Process Supervision Computer
Process Supervision Computer
System Bus
MC
MC
MC
PLANT COMPLEX

Figure 6 Bus-oriented hierarchical system.

decrease, productivity and work conditions improvement, etc.

Based on the above premises, the distributed computer control system to be selected should include:

A rich library of special software packages for each control, supervisory, production and management level, particularly

At control level: a full set of preprocessing, control, alarm, and calculation algorithms for measured process variables that is applicable to a wide repertoire of sensing and actuating elements, as well as a versatile display concept with a large number of operator friendly facilities and screen mimics

At supervisory level: wide alarm survey and tracing possibilities, instantaneous, trend, and short historical reporting features that include the process and plant files management, along with special software packages and block-oriented languages for continuous and batch process control and for configuration of plant mimic diagrams, model building and parameter estimation options, etc.

At production level: efficient software for online production scheduling and rescheduling, for performance monitoring and quality control, for recipe handling, and for transparent and exhaustive production data collection and structured reporting

At management level: abundant stock of professional software for production planning and supervision, order dispatch and terms check, order and sales surveys and financial balancing, market analysis and customer statistics, etc.

A variety of hardware features

At control level: attachment possibility for most types of sensors, transducers, and actuators, reliable and explosion-proof installation, hard-duty and failsafe version of control units, online system reconfiguration with a high degree of systems expandability, guaranteed further development of control hardware in the future by the same vendor, extensive provision of online diagnostic and preventive maintenance features

At supervisory and production level: wide program of interactive monitoring options designed to meet the required industrial standards, multiple computer interfaces to integrate different kinds of servers and workstations using internationally standardized bus systems and local area networks, interfacing possibilities for various external data storage media

At management level: wide integration possibilities of local and remote terminals and workstations.

It is extremely difficult to completely list all items important for planning a widespread multicomputer system that is supposed to enable the implementation of various operational functions and services. However, the aspects summarized here represent the majority of essential guiding aids to the system planner.

1.5.2 Hierarchical Levels

In order to appropriately lay out a distributed computer control system, the problems it is supposed to solve have to be specified [21]. This has to be done after a detailed *plant analysis* and by knowledge elicitation from the plant experts and the experts of different enterprise departments to be integrated into the automation system [22]. Should the distributed system cover automation functions of all hierarchical levels, a detailed analysis of all functions and services should be carried out, to result in an implementation report, from which the hardware and software of the system are to be planned. In the following, a short review of the most essential functions to be implemented is given for all hierarchical levels.

At *plant instrumentation level* [23], the details should be listed concerning the

Sensors, actuators, and field controllers to be connected to the system, their type, accuracy, grouping, etc.
Alarm occurrences and their locations
Backup concept to be used
Digital displays and binary indicators to be installed in the field
Completed plant mimic diagrams required
Keyboards and local displays, hand pads, etc. available
Field bus to be selected.

At this lowest hierarchical level of the system the field-mounted instrumentation and the related interfaces for data collections and command distribution for open- and closed-loop control are situated, as well as the electronic circuits required for adaptation of terminal process elements (sensors and actuators) to the computer input/output channels, mainly by signal conditioning using:

Voltage-to-current and current-to-voltage conversion
Voltage-to-frequency and frequency-to-voltage conversion
Input signal preprocessing (filtering, smoothing, etc.)
Signal range switching
Input/output channel selection
Galvanic isolation.

In addition, the signal format and/or digital signal representation has also to be adapted using:

Analog-to-digital and digital-to-analog conversion
Parallel-to-serial and serial-to-parallel conversion
Timing, synchronization, triggering, etc.

The recent development of FIELDBUS, the international *process data transfer standard*, has directly contributed to the standardization of *process interface* because the FIELDBUS concept of data transfer is a universal approach for interfacing the final field control elements to the programmable controllers and similar digital control facilities.

The search for the "best" FIELDBUS standard proposal has taken much time and has created a series of "good" bus implementations that are at least de facto accepted standards in their application areas, such as Bitbus, CiA, FAIS, FIP, IEC/ISA, Interbus-S, mISP, ISU-Bus, LON, Merkur, P-net, PROFIBUS, SERCOS, Signalbus, TTP, etc. Although an internationally accepted FIELDBUS standard is still not available, some proposals have widely been accepted but still not standardized by the ISO or IEC. One of such proposals is the PROFIBUS (PROcess FIeld BUS) for which a user group has been established to work on implementation, improvement, and industrial application of the bus.

In Japan, the interest of users has been concentrated on the FAIS (Factory Automation Interconnection System) Project, which is expected to solve the problem of a time-critical communication architecture, particularly important for production engineering. The final objective of the bus standardization work is to support the commercial process instrumentation with the built-in field bus interface. However, also here, finding a unique or a few compatible standard proposals is extremely difficult.

The FIELDBUS concept is certainly the best answer to the increasing cabling complexity at sensor and actuator level in production engineering and processing industries, which was more difficult to manage using the point-to-point links from all sensors and actuators to the central control room. Using the FIELDBUS concept, all sensors and actuators are interfaced to the distributed computer system in a unique way, as any external communication facility. The benefits resulting from this are multiple, some of them being:

Enormous decrease of cabling and installation costs.
Straightforward adaptation to any future sensor and actuator technology.
Easy configuration and reconfiguration of plant instrumentation, automatic detection of transmission errors and cable faults, data transmission protocol.
Facilitated implementation and use of hot backup by the communication software.
The problem of common-mode rejection, galvanic isolation, noise, and crosstalk vanishes due to digitalization of analog values to be transmitted.

Plant instrumentation includes all field instrumentation elements required for plant monitoring and control. Using the process interface, plant instrumentation is adapted to the input–output philosophy of the computer used for plant automation purposes or to its data collection bus.

Typical plant instrumentation elements are:

Physical transducers for process parameters
On/off drivers for blowers, power supplies, pumps, etc.
Controllers, counters, pulse generators, filters, and the like
Display facilities.

Distributed computer control systems have provided a high motivation for extensive development of plant instrumentation, above all with regard to incorporation of some intelligent functions into the sensors and actuators.

Sensors and actuators [24,25] as terminal control elements are of primary interest to control engineers, because the advances of sensor and actuator technology open new perspectives in further improvement of plant automation. In the past, the development of special sensors has always enabled solving control problems that have not been solvable earlier. For example, development of special sensors for online measurement of moisture and specific weight of running paper sheet has enabled high-precision control of the paper-making process. Similar progress in the processing industry is expected with the development of new electromagnetic, semiconductor, fiber-optic, nuclear, and biological sensors.

The VLSI technology has definitely been a driving agent in developing new sensors, enabling the extremely small microchips to be integrated with the sensors or the sensors to be embedded into the microchips. In this way *intelligent sensors* [26] or *smart transmitters* have been created with the data preprocessing and digital communication functions implemented in the chip. This helps increase the measurement accuracy of the sensor and its direct interfacing to the field bus. The most preferable preprocessing algorithms implemented within intelligent sensors are:

Calibration and recalibration in the field
Diagnostic and troubleshooting
Reranging and rescaling
Ambient temperature compensation
Linearization
Filtering and smoothing
Analog-to-digital and parallel-to-serial conversion
Interfacing to the field bus.

Increasing the intelligence of the sensors is simply to be viewed as a shift of some functions, originally implemented in a microcomputer, to the sensor itself. Much more technical innovation is contained in the emerging semiconductor and magnetic sensors, biosensors and chemical sensors, and particularly in fiber-optic sensors.

Fiber devices have for a long time been one of the most promising development fields of fiber-optic technology [27,28]. For instance, the sensors developed in this field have such advantages as:

High noise immunity
Insensitivity to electromagnetic interfaces
Intrinsic safety (i.e., they are explosion proof)
Galvanic isolation
Light weight and compactness
Ruggedness
Low costs
High information transfer capacity.

Based on the phenomena they operationally rely on, the optical sensors can be classified into:

Refractive index sensors
Absorption coefficient sensors
Fluorescence constant sensors.

On the other hand, according to the process used for sensing of physical variables, the sensors could be:

Intrinsic sensors, in which the fiber itself carries light to and from a miniaturized optical sensor head, i.e., the optical fiber forms here an intrinsic part of the sensor.
Extrinsic sensors, in which the fiber is only used as a transmission.

It should, nevertheless, be pointed out that—in spite of a wealth of optical phenomena appropriate for sensing of process parameters—the elaboration of industrial versions of sensors to be installed in the instrumentation field of the plant will still be a matter of hard work over the years to come. The initial enormous enthusiasm, induced by the discovery that fiber-optic sensing is viable, has overlooked some considerable implementation obstacles of sensors to be designed for use in industrial environments. As a consequence, there are relatively few commercially available fiber-optic sensors applicable to the processing industries.

At the end of the 1960s, the term *integrated optics* was coined, a term analogous to *integrated circuits*. The new term was supposed to indicate that in the future LSI chips, photons should replace electrons. This, of course, was a rather ambitious idea that was later amended to become *optoelectronics*, indicating the physical merger of photonic and electronic circuits, known as *optical integrated circuits*. Implementation of such circuits is based on thin-film waveguides, deposited on the surface of a substrate or buried inside it.

At the *process control level*, details should be given (Fig. 7) concerning:

Individual control loops to be configured, including their parameters, sampling and calculation time intervals, reports and surveys to be prepared, fault and limit values of measured process variables, etc.
Structured content of individual logs, trend records, alarm reports, statistical reviews, and the like
Detailed mimic diagrams to be displayed
Actions to be effected by the operator
Type of interfacing to the next higher priority level exceptional control algorithms to be implemented.

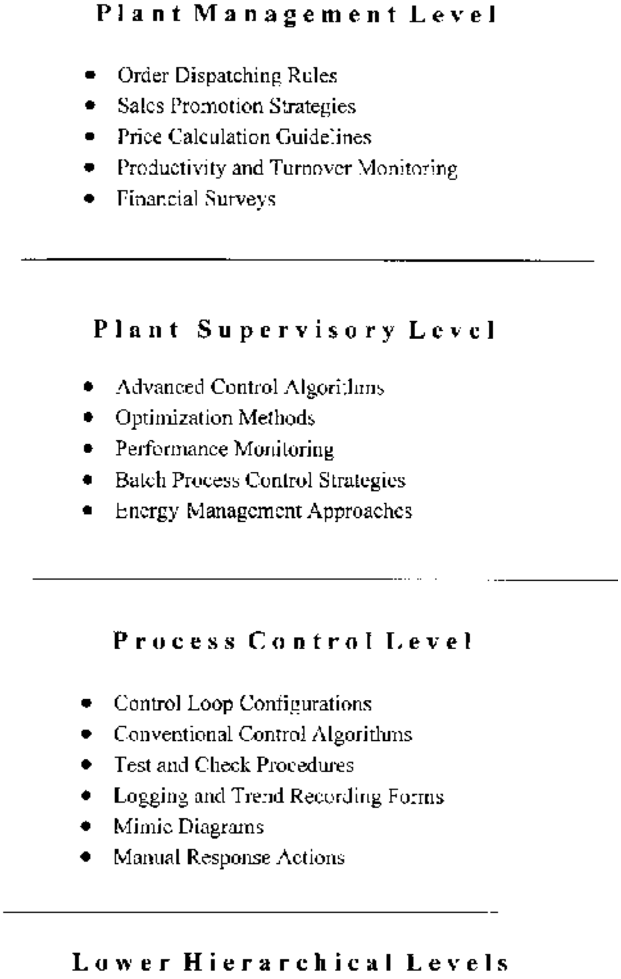

Plant Management Level
• Order Dispatching Rules
• Sales Promotion Strategies
• Price Calculation Guidelines
• Productivity and Turnover Monitoring
• Financial Surveys
Plant Supervisory Level
• Advanced Control Algorithms
• Optimization Methods
• Performance Monitoring
• Batch Process Control Strategies
• Energy Management Approaches
Process Control Level
• Control Loop Configurations
• Conventional Control Algorithms
• Test and Check Procedures
• Logging and Trend Recording Forms
• Mimic Diagrams
• Manual Response Actions
Lower Hierarchical Levels

Figure 7 Functional hierarchical levels.

At this level the functions required for collection and processing of sensor data, for process control algorithms, as well as the functions required for calculation of command values to be transferred to the plant are stored. Examples of such functions are functions for *data acquisition functions* include the operations needed for sensor data collection. They usually appear as initial blocks in an open- or closed-loop control chain, and represent a kind of interface between the system hardware and software. In the earlier process control computer systems, the functions were known as *input device drivers* and were usually a constituent part of the *operating system*. To the functions belong:

Analog data collection
Thermocouple data collection
Digital data collection
Binary/alarm data collection
Counter/register data collection
Pulse data collection.

As parameters, usually the input channel number, amplification factor, compensation voltage, conversion factors, and others are to be specified. The functions can be triggered *cyclically* (i.e., *program controlled*) or *event-driven* (i.e., *interrupt controlled*).

Input signal-*conditioning* algorithms are mainly used for preparation of acquired plant data, so that the data can—after being checked and tested—be directly used in computational algorithms. Because the measured data have to be extracted from a noisy environment, the algorithms of this group must include features like separation of signal from noise, determination of physical values of measured process variable, decoding of digital values, etc.

Typical signal-conditioning algorithms are:

Local linearization
Polynomial approximation
Digital filtering
Smoothing
Bounce suppression of binary values
Root extraction for flow sensor values
Engineering unit conversion
Encoding, decoding, and code version.

Test and check functions are compulsory for correct application of control algorithms that always have to operate on true values of process variables. Any error in sensing elements, in data transfer lines, or in input signal circuits delivers a false measured value which—when applied to a control algorithm—can lead to a false or even to a catastrophic control action. On the other hand, all critical process variables have to be continuously monitored, e.g., checked against their *limit values* (or *alarm values*), whose crossing certainly indicates the *emergency status* of the plant.

Usually, the *test and check algorithms* include:

Plausibility test
Sensor/transmitter test
Tolerance range test
Higher/lower limit test
Higher/lower alarm test
Slope/gradient test
Average value test.

As a rule, most of the anomalies detected by the described functions are, for control and statistical purposes, automatically stored in the system, along with the instant of time they have occurred.

Dynamic compensation functions are needed for specified implementation of control algorithms. Typical functions of this group are:

Lead/lag
Dead time
Differentiate
Integrator
Moving average
First-order digital filter
Sample-and-hold
Velocity limiter.

Basic control algorithms mainly include the PID algorithm and its numerous versions, e.g.:

PID-ratio
PID-cascade
PID-gap
PID-auto-bias
PID-error squared
I, P, PI, PD

As parameters, the values like proportional gain, integral reset, derivative rate, sampling and control intervals, etc. have to be specified.

Output signal condition algorithms adapt the calculated output values to the final or actuating elements to be influenced. The adaptation includes:

Calculation of full, incremental, or percentage values of output signals
Calculation of pulse width, pulse rate, or number of pulses for outputting
Book-keeping of calculated signals, lower than the sensitivity of final elements
Monitoring of end values and speed saturation of mechanical, pneumatic, and hydraulic actuators.

Output functions corresponds, in the reversed sense, to the input functions and include the analog, digital, and pulse output (e.g., pulse width, pulse rate, and/or pulse number).

At *plant supervisory level* (Fig. 7) the functions are concentrated, required for optimal process control, process performance monitoring, plant alarm management, and the like. For optimal process control, advanced, model-based control strategies are used such as:

Feed-forward control
Predictive control
Deadbeat control
State-feedback control
Adaptive control
Self-tuning control.

When applying the advanced process control, the:

Mathematical process model has to be built.

Optimal performance index has to be defined, along with the restriction on process or control variables.
Set of control variables to be manipulated for the automation purposes has to be identified.
Optimization method to be used has to be selected.

In engineering practice, the *least-squares error* is used as *performance index* to be minimized, but a number of alternative indices are also used in order to attain:

Time optimal control
Fuel optimal control
Cost optimal control
Composition optimal control.

Adaptive control [29] is used for implementation of optimal control that automatically accommodates the unpredictable environmental changes or signal and system uncertainties due to the parameter drifts or minor component failures. In this kind of control, the dynamic systems behavior is repeatedly traced and its parameters estimated which—in the case of their deviation from the given optimal values—have to be compensated in order to retain their constant values.

In modern control theory, the term *self-tuning control* [30] has been coined as alternative to adaptive control. In a self-tuning system control parameters are, based on measurements of system input and output, automatically tuned to result into a sustained optimal control. The tuning itself can be affected by the use of measurement results to:

Estimate actual values of system parameters and, in the sequence, to calculate the corresponding optimal values of control parameters, or to
Directly calculate the optimal values of control parameters.

Batch process control is basically a sequential, well-timed *stepwise control* that in addition to a preprogrammed time interval generally includes some *binary state indicators*, the status of which is taken at each *control step* as a decision support for the next control step to be made. The functional modules required for configuration of batch control software are:

Timers, to be preset to required time intervals or to the real-time instants
Time delay modules, time- or event-driven, for delimiting the control time intervals
Programmable up-count and *down-count timers* as time indicators for triggering the preprogrammed operational steps
Compactors as decision support in initiation of new control sequences
Relational blocks as internal message elements of control status
Decision tables, defining—for specified input conditions—the corresponding *output conditions* to be executed.

In a similar way the *recipe handling* is carried out. It is also a batch-process control, based on stored recipes to be downloaded from a mass storage facility containing the completed recipes library file. The handling process is under the competence of a *recipe manager*, a batch-process control program.

Energy management software takes care that all available kinds of energy (electrical, fuel, steam, exothermic heat, etc.) are optimally used, and that the short-term (daily) and long-term energy demands are predicted. It continuously monitors the generated and consumed energy, calculates the efficiency index, and prepares the relevant *cost reports*. In optimal energy management the strategies and methods are used, which are familiar in optimal control of stationary processes.

Contemporary distributed computer control systems are equipped with a large quantity of different *software packages* classified as:

System software, i.e., the computer-oriented software containing a set of tools for development, generation, test, run, and maintenance of programs to be developed by the user
Application software, to which the monitoring, control loop configuration, and communication software belong.

System software is a large aggregation of different *compilers* and *utility programs*, serving as systems development tools. They are used for implementation of functions that could not be implemented by any combination of program modules stored in the library of functions. When developed and stored in the library, the application programs extend its content and allow more complex control loops to be configured. Although it is, at least in principle, possible to develop new programmed functional modules in any languages available in process control systems, high-level languages like:

Real-time languages
Process-oriented languages

are still preferred for such development.

Real-time programming languages are favored as support tools for implementation of control software because they provide the programmer with the necessary features for sensor data collection, actuator data distribution, interrupt handling, and programmed real-time and difference-time triggering of actions. Real-time FORTRAN is an example of this kind of high-level programming language.

Process-oriented programming languages go one step further. They also support planning, design, generation, and execution of application programs (i.e., of their tasks). They are higher-level languages with *multitasking* capability, that enables the programs, implemented in such languages, to be simultaneously executed in an interlocked mode, in which a number of real-time tasks are executed synchronously, both in time- or event-driven mode. Two outstanding examples of process-oriented languages are:

Ada, able to support implementation of complex, comprehensive system automation software in which, for instance, the individual software packages, generated by the members of a programming team, are integrated in a cooperative, harmonious way

PEARL (Process and Experiment Automation Real-Time Language), particularly designed for laboratory and industrial plant automation, where the acquisition and real-time processing of various sensor data are carried out in a multitasking mode.

In both languages, a large number of different kinds of data can be processed, and a *large-scale plant* can be controlled by decomposing the global plant control problem into a series of small, well-defined *control tasks* to run *concurrently*, whereby the start, suspension, resumption, repetition, and stop of individual tasks can be preprogrammed, i.e., planned.

In Europe, and particularly in Germany, PEARL is a widespread automation language. It runs in a number of distributed control systems, as well as in diverse mainframes and personal computers like PDP-11, VAX 11/750, HP 3000, and Intel 80x86, Motorola 68000, and Z 8000.

Besides the general purpose, real-time and process-oriented languages discussed here, the majority of commercially available distributed computer control systems are well equipped with their own, *machine-specific*, high-level programming languages, specially designed for facilitation of development of *user-tailored* application programs.

At the *plant management* level (Fig. 7) a vast quantity of information should be provided, not familiar to the control engineer, such as information concerning:

Customer order files
Market analysis data
Sales promotion strategies
Files of planned orders along with the delivery terms
Price calculation guidelines
Order dispatching rules
Productivity and turnover control
Financial surveys

Much of this is to be specified in a structured, alphanumeric or graphical form, this because—apart from the data to be collected—each operational function to be implemented needs some *data entries* from the lower neighboring layer, in order to deliver some *output data* to the higher neighboring layer, or vice versa. The data themselves have, for their better management and easier access, to be well-structured and organized in *data files*. This holds for data on all hierarchical levels, so that in the system at least the following databases are to be built:

Plant databases, containing the parameter values related to the plant

Instrumentation databases, where the data are stored related to the individual final control elements and the equipment placed in the field

Control databases, mainly comprising the configuration and parametrization data, along with the nominal and limit values of the process variable to be controlled

Supervisory databases required for plant performance monitoring and optimal control, for plant modeling and parameter estimation, as well as production monitoring data

Production databases for accumulation of data relevant to raw material supplies, energy and products stock, production capacity and actual product priorities, for specification of product quality classes, lot sizes and restrictions, stores and transport facilities, etc.

Management databases, for keeping trace of customer orders and their current status, and for storing the data concerning the sales planning, raw material and energy resources status and demands, statistical data and archived long-term surveys, product price calculation factors, etc.

Before the structure and the required volume of the distributed computer system can be finalized, a large number of plant, production, and management-relevant data should be collected, a large number of appropriate algorithms and strategies selected, and a considerable amount of specific knowledge by interviewing various experts elucidated through the system analysis. In addition, a good system design demands a good cooperation between the user and the computer system vendor because at this stage of the project planning the user is not quite familiar with the vendor's system, and because the vendor should—on the user's request—implement some particular application programs, not available in the standard version of system software.

After finishing the system analysis, it is substantial to entirely *document* the results achieved. This is particularly important because the plants to be automated are relatively complex and the functions to be implemented distributed across different hierarchical levels. For this purpose, the detailed instrumentation and installation plans should be worked out using standardized symbols and labels. This should be completed with the *list of control and display flow charts* required. The programmed functions to be used for *configuration* and *parametrization* purposes should be summarized in a tabular or matrix form, using the *fill-in-the-blank* or *fill-in-the-form* technique, *ladder diagrams*, *graphical function charts*, or in special *system description languages*. This will certainly help the system designer to better tailor the hardware and the system programmer to better style the software of the future system.

To the *central computer system* a number of computers and computer-based terminals are interconnected, executing specific automation functions distributed within the plant. Among the distributed facilities only those directly contributing to the plant automation are important, such as:

Supervisory stations
Field control stations

Supervisory stations are placed at an intermediate level between the central computer system and the field control stations. They are designed to operate as autonomous elements of the distributed computer control system executing the following functions:

State observation of process variables
Calculation of optimal set-point values
Performance evaluation of the plant unit they belong to
Batch process control
Production control
Synchronization and backup of subordinated field control stations

Because they belong to some specific plant units, the supervisory stations are provided with special application software for *material tracking*, *energy balancing*, *model-based control*, *parameter tuning* of control loops, *quality control*, *batch control*, *recipe handling*, etc.

In some applications, the supervisory stations figure as *group stations*, being in charge of supervision of a group of controllers, aggregates, etc. In the small-scale to middle-scale plants also the functions of the central computer system are allocated to such stations.

A brief review of commercially available systems shows that the following functions are commonly implemented in supervisory stations:

Parameter tuning of controllers: CONTRONIC (ABB), DCI 5000 (Fisher and Porter), Network 90 (Bailey Controls), SPECTRUM (Foxboro), etc.
Batch control: MOD 300 (Taylor Instruments), TDC 3000 (Honeywell), TELEPERM M (Siemens), etc.
Special, *high-level control*: PLS 80 (Eckhardt), SPECTRUM, TDC 3000, CONTRONIC P, NETWORK 90, etc.
Recipe handling: ASEA-Master (ABB), CENTUM and YEWPACK II (Yokogawa), LOGISTAT CP-80 (AEG Telefunken), etc.

The supervisory stations are also provided with the real-time and process-oriented general or specific high-level programming languages like FORTRAN, RT-PASCAL, BASIC, CORAL [PMS (Ferranti)], PEARL, PROSEL [P 4000 (Kent)], PL/M, TML, etc. Using the languages, higher-level application programs can be developed.

At the lowest hierarchical level the *field control stations*, i.e., the programmable controllers are placed, along with some process monitors. The stations, as autonomous subsystems, implement up to 64 control loops. The software available at this control level includes the modules for

Process data acquisition
Process control
Control loop configuration

Process data acquisition software, available within the contemporary distributed computer control systems, is *modular software*, comprising the algorithms [31] for sensors, data collection, and preprocessing, as well as for actuator data distribution [31,32]. The software modules implement functions like:

Input device drivers, to serve the programming of analog, digital, pulse, and alarm or interrupt inputs, both in event drivers or in cyclic mode

Input signal conditioning, to preprocess the collected sensor values by applying the linearization, digital filtering and smoothing, bounce separation, root extraction, engineering conversion, encoding, etc.

Test and check operations, required for signal plausibility and sensor/transmitter test, high and low value check, trend check, etc.

Output signal conditioning, needed for adapting the output values to the actuator driving signals, like calculation of full and incremental output values, based on the results of the control algorithm used, or the calculation of pulse rate, pulse width, or the total number of pulses for outputting

Output device drivers, for execution of calculated and conditioned output values.

Process control software, also organized in modular form, is a collection of control algorithms, containing:

Basic control algorithms, i.e., the PID algorithm and its various modifications (PID ratio, cascade, gap, autobias, adaptive, etc.)

Advanced control algorithms like feed-forward, predictive, deadbeat, state feedback, self-tuning, nonlinear, and multivariable control.

Control loop configuration [33] is a two-step procedure, used for determination of:

Structure of individual control loops in terms of functional modules used and of their interlinkage, required for implementation of the desired overall characteristics of the loop under configuration, thus called the loop's configuration step

Parameter values of functional modules involved in the configuration, thus called the loop's parametrization step.

Once configured, the control loops are stored for their further use. In some situations also the parameters of the block in the loop are stored.

Generally, the functional blocks available within the field control stations—in order not to be destroyed—are stored in ROM or EPROM as a sort of *firmware module*, whereas the data generated in the process of configuration and parametrization are stored in RAM, i.e., in the memory where the configured software runs.

It should be pointed out that every block required for loop configurations is stored *only once* in ROM, to be used in *any numbers* of loops configured by simply addressing it, along with the pertaining parameter values in the block linkage data. The approach actually represents a kind of soft wiring, stored in RAM.

For multiple use of functional modules in ROM, their subroutines should be written in *re-entrant form*, so that the start, interruption, and continuation of such a subroutine with different initial data and parameter values is possible at any time.

It follows that once having all required functional blocks as a library of subroutine modules, and the tool for their mutual patching and parameterization, the user can program the control loops *in the field* in a ready-to-run form. The programming is here a relatively easy task because the loop configuration means that, to implement the desired control loop, the required subroutine modules should be taken from the library of functions and linked together.

1.5.3 Data File Organization

The functions, implemented within the individual functional layers, need some entry data in order to run and generate some data relevant to the closely related functions at the "neighboring" hierarchical levels. This means that the automation functions implemented should directly access some relevant initial data to generate some data of interest to the neighboring hierarchical levels. Consequently, the system functions and the relevant data should be allocated according to their tasks; this represents the basic concept of distributed, hierarchically organized automation systems: automation functions should be stored where they are needed, and the data where they are generated, so that only some selected data have to be transferred to the adjacent hierarchical levels. For instance, data required for direct control and plant supervision should be allocated in the field, i.e., next to the plant instrumentation and data, required for higher-level purposes, should be allocated near to the plant operator.

Of course, the organization of data within a hierachically structured system requires some specific considerations concerning the generation, access, updating, protection, and transfer of data between different files and different hierarchical levels.

As common in information processing systems, the data are basically organized in files belonging to the relevant database and being distributed within the sys-

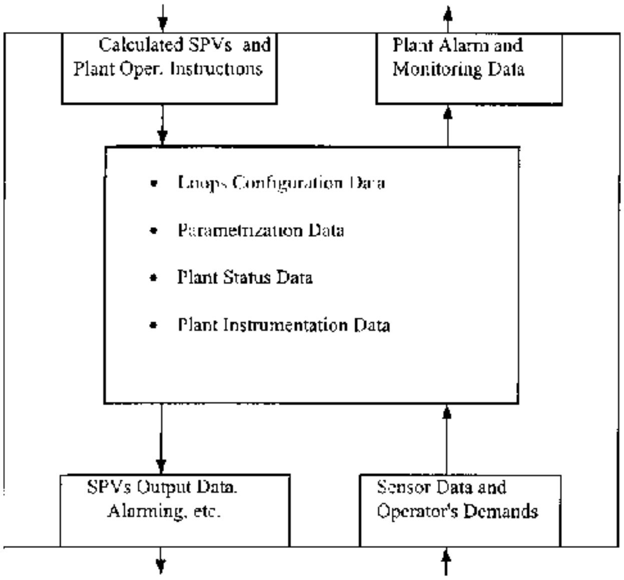

Calculated SPVs and Plant Oper. Instructions
Plant Alarm and Monitoring Data
• Loops Configuration Data
• Parametrization Data
• Plant Status Data
• Plant Instrumentation Data
SPVs Output Data. Alarming, etc.
Sensor Data and Operator's Demands

Figure 9 Process DB.

tem, so that the problem of data structure, local and global data relevance, data generation and access, etc. is the foreground one. In a distributed computer control system data are organized in the same way as their automation functions: they are attached to different hierarchical levels [4]. At each hierarchical level, only the selected data are received from other levels, whereby the intensity of the data flow "upward" through the system decreases, and in the opposite direction increases. Also, the communication frequency between the "lower" hierarchical levels is higher, and the response time shorter than between the "higher" hierarchical levels. This is due to the automation functions of lower levels servicing the real-time tasks, whereas those of the higher levels service some long-term planning and scheduling tasks.

The content of individual database units (DB) (Fig. 8) basically depends on their position within the hierarchical system. So, the process database (Fig. 9), situated at process control level, contains the data necessary for data acquisition, preprocessing, checking, monitoring and alarm, open- and closed-loop control, positioning, reporting, logging, etc. The database unit also contains, as long-term data, the specifications concerning the loop configuration and the parameters of individual functional blocks used. As short-term data it contains the measured actual values of process variables, the set-point values, calculated output values, and the received plant status messages. Depending on the nature of the implemented functions, the origin of collected data, and the destination of generated data, the database unit at process control level has—in order to handle a large number of short-life data having a very fast access—to be efficient under real-time conditions. To the next "higher" hierarchical level only some actual process values and plant status messages are forwarded, along with short history of some selected process variables. In the reverse direction, calculated optimal set-point values for controllers are respectively to be transferred.

The *plant database*, situated at supervision control level, contains data concerning the plant status, based on which the monitoring, supervision, and operation of plant is carried out (Fig. 10). As long-term data, the database unit contains the specifications concerning the available standard and user-made displays, as well as data concerning the mathematical model of the plant. As short-term data the database contains the actual status and alarm messages, calculated values of process variables, process parameters, and optimal set-point values for controllers. At the hierarchical

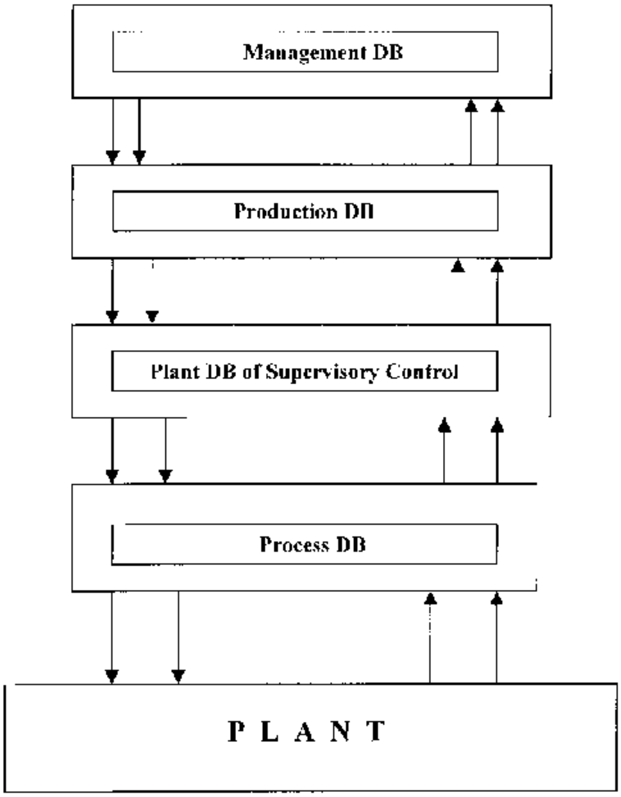


Figure 8 Individual DB units.

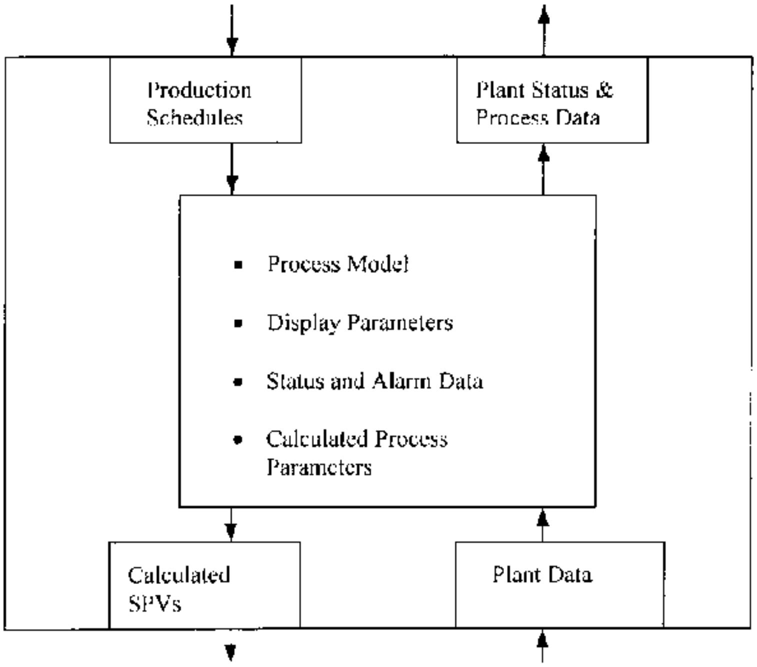


Figure 10 Database of supervisory control level.

level, a large number of data are stored whose access time should be within a few seconds. Here, some calculated data have to be stored for a longer time (historical, statistical, and alarm data), so that for this purpose hard disks are used as backup storage. To the "higher" hierarchical level, only selected data are transferred for production scheduling, and directives are received.

The *production database*, situated at production scheduling and control level (Fig. 11), contains data concerning the products and raw material stocks, production schedules, production goals and priorities, lot sizes and restrictions, quality control as well as the store and transport facilities. As long-term data the archived statistical and plant alarm reports are stored in bulk memories. The data access time is here in no way critical. To the "higher" hierarchical level, the status of the production and order processing, as well as of available facilities necessary for production replanning is sent, and in the reverse direction the target production data.

Finally, the *management database*, stored at corporate or enterprise management level (Fig. 12), contains data concerning the customer orders, sales planning, product stocks and production status, raw material and energy resources and demands, status of store and transport facilities, etc. Data stored here are long-term, requiring access every few minutes up to many weeks. For this reason, a part of the database can be stored on portable magnetic media, where it can be deposited for many years for statistical or administrative purposes.

The fact that different databases are built at different hierarchical levels and possibly stored in different computers, administrated by different database management or operating systems, makes the access of any hierarchical level difficult. Inherent problems here are the problems of formats, log output procedures, concurrency control, and other logical differences concerning the data structures, data management languages, label incompatibilities, etc. In the meantime, some appraoches have been suggested for solving some of the problems, but there is still much creative work to be done in this field in order to implement flexible, level-independent access to any database in a distributed computer system.

Another problem, typical for all time-related databases, such as the real-time and production management databases, is the representation of *time-related data*. Such data have to be integrated into the context of time, a capability that the conventional database management systems do not have. In the meantime, numerous proposals have been made along this line which include the time to be stored as a universal attri-

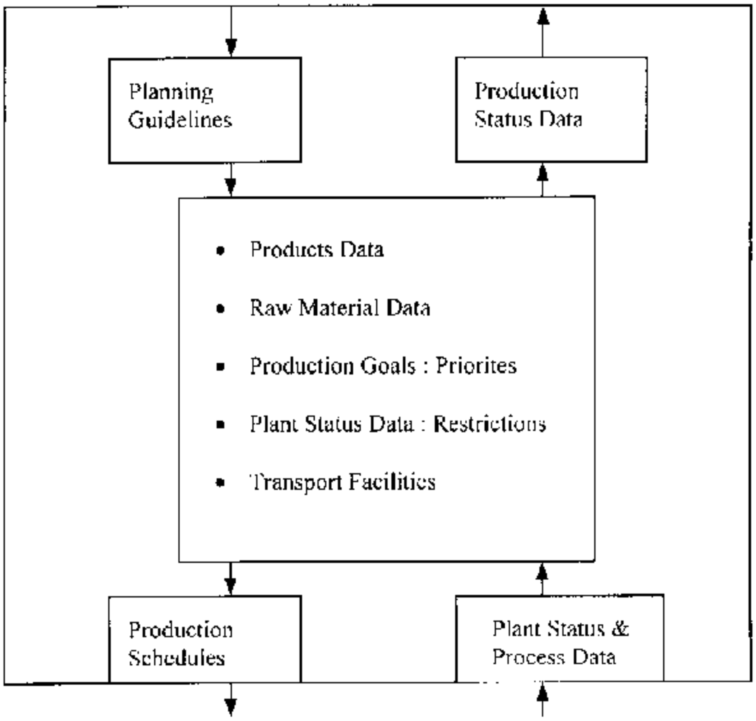


Figure 11 Database of production scheduling and control level.

bute. The attribute itself can, for instance, be transaction time, valid time, or any user-defined time. Recently, four types of time-related databases have been defined according to their ability to support the time concepts and to process temporal information:

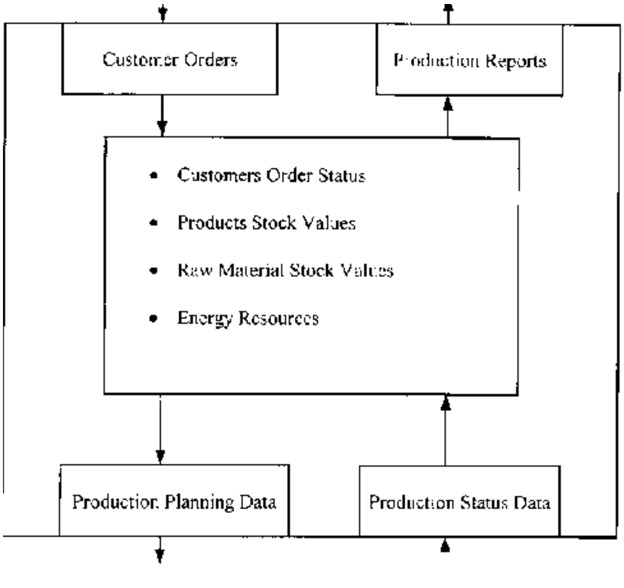


Figure 12 Management database.

Snapshot databases, i.e., databases that give an instance or a state of the data stored concerning the system (plant, enterprise) at a certain instant of time, but not necessarily corresponding to the current status of the system. By insertion, deletion, replacement, and similar data manipulation a new snapshot database can be prepared, reflecting a new instance or state of the system, whereby the old one is definitely lost.

Rollback databases, e.g., a series of snapshot databases, simultaneously stored and indexed by transaction time, that corresponds to the instant of time the data have been stored in the database. The process of selecting a snapshot out of a rollback database is called rollback. Also here, by insertion of new and deletion of old data (e.g., of individual snapshots) the rollback databases can be updated.

Historical databases, in fact snapshot databases in valid time, i.e., in the time that was valid for the systems as the databases were built. The content of historical databases is steadily updated by deletion of invalid data, and insertion of actual data acquired. Thus, the databases always reflect the reality of the system they are related to. No data belonging to the past are kept within the database.

Temporal databases are a sort of combination of rollback and historical databases, related both to the transition time and the valid time.

1.6 COMMUNICATION LINKS REQUIRED

The point-to-point connection of field instrumentation elements (sensors and actuators) and the facilities located in the *central control room* is highly inflexible and costly. This total reduction of wiring and cable-laying expenses remains the most important objective when installing new, centralized automation systems. For this purpose, the placement of a remote process interface in the *field multiplexers* and *remote terminal units* (RTUs) was the initial step in partial system decentralization. With the availability of microcomputers the remote interface and remote terminal units have been provided with the due intelligence so that gradually some data acquisition and preprocessing functions have been also transferred to the frontiers of the plant instrumentation.

Yet, data transfer within the computer-based, distributed hierarchical system needs an efficient, universal communication approach for interconnecting the numerous intelligent, spatially distributed subsystems at all automation levels. The problems to be solved in this way can be summarized as follows:

At *field level*: interconnection of individual *final elements* (sensors and actuators), enabling their *telediagnostics* and *remote calibration* capability

At *process control level*: implementation of individual *programmable control loops* and provision of *monitoring, alarms*, and *reporting* of data

At *production control level*: collection of data required for *production planning, scheduling, monitoring*, and *control*

At *management level*: integration of the production, sales, and other *commercial data* required for order processing and *customer services*.

In the last two or more decades much work has been done on standardization of a data communication links, particularly appropriate for transfer of process data from the field to the central computer system. In this context, Working Group 6 of Subcommittee 65C of the International Electrotechnical Commission (IEC), the scope of which concern the Digital Data Communications for Measurement and Control, has been working on PROWAY (Process Data Highway), an international standard for a high-speed, reliable, noise immune, low-cost data transfer within the plant automation systems. Designed as a bus system, PROWAY was supposed to guarantee the data transfer rate of 1 Mbps over a distance of 3 km, with up 100 participants attached along the bus. However, due to the IEEE work on project 802 on local area networks, which at the time of standardization of PROWAY had already been accepted by the communication community, the implementation of PROWAY was soon abandoned.

The activity of the IEEE in the field of local area networks was welcomed by both the IEC and the International Organization for Standardization (ISO) and has been converted into corresponding international standards. In addition, the development of modern intelligent sensors and actuators, provided by telediagnostics and remote calibration capabilities, has stimulated the competent professional organizations (IEC, ISA, and the IEEE itself) to start work on the standardization of a special communication link, appropriate for direct transfer of field data, the FIELDBUS. The bus standard was supposed to meet at least the following requirements:

Multiple drop and redundant topology, with a total length of 1.5 km or more.

For data transmission twisted pair, coax cable, and optical fiber should be applicable.

Single-master and multiple-master bus arbitration must be possible in multicast and broadcast transmission mode.

Access time of 5–20 sec or a scan rate of 100 samples per second should be guaranteed.

High-reliability with the error detection features built in the data transfer protocol.

Galvanic and electrical (> 250 V) isolation.

Mutual independence of bus participants.

Electromagnetic compatibility.

The requirements have simultaneously been worked out by IEC TC 65C, ISA SP 50, and IEEE P 1118. However, no agreement has been achieved on final standard document because four standard candidates have been proposed:

BITBUS (Intel)
FIP (Factory Instrumentation Protocol) (AFNOR)
MIL-STD-1533 (ANSI)
PROFIBUS (Process Field Bus) (DIN).

The standardization work in the area of local area networks, however, has in the last more than 15 years been very successful. Here, the standardization activities have been concentrated on two main items:

ISO/OSI Model
IEEE 802 Project.

The ISO has within its Technical Committee 97 (Computers and Information Processing) established the subcommittee 16 (Open Systems Interconnection) to work on architecture of an international standard of what is known as the OSI (Open Systems Interconnection) model [34], which is supposed to be a reference model of future communication systems. In the model, a hierarchically layered structure is used to include all aspects and all operating functions essential for compatible information transfer in all application fields concerned. The model structure to be standardized defines individual layers of the communication protocol and their functions there. However, it should not deal with the protocol implementation technology.

The work on the OSI model has resulted in a recommendation that the future open system interconnection standard should incorporate the following functional layers (Fig. 13):

Physical layer, the layer closed to the data transfer medium, containing the physical and procedural fucntions related to the medium access, such as switching of physical connections, physical message transmission, etc., without any prescription of any specific medium

Data link layer, responsible for procedural functions related to link establishment and release, transmission framing and synchronization, sequence and flow control, error protection

Network layer, required for reliable, cost-effective, and transparent transfer of data along the transmission path between the end stations by adequate routing, multiplexing, internetworking, segmentation, and block building

Transport layer, designed for establishing, supervision, and release of logic transport connections between the communication participants, aiming at optimal use of network layer services

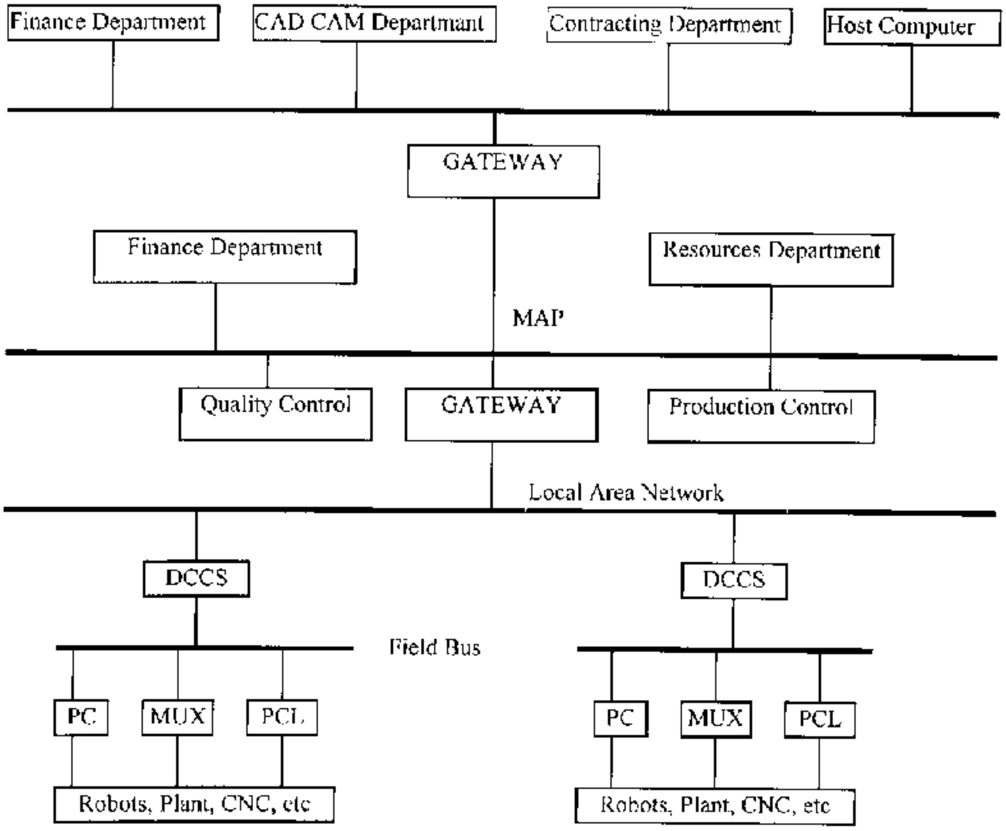


Figure 13 Integrated computer-aided manufacturing.

Session layer, in charge of opening, structuring, control, and termination of a communication session by establishing the connection to the transport layer

Presentation layer, which provides independence of communication process on the nature and the format of data to be transferred by adaptation and transformation of source data to the internal system syntax conventions understandable to the session layer

Application layer, the top layer of the model, serving the realization and execution of user tasks by data transfer between the application processes at semantic level.

Within distributed computer control systems, usually the physical, logic link, and application layers are required, other layers being needed only when internetworking and interfacing the system with the public networks.

As mentioned before, the first initiative of IEEE in standardization of local area networks [18,35] was undertaken by establishing its Project 802. The project work has resulted in release of the Draft Proposal Document on Physical and Data Link Layers, that still was more a complication of various IBM Token Ring and ETHERNET specifications, rather than an entirely new standard proposal. This was, at that time, also to be expected because in the past the only commercially available and technically widely accepted de facto communication standard was ETERNET and the IBM Internal Token Ring Standard. The slotted ring, developed at the University of Cambridge and known as the Cambridge Ring, was not accepted as a standard candidate.

Real standardization work within the IEEE has in fact started by shaping the new bus concept based on CSMA/CD (Carrier Sense Multiple Access/Contention Detection) principle of MAC (Medium Access Control). The work has later been extended to standardization of a token passing bus and a token passing ring, that have soon been identified as future industrial standards for building complex automation systems.

In order to systematically work on standardization of local area networks [36], the IEEE 802 Project has been structured as follows:

802.1 Addressing, Management, Architecture
802.2 Logic Link Control
802.3 CSMA/CD MAC Sublayer
802.4 Token Ring MAC Sublayer
802.5 Token Ring MAC Sublayer
802.6 Metropolitan Area Networks
802.7 Broadband Transmission
802.8 Fiber Optics
802.9 Integrated Voice and Data LANs.

The CSMA/CD standard defines a *bit-oriented* local area network, most widely used in implementation of the ETHERNET system as an improved ALOHA concept. Although being very reliable, the CSMA/CD medium access control is really efficient when the aggregate channel utilization is relatively low, say lower than 30%.

The *token ring* is a priority type, medium access control principle in which a symbolic *token* is used for setting the priority within the individual ring participants. The token is passed around the ring, interconnecting all the stations. Any station intending to transmit data should wait for the free token, declare it by encoding for a busy token, and start sending the message frames around the ring. Upon completion of its transmission, the station should insert the free token back into the ring for further use.

In the token ring, a special 8-bit pattern is used, say 11111111 when free, and 11111110 when busy. The pattern is passed without any addressing information. In the token bus, the token, carrying an addressing information related to the next terminal unit permitted to use the bus, is used. Each station, after finishing its transmission, inserts the address of the next user into the token and sends it along the bus. In this way, after circulating through all participating stations the token again returns to the same station so that actually a logic ring is virtually formed into which all stations are included in the order they pass the token to each other.

In distributed computer control systems, communication links are required for exchange of data between individual system parts in the range from the process instrumentation up to the central mainframe and the remote intelligent terminals attached to it. Moreover, due to the hierarchical nature of the system, different types of data communication networks are needed at different hierarchical levels. For instance:

The *field level* requires a communication link designed to collect the sensor data and to distribute the actuator commands.

The *process control level* requires a high-performance bus system for interfacing the programmable controllers, supervisory computers, and the relevant monitoring and command facilities.

The *production control* and *production management* level requires a real-time local area network as a system interface, and a long-distance communication link to the remote intelligent terminals belonging to the system.

Presently, almost all commercially available systems use at all communication levels very well-known interntional bus and network standards. This facilitates the products compatibility of different computer and instrumentation manufacturers, giving the user's system planner to work out a powerful, low-cost multicomputer system by integrating the subsystems with highest performance-to-price ratio.

Although there is a vast number of different communication standards used in design of different commercially available distributed computer control systems, their comparative analysis suggests their general classification into:

Automation systems for *small-scale plants* and *medium-scale plants*, having only the field and the process control level. They are basically *bus-oriented systems* requiring not more than two buses. The systems can, for higher level automation purposes, be interfaced via any suitable communication link to a *mainframe*.

Automation systems for medium-scale to *large-scale plants* additionally having the production planning and control level. They are *area network oriented* and can require a *long distance bus* or a *bus coupler* (Fig. 1).

Automation systems for large-scale plants with the *integrated automation* concept, requiring more or less all types of communication facilities: buses, rings, local area networks, public networks, and a number of bus couplers, network bridges, etc. Manufacturing plant automation could even involve different *backbone buses* and local area networks, *network bridges* and *network gateways*, etc. (Fig. 13). Here, due to the *MAP/TOP standards*, a broad spectrum of processors and programmable controllers of different vendors (e.g. Allen Bradley, AT&T, DEC, Gould, HP, Honeywell, ASEA, Siemens, NCR, Motorola, SUN, Intel, ICL, etc.) have been mutually interfaced to directly exchange the data via a MAP/TOP system.

The first distributed control system launched by Honeywell, the TDC 2000 system, was a multiloop controller with the controllers distributed in the field, and was an encouraging step, soon to be followed by a number of leading computer and instrumentation vendors such as Foxboro, Fisher and Porter, Taylor Instruments, Siemens, Hartman and Braun, Yokogawa, Hitachi, and many others. Step by step, the system has been improved by integrating powerful supervisory and monitoring facilities, graphical processors, and general purpose computer systems, interconnected via high-performance buses and local area networks. Later on, programmable logic controllers, remote terminal units, SCADA systems, smart sensors and actuators, intelligent diagnostic and control software, and the like was added to increase the system capabilities.

For instance, in the LOGISTAT CP 80 System of AEG, the following hierarchical levels have been implemented (Fig. 14):

Process level, or process instrumentation level

Direct control level or DDC level for signal data processing, open- and closed-loop control, monitoring of process parameters, etc.

Group control level for remote control, remote parametrizing, status and fault monitoring logic, process data filling, text processing, etc.

Process control level, for plant monitoring, production planning, emergency interventions, production balancing and control, etc.

Operational control levels, where all the required calculations and administrative data processing

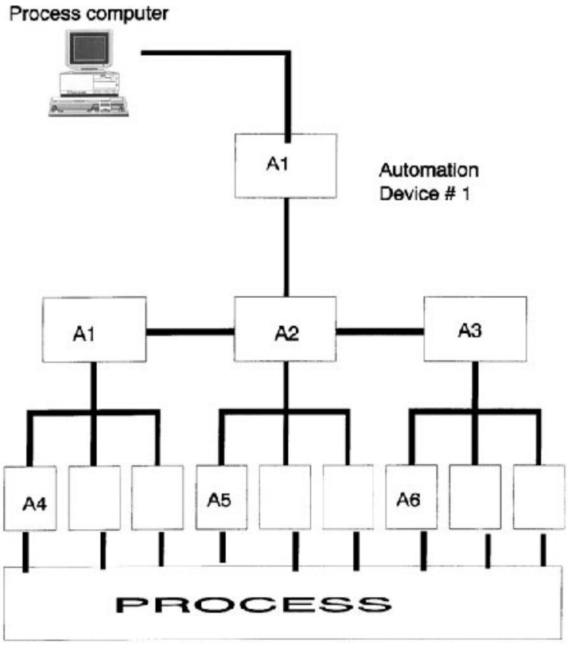


Figure 14 LOGISTAT CP 80 system.

are carried out, statistical reviews prepared, and market prognostic data generated.

In the system, different computer buses (K 100, K 200, and K 400) are used along with the basic controller units A 200 and A 500. At each hierarchical level, there are corresponding monitoring and command facilities B 100 and B 500.

A *multibus system* has also been applied in implementing the ASEA MASTER system, based on Master Piece Controllers for continuous and discrete process control. The system is widely extendable to up to 60 controllers with up to 60 loops each controller. For plant monitoring and supervision up to 12 color display units are provided at different hierarchical levels. The system is straightfowardly designed for integrated plant control, production planning, material tracking, and advanced control. In addition, a *twin bus* along with the ETHERNET Gateway facilitates direct system integration into a large multicomputer system.

The user benefits from a well-designed backup system that includes the ASEA compact backup controllers, manual stations, twin bus, and various internal redundant system elements.

An original idea is used in building the integrated automation system YEW II of Yokogawa in which the main modules:

YEWPAC (packaged control system)
CENTUM (system for distributed process control)
YEWCOM (process management computer system)

have been integrated via the fiber-optic data link.

Also in the distributed computer control system DCI 5000 of Fisher and Porter, some subsystems are mutually linked via fiber-optic data transfer paths that, along with the up to 50 km long ETHERNET coax cable, enable the system to be widely interconnected and serve as a physically spread out data management system. For longer distances, if required, fiber-optic bus repeaters can also be used.

A relatively simple but highly efficient concept underlines the implementation of the MOD 300 system of Taylor, where a communication ring carries out the integrating function of the system.

Finally, one should keep in mind that not always the largest distributed installations are required to solve the plant automation problems. Also simple, multiloop programmable controllers, interfaced to an IBM-compatible PC with its monitor as the operator's station are sufficient in automation practice. In such a configuration the RS 232 can be used as a communication link.

1.7 RELIABILITY AND SAFETY ASPECTS

Systems reliability is a relatively new aspect that design engineers have to take into consideration when designing the system. It is defined in terms of the probability that the system, for some specified conditions, normally performs its operating function for a given period of time. It is an indicator of how well and how long the system will operate in the sense of its design objectives and its functional requirements before it fails. It is supposed that the system works permanently and is subject to random failures, like the electronic or mechanical systems are.

Reliability of a computer-based system is generally determined by the reliability of its hardware and software. Thus, when designing or selecting a system from reliability point of view, both reliability components should be taken into consideration.

With regard to the reliability of systems hardware, the overall system reliability can be increased by increasing the reliability of its individual components and by system design for reliability, using multiple, redundant structures. Consequently, the design of distributed control systems can increase the overall system reliability by selecting highly reliable system components (computers, display facilities, communication links, etc.) and implementing with them a highly reliable system structure, whereby first the question should be answered as to how redundant a multicomputer system should be in order to still be operational and affordable, and to still operate in the worst case when a given number of its components fail.

Another aspect to be considered is the system features of automatic component-failure detection and failure isolation. In automation systems this particularly concerns the sensing elements working in severe industrial environments. The solution here consists of a *majority voting* or "*m from n*" approach, possibly supported by the *diversity principle*, i.e., using a combination of sensing elements of different manufacturers, connected to the system interface through different data transfer channels [37].

The majority voting approach and the diversity principle belong to the category of *static* redundancy implementations. In systems with repair, like electronic systems, *dynamic* redundancy is preferred, based on the *backup* and *standby* concept. In a highly reliable, dynamically redundant, failure-tolerant system additional "parallel" elements are assigned to each outmost critical active element, able to take over the function of the active element in case it fails. In this way, alternatively can be implemented:

Cold standby, where the "parallel" elements are switched *off* while the active element is running properly and switched *on* when the active element fails.

Hot standby, where the "parallel" element is permanently switched *on* and repeats in offline, open-loop mode the operations of the active elements and is ready and able to take over online the operations from the active element when the element fails.

Reliability of software is closely related to the reliability of hardware, and introduces some additional features that can deteriorate the overall reliability of the system. As possible software failures the coding and conceptual errors of subroutines are addressed, as well as the nonpredictability of total execution time of critical subroutines under arbitrary operating conditions. This handicaps the interrupt service routines and the communication protocol software to guarantee the required *time-critical responses*. Yet, being intelligent enough, the software itself can take care of automatic *error detection*, *error location*, and *error correction*. In addition, a simulation test of software before it is online used can reliably estimate the worst-case execution time. This is in fact a standard procedure because the preconfigured software of distributed computer control systems is well tested and evaluated offline and online by simulation before being used.

Systems safety is another closely related aspect of distributed control systems application in the automation of industrial plants, particularly of those that are critical with regard to possible explosion consequences in the case of malfunction of the control system installed in the plant. For long period of time, one of the major difficulties in the use of computer-based automation structures was that the safety authorization agencies refused to licence such structures as *safe enough*. The progress in computer and instrumentation hardware, as well as in monitoring and diagnostic software, has enabled building computer-based automation systems acceptable from the safety point of view because it can be demonstrated that for such systems:

Failure of instrumentation elements in the field, including the individual programmable controllers and computers at direct control level, does not create hazardous situations.

Such elements, used in critical positions, are self-inspected entities containing failure detection, failure annunciation, and failure safety through redundancy.

Reliability and fail-safety aspects of distributed computer control systems demand some specific criteria to be followed in their design. This holds for the overall systems concept, as well as for the hardware elements and software modules involved. Thus, when designing the system hardware [37]:

Only the well-tested, long-time checked, highly reliable heavy-duty elements and subsystems should be selected.

A modular, structurally transparent hardware concept should be taken as the design base.

Wherever required, the reliability provisions (majority voting technique and diversity principle) should be built in and supported by error check and diagnostic software interventions.

For the most critical elements the cold standby and/or hot standby facilities should be used along with the noninterruptible power supply.

Each sensor's circuitry, or at least each sensor group, should be powered by independent supplies.

A variety of sensor data checks should be provided at signal preprocessing level, such as plausibility, validity, and operability check.

Similar precautions are related to the design of software, e.g. [38]:

Modular, free configurable software should be used with a rich library of well-tested and online-verified modules.

Available loop and display panels should be relatively simple, transparent, and easy to learn.

A sufficient number of diagnostic, check, and test functions for online and offline system monitoring and maintenance should be provided.

Special software provisions should be made for *bump-free* switch over from *manual* or *automatic to computer* control.

These are, of course, only some of the most essential features to be implemented.

REFERENCES

1. JE Rijnsdrop. Integrated Process Control and Automation. Amsterdam: Elsevier, 1991.
2. G Coulouris, J Dollimore, T Kindberg. Distributed systems—concepts and design. ISA International Conference, New York, 2nd ed, 1994.

3. D Johnson. Programmable Controllers for Factory Automation. New York: Marcel Dekker, 1987.
4. D Popovic, VP Bhatkar. Distribution Computer Control for Industrial Automation. New York: Marcel Dekker, 1990.
5. PN Rao, NK Tewari, TK Kundra. Computer-Aided Manufacturing. New York: McGraw-Hill, 1993; New Delhi: Tata, 1990.
6. GL Batten Jr. Programmable Controllers. TAB Professional and Reference Books, Blue Ridge Summit, PA, 1988.
7. T Ozkul. Data Acquisition and Process Control Using Personal Computers. New York: Marcel Dekker, 1996.
8. D Popovic, VP Bhaktkar. Methods and Tools for Applied Artificial Intelligence. New York: Marcel Dekker, 1994.
9. DA White, DA Sofge, eds. Handbook of Intelligent Control—Neural, Fuzzy and Adaptive Approaches. New York: Van Nostrand, Reinhold, 1992.
10. J Litt. An expert system to perform on-line controller tuning. IEEE Control Syst Mag 11(3): 18–33, 1991.
11. J McGhee, MJ Grandle, P Mowforth, eds. Knowledge-Based Systems for Industrial Control. London: Peter Peregrinus, 1990.
12. PJ Antsaklis, KM Passino, eds. An Introduction to Intelligent and Autonomous Control. Boston, MA: Kluwer Academic Publishers, 1993.
13. CH Chen. Fuzzy Logic and Neural Network Handbook. New York: McGraw-Hill, 1996.
14. D Driankov, H Helleudoorn, M Reinfrank. An Introduction to Fuzzy Control. Berlin: Springer-Verlag, 1993.
15. RJ Markus, ed. Fuzzy Logic Technology and Applications. New York: IEEE Press, 1994.
16. CH Dagel, ed. Artificial Neural Networks for Intelligent Manufacturing. London: Chapman & Hall, 1994.
17. WT Miller, RS Sutton, PJ Werfos, eds. Neural Networks for Control. Cambridge, MA: MIT Press, 1990.
18. PJ Werbros. Neurocontrol and related techniques. In: A Maren, C Harston, R Pap, eds. Handbook of Neural Computing Applications. New York: Academic Press, 1990.
19. A Ray. Distributed data communication networks for real-time process control. Chem Eng Commun 65(3): 139–154, 1988.
20. D Popovic, ed. Analysis and Control of Industrial Processes. Braunschweig, Germany: Vieweg-Verlag, 1991.
21. PH Laplante. Real-time Systems Design and Analysis. New York: IEEE Press, 1993.
22. KD Shere, RA Carlson. A methodology for design, test, and evaluation of real-time systems. IEEE Computer 27(2): 34–48, 1994.
23. L Kane, ed. Advanced Process Control Systems and Instrumentation. Houston, TX: Gulf Publishing Co., 1987.
24. CW De Silvar. Control Sensors and Actuators. Englewood Cliffs, NJ: Prentice Hall, 1989.
25. RS Muller et al. eds. Microsensors. New York: IEEE Press, 1991.
26. MM Bob. Smart transmitters in distributed control—new performances and benefits, Control Eng 33(1): 120–123, 1986.
27. N Chinone and M Maeda. Recent trends in fiber-optic transmission technologies for information and communication networks. Hitachi Rev 43(2): 41–46, 1994.
28. M Maeda, N Chinone. Recent trends in fiber-optic transmission technologies. Hitachi Rev 40(2): 161–168, 1991.
29. T Hägglund, KJ Aström. Industrial adaptive controllers based on frequency response techniques. Automatica 27(4): 599–609, 1991.
30. PJ Gawthrop. Self-tuning pid controllers—algorithms and implementations. IEEE Trans Autom Control 31(3): 201–209, 1986.
31. L Sha, SS Sathaye. A systematic approach to designing distributed real-time systems. IEEE Computer 26(9): 68–78, 1993.
32. MS Shatz, JP Wang. Introduction to distributed software engineering. IEEE Computer 20(10): 23–31, 1987.
33. D Popovic, G Thiele, M Kouvaras, N Bouabdalas, and E Wendland. Conceptual design and C-implementation of a microcomputer-based programmable multi-loop controller. J Microcomputer Applications 12: 159–165, 1989.
34. MR Tolhurst, ed. Open Systems Interconnections. London: Macmillan Education, 1988.
35. L Hutchison. Local Area Network Architectures. Reading, MA: Addison-Wesley, 1988.
36. W Stalling. Handbook of Computer Communications Standards. Indianapolis, In: Howard W. Sams & Company, 1987.
37. S Hariri, A Chandhary, B Sarikaya. Architectural support for designing fault-tolerant open distributed systems. IEEE Computer 25(6): 50–62, 1992.
38. S Padalkar, G Karsai, C Biegl, J Sztipanovits, K Okuda, and Miyasaka. Real-Time Fault Diagnostics. IEEE Expert 6: 75–85, 1991.

Chapter 3.2

Stability

Allen R. Stubberud
University of California Irvine, Irvine, California

Stephen C. Stubberud
ORINCON Corporation, San Diego, California

2.1 INTRODUCTION

The stability of a system is that property of the system which determines whether its response to inputs, disturbances, or initial conditions will decay to zero, is bounded for all time, or grows without bound with time. In general, stability is a binary condition: either yes, a system is stable, or no, it is not; both conditions cannot occur simultaneously. On the other hand, control system designers often specify the relative stability of a system, that is, they specify some measure of how close a system is to being unstable. In the remainder of this chapter, stability and relative stability for linear time-invariant systems, both continuous-time and discrete-time, and stability for nonlinear systems, both continuous-time and discrete-time, will be defined. Following these definitions, criteria for stability of each class of systems will be presented and tests for determining stability will be presented. While stability is a property of a system, the definitions, criteria, and tests are applied to the mathematical models which are used to describe systems; therefore before the stability definitions, criteria, and tests can be presented, various mathematical models for several classes of systems will first be discussed. In the next section several mathematical models for linear time-invariant (LTI) systems are presented, then in the following sections the definitions, criteria, and tests associated with these models are presented. In the last section of the chapter, stability of nonlinear systems is discussed.

2.2 MODELS OF LINEAR TIME-INVARIANT SYSTEMS

In this section it is assumed that the systems under discussion are LTI systems, and several mathematical relationships, which are typically used to model such systems, are presented.

2.2.1 Differential Equations and Difference Equations

The most basic LTI, continuous-time system model is the nth order differential equation given by

$$\sum_{i=0}^{n} a_i \frac{d^i y}{dt^i} = \sum_{l=0}^{m} b_l \frac{d^l u}{dt^l} \tag{1}$$

where the independent variable t is time, $u(t)$, $t \geq 0$, is the system input, $y(t)$, $t \geq 0$, is the system output, the parameters a_i, $i = 0, 1, \ldots, n$, $a_n \neq 0$, and b_l, $l = 0, 1, \ldots, m$, are constant real numbers, and m and n are positive integers. It is assumed that $m \leq n$. The condition that $m \leq n$ is not necessary as a mathematical requirement; however, most physical systems satisfy this property. To complete the input–output relationship for this system, it is also necessary to specify n boundary conditions for the system output. For the purposes of this chapter, these n conditions will be n initial conditions, that is, a set of fixed values of $y(t)$ and its first $n-1$ derivatives at $t=0$. Finding the solution of this differential equation is then an initial-value problem.

A similar model for LTI, discrete-time systems is given by the nth-order difference equation

$$\sum_{i=0}^{n} a_i y(k+i) = \sum_{l=0}^{m} b_l u(k+l) \tag{2}$$

where the independent variable k is a time-related variable which indexes all of the dependent variables and is generally related to time through a fixed sampling period, T, that is, $t = kT$. Also, $u(k)$ is the input sequence, $y(k)$ is the output sequence, the parameters a_i, $i = 0, 1, \ldots, n$, $a_n \neq 0$, and b_l, $l = 0, 1, \ldots, m$, are constant real numbers, and m and n are positive integers with $m \leq n$. The condition $m \leq n$ guarantees that the system is causal. As with the differential equation, a set of n initial conditions on the output sequence completes the input–output relationship and finding the solution of the difference equation is an initial-value problem.

2.2.2 Transfer Functions

From the differential equation model in Eq. (1), another mathematical model, the transfer function of the system, is obtained by taking the one-sided Laplace transform of the differential equation, discarding all terms containing initial conditions of both the input $u(t)$ and output $y(t)$, and forming the ratio of the Laplace transform $Y(s)$ of the output to the Laplace transform $U(s)$ of the input. The final result is $H(s)$, the transfer function, which has the form

$$H(s) \equiv \left.\frac{Y(s)}{U(s)}\right|_{\text{ICs}=0} = \frac{\sum_{l=0}^{m} b_l s^l}{\sum_{i=0}^{n} a_i s^i} \tag{3}$$

where s is the Laplace variable and the parameters a_i, $i = 0, 1, \ldots, n$, and b_l, $l = 0, 1, \ldots, m$, and the positive integers m and n are as defined in Eq. (1).

For a discrete-time system modeled by a difference equation as in Eq. (2), a transfer function can be developed by taking the one-sided z-transform of Eq. (2), ignoring all initial-value terms, and forming the ratio of the z-transform of the output to the z-transform of the input. The result is

$$H(z) \equiv \left.\frac{Y(z)}{U(z)}\right|_{\text{ICs}=0} = \frac{\sum_{l=0}^{m} b_l z^l}{\sum_{i=0}^{n} a_i z^i} \tag{4}$$

where z is the z-transform variable, $Y(z)$ is the z-transform of the output $y(k)$, $U(z)$ is the z-transform of the input $u(k)$, and the other parameters and integers are as defined for Eq. (2).

2.2.3 Frequency Response Functions

Another mathematical model which can be used to represent a system defined by Eq. (1) is the frequency response function which can be obtained by replacing s by $j\omega$ in Eq. (3), thus forming

$$H(j\omega) = H(s)|_{s=j\omega} = \frac{\sum_{l=0}^{m} b_l (j\omega)^l}{\sum_{i=0}^{n} a_i (j\omega)^i} \tag{5}$$

where $j = \sqrt{-1}$ and ω is a real frequency variable measured in rad/sec. All of the other parameters and variables are as defined for Eq. (3).

The frequency response function for an LTI, discrete-time system defined by Eqs. (2) and (4) is obtained by replacing z by $e^{j\omega T}$ in Eq. (3) thus forming

$$H(e^{j\omega T}) = H(z)|_{z=e^{j\omega T}} = \frac{\sum_{l=0}^{m} b_l (e^{j\omega T})^l}{\sum_{i=0}^{n} a_i (e^{j\omega T})^i} \tag{6}$$

where the parameters and integers are as defined in Eq. (2). T is the sampling period as discussed for Eq. (2).

2.2.4 Impulse Responses

Transfer functions and frequency response functions are called frequency domain models, since their independent variables s, z, and ω are related to sinusoids and exponentials which are periodic functions and generalizations of periodic functions. Systems also have time domain models in which the independent variable is time. The most common of these is the impulse response function. For this chapter, the most general impulse response which will be considered is that which results if the inverse Laplace transform is taken of the transfer function. Systems of a more general nature than this also have impulse responses but these will not be considered here. Thus the impulse response function will be defined by

$$h(t) = L^{-1}[H(s)]$$

where $L^{-1}[\cdot]$ represents the inverse Laplace transform operator. The input-output relation (assuming all initial conditions are zero) for a system defined by an impulse response function is given by

$$y(t) = \int_{\tau=0}^{\tau=t} h(t-\tau)u(\tau)d\tau \tag{7}$$

where the limits on the integral result from earlier assumptions on the differential equation model.

For an LTI, discrete-time system, the impulse response sequence (in reality, an impulse does not exist in the discrete-time domain) will be defined as the inverse z transform of the transfer function in Equation (4), that is,

$$h(k) = Z^{-1}[H(z)]$$

where $Z^{-1}[\cdot]$ represents the inverse z-transform. The input–output relation (assuming all initial conditions are zero) of a discrete-time system defined by an impulse response sequence is given by

$$y(k) = \sum_{j=0}^{k} h(k-j)u(j) \tag{8}$$

where the limits on the summation result from earlier assumptions on the difference equation model.

2.2.5 State Space Models

A more general representation of LTI, continuous-time systems is the state space model, which consists of a set of n first-order differential equations which are linear functions of a set of n state variables, x_i, $i = 1, 2, \dots, n$, and their derivatives and a set of r inputs, u_i, $i = 1, 2, \dots, r$, and a set of m output equations which are linear functions of the n state variables and the set of r inputs. A detailed discussion of these models can be found in Santina et al. [1]. These equations can be written in vector–matrix form as

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \end{aligned} \tag{9}$$

where $\mathbf{x}$ is a vector composed of the n state variables, $\mathbf{u}$ is a vector composed of the r inputs, $\mathbf{y}$ is a vector composed of the m outputs, $\mathbf{A}$ is an $n \times n$ matrix of constant real numbers, $\mathbf{B}$ is an $n \times r$ matrix of constant real numbers, $\mathbf{C}$ is an $m \times n$ matrix of constant real numbers, and $\mathbf{D}$ is an $m \times r$ matrix of constant real numbers. As with Eq. (1) the independent variable t is time and n initial values of the state variables are assumed to be known, thus this is also an initial-value problem. Note that Eq. (1) can be put into this state space form.

For LTI, discrete-time systems there also exist state space models consisting of a set of n first-order difference equations and a set of m output equations. In vector–matrix form these equations can be written as

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k) \\ \mathbf{y}(k) &= \mathbf{C}\mathbf{x}(k) + \mathbf{D}\mathbf{u}(k) \end{aligned} \tag{10}$$

where k is the time variable, $\mathbf{x}(k)$ is the n-dimensional state vector, $\mathbf{u}(k)$ is the r-dimensional input vector, $\mathbf{y}(k)$ is the m-dimensional output vector, and $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ are constant matrices with the same dimension as the corresponding matrices in Eq. (9). Note that Eq. (2) can be put into this state space form.

2.2.6 Matrix Transfer Functions

As the differential equation model in Eq. (1) was Laplace transformed to generate a transfer function, the state space model in Eq. (9) can be Laplace transformed, assuming zero initial conditions, to form the matrix input–output relationship

$$\mathbf{Y}(s) = [\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}]\,\mathbf{U}(s) = \mathbf{T}(s)\,\mathbf{U}(s)$$

where $\mathbf{U}(s)$ is the Laplace transform of the input vector $\mathbf{u}(t)$, $\mathbf{Y}(s)$ is the Laplace transform of the output vector $\mathbf{y}(t)$, and the transfer function matrix of the system is given by

$$\mathbf{T}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D} \tag{11}$$

By similar application of the z-transform to the discrete-time state Eq. (10), the discrete-time matrix transfer function is given by

$$\mathbf{T}(z) = \mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D} \tag{12}$$

2.2.7 Matrix Impulse Responses

For an LTI continuous-time system defined by a state space model, the inverse Laplace transform of the matrix transfer function in Eq. (11) produces a matrix impulse response of the form

$$\mathbf{T}(t) = \mathbf{C}\Phi(t)\,\mathbf{B} + \mathbf{D} \tag{13}$$

where $\Phi(t) = e^{\mathbf{A}t}$ is called the state transition matrix.

The input–output relationship, excluding initial conditions, for a system described by this matrix impulse response model is given by the convolution integral given by

$$\mathbf{y}(t) = \mathbf{C} \int_{\tau=0}^{\tau=t} e^{\mathbf{A}(t-\tau)} \mathbf{B}\mathbf{u}(\tau)\, d\tau + \mathbf{D}\mathbf{u}(t)$$

where all of the parameters and variables are as defined for the state space model in Eq. (9) and the limits on the integral result from the conditions for the state space model.

Similarly, for an LTI, discrete-time system, a matrix impulse response is generated by the inverse z-transform of the matrix transfer function of Eq. (12), thus forming

$$\mathbf{T}(k) = \mathbf{C}\Phi(k)\,\mathbf{B} + \mathbf{D} \tag{14}$$

where $\Phi(k) = \mathbf{A}^k$ is called the state transition matrix. the input–output relationship, excluding initial conditions, for a system described by this matrix impulse response model is given by the convolution summation given by

$$\mathbf{y}(k) = \mathbf{C} \sum_{j=0}^{k} \mathbf{A}^{k-j} \mathbf{B}\mathbf{u}(j) + \mathbf{D}\mathbf{u}(j)$$

2.2.8 Summary of Section

In this section, a total of 14 commonly used models for linear time-invariant systems have been presented. Half of these models are for continuous-time systems and the rest are for discrete-time systems. In the next section, criteria for stability of systems represented by these models and corresponding tests for stability are presented.

2.3 DEFINITIONS OF STABILITY

When we deal with LTI systems, we usually refer to three main classes of stability: absolute stability, marginal stability, and relative stability. While each class of stability can be considered distinct, they are interrelated.

2.3.1 Absolute Stability

Absolute stability is by far the most important class of stability. Absolute stability is a binary condition: either a system is stable or it is unstable, never both. In any sort of operational system, this is the first question that needs to be asked. In order to answer this question, we need to determine whether or not the system has an input.

A zero-input system is said to be absolutely stable if, for any set of initial conditions, the system output:

1. Is bounded for all time $0 < t < \infty$
2. Returns to the equilibrium point 0 as time approaches infinity.

This type of stability is also referred as asymptotic stability. As described in Hostetter et al. [2], this nomenclature is used because the impulse response of a stable system asymptotically decays to zero.

For the systems considered in this chapter, the definition for asymptotic stability can be mathematically defined in terms of the impulse response function as

$$|h(t)| < \infty \qquad \forall\, t \geq t_0$$

and

$$\lim_{t\to\infty} |h(t)| = 0$$

or for the discrete-time system as

$$|h(kT)| < \infty \qquad \forall\, k \geq 0$$

and

$$\lim_{k\to\infty} |h(kT)| = 0$$

When the system in question has an input, whether an external input, control input, or a disturbance, we change the definition of stability to that of bounded-input–bounded-output stability, also referred to as BIBO stable or simply BIBO.

A system is said to be BIBO stable if every bounded input results in a bounded output. Mathematically, given an arbitrary bounded input $u(t)$, that is,

$$|u(t)| \leqslant N < +\infty \qquad \forall\, t \geq 0$$

the resulting output is bounded, that is, there exists a real finite number M such that

$$|y(t)| \leqslant M < +\infty \qquad \forall\, t > 0$$

The mathematical definition for the discrete-time case is identical.

The important consequence of the two previous definitions for stability is the location of the poles of the transfer function of a stable system. The transfer function is in the form of a ratio of polynomials as in Eq. (3). The denominator polynomial is called the characteristic polynomial of the system and if the characteristic polynomial is equated to zero, the resulting equation is called the characteristic equation. The roots of the characteristic equation are called the poles of the transfer function. As shown in Kuo [3], a necessary and sufficient condition for a system to be absolutely stable is that all of its poles must lie in the left half plane (the poles have negative real part) for continuous-time systems, or lie within the unit circle (pole magnitude less than 1) for discrete-time systems. For systems defined by the state space model, the poles are the eigenvalues of the state transition matrix **A**.

2.3.2 Marginal Stability

As with all definitions in engineering, there exist some exceptions to the rules of stability. Several important systems, such as the differentiation operator and the pure integrator as continuous-time systems, violate the rules of asymptotic stability and BIBO stability, respectively. Other cases exist in the set of systems that have resonant poles along the $j\omega$-axis (imaginary axis), for continuous-time systems, or on the unit circle, for discrete-time systems.

While the differentiation operator violates the asymptotic stability definition, since the impulse response is not bounded, it does satisfy the BIBO definition. In any case, it is generally considered stable. For an integrator, the impulse response is a constant and thus bounded and in the limit is a constant, as described in Oppenheim et al. [4]; however, for a step input, which is bounded, the output grows without bound. The same would occur if the input is a sinusoid of the same frequency as any imaginary axis (unit circle for discrete-time systems) poles of a system. Since such systems only "blow up" for a countable finite number of bounded inputs, such systems are often considered stable. However, for any input along the imaginary axis (unit circle), the output of these systems will neither decay to zero nor even to a stable value. Systems such as these are referred to as marginally stable. For these systems the roots of the polynomial, or eigenvalues of the state transition matrix, that do not meet the criteria for absolute stability, lie on the imginary axis (zero real-value part) for a continuous-time system or lie on the unit circle (magnitude equal to 1) for a discrete-time system. While some consider such systems as stable, others consider them unstable because they violate the definition of absolute stability.

2.3.3 Relative Stability

Once we have determined that a system is absolutely stable, usually we desire to know "How stable is it?" To a design engineer such a measure provides valuable information. It indicates the allowable variation or uncertainty that can exist in the system. Such a measure is referred to as relative stability. Many different measures for relative stability are available so it is important to discuss desirable measures.

2.4 STABILITY CRITERIA AND TESTS

Now that we have defined stability, we need tools to test for stability. In this section, we discuss various criteria and tests for stability. This section follows the format from the preceding section in that we first discuss the techniques for determining absolute stability, followed by those for marginal stability, and finally those for relative stability.

2.4.1 Absolute Stability Criteria

There exist two approaches for determining absolute stability. The first is to use time-domain system models. The second, and the most usual, is to deal with the transfer function. Both types of techniques are presented here.

2.4.1.1 Zero-Input Stability Criteria

Given a zero-input system in the time-domain form of Eq. (1) with all terms on the right-hand side equal to zero, stability is defined in terms of the impulse response function which, for stability, must satisfy the following conditions:

1. There exists a finite real number M such that $|h(t)| \leq M$ for all $t \leq t_0$.
2. $\lim_{l\to\infty} |h(t)| = 0$.

Similar criteria for the impulse response sequence $h(kT)$ can be used to determine stability for the discrete-time case.

If a system is modeled by the state-space form of Eq. (9), the criteria become:

1. There exists a value M such that $\|x(t)\| \leq M$ for all $t \leq t_0$.
2. $\lim_{t\to\infty} \|x(t)\| = 0$.

2.4.1.2 Bounded-Input–Bounded-Output Time Domain Criteria

When systems such as the continuous-time system, Eq. (1), or the discrete-time system, Eq. (2), are modeled by their impulse response functions, BIBO stability can be demonstrated directly from the definition and the property of convolution. Since the input is bounded there exists a value N such that

$$|u(t)| \le N < \infty$$

If the output $y(t)$ is bounded, then there exists a value M such that

$$|y(t)| \le M < \infty$$

which implies that

$$\begin{aligned} |y(t)| &\le \int_0^\infty |u(t-\tau)\,h(\tau)|\,d\tau \le \int_0^\infty |u(t-\tau)|\,d\tau \\ &\le \int_0^\infty N|h(\tau)|\,d\tau < \infty \end{aligned}$$

Since the input is bounded by a constant, all we need to show is that

$$\int_0^\infty |h(\tau)|\,d\tau < \infty$$

For the discrete-time case, we similarly need to show

$$\sum_{k=0}^{\infty} |h(kT)| < \infty$$

For the state space form of the problem, Eq. (9), we use the formulation

$$x(t) = \int_0^t \Phi(t-\tau)\,u(\tau)\,d\tau$$

which results in the following tests, for continuous-time and discrete-time systems, respectively:

$$\int_0^\infty \|\Phi(\tau)\|\,d\tau < \infty$$

or

$$\sum_{k=0}^{\infty} \|\Phi(kT)\| < \infty$$

where $\Phi(t)$ is the state-transition matrix of the continuous-time system and $\|\cdot\|$ represents a matrix norm and similarly for the discrete-time system.

2.4.1.3 Polynomial Coefficient Test (Continuous-Time Systems)

Polynomial coefficient tests provide a quick method to determine if a system is unstable by looking for poles of the system's characteristic polynomial in the right half plane. These are typically the first of several tests used to determine whether or not the roots of the characteristic equation, or poles, of the system lie in the right half plane or along the $j\omega$-axis. While they provide sufficient conditions for unstable poles, they do not provide necessary conditions. However, they are fast and require no computation.

Given a polynomial

$$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0$$

we may be able to determine if there exist any roots that lie outside the left half plane using the following table.

Properties of polynomial coefficients	Conclusion about roots from the coefficient test
Differing algebraic signs	At least one root in right half plane
Zero-valued coefficients	Imaginary axis root and/or a pole in right half plane
All algebraic signs same	No information

Example 1. *Given the polynomial*

$$4x^3 + 7x^2 - 3x + 1 = 0$$

we know that at least one root lies in the right half plane because there is at least one sign change in the coefficients. If this were the characteristic equation of a system, the system would be unstable.

However, we would not know about any of the root locations for the equation

$$x^5 + 3x^4 + 12x^3 + x^2 + 92x + 14 = 0$$

because all of the signs are the same and none of the coefficients are zero.

2.4.1.4 Routh Test (Continuous-Time Systems)

While the coefficient tests can tell you if you have an unstable system, it cannot inform you whether or not you have a stable system or how many poles lie outside the left half plane. In order to overcome this difficulty, we apply the Routh test. In much of the literature, this is also referred to as the Routh–Hurwitz test.

Given an equation of the form

$$a_n s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0 = 0$$

the Routh test is performed using the Routh table:

$$\begin{array}{c|cccc}
s^n & a_n & a_{n-2} & a_{n-4} & \dots \\
s^{n-1} & a_{n-1} & a_{n-3} & a_{n-5} & \dots \\
s^{n-2} & b_1 & b_2 & b_3 & \dots \\
s^{n-3} & c_1 & c_2 & c_3 & \dots \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
s^0 & & & &
\end{array}$$

where the a_is are the coefficients of the polynomial and

$$b_l = \frac{a_{n-1}a_{n-2i} - a_n a_{n-(2i+1)}}{a_{n-1}} \quad \text{and}$$

$$c_i = \frac{b_1 a_{n-(2i+1)} - a_{n-1} b_{i+1}}{b_1}$$

A polynomial has all of its roots in the left half plane if, and only if, all of the elements in the first column of Routh table have the same sign. If sign changes are present, then the number of roots with positive real parts is equal to the number of changes of sign of the elements in the first column.

Example 2. *Given the polynomial equation*

$$s^4 + 4s^3 + 101s^2 + 494s + 600 = 0$$

we create the Routh table

$$\begin{array}{c|ccc}
s^4 & 1 & 101 & 600 \\
s^3 & 4 & 494 & 0 \\
s^2 & & & \\
s^1 & & & \\
s^0 & & &
\end{array}$$

We compute the third row from

$$b_1 = \frac{4 \times 101 - 1 \times 494}{4} \quad \text{and}$$

$$b_2 = \frac{4 \times 600 - 1 \times 0}{4}$$

The resulting Routh table is

$$\begin{array}{c|ccc}
s^4 & 1 & 101 & 600 \\
s^3 & 4 & 494 & 0 \\
s^2 & -90/4 & 600 & \\
s^1 & & & \\
s^0 & & &
\end{array}$$

Row four is computed similarly from rows two and three:

$$\begin{array}{c|ccc}
s^4 & 1 & 101 & 600 \\
s^3 & 4 & 494 & 0 \\
s^2 & -90/4 & 600 & \\
s^1 & 1802/3 & & \\
s^0 & & &
\end{array}$$

The completed Routh table becomes

$$\begin{array}{c|ccc}
s^4 & 1 & 101 & 600 \\
s^3 & 4 & 494 & 0 \\
s^2 & -90/4 & 600 & \\
s^1 & 1802/3 & & \\
s^0 & 600 & &
\end{array}$$

We note that we have two sign changes, rows two to three and three to four, which implies that two of our roots are in the right half plane and three are in the left half plane.

As with all techniques there are problems that can occur with the Routh table. Polynomials do exist that can result in the computation of a zero in the left column in the Routh table. An excellent example can be found in Hostetter et al. [2]. A polynomial equation of the form

$$s^4 + 3s^3 + 2s^2 + 6s + 4 = 0$$

will result in the following Routh table:

$$\begin{array}{c|ccc}
s^4 & 1 & 2 & 4 \\
s^3 & 3 & 6 & 0 \\
s^2 & 0 & 4 & \\
s^1 & & & \\
s^0 & & &
\end{array}$$

Note the zero in the third row of the table. There are two methods to alleviate this problem. The first is to multiply the original polynomial by a first-order known-root polynomial such as $s + 1$. This produces a new polynomial

$$s^5 + 4s^4 + 5s^3 + 8s^2 + 10s + 4 = 0$$

whose computed Routh table is

$$\begin{array}{c|ccc}
s^5 & 1 & 5 & 10 \\
s^4 & 4 & 8 & 4 \\
s^3 & 3 & 9 & \\
s^2 & -4 & 4 & \\
s^1 & 12 & 0 & \\
s^0 & 4 & &
\end{array}$$

The second technique is to realize that a minor perturbation in any coefficient would result in a nonzero entry in the third element of the first column. For our example, the perturbed Routh table would be

$$\begin{array}{c|ccc} s^4 & 1 & 2 & 4 \\ s^3 & 3 & 6 & 0 \\ s^2 & \varepsilon & 4 & \\ s^1 & \dfrac{6\varepsilon - 12}{\varepsilon} & & \\ s^0 & 4 & & \end{array}$$

Once the Routh table is completed, we let ε go to zero. This results in the Routh table of

$$\begin{array}{c|ccc} s^4 & 1 & 2 & 4 \\ s^3 & 3 & 6 & 0 \\ s^2 & 0 & 4 & \\ s^1 & -\infty & & \\ s_0 & 4 & & \end{array}$$

A final note is that while ε can be of any sign it should be taken to have the same sign as the previous column element.

2.4.1.5 Nyquist Stability Criterion

The Routh test provides information pertaining to stability and the number of unstable poles of a system. However, often when we are designing a control system, we want even more information. The Nyquist stability criterion is a quasigraphical frequency domain method for determining stability and can be used for design as well.

To implement this technique, we start with the term of the general input–output gain rule or Mason's gain formula [3,5]:

$$1 + F(s) = 1 + \frac{\text{Num}(s)}{\text{Den}(s)} \tag{15}$$

This is the denominator of the closed-loop system transfer function

$$\frac{F(s)}{1 + F(s)}$$

where we note that $F(s)$ is the ratio of the polynomials Num(s) to Den(s). To generate a Nyquist plot, we let $s = j\omega$, calculate the real and imaginary parts of Eq. (15) for all of the values along the $j\omega$-axis, and then plot the results on the complex plane.

Example 3. *The curve for*

$$1 + F(s) = 1 + \frac{s^2 + 3s + 2}{s^3 + 3s^2 - 4}$$

when $s = j\omega$, $-\infty < \omega < \infty$ is seen in Fig. 1. The dashed line indicates results for the negative frequency values.

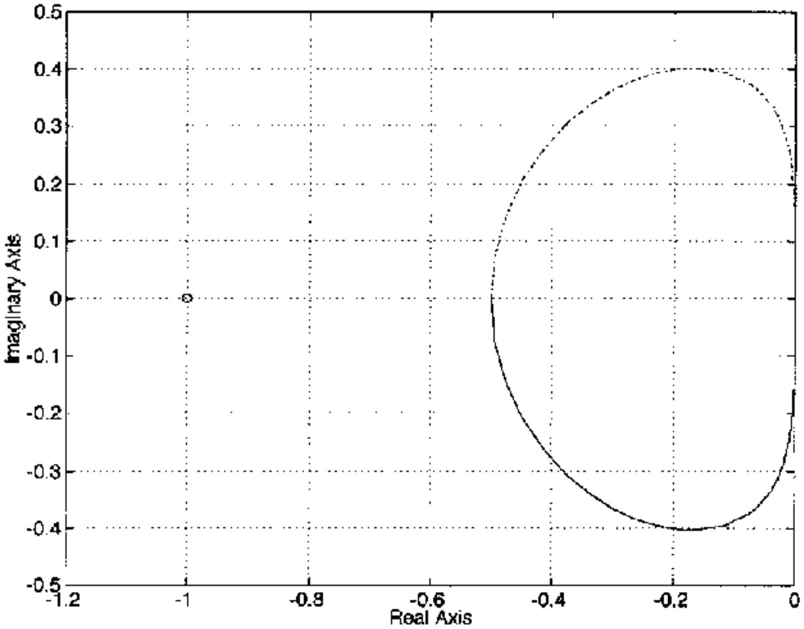


Figure 1 The Nyquist curve does not encircle the point $(-1, 0)$.

A closed-loop system with

$$\mathrm{Num}(s) + \mathrm{Den}(s) = 0$$

as the characteristic equation of its open-loop transfer function is stable if, and only if, the Nyquist plot encircles in the counterclockwise direction the point $(-1, 0)$ in the complex plane the same number of times as the number of poles of $F(s)$ which have a positive real parts. If there are no poles of $F(s)$ with positive real parts, then the Nyquist plot does not encircle the point $(-1, 0)$.

The number of the unstable roots of the numerator of $1 + F(s)$, which are the poles of the closed-loop system, is equal to the difference between the number of right half plane poles and the number of counterclockwise encirclements.

For the example above, we have no encirclements and one right half plane pole. Therefore, we have one right half plane zero, which translates into one right half plane pole in our closed-loop system.

For discrete-time systems, we change from using the imaginary axis for computing the Nyquist plot to the frequency values along the unit circle. The number of zeros in the system are related to those outside the unit circle.

2.4.1.6 Schur–Cohn Test (Discrete-Time Systems)

One of the earliest techniques used by control engineers to determine the stability of a discrete-time system is the Schur–Cohn test. For a given characteristic polynomial,

$$a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$$

we can determine the system stability by examining the determinants

$$\Delta_k = \begin{vmatrix} a_0 & 0 & 0 & \dots & 0 & a_n & a_{n-1} & \dots & a_{n-k+1} \\ a_1 & a_0 & 0 & \dots & 0 & 0 & a_n & \dots & a_{n-k+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{k-1} & a_{k-2} & a_{k-3} & \dots & 0 & 0 & 0 & \dots & a_n \\ \bar{a}_n & 0 & 0 & \dots & 0 & \bar{a}_0 & \bar{a}_1 & \dots & \bar{a}_{k-1} \\ \bar{a}_{n-1} & \bar{a}_n & 0 & \dots & 0 & 0 & \bar{a}_0 & \dots & \bar{a}_{k-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \bar{a}_{n-k+1} & \bar{a}_{n-k+2} & \bar{a}_{n-k+3} & \dots & \bar{a}_n & 0 & 0 & \dots & \bar{a}_0 \end{vmatrix}$$

where

$$k = 1, 2, \dots, n$$

and

$$\bar{a}_k = \text{complex conjugate } a_k$$

The system is said to be stable if, and only if,

$$\begin{array}{ll} \Delta_k < 0 & k \text{ odd} \\ \Delta_k > 0 & k \text{ even} \end{array}$$

Example 4. *Determine if the system with characteristic polynomial*

$$4z^2 + 2z + 2$$

is stable. We form the determinants

$$\Delta_1 = \begin{vmatrix} 2 & 4 \\ 4 & 2 \end{vmatrix} = 4 - 16 = -12$$

$$\Delta_2 = \begin{vmatrix} 2 & 0 & 4 & 2 \\ 2 & 2 & 0 & 4 \\ 4 & 0 & 2 & 2 \\ 2 & 4 & 0 & 2 \end{vmatrix}$$
$$= (4 - 16)^2 - 4(2 - 4)^2 144 - 16 = 128$$

Since $\Delta_1 < 0$ *and* $\Delta_2 > 0$, *the system is stable.*

2.4.1.7 Jury Test (Discrete-Time Systems)

The Jury test for determining the stability of discrete-time systems is similar to the Routh criterion for continuous-time systems and is much simpler to implement than the Schur–Cohn test. Where the Routh test determines if the roots of a polynomial equation are in the left half plane, the Jury test determines whether the roots of a polynomial equation are inside the unit circle.

The Jury test begins with the development of the Jury array:

$$\begin{array}{r|lllllll} 1 & a_0 & a_1 & a_2 & \dots & \dots & a_{n-1} & a_n \\ 2 & a_n & a_{n-1} & a_{n-2} & \dots & \dots & a_1 & a_0 \\ 3 & b_0 & b_1 & b_2 & \dots & \dots & b_{n-1} & \\ 4 & b_{n-1} & b_{n-2} & b_{n-3} & \dots & \dots & b_0 & \\ 5 & c_0 & c_1 & c_2 & \dots & c_{n-2} & & \\ 6 & c_{n-2} & c_{n-3} & c_{n-4} & \dots & c_0 & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & & & \\ 2n-5 & r_0 & r_1 & r_2 & r_3 & & & \\ 2n-4 & r_3 & r_2 & r_1 & r_0 & & & \\ 2n-3 & s_0 & s_1 & s_2 & & & & \end{array}$$

where

$$b_k = \begin{vmatrix} a_0 & a_{n-k} \\ a_n & a_k \end{vmatrix}$$

$$c_k = \begin{vmatrix} b_0 & b_{n-1-k} \\ b_{n-1} & b_k \end{vmatrix}$$

$$s_0 = \begin{vmatrix} r_0 & r_3 \\ r_3 & r_0 \end{vmatrix}$$

$$s_1 = \begin{vmatrix} r_0 & r_2 \\ r_2 & r_1 \end{vmatrix}$$

and

$$s_2 = \begin{vmatrix} r_0 & r_1 \\ r_3 & r_2 \end{vmatrix}$$

The terms a_i are the coefficients of the polynomial

$$H(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$$

The elements of the next two rows are computed using the determinants defined above. The process continues, with each new pair of rows having one less column than the prior pair of rows until row $2n-3$ is computed. This last row has only three elements and does not have its reversed-order pair. Thus a seventh-order polynomial would have 11 rows.

Once the Jury array is complete, we can perform the entire Jury test which provides necessary and sufficient conditions for the roots of the polynomial equation to have magnitudes less than 1. Note that the first two parts of the test can be completed without construction of the Jury array.

1. $H(z)|_{z=1} > 0.$
2. $H(z)|_{z=-1} \begin{cases} > 0 & \text{for } n \text{ even,} \\ < 0 & \text{for } n \text{ odd.} \end{cases}$
3. $|a_0| < a_n$

   $|b_0| > |b_{n-1}|$
   $|c_0| > |c_{n-2}|$

   $|r_0| > |r_3|$
   $|s_0| > |s_2|.$

Example 5. *Given the characteristic polynomial*

$$H(z) = z^4 + 0.6z^3 + 0.3z^2 - 0.5z + 0.25$$

is the system stable?

Test 1:

$$H(1) = 1 + 0.6 + 0.3 - 0.5 + 0.25 = 1.65 > 0$$

Test 2:

$$H(-1) = 1 - 0.6 + 0.3 + 0.5 + 0.25 = 1.75 > 0$$

Construct the Jury array:

1	0.25	−0.5	0.3	0.6	1
2	1	0.6	0.3	−0.5	0.25
3	−0.9375	−0.725	−0.225	0.65	
4	0.65	−0.225	−0.725	−0.9375	
5	0.4654	0.8259	0.6821		

Test 3:

$$0.25 < 1$$
$$0.9375 > 0.65$$
$$0.4654 < 0.6821$$

Since the final inequality violates Test 3, the system is unstable.

We note that the element a_n is assumed to be positive and can always be made so without changing the roots of the system by multiplying the polynomial by -1.

2.4.1.8 w-Plane Transformation/Routh Test for Discrete-Time Systems

While the Jury test informs us of stability, it does not tell us how many poles are outside the unit circle. If this is desired, we can employ the Routh test. We cannot perform the test directly on the polynomial because again we would only be able to determine whether or not the poles of the discrete-time system were in the right or left half planes, which is basically useless. In order to invoke this test, we perform the bilinear transformation [5] on the discrete-time based polynomial. The complex variable z is then transformed into the new complex variable w which is similar to the complex variable s in the familiar s-plane.

The bilinear transformation is given by the equivalent expressions:

$$z = \frac{1+w}{1-w} \qquad w = \frac{z-1}{z+1}$$

This transformation transforms roots of a polynomial equation inside the unit circle of the z-plane into the left half of the s-plane, transforms the roots outside the unit circle into the right half plane, and the roots on the unit circle onto the j-axis.

We can examine the location of the roots of the polynomial equation

$$H(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0 = 0$$

by letting $z = (1+w)/(1-w)$, thus generating the equation

$$H(w) = a_n\left(\frac{1+w}{1-w}\right)^n + a_{n-1}\left(\frac{1+w}{1-w}\right)^{n-1} + \cdots + a_1\frac{1+w}{1-w} + a_0 = 0$$

We need only concern ourselves with the numerator of this equation to find root locations:

$$a_n(1+w)^n - a_{n-1}(1+w)^{n-1}(1-w) + \cdots + a_1(1+w)(1-w)^{n-1} + a_0(1-w)^n = 0$$

and apply the Routh test to determine root locations.

Example 6. *Given the discrete-time characteristic polynomial*

$$H(z) = 4z^2 - 4z + 1$$

we want to determine stability and the number of any unstable poles.

The transformed equation is given by

$$w^2 + 6w + 9 = 0$$

Now we apply the Routh test which indicates that this is the characteristic polynomial of an absolutely stable system.

2.4.1.9 Eigenvalue Computation

If a system is modeled in the state-space form

$$\dot{\mathbf{x}}(t) = \mathbf{Ax}(t) + \mathbf{Bu}(t)$$
$$\mathbf{y}(t) = \mathbf{Cx}(t) + \mathbf{Du}(t)$$

the stability is determined by the location of the eigenvalues of the matrix **A**. For continuous time systems, the eigenvalues must be in the left half plane. Similarly, for discrete-time systems, the magnitude of the eigenvalues must be less than one. The question becomes how do we find the eigenvalues. There are many techniques to compute the eigenvalues of a matrix. Several can be found in Wilkinson [6] and Golub and Van Loan [7]. New techniques are probably being developed as you read this. A computer implementation can be found in any numerical linear algebra package such as EISPACK. In this section we outline one technique, the real Schur decomposition.

The real Schur form is a block triangular form

$$\begin{bmatrix} D_{11} & X & X & X & X \\ & D_{22} & X & X & X \\ & & \ddots & \vdots & \vdots \\ & 0 & & D_{(n-1)(n-1)} & X \\ & & & & D_{nn} \end{bmatrix}$$

where the diagonal block elements, D_{ii}, are either 1×1 or 2×2 element blocks. The single-element blocks are the real eigenvalues of the system, while the 2×2 blocks represent the complex and imaginary eigenvalues via

$$\Lambda(D_{ii}) = \lambda_i \pm j\mu_i$$

where

$$D_{ii} = \begin{bmatrix} \lambda_i & \mu_i \\ -\mu_i & \lambda_i \end{bmatrix}$$

The algorithm begins by reducing the matrix **A** to what is referred to as an upper Hessenberg form

$$\begin{bmatrix} X & X & X & X & X \\ X & X & X & X & X \\ 0 & X & X & X & X \\ 0 & 0 & X & X & X \\ 0 & 0 & 0 & X & X \end{bmatrix}$$

We then use the iteration

```
for $k = 1, 2, 3, \ldots$
$H_{k-1} = U_k R_k$
$H_k = R_k U_k$
end
```

where $H_{k-1} = U_k R_k$ is a QR factorization, a technique that reduces a matrix to a product of an orthogonal matrix postmultiplied by an upper triangular matrix [7].

Once the algorithm is completed, you check for any eigenvalues whose real part is nonnegative. Each such eigenvalue is an unstable pole of the transfer function.

2.4.1.10 Kharatonov Polynomials

When actually designing a real system, you may ask questions about variations in the parameters of the physical system compared to the design parameters. Resistors and motors may be the "same" but no two are identical. Operating conditions and/or age can cause changes in operating parameters. Will these changes affect the system's stability? Hopefully not. However, in today's litigious society, we need a little more than hope.

To check the stability for your system over a range of values for each coefficient, we can use the Kharatonov polynomials [8]. Given a polynomial with a range of values for each coefficient

$$[a_n^-, a_n^-)s^n + [a_{n-1}^+, a_{n-1}^-]s^{n-1} + \cdots + [a_1^+, a_1^-]s + [a_0^+, a_0^-] = 0$$

where $[a_i^+, a_i^-]$ indicates the bounds of a coefficient, we can determine the stability of the system by determining the stability of the following four polynomials:

$$p_1(s) = a_0^+ + a_1^+ s + a_2^- s^2 + a_3^- s^3 + a_4^+ s^4 + a_5^+ s^5 + a_6^- s^6 + a_7^- s^7 + \cdots$$

$$p_2(s) = a_0^- + a_1^- s + a_2^+ s^2 + a_3^+ s^3 + a_4^- s^4 + a_5^- s^5 + a_6^- s^6 + a_7^+ s^7 + \cdots$$

$$p_3(s) = a_0^+ + a_1^- s + a_2^- s^2 + a_3^+ s^3 + a_4^+ s^4 + a_5^- s^5 + a_6^- s^6 + a_7^+ s^7 + \cdots$$

$$p_4(s) = a_0^- + a_1^+ s + a_2^+ s^2 + a_3^- s^3 + a_4^- s^4 + a_5^+ s^5 + a_6^- s^6 + a_7^- s^7 + \cdots$$

Now all that needs to be shown is that the roots of each of these four equations are in the left half plane, and we have guaranteed stability over the entire range of all the coefficients given.

2.4.2 Marginal Stability

2.4.2.1 Polynomial Test (Continuous-Time Systems)

If our interest is not in absolute stability, the coefficient test results change. If a coefficient is zero, then we know that at least one root can lie on the imaginary axis. However, the location of the other roots, if the signs do not change, are not known. Thus, the result of a zero coefficient is necessary but not sufficient for marginal stability. The table below may give us information about relative stability.

Properties of polynomial coefficients	Conclusion about roots from the coefficient test
Differing algebraic signs	At least one root in right half plane
Zero-valued coefficients	No information
All algebraic signs same	No information

2.4.2.2 Routh Test (Continuous-Time Systems)

In the earlier section on the Routh test, we avoided asking the question what happens if the roots of the polynomial lie on the imaginary axis. If a system is marginally stable or just has imaginary roots, the Routh table can terminte prematurely. In this section, we provide a technique for dealing with this problem.

Given the polynomial

$$H(s) = s^4 + 5s^3 + 10s^2 + 20s + 24$$

the computed Routh table is

$$\begin{array}{c|ccc} s^4 & 1 & 10 & 24 \\ s^3 & 5 & 20 & 0 \\ s^2 & 6 & 24 & \\ s^1 & 0 & 0 & \\ s^0 & & & \end{array}$$

As expected, it has terminated prematurely with a row of zeros. This implies that

$$6s^2 + 24$$

is a factor of the original polynomial. We replace the zero row of the Routh table with the derivative of this factor, that is, $(d/ds)(6s^2 + 24) = 12s$,

$$\begin{array}{c|ccc} s^4 & 1 & 10 & 24 \\ s^3 & 6 & 20 & 0 \\ s^2 & 6 & 24 & \\ s^1 & 12 & 0 & \\ s^0 & 24 & & \end{array}$$

and continue computing the Routh table. The result implies that we have two roots in the left half plane, and two imaginary roots, thus our system is marginally stable. If there were a change in signs between any row, then we would have a pole in the right half plane.

Any time that an imaginary pair of roots exists, then the Routh table will contain a zero row. All of the roots will be contained in the factor polynomial.

2.4.2.3 Other Algorithms

The w-plane, eigenvalue, and Kharatonov techniques can be expanded to look for marginally stable poles just by changing what we are looking for, nonpositive poles instead of nonnegative poles.

2.4.3 Relative Stability

Our final discussion on stability for linear time-invariant systems is about relative stability. We have presented several techniques to determine whether or not a system is stable. However, we often like to know how stable a system is. To what degree can the system be changed before stability is lost. Relative stability techniques give this measure of the degree of stability.

2.4.3.1 Distance Measure of the Poles

Relative stability is important in design because the locations of the poles have a great deal of effect on the performance of the system. For instance, complex conjugate pole pairs that are close to the imaginary axis can cause ringing behavior in the system. Poles that have a real part whose magnitude is less than the imaginary part can show resonance behavior as the input frequency gets closer to the resonant frequency. Therefore, we should use a measure of distance from the imaginary axis as a measure of relative stability, right? *Wrong*!

As seen in Oppenheim et al. [4], Butterworth poles can be close to the imginary axis but the system behavior is quite stable and without a resonant frequency. Also, in state space problems small changes in particular elements can cause major changes in the system behavior. For example, the state transition matrix

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & a & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ \varepsilon & 0 & 0 & 0 & 0 \end{bmatrix}$$

has its poles located at the origin if a and ε are set to zero. If a is set to 100, the poles are still located at the origin. However, if ε is set to 1, the system poles are distributed on the unit circle, which for both discrete-time and continuous-time systems prevents absolute stability. The change of ε is small compared to that of a, yet it changes the stability of the system substantially. The same can happen when the parameters of the characteristic polynomial change. This is one reason for the development of Kharatonov's stability test.

Pole location can tell us a great deal about the system behavior, but the simple measure of distance from the j-axis should not be used as a measure of relative stability.

2.4.3.2 Gain and Phase Margin

Gain and phase margin have long been used as a useful measure of relative stability. Both of these quantities are computed using the open-loop transfer function

$$1 + F(s) = 1 + \frac{\mathrm{Num}(s)}{\mathrm{Den}(s)}$$

the same that was used for the Nyquist stability criterion. As with the Nyquist stability criterion, we note that the technique works as well for discrete-time systems. Simply replace all references to the imaginary axis with references to the unit circle.

We define gain margin as the magnitude of the reciprocal of the open-loop transfer function at the phase crossover frequency, $\omega_\pi(\text{phase} = -180^\circ)$.

Phase margin is defined as 180° plus the phase angle of the open-loop transfer function at the frequency where the gain is equal to unity.

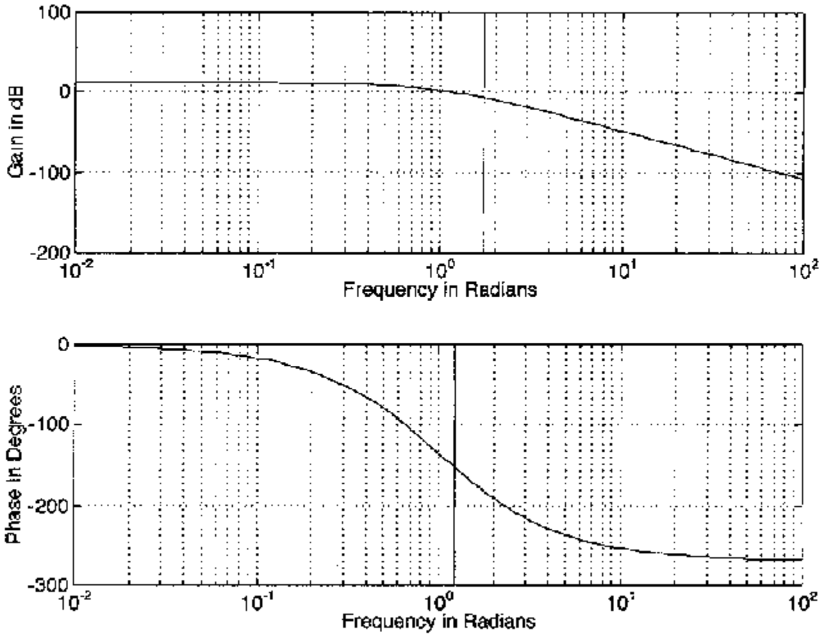


Figure 2 Magnitude and phase Bode plots demonstrate gain and phase margins.

Mathematically, we write gain margin as

$$\text{gain margin} \equiv \frac{1}{|F(\omega_\pi)|}$$

and phase margin as

$$\phi_{PM} \equiv [180 + \arg(F(\omega_1)] \text{ degrees}$$

Note that we can also define gain margin in decibels

$$\text{gain margin} \equiv -20 \log |F(\omega_\pi)|$$

To use these two quantities, we need to interpret them. Gain margin is measured as the number of decibels below 0 dB that the open-loop transfer function is at the phase crossover frequency. Phase margin is measured as the number of degrees above −180° that the phase of the open-loop transfer is when its gain is equal to unity.

While both a positive phase and gain margin can usually indicate stability of a system, there do exist cases where this is not true, thus care should be taken when determining absolute stability. If a system is not absolutely stable, then relative stability has no meaning.

Example 7. *Given the open-loop transfer function*

$$F(s) = \frac{4}{s^3 + 3s^2 + 3s + 1}$$

determine the phase and gain margin. We shall use a Bode plot [2,5,8,9] to perform the analysis. As seen in Fig. 2, the phase crossover frequency is at 1.7348 rad/sec. This implies that the gain margin is

$$\text{gain margin} = 1.9942 = 5.9954\,\text{dB}$$

The phase margin is measured at a frequency of 1.234 rad/sec. The phase margin is

$$\text{phase margin} = 27.0882^\circ$$

2.5 STABILITY OF NONLINEAR SYSTEMS

In this section we discuss the stability of nonlinear systems, both continuous-time and discrete-time. As for LTI systems, stability is a binary concept; however, beyond that, stability of nonlinear systems is much more complex, thus the stability criteria and tests are more difficult to apply than those for LTI systems. Two models will be used to represent nonlinear systems. For nonlinear, continuous-time systems the model is

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{f}[\mathbf{x}(t), \mathbf{u}(t)] \\ \mathbf{y}(t) &= \mathbf{g}[\mathbf{x}(t), \mathbf{u}(t)] \end{aligned} \tag{16}$$

where the nonlinear differential equation is in state variable form and the second equation is the output equation of the system. For nonlinear, discrete-time systems the model is

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{f}[\mathbf{x}(k), \mathbf{u}(k)] \\ \mathbf{y}(k) &= \mathbf{g}[\mathbf{x}(k), \mathbf{u}(k)] \end{aligned} \tag{17}$$

where the nonlinear difference equation is in state variable form and the second equation is the output equation of the system. In the following two sections, two different stability concepts will be presented for the nonlinear systems models defined above.

2.5.1 Linearization and Small Perturbation Stability

The small perturbation stability of a nonlinear, continuous-time system is defined in a small region near a "point" defined by a particular input vector $\bar{\mathbf{u}}(t)$ and the corresponding output vector $\bar{\mathbf{x}}(t)$, the ordered pair $\{\bar{\mathbf{x}}(t), \bar{\mathbf{u}}(t)\}$ is called an operating point. The nonlinear continuous-time system defined in Eq. (16) is linearized about the operating point by defining the linear perturbations $\delta\mathbf{x}(t) = \mathbf{x}(t) - \bar{\mathbf{x}}(t)$, $\delta\mathbf{u}(t) = \mathbf{u}(t) - \bar{\mathbf{u}}(t)$, and $\delta\mathbf{y}(t) = \mathbf{y}(t) - \bar{\mathbf{y}}(t)$, then expanding the functions $\mathbf{f}[\mathbf{x}(t), \mathbf{u}(t)]$ and $\mathbf{g}[\mathbf{x}(t), \mathbf{u}(t)]$ in a Taylor series expansion about the operating point $\{\bar{\mathbf{x}}(t), \bar{\mathbf{u}}(t)\}$, retaining only the first two terms of the Taylor series, and recognizing that $\dot{\bar{\mathbf{x}}}(t) = \mathbf{f}[\bar{\mathbf{x}}(t), \bar{\mathbf{u}}(t)]$ and $\bar{\mathbf{y}}(t) = \mathbf{g}[\bar{\mathbf{x}}(t), \bar{\mathbf{u}}(t)]$, the following two small perturbation equations result:

$$\begin{aligned} \delta\dot{\mathbf{x}}(t) &= \left.\frac{\partial\mathbf{f}}{\partial\mathbf{x}}\right|_{\substack{\mathbf{x}=\bar{\mathbf{x}}(t)\\ \mathbf{u}=\bar{\mathbf{u}}(t)}} \delta\mathbf{x}(t) + \left.\frac{\partial\mathbf{f}}{\partial\mathbf{u}}\right|_{\substack{\mathbf{x}=\bar{\mathbf{x}}(t)\\ \mathbf{u}=\bar{\mathbf{u}}(t)}} \delta\mathbf{u}(t) \\ \delta\mathbf{y}(t) &= \left.\frac{\partial\mathbf{g}}{\partial\mathbf{x}}\right|_{\substack{\mathbf{x}=\mathbf{x}(t)\\ \mathbf{u}=\mathbf{u}(t)}} \delta\mathbf{x}(t) + \left.\frac{\partial\mathbf{g}}{\partial\mathbf{u}}\right|_{\substack{\mathbf{x}=\mathbf{x}(t)\\ \mathbf{u}=\mathbf{u}(t)}} \delta\mathbf{u}(t) \end{aligned} \tag{18}$$

where

$$\frac{\partial\mathbf{f}}{\partial\mathbf{x}} = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_n}{\partial x_1} & \cdots & \dfrac{\partial f_n}{\partial x_n} \end{bmatrix} \qquad \frac{\partial\mathbf{f}}{\partial\mathbf{u}} = \begin{bmatrix} \dfrac{\partial f_1}{\partial u_1} & \cdots & \dfrac{\partial f_1}{\partial u_r} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_n}{\partial u_1} & \cdots & \dfrac{\partial f_n}{\partial u_r} \end{bmatrix}$$

$$\frac{\partial \mathbf{g}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_m}{\partial x_1} & \cdots & \frac{\partial g_m}{\partial x_n} \end{bmatrix} \quad \frac{\partial \mathbf{g}}{\partial \mathbf{u}} = \begin{bmatrix} \frac{\partial g_1}{\partial u_1} & \cdots & \frac{\partial g_1}{\partial u_r} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_m}{\partial u_1} & \cdots & \frac{\partial g_m}{\partial u_r} \end{bmatrix}$$

Note that these equations are linear equations in the small perturbations, and further note that if the elements of the operating point are constants, that is, if $\bar{\mathbf{u}}(t) = \bar{\mathbf{u}} =$ a constant vector and $\bar{\mathbf{x}}(t) = \bar{\mathbf{x}} =$ a constant vector, then these equations are time invariant and Eq. (18) is an LTI, continuous-time system as given in Eq. (9). When these equations are time invariant, all of the criteria and tests for stability that are applicable to LTI, continuous-time systems in Sec. 2.4 are directly applicable to these equations. It should be remembered that stability of this type is valid only when the linear perturbations $\delta\mathbf{x}(t)$, $\delta\mathbf{u}(t)$, and $\delta\mathbf{y}(t)$ are "small." The problem with this requirement is that it is, in general, very difficult, if not impossible, to determine how small they must be. In spite of this, the stability of the linearized equations is a valuable tool in nonlinear control system design.

Example 8. *The nonlinear system*

$$\begin{aligned} \dot{\mathbf{x}} &= \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \mathbf{f}(\mathbf{x}, \mathbf{u}) = \begin{bmatrix} x_2 \\ \sin x_1 + u \end{bmatrix} \\ \mathbf{y} &= y = x_1 \end{aligned}$$

is a simple model of a pendulum driven by a torque u. This system has two operating points of interest: $\{\bar{x}_1 = 0,\ \bar{x}_2 = 0,\ \bar{u} = 0\}$*, which represents the case when the pendulum is at rest and hanging straight down, and* $\{\bar{x}_1 = \pi,\ \bar{x}_2 = 0,\ \bar{u} = 0\}$*, which represents the case when the pendulum is at rest and standing straight up. The linearized equations for the first case are*

$$\begin{aligned} \delta\dot{\mathbf{x}} &= \begin{bmatrix} \delta\dot{x}_1 \\ \delta\dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \delta u \\ &= \mathbf{A}\,\delta\mathbf{x} + \mathbf{b}\,\delta\mathbf{u} \\ \delta y &= \delta x_1 = c\,\delta x_1 \end{aligned}$$

The small perturbation stability is determined by the eigenvalues of the matrix **A** which are located at $s = \pm j$. Thus the system is marginally stable about the operating point $\{\bar{x}_1 = 0, \bar{x}_2 = 0, \bar{u} = 0\}$. For the operating point $\{\bar{x}_1 = \pi, \bar{x}_2 = 0, \bar{u} = 0\}$, the linearized equations are

$$\begin{aligned} \delta\dot{\mathbf{x}} &= \begin{bmatrix} \delta\dot{x}_1 \\ \delta\dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u \\ &= \mathbf{A}\,\delta\mathbf{x} + \mathbf{b}u \\ \delta y &= \delta x_1 = c\,\delta x_1 \end{aligned}$$

For this case, the eigenvalues of the matrix **A** are at $s = \pm 1$. The pole in the right half plane indicates the system is unstable, which certainly satisfies our intuition that a pendulum which is standing straight up is in an unstable position.

Nonlinear, discrete-time systems described by Eq. (17) can be linearized similarly with the resulting linear perturbation equations given by

$$\begin{aligned} \delta\mathbf{x}(k+1) &= \left.\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right|_{\substack{\mathbf{x}=\bar{\mathbf{x}}(k) \\ \mathbf{u}=\bar{\mathbf{u}}(k)}} \delta\mathbf{x}(k) + \left.\frac{\partial \mathbf{f}}{\partial \mathbf{u}}\right|_{\substack{\mathbf{x}=\bar{\mathbf{x}}(k) \\ \mathbf{u}=\bar{\mathbf{u}}(k)}} \delta\mathbf{u}(k) \\ \delta\mathbf{y}(k) &= \left.\frac{\partial \mathbf{g}}{\partial \mathbf{x}}\right|_{\substack{\mathbf{x}=\bar{\mathbf{x}}(k) \\ \mathbf{u}=\bar{\mathbf{u}}(k)}} \delta\mathbf{x}(k) + \left.\frac{\partial \mathbf{g}}{\partial \mathbf{u}}\right|_{\substack{\mathbf{x}=\bar{\mathbf{x}}(k) \\ \mathbf{u}=\bar{\mathbf{u}}(k)}} \delta\mathbf{u}(k) \end{aligned} \tag{19}$$

where $\{\bar{\mathbf{x}}(k), \bar{\mathbf{u}}(k)\}$ is the operating point and the notation is the same as in Eq. (18). As with the linearized equations for continuous-time systems, these equations are valid only for small perturbations, that is, $\delta\mathbf{x}(k)$, $\delta\mathbf{u}(k)$, and $\delta\mathbf{y}(k)$ must be "small." Even though determining how small is generally impossible, the stability analysis obtained from this linearized model can be a valuable tool in control system design. As is the case for the small-perturbation continuous-time model in Eq. (18), when $\bar{\mathbf{u}}(k) = \bar{\mathbf{u}} =$ a constant vector and $\bar{\mathbf{x}}(k) = \bar{\mathbf{x}} =$ a constant vector, the small perturbation system in Eq. (19) is an LTI, discrete-time system and all of the stability criteria and tests in Sec. 2.4 are applicable to this system.

2.5.2 Lyapunov Stability for Nonlinear Systems

In this section the stability of nonlinear systems with zero input will be examined using the Lyapunov stability criterion. Since $\mathbf{u} = \mathbf{0}$, the equations defining nonlinear systems, Eqs. (16) and (17), will be rewritten, respectively, as

$$\begin{aligned} \dot{\bar{\mathbf{x}}}(t) &= \mathbf{f}[\bar{\mathbf{x}}(t)] \\ \bar{\mathbf{y}}(t) &= \mathbf{g}[\bar{\mathbf{x}}(t)] \end{aligned}$$

and

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{f}[\mathbf{x}(k)] \\ \mathbf{y}(k) &= \mathbf{g}[\mathbf{x}(k)] \end{aligned}$$

The stability for each of these systems is determined by the first equation only, thus only the first equations need to be considered, that is, the equations

$$\dot{\mathbf{x}} = \mathbf{f}[\mathbf{x}(t)] \tag{16'}$$

and

$$\mathbf{x}(k+1) = \mathbf{f}[\mathbf{x}(k)] \tag{17'}$$

will be examined for stability. For both of these equations, a singular point is defined as a solution $\mathbf{x}_0$ for the equation $\mathbf{f}[\mathbf{x}_0] = \mathbf{0}$. Note that a solution is generally not unique for nonlinear systems. The stability of the system, whether it is continuous-time or discrete-time, is determined with respect to one or more of the singular points. A singular point is said to be stable if there exist two n-dimensional spheres of finite radii, r and R, each centered at the singular point and such that for any solution $\mathbf{x}(t)$ of the differential equation (any solution $\mathbf{x}(k)$ of the difference equation) that starts in the sphere of radius r remains in the sphere of radius R forever. A stable singular point is called asymptotically stable if all solutions $\mathbf{x}(t)$ of the differential equation [$\mathbf{x}(k)$ for difference equation] approach the singular point as time approaches infinity.

If the origin of the state space is a singular point, that is, one solution of the equation $\mathbf{f}[\mathbf{x}_0] = \mathbf{0}$ is $\mathbf{x}_0 = \mathbf{0}$, then the Lyapunov stability criterion states that the origin is a stable singular point if a Lyapunov function (a scalar function) can be found such that:

1. $V(\mathbf{x}) > 0$ for all $\mathbf{x} \neq \mathbf{0}$.
2. For
   a. Continuous-time systems, $\dot{V} \leq 0$ for all $\mathbf{x}$.
   b. Discrete-time systems, $\Delta V(k) = V(k+1) - V(k) \leq 0$ for all $\mathbf{x}$.

For continuous-time systems, if in addition to the conditions above, $\dot{V} = 0$ if, and only if, $\mathbf{x} = 0$ then the origin is called asymptotically stable. For discrete-time systems if in addition to the conditions above, $\Delta V(k) = \mathbf{0}$ if, and only if, $\mathbf{x} = \mathbf{0}$, then the origin is called asymptotically stable. Note that these conditions are sufficient, but not necessary, for stability.

Example 9. *Consider the nonlinear differential equation*

$$\dot{\mathbf{x}} = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \mathbf{f}(\mathbf{x}) = \begin{bmatrix} x_2 \\ -x_2 - x_2^3 - x_1 \end{bmatrix}$$

Obviously, the origin, $x_1 = x_2 = 0$, is a singular point and the Lyapunov stability criterion might be used to determine its stability. Consider the Lyapunov function defined by $V(x_1, x_2) = x_1^2 + x_2^2$, which is positive unless $x_1 = x_2 = 0$. Its derivative is given by $\dot{V}(x_1, x_2) = 2x_1\dot{x}_1 + 2x_2\dot{x}_2 = -2x_2^2 - 2x_2^4$, which is never positive, thus the origin is stable. Note that since the derivative can be zero for $x_1 \neq 0$, then the condition for asymptotic stability is not satisfied. This does not mean that the system is not asymptotically stable, only that this Lyapunov function does not guarantee asymptotic stability. Another Lyapunov function might satisfy the condition for asymptotic stability.

In using the Lyapunov stability theory, it should be noted that there is no suggestion as to the form of Lyapunov function for any particular system. Generally, the choice of a suitable Lyapunov function is left to the system analyst.

REFERENCES

1. MS Santina, AR Stubberud, GH Hostetter. Digital Control System Design, Second Edition. Fort Worth, TX: Saunders College Publishing, 1994.
2. GH Hostetter, CJ Savant, Jr, RT Stefani. Design of Feedback Control Systems, Second Edition. New York: Saunders College Publishing, 1989.
3. BC Kuo. Automatic Control Systems, 4th ed. Englewood Cliffs, NJ: Prentice-Hall, 1982.
4. AV Oppenheim, AS Willsky, IT Young. Signals and Systems. Englewood Cliffs, NJ: Prentice-Hall, 1983.
5. JJ Di Stefano, AR Stubberud, IJ Williams. Feedback and Control Systems, 2nd ed. New York: McGraw-Hill, 1990.
6. JH Wilkinson. The Algebraic Eigenvalue Problem. Oxford: Oxford University Press, 1992.
7. GH Golub, CF Van Loan. Matrix Computations, 2nd ed. Baltimore, MD: The Johns Hopkins University Press, 1989.
8. W Levine, ed. The Control Handbook. New York: CRC Press, 1996.
9. RC Dorf. Modern Control Systems, 3rd ed. Reading, MA: Addison-Wesley, 1980.

Chapter 3.3

Digital Signal Processing

Fred J. Taylor
University of Florida, Gainesville, Florida

3.1 INTRODUCTION

Signal processing is as old as history itself. Early man relied on acoustic and optical signal processing for his very existence. Man is, in some respects, the quintessential signal processing machine. With a few exceptions, prior to the advent of digital electronics, signal processing technology was called *analog* (*continuous-time*). Analog electronic signal processing systems were historically designed using resistors, capacitors, inductors, and operational amplifiers. By mid-century another technology emerged called *sampled-data* (*discrete-time*) systems (see Chap. 3.4). In general, all these technologies are in the process of being replaced by *digial signal processing* (*DSP*) systems. DSP is a relatively young branch of engineering which can trace its origins back to the mid-1960s with the introduction of the now-celebrated Cooley–Tukey *fast Fourier transform* (FFT) algorithm. The FFT algorithm was indeed a breakthrough in that it recognized both the strengths and weaknesses of a general-purpose digital computer and used this knowledge to craft an efficient computer algorithm for computing Fourier transforms. The popularity and importance of DSP has continued to grow ever since.

Contemporary DSP applications areas include:

1. *General purpose*
   Filtering (convolution)
   Detection (correlation)
   Spectral analysis (Fourier transforms)
   Adaptive filtering
   Neural computing
2. *Instrumentation*
   Waveform generation
   Transient analysis
   Steady-state analysis
   Biomedical instrumentation
3. *Information systems*
   Speech processing
   Audio processing
   Voice mail
   Facsimile (fax)
   Modems
   Cellular telephones
   Modulators, demodulators
   Line equalizers
   Data encryption
   Spread-spectrum
   Digital and LAN communications
4. *Graphics*
   Rotation
   Image transmission and compression
   Image recognition
   Image enhancement
5. *Control*
   Servo control
   Disk control
   Printer control
   Engine control
   Guidance and navigation

Vibration (modal) control
Power systems monitors
Robots

6. *Others*
Radar and sonar
Radio and television
Music and speech synthesis
Entertainment

The study of analog systems remains closely related to DSP at many levels. Classical digital filters are, in fact, simply digital manifestations of analog radio filters whose structures have been known for nearly 75 years. One of the principal differences between an analog and digital system is found in how they interface to the external world. Analog systems import analog signals and export the same without need of a domain conversion. Digital systems, alternatively, must change the domain of any analog signal to digital before processing and return the signal to the analog domain in some cases. A typical DSP signal processing stream is shown in Fig. 1. An analog *antialiasing filter* is introduced to eliminate aliasing (see Chap. 3.4) by heavily attenuating input signal energy above the Nyquist frequency $f_s/2$, where f_s is the sampling frequency. The conditioned signal is then passed to an *analog-to-digital converter* (ADC). Following the ADC is the DSP system which typically implements a set of instructions which are defined by a *DSP algorithm* (e.g., filter) whose output may or may not be converted back into the analog domain, depending on the application. An analog signal can be reconstructed from a digital signal using *a digital-to-analog converter* (*DAC*). The typical DSP system is characterized in Fig. 1.

Digital filters initially made their appearance in the mid-1960s using discrete logic. Their expense and limited programmability restricted their use to narrowly defined applications. Digital filters are now regularly developed using commonly available commercial off-the-shelf (COTS) DSP microprocessors and application-specific integrated circuits (ASICs). A vast array of CAD tools and products can now be found to support this technology. The struggle between analog and DSP will continue into the future with the race increasingly favoring DSP well into the 21st century. It is commonly assumed that the attributes of analog and digital signal processing systems compare as follows:

The continued evolution of the semiconductor is being driven by digital devices and digital signal processing systems which provide a technological advantage over analog systems. This gap between digital and analog performance and price points is increasingly favoring digital.

Digital systems can operate at extremely low frequencies which are unrealistic for an analog system.

Digital systems can be designed with high precision and dynamic range, far beyond the ability of analog systems.

Digital systems can be easily programmed to change their function; reprogramming analog systems is extremely difficult.

Digital signals can easily implement signal delays which are virtually impossible to achieve in analog systems.

Digital signals can easily implement nonlinear signal operations (e.g., compression), which are virtually impossible to implement with analog technology.

Digital systems remain stable and repeatable results, whereas analog systems need periodic adjustment and alignment.

Digital systems do not have impedance-matching requirements; analog systems do.

Digital systems are less sensitive to additive noise as a general rule.

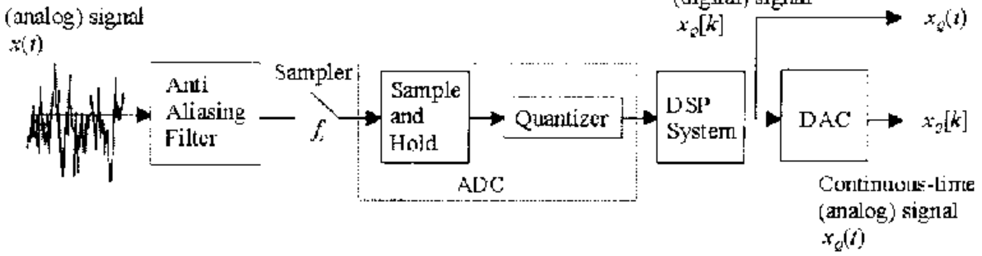


Figure 1 DSP signal train.

There are a few areas in which analog signal processing will remain competitive, if not supreme, for the following reasons:

Analog systems can operate at extremely high frequencies [e.g., radio frequencies (RF)], whereas digital systems are limited by the maximum frequency of an ADC.

Some low-level signal processing solutions can be achieved for the cost of a resistor, capacitor, and possibly operational amplifier, which would establish a price point below the current minimum DSP solution of around $5.

3.2 ANALOG SIGNALS AND SYSTEMS

The study of DSP usually begins with its progenitor, analog signal processing. *Continuous-time* or *analog* signals are defined on a continuum of points in both the independent and dependent variables. Electronic analog filters have existed throughout the 20th century and generally are assumed to satisfy an ordinary differential equation (ODE) of the form

$$\sum_{m=0}^{N} a_m \frac{d^m y(t)}{dt^m} = \sum_{m=0}^{M} b_m \frac{d^m x(t)}{dt^m} \tag{1}$$

The classic analog filter types, called Cauer, Butterworth, Bessel, and Chebyshev are well studied and have been reduced to standard tables. Analog filters are historically low order (≤ 4) and are often physically large devices. High-precision high-order analog filters are notoriously difficult to construct due to the inexactness of the analog building-block elements and inherent parameter sensitivity problems. Currently analog filters have been routinely reduced to electronic integrated circuits (IC) which adjust their frequency response using external resistors and capacitors.

3.3 DIGITAL SYSTEMS

Digital systems are generally modeled to be *linear shift-invariant* (*LSI*) systems (see Chap. 3.4). The output of an LSI system, say $y[k]$, of an LSI having an impulse response $h[k]$ to an input $x[k]$, is given by the convolution sum

$$\sum_{m=0}^{N} a_m y[k-m] = \sum_{m=0}^{M} b_m x[k-m] \tag{2}$$

Computing the convolution sum can be side-stepped by using the z-transform (see Chap. 3.4) and the *convolution theorem*, which states that if

$$\begin{aligned} h[k] &\overset{Z}{\longleftrightarrow} H(z) \\ x[k] &\overset{Z}{\longleftrightarrow} X(z) \\ y[k] &\overset{Z}{\longleftrightarrow} Y(z) \end{aligned} \tag{3}$$

then

$$Y(z) = Z(y[k]) = Z(h[k]) * x[k]) = X(z)\,H(z) \tag{4}$$

The advantage provided by the convolution theorem is that the computationally challenging convolution sum can be replaced by a set of simple algebraic operations. A comparison of the computation requirements to produce a *convolution sum* using time- and z-transform-domain methods is shown in Fig. 2.

The z-transform of the convolution sum defined in Eq. (2) is given by

$$\left(\sum_{m=0}^{N} a_m z^{-m}\right) Y(z) = \left(\sum_{m=0}^{M} b_m z^{-m}\right) X(z) \tag{5}$$

The ratio of input and output transforms, namely $H(z) = Y(z)/X(z)$, is formally called the *transfer function*. Algebraically the transfer function of an LSI system satisfies

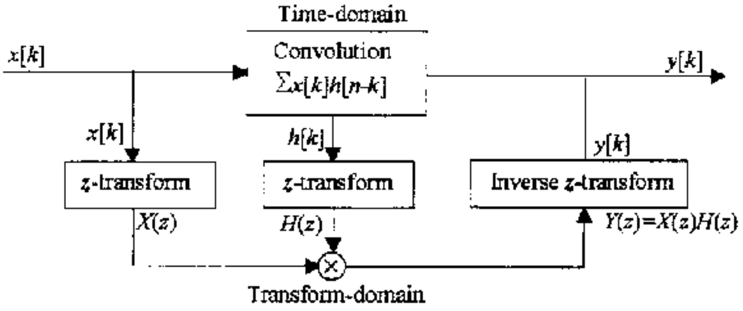


Figure 2 Convolution theorem.

$$H(z) = \frac{Y(z)}{X(z)} = \frac{N(z)}{D(z)} = \frac{\left(\sum_{m=0}^{M} b_m z^{-m}\right)}{\left(\sum_{m=0}^{N} a_m z^{-m}\right)} \tag{6}$$

The *poles* of the digital system are given by the roots of $D(z)$ found in Eq. (6), namely

$$\sum_{m=0}^{N} a_m z^m = \prod_{m=0}^{N} (p_m - z) = 0 \tag{7}$$

and are denoted p_m. The *zeros* of a digital system are given by the roots of $N(z)$ found in Eq. (6), namely

$$\sum_{m=0}^{M} b_m z^m = \prod_{m=0}^{N} (z_m - z) = 0 \tag{8}$$

and are dentoed z_m. The location of the poles and zeros, relative to the periphery of the unit circle in the z-plane are important indicators of system performance. A class of stability, for example, can be assured if the poles are interior to the unit circle (see Chap. 3.4).

If the system is *asymptotically stable*, then after a period of time any transient signal components (due to possible nonzero initial conditions) will decay to zero, leaving only externally forced (inhomogeneous) signal components at the output. If the input is a sinusoid, then after the transients have decayed, the signal found at the filter's output is called the *steady-state sinusoidal response*. If the input frequency is slowly swept from DC to the Nyquist frequency, the steady-state frequency response can be measured. Mathematically, the steady-state frequency response is equivalently given by

$$A(\omega) = |H(e^{j\omega})| = |H(z)|_{z=e^{j\omega}} \quad \text{(magnitude frequency response)} \tag{9}$$

$$\phi(e^{j\omega}) = \arg(H(e^{j\omega})) = \arctan\left(\frac{\mathrm{Im}(H(e^{j\omega}))}{\mathrm{Re}(H(e^{j\omega}))}\right) \quad \text{(phase response)} \tag{10}$$

where $\omega \in [-\pi, \pi]$. The amplitude response corresponds to the gain added to the input at frequency ω and the phase response specifies what phase shift, or delay, has been applied to the input. Therefore, if the input is assumed to be given by $x[k] = Ve^{j\omega k}$, then the output (after any transients have decayed) would be given by $y[k] = VA(\omega)e^{j\omega k+\phi}$. This simply restates a fundamental property of linear systems, namely that an LSI cannot create any new frequencies, but can simply alter the magnitude and phase of the signal presented to the input.

Another important steady-state property of an LSI is called the *group delay*. It has importance in communications and control systems where it is desired that a signal have a well-behaved propagation delay within a filter. In many design cases, it is important that the propagation delay through the system be frequency invariant. Such systems are said to be *linear phase*. The frequency-dependent propagation delay of an LSI is defined by the group delay measure which is given by

$$\tau_g = -\frac{d\phi(e^{j\omega})}{d\omega} \qquad \text{(group delay)} \tag{11}$$

From Eqs. (9) and (10) is can be noted that the spectral properties of $H(z)$ can be analytically computed if $H(z)$ is known in closed form. However, in many cases, signals and systems are only known from direct measurement or observation. In such cases the spectrum of a signal or system must be computed directly from time-series data. Historically, this is the role of the *discrete Fourier transform* (*DFT*).

3.4 FOURIER ANALYSIS

The frequency-domain representation of a continuous-time signal is defined by the *continuous-time Fourier transform* (CTFT). The CTFT *analysis equation* satisfies

$$X(j\Omega) = \int_{-\infty}^{\infty} x(t)e^{-j\Omega t}\, dt \tag{12}$$

and the synthesis equation is given by

$$x(t) = \int_{-\infty}^{\infty} X(j\Omega)e^{j\Omega t}\, d\Omega \tag{13}$$

where Ω is called the *analog frequency* in radians per second and $X(j\Omega)$ is called the *spectrum* of $x(t)$. Computing a CTFT with infinite limits of integration with a digital computer is virtually impossible. A modification of the Fourier transform, called the *continuous-time Fourier series* (CTFS) simplified the computational problem by restricting the study to periodic continuous-time signals $x_p(t)$ where $x_p(t) = x_p(t+T)$ for all time t. Regardless of the form that a continuous-time Fourier transform takes, it is again impractical to compute using a general-purpose digital computer. A computer expects data to be in a digital

Table 1 Properties of a DFT

Discrete-time series	Discrete Fourier transform	Remark
$x[k] = \sum_{m=0}^{L} a_m x_m[k]$	$X[n] = \sum_{m=0}^{L} a_m X_m[n]$	Linearity
$x_N[k] = x[((k-q) \bmod N)]$	$X_N[n] = X[n]W_N^{qn}$	Circular time shift
$x_N[k] = x^*([k] \bmod N)$	$X_N[n] = X^*([-n] \bmod N)$	Time reversal
$x_N[k] = x[k]W_N^{-k}$	$X_N[n] = X([n-q] \bmod N)$	Modulation
$x_N[k] = \sum_{m=0}^{L} a_m x_m[k \bmod N]$	$X_N[n] = \sum_{m=0}^{L} a_m X_m[n]$	Linearity
$x_N[k] = \sum_{k=0}^{N-1} x[k \bmod N] y[k \bmod N]$	$X_N[n] = \frac{1}{N}\sum_{n=0}^{N-1} X[n]\, Y[k]^*$	Parseval (power)

sampled format and be of finite duration. What is therefore needed is an algorithm which can operate on a time series. The *discrete Fourier transform* (*DFT*) is such a tool in that it maps an N-sample time series (possibly complex) into an N-harmonic array in the frequency domain. Since the harmonics are, in general, complex, the DFT is a mapping of complex space into a complex space (i.e., $C^N \leftrightarrow C^N$). The DFT of an N-sample time series, denoted $x_N[k]$, is given by

$$X[n] = \sum_{k=0}^{N-1} x_N[k] W_N^{nk} \tag{14}$$

for $0 \leq n < N$, $W_N = e^{-j2\pi/N}$, and $X[n]$ is called the nth harmonic. The complex exponential W_N is seen to be periodic with period N, which also defines the periodicity of the DFT. Therefore $X[n] = X[n \pm kN]$ for any integer N. Equation (14) is called the *DFT analysis equation* and defines the N harmonics of $x_N[k]$ for $0 \leq n < N$. The inverse transform, called the DFT *synthesis equation*, is given by

$$x_N[k] = \frac{1}{N}\sum_{k=0}^{N-1} X[n] W_N^{-nk} \tag{15}$$

for $0 \leq k < N$. The advantage of the DFT is its ability to compute a spectrum from the bounded sample values of $x_N[k]$ without regard to the established mathematical properties of $x[k]$. The DFT algorithm presented in Eq. (14) defines what is computationally called the *direct method* of producing a spectrum. The direct method computes the N harmonics of $X[n]$ by repeatedly performing complex multiply–accumulate (MAC) operations on the elements of the N-sample time series $x_N[k]$. The MAC complexity of the direct method is classified as being order N^2. This translates to a finite, but possibly long, computation time. This condition was radically altered with the advent of the *fast Fourier transform* (*FFT*) algorithm. It should be appreciated, however, that "fast" has a relative meaning in this case. The FFT is a well known computer algorithm which converts an order N^2 calculation to an order $N\log_2(N)$ computation. The FFT, while being faster than a direct method, still remains computationally intensive. In addition, the FFT incurs some overhead penalty. Typically, as a general rule, the advantage of a software-based FFT over a direct DFT is not realized unless $N \geq 32$. For high-speed applications, *application-specific integrated circuits* (*ASICs*), dedicated DFT chips, have been developed for general use.

The list of DFT properties is found in Table 1 and the parameters of a DFT reviewed in Table 2. The fundamental parameters which define the precision of a DFT are defined in terms of the sample rate f_s and N, the number of samples to be transformed.

The performance of a DFT (usually implemented as an FFT) is well known and understood. Variations of the basic algorithm have been developed to efficiently handle the case where the input data are known to be real. Called *real FFTs*, they offer a speed-up of a factor of two over their more general counterparts. Various methods have been developed to integrate short DFT units together to create a long DFT (viz., Cooley–Tukey, Good–Thomas, etc.), which can be useful in the hands of a skilled DSP engineer. Nevertheless, a DFT or FFT is rarely designed in a contemporary setting. Instead they are simply extracted from an abundance of math software libraries, CAD packages, or from a runtime executable supplied by a technology vendor.

Table 2 DFT Parameters

Parameter	Notation or units
Sample size	N samples
Sample period	T_s sec
Record length	$T = NT_s$ sec
Number of harmonics	N harmonics
Number of positive (negative) harmonics	$N/2$ harmonics
Frequency spacing between harmonics	$\Delta f = 1/T = 1/NT_s = f_s/N$ Hz
DFT frequency (one-sided baseband range)	$f \in [0, f_s/2)$ Hz
DFT frequency (two-sided baseband range)	$f \in [-f_s/2, f_s/2)$ Hz
Frequency of the kth harmonic	$f_k = kf_s/N$ Hz

While the DFT is well known, the interpretation of a DFT spectrum requires some care and experience. The DFT is a baseband signal analysis which means that it maps all spectral information into the frequency range $f \in [-f_s/2, f_s/2)$ Hz which is centered about DC. Obviously if a signal is not properly sampled in the context of the Nyquist sampling theorem, *aliasing errors* will occur introducing what are called *artifacts*. Artifacts are spectral lines which appear in the baseband spectra can be due to signal energy lying outside the baseband spectrum. If a signal at frequency f is sampled at a rate f_s, and $f > f_s/2$ (below the Nyquist rate), then an artifact will occur at frequency $f' = f\mathrm{mod}(f_s)$ in the DFT spectrum where $f' \in [-f_s/2\ f_s/2)$ Hz. An example of aliasing is found in Fig. 3. Artifacts due to aliaising can be eliminated, or controlled, by placing an *antialiasing filter* before the ADC which will limit the highest frequency presented to the ADC to be bound by $f_s/2$ Hz (i.e., Nyquist limiter).

Artifacts can also result from an effect called *leakage*. Recall that a DFT assumes the signal to be trans-

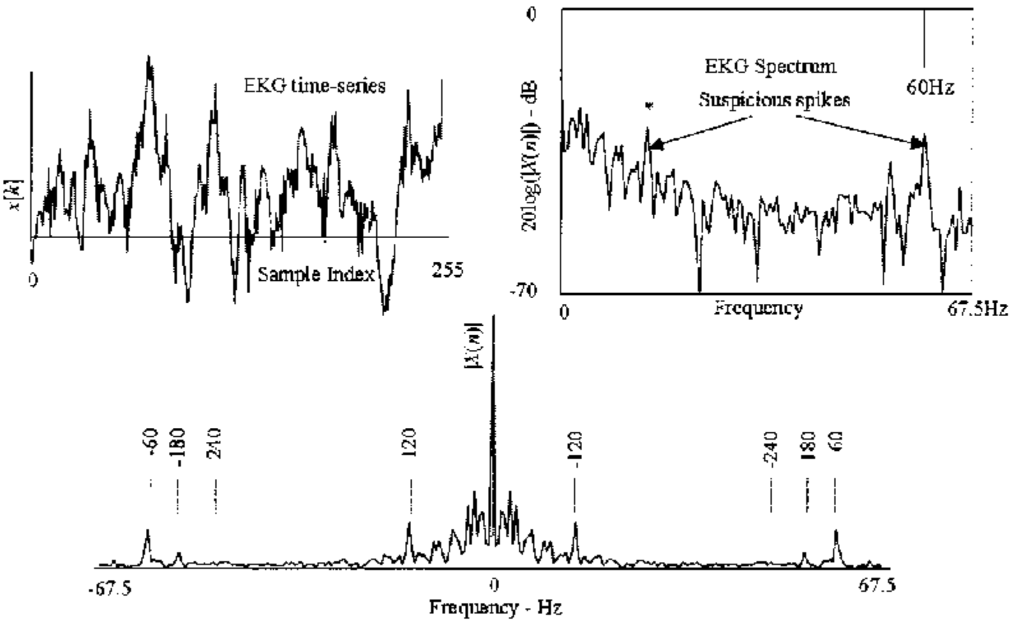


Figure 3 Top left shows an EKG signal with 60 Hz distortion which is sampled at a 135 Hz rate. Top right is the one-sided (positive) spectrum showing a number of peaks over the baseband frequency range $f \in [0, 67.5]$ Hz. The prominent high-frequency peak is attributed to 60 Hz contamination. Bottom center is the two-sided spectrum showing the location of the aliased contaminates at ±120 Hz and ±180 Hz. Notice that the aliased spectral components wrap the folding (Nyquist) frequency as defined by the rule $f_a = f\,\mathrm{mod}\,(f_s)$ where $f_a \in [-67.5, 67.5]$ Hz, $f = 120$ and 180 Hz, and $f_s = 135$ Hz.

formed, namely $x_N[k]$, is *periodic* N-sample time series with period N. Suppose the actual signal $x[k]$ is not periodic. Then the DFT of $x_N[k]$, which assumes periodicity, will differ from an infinitely long DFT of the aperiodic parent $x[k]$. The difference between the N-sample spectra is due to energy found at the boundary of N-sample intervals leaking into the DFT spectrum. This phenomenon can be motivated by analyzing the data shown in Fig. 4. Shown are two time series of length N, along with their periodic extension. One time series completes an integer number of cycles in N samples and the other does not. The difference in their spectra is also shown in Fig. 4. The DFT of a signal completing an integer number of oscillations in N samples is seen to possess a well-defined and localized line spectrum. The other spectrum exhibits "spreading" of spectral energy about local spectral lines. The leaked energy from the jump discontinuity found at the N-sample boundary can be reduced by increasing the length of the time series (i.e., N) or through the use of a data window smoothing function.

3.5 WINDOWING

Figure 5 describes an arbitrary signal of infinite length $x[k]$ and its assumed spectrum. Also shown is a *gating*, or *window* function of length N denoted $w[k]$. The object of the window function is to reduce the presence of *artifacts* introduced by creating a finite duration signal $x_N[k]$ from an arbitrary parent time series $x[k]$. The potential dilatory effects of such action were graphically interpreted in Fig. 4. The finite-duration signal produced by the N-sample gating function is given by $x_N[k] = x[k]\,w[k]$. The leakage artifacts can be suppressed by reducing the influence of jump discontinuities at the window boundary. This can be achieved by having the leading and trailing tails of $w[k]$ take on values at or near zero. A rectangular window, or gating function, obviously does not satisfy this criterion and, as previously seen, can introduce artifacts into the spectrum. Popular windows which do meet this criterion are shown below (rectangular included for completeness).

Rectangular

$$w[k] = 1 \qquad k \in [0, N-1] \tag{17}$$

Bartlett (triangular)

$$w[k] = \begin{cases} \dfrac{2k}{N-1} & k \in \left[0, \dfrac{N-1}{2}\right] \\ 2 - \dfrac{2k}{N-1} & k \in \left[\dfrac{n-1}{2}, N-1\right] \end{cases} \tag{18}$$

Hann

$$w[k] = \frac{1}{2}\left(1 - \cos\left(\frac{2\pi k}{N-1}\right)\right) \quad k \in [0, N-1] \tag{19}$$

Hamming

$$w[k] = 0.54 - 0.46\cos\left(\frac{2\pi k}{N-1}\right) \quad k \in [0, N-1] \tag{20}$$

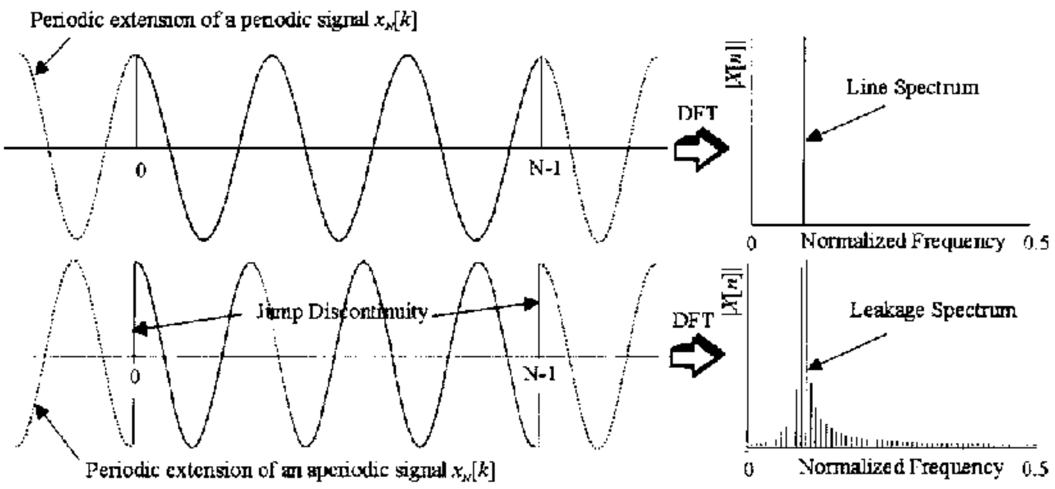


Figure 4 Example of leakage and its cause.

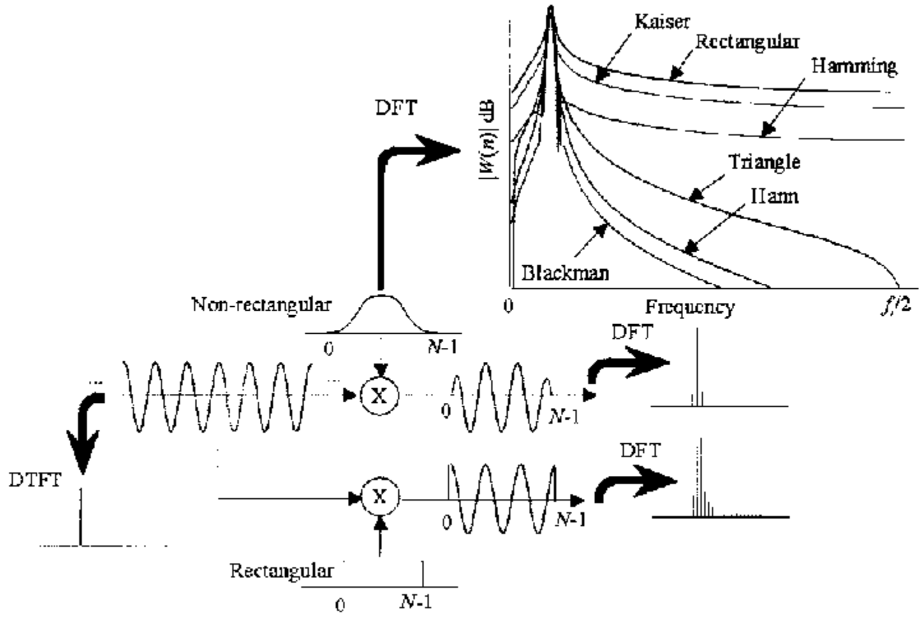


Figure 5 The mechanics of windowing is modeled as a time-domain gating operating. Notice how the nonrectangular window suppresses leakage artifacts. Shown in the upper right is the DFT of a leaky sinusoid processed by a number of popular window functions. Notice that the resulting magnitude spectrum (in dBs) exhibits various tradeoffs between the width of the main lobe and sideband attenuation.

Blackman

$$w[k] = 0.42 - 0.5\cos\left(\frac{2\pi k}{N-1}\right) + 0.08\cos\left(\frac{4\pi k}{N-1}\right)$$
$$k \in [0, N-1] \tag{21}$$

Kaiser

$$w[k] = I_0(\beta) \qquad k \in [0, N-1]$$
$$(I_0 \text{ is a 0th-order Bessel function}) \tag{22}$$

All the above windows, except the rectangular window, have values near zero locally about $k = 0$ and $k = N - 1$. The shape and spectrum of these windows are reported in Fig. 6. A good window is considered to be one which has a spectrum exhibiting a narrow main lobe along with deep sideband attenuation. The presented windows (except the rectangular window) achieve this to various degrees and with different tradeoffs.

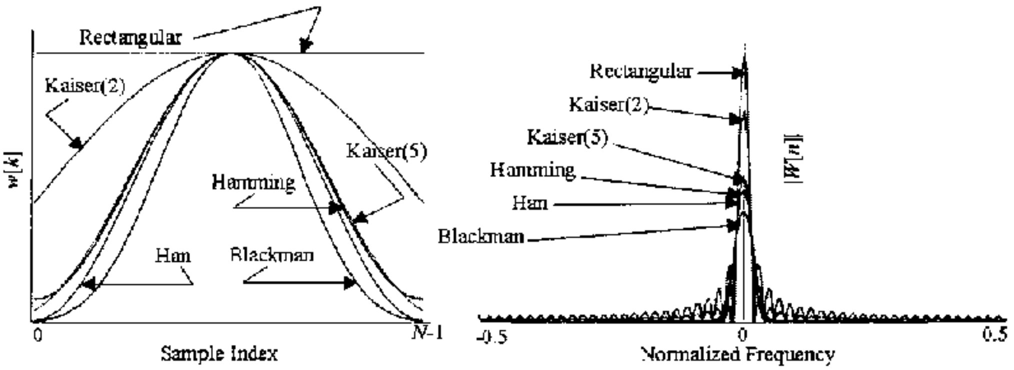


Figure 6 Experimental study of windowing on a DFT showing the time-domain and frequency-domain envelope of basic window functions.

3.6 DIGITAL FILTERS

Digital filters can be grouped into three broad classes called (1) *finite impulse response* (*FIR*) filters, (2) *infinite impulse response* (*IIR*) filters, and (3) *multirate* filters. Filters are also historically classified in terms of their function with their most common being *lowpass*, *highpass*, *bandpass*, or *bandstop* filtering. However, it should not be forgotten that all digital filters which are based on common LSI models share a common mathematical framework and are often implemented with a common technology (i.e., DSP microprocessors).

3.6.1 Infinite Impulse Response Filters (IIR)

An IIR filter is sometimes called a *recursive filter* due to the fact that it contains feedback paths. An IIR filter is generally modeled by the LSI transfer function.

$$H(z) = \frac{N(z)}{D(z)} = \frac{\sum_{i=0}^{M} b_i z^{-i}}{\sum_{i=0}^{N} a_i z^{-i}} = K z^{N-M} \frac{\prod_{i=0}^{M-1} (z - z_i)}{\prod_{i=0}^{N-1} (z - p_i)} \qquad (23)$$

The presence of the denominator terms [i.e., $D(z)$] establishes the fact that the IIR contains feedback data paths. The numerator terms [i.e., $N(z)$] in turn define the filter's feedforward data paths. It is the presence of feedback, however, which allows IIRs to achieve high-frequency selectivity and near resonate behavior. The frequency response of an IIR is determined by evaluating $H(z)$ for $z = e^{j\omega}$. This act scans a continuous range of frequencies which is normally assumed to be bounded between plus or minus the Nyquist frequency or $-f_s/2 \leq f \leq f_s/2$. It is often more convenient to interpret this frequency range to be *normalized* to $-\pi \leq \omega \leq \pi$ rad/sec or $-0.5 \leq f < 0.5$ Hz. Upon evaluation, one obtains

$$H(e^{j\omega}) = \frac{\sum_{i=0}^{M} b_i e^{-j\omega}}{\sum_{i=0}^{N} a_i e^{-j\omega}} = K e^{j\omega(N-M)} \frac{\prod_{i=0}^{M-1} (e^{j\omega} - z_i)}{\prod_{i=0}^{N-1} (e^{j\omega} - p_i)} \qquad (24)$$

where $-\pi \leq \omega \leq \pi$. As a general rule, an IIR can meet very demanding magnitude frequency-response specifications with a reasonable filter order (i.e., $N \leq 8$). The design of such filters has been highly refined and much is known about *classical digital filters*. The origins of classical filters can be traced back nearly a century to the early days of radio engineering. From the beginning of the radio era to today, frequency-selective filters have been extensively used to isolate the radio broadcast spectrum into distinct information bands. Radio engineers historically used tables and graphs to determine the parameters of a filter. The designer of digital filters relies on the use of computer program to support the design process.

The task of the classical filter designer is one of creating a system whose magnitude frequency response emulates that of an ideal filter. Historically, classical design paradigms are based on the well-known models of Bessel, Butterworth, Chebyshev, and Cauer (elliptical). To standardize the design procedure, a set of normalized lowpass filter models for each of these classes was agreed upon and reduced to a standardized design model. The models, called *analog prototypes*, assumed a −1 dB or −3 dB passband deviation from an ideal flat passband which extends from 0 to 1 rad/sec. In a classical design environment, the analog prototype, denoted $H_p(s)$, is read from prepared tables, charts, and graphs and then mapped into the desired analog filter which has the magnitude frequency-response shape but a cutoff frequency other than 1 rad/sec. The resulting scaled filter is called the (desired) *analog filter* and is denoted $H(s)$. The filter $H(s)$ meets or exceeds the magnitude frequency-response design constraints posed for an acceptable analog filter solution. The mapping rule which will take an analog prototype into its final analog form is called a *frequency-to-frequency transform*, summarized in Table 3 and interpreted in Fig. 7.

The analog prototype magnitude-squared frequency response, measured at the preagreed analog passband cutoff frequency of $\Omega = 1$, is often interpreted as

$$|H(s)|^2_{s=j1} = \frac{1}{1+\varepsilon^2} \qquad (25)$$

If $\varepsilon^2 = 1.0$, the prototype is said to be a −3 dB filter. Referring to Fig. 8, observe that the analog filter is to be mapped to an analog filter having target frequencies Ω_p, Ω_{p1}, Ω_{p2}, Ω_{a_1}, and Ω_{a_2}, called *critical frequencies*. The passband and stopband gains are specified in terms of the parameters ε and A. The steepness of the filter skirt is measured in terms of the *transition gain ratio* which is given by $\eta = \varepsilon/(A^2 - 1)^{1/2}$. The *frequency transition ratio*, denoted k_d, measures the transition bandwidth. The possible values of k_d, are given by

Table 3 Frequency-to-Frequency Transforms

N^{th}-order prototype	Transform	Order
Lowpass to lowpass	s/Ω_p	N
$\vdots$	$\vdots$	$\vdots$
Lowpass to highpass	$s \leftarrow \Omega_p/s$	N
Lowpass to bandpass	$\leftarrow (s2 + \Omega_H\Omega_L)/(s + \Omega_H\Omega_1))$	$2N$
Lowpass to bandstop	$s \leftarrow (s + \Omega_H\Omega_L))/(s^2 + \Omega_H\Omega_L)$	$2N$

Lowpass: $$k_d = \frac{\Omega_p}{\Omega_a} \tag{26}$$

Highpass: $$k_d = \frac{\Omega_a}{\Omega_p} \tag{27}$$

Bandpass: $$k_d = \begin{cases} k_1 & \text{if } \Omega_u^2 \geq \Omega_v^2 \\ k_2 & \text{if } \Omega_u^2 < \Omega_v^2 \end{cases} \tag{28}$$

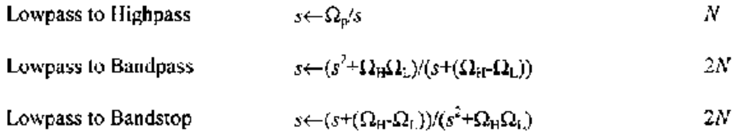
Bandstop: $$k_d = \begin{cases} 1/k_1 & \text{if } \Omega_u^2 \geq \Omega_v^2 \\ 1/k_2 & \text{if } \Omega_u^2 < \Omega_v^2 \end{cases} \tag{29}$$

where

$$\Omega_c = \frac{\Omega_{p_1}\Omega_{p_2}}{\Omega_{p_2} - \Omega_{p_2}} \tag{30}$$

$$k_1 = \left(\frac{\Omega_{a_1}}{\Omega_c}\right)\frac{1}{\left(1 - \dfrac{\Omega_{a1}^2}{\Omega_v^2}\right)} \tag{31}$$

$$k_2 = \left(-\frac{\Omega_{a_2}}{\Omega_c}\right)\frac{1}{\left(1 - \dfrac{\Omega_{a_2}^2}{\Omega_v^2}\right)} \tag{32}$$



Lowpass to Highpass	$s \leftarrow \Omega_p/s$	N
Lowpass to Bandpass	$s \leftarrow (s^2 + \Omega_H\Omega_L)/(s + (\Omega_H - \Omega_L))$	$2N$
Lowpass to Bandstop	$s \leftarrow (s + (\Omega_H - \Omega_L))/(s^2 + \Omega_H\Omega_L)$	$2N$

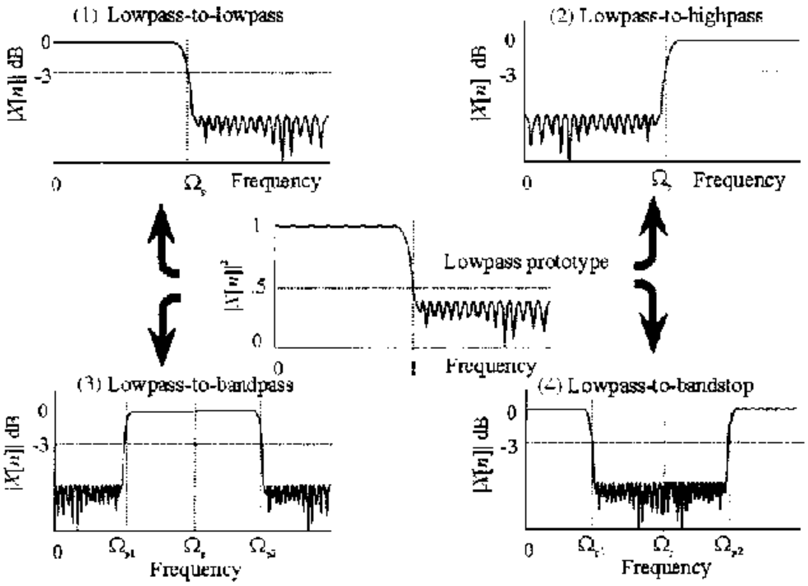


Figure 7 Frequency-to-frequency transforms showing the mapping of a prototype lowpass to (1) lowpass, (2) highpass, (3) bandpass, and (4) bandstop mappings.

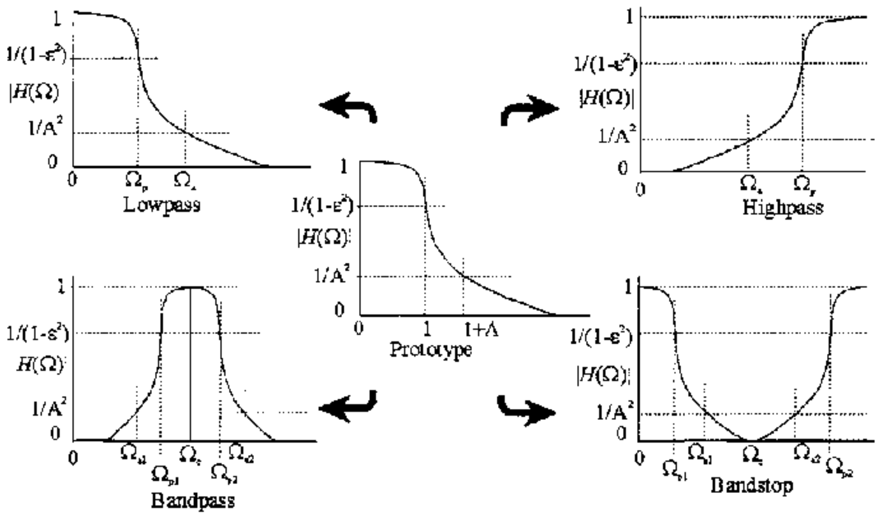


Figure 8 Frequency-response specifications of a lowpass analog prototype filter.

$$\Omega_u^2 = \Omega_{a_1}\Omega_{a_2} \tag{33}$$

$$\Omega_v^2 = \Omega_{p_1}\Omega_{p_2} \tag{34}$$

From $\delta^2 = A^2 - 1$, the *gain ratio* may be defined to be $d = \delta/\varepsilon$. From these parameters the order and transfer function of classical analog prototype filters can be determined.

3.6.2 Classic Butterworth Filter

An Nth-order *Butterworth filter* has a magnitude-squared frequency response given by

$$|H(s)|^2 = \frac{1}{1+\varepsilon^2 s^{2N}} \tag{35}$$

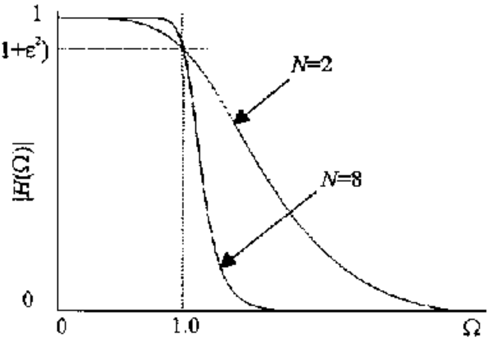


Figure 9 Typical Butterworth lowpass filter showing a magnitude frequency response.

The order is estimated to be

$$N = \log\left(\frac{A^2-1}{\varepsilon^2}\right)\frac{1}{2\log(1/k_d)} \tag{36}$$

A typical magnitude frequency response for a lowpass Butterworth filter is shown in Fig. 9. It is characterized by a smooth transition from pass- to stopband.

3.6.3 Classical Chebyshev Filter

An Nth-order lowpass *Chebyshev-I filter* has a magnitude-squared frequency response given by

$$|H(s)|^2 = \frac{1}{1+\varepsilon^2 C_N^2(s)} \tag{37}$$

where $C_N(s)$ is a Chebyshev polynomial of order N. The order of a Chebyshev-I filter is estimated to be

$$N = \frac{\log\left[\left(1+\frac{\sqrt{1-\eta^2}}{\eta}\right)\right]}{\log\left[\frac{1}{k_d}+\sqrt{\left(\frac{1}{k_d}\right)^2-1}\right]} \tag{38}$$

A variation on the Chebyshev-I model is called the *Chebyshev-II filter* and is given by

$$|H(s)|^2 = \frac{1}{1+[\varepsilon^2 C_N^2(j\Omega_a/s)]} \tag{39}$$

The order estimation formula is that for the Chebyshev-I model. A typical magnitude frequency response for a Chebyshev-I or -II filter is displayed in Fig. 10. The Chebyshev-I filter is seen to exhibit ripple in the passband and have a smooth transition into the stopband. The Chebyshev-II filter is seen to have ripple in the stopband and smooth transition into the passband.

3.6.4 Classical Elliptical Filters

The attenuation of an Nth-order *elliptical filter* is given by the solution to an elliptical integral equation. The order of an elliptical filter is estimated to be

$$N \geq \frac{\log(16D)}{\log(1/q)} \tag{40}$$

where

$$k' = \sqrt{(1-k_d^2)} \tag{41}$$

$$q_0 = \frac{1-\sqrt{k'}}{2(1+\sqrt{k'})} \tag{42}$$

$$q = q_0 + 2q_0^5 + 15q_0^9 + 15_q^{13} \tag{43}$$

$$D = d^2 \tag{44}$$

The typical magnitude frequency response of an elliptical lowpass filter is shown in Fig. 11. It can be seen that an elliptical filter has ripple in both the pass- and stopbands.

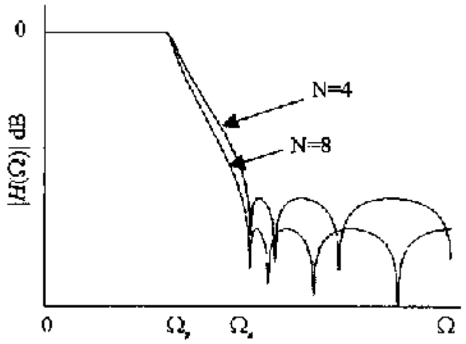


Figure 11 Magnitude frequency response of a typical elliptical IIR filter.

3.6.5 Other IIR Forms

Analog filter models, other than the classical Butterworth, Chebyshev, and elliptical filter models are also routinely encountered. Filters with an arbitrary magnitude frequency response can be defined by the invention of an engineer or synthesized from measured data using spectral estimation tools such as *autoregressive* (*AR*) or *autoregressive moving-average* (*ARMA*) models. In all cases, the design objective is to create a model of an Nth-order transfer function $H(z)$.

If the filter design process begins with a legacy analog filter model, then the designer of a digital filter replacement of an analog system must convert $H(s)$ into a discrete-time filter model $H(z)$. The basic domain conversion methods [i.e., $H(s) \rightarrow H(z)$] in common use are:

1. Impulse invariant method
2. Bilinear z-transform method.

3.6.5.1 Impulse-Invariant IIR Design

The *impulse-invariant filter* design method produces a sampled-data system which is based on a continuous-

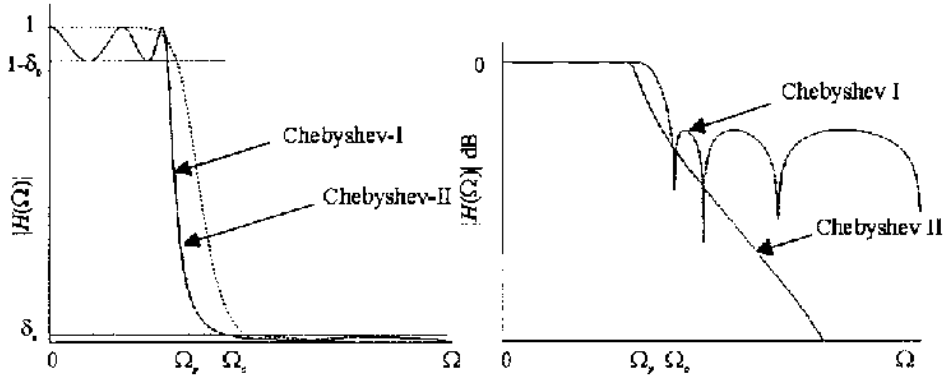


Figure 10 Typically Chebyshev-I and -II lowpass filter magnitude frequency response (linear on the left, logarithmic on the right).

time system model. The impulse response of a discrete-time impulse-invariant system, denoted $h_d[k]$, is related to the sampled values of continuous-time system's impulse response $h_a(t)$ through the defining relationship

$$h_d[k] = T_s h_a(kT_s) \tag{45}$$

That is, if a system is impulse invariant, then the discrete- and continuous-time impulse responses agree, up to a scale factor T_s, at the sample instances. The standard z-transform possesses the *impulse-invariant property*. This can be of significant importance in some application areas, such as control, where knowledge of the envelope of a signal in the time domain is of more importance than knowledge of its frequency response. Specifically, if a controller's specifications are defined in terms of risetime of overshoot, then an impulse-invariant solution is called for, since the frequency response of the controller is immaterial.

Consider an analog filter having a known impulse response $h_a(t)$ with a known transfer function. For the sake of development, consider the Nth-order system described by the transfer function $H(s)$ having N distinct poles. Then, upon taking the impulse-invariant z-transform of $H(s)$, the following results:

$$\begin{aligned} h_a(t) \Leftrightarrow H_a(s) &= \sum_{i=1}^{N} \frac{a_i}{s+p_i} \overset{Z}{\longleftrightarrow} \frac{1}{T_s} \sum_{i=1}^{N} \frac{a_i}{1+e^{p_i T_s} z^{-1}} \\ &= \frac{1}{T_s} H(z) \Leftrightarrow \frac{1}{T_s} h[k] \end{aligned} \tag{46}$$

which mathematically restates the impulse-invariant property of the standard z-transform. As a direct consequence, the frequency response of a discrete-time having a transfer function $H(z)$ can be computed to be

$$H(e^{j\Omega}) = \frac{1}{T_s} \sum_{k=-\infty}^{\infty} H_a\left(j\left(\frac{\Omega}{T_s} - \frac{2k\pi}{T_s} \right) \right) \tag{47}$$

Equation (47) states that under the z-transform, the frequency response of the resulting system, namely $H(e^{j\Omega})$, is periodic on centers separated by $2\pi/T_s = f_s$ radians (see Fig. 12) in the frequency domain. Observe that the spectral energy from any one frequency band, centered about $\omega = k\omega_s$, can potentially overlap the neighboring spectral image of $H_a(j\omega)$ centered about $\omega = m\omega_s$, $m \neq k$. This overlap is called *aliasing*. Aliasing was noted to occur when a signal was sampled at too low a rate (see Chap. 3.4). Unfortunately, analog filters generally have a frequency response which can technically persist for all finite frequencies and therefore *can* naturally introduce aliasing errors for any finite sampling frequency.

Example 1. First-Order Impulse-Invariant System: *Consider the first-order RC circuit having an input forcing function $v(t)$ developing an output voltage $v_o(t)$, defined by the solution to the ordinary differential equation*

$$v_0(t) + RC\frac{dv_0(t)}{dt} = v(t)$$

The transfer function associated with the RC circuit model is

$$H(s) = \frac{1}{1 - RCs}$$

It then immediately follows that the impulse response is given by

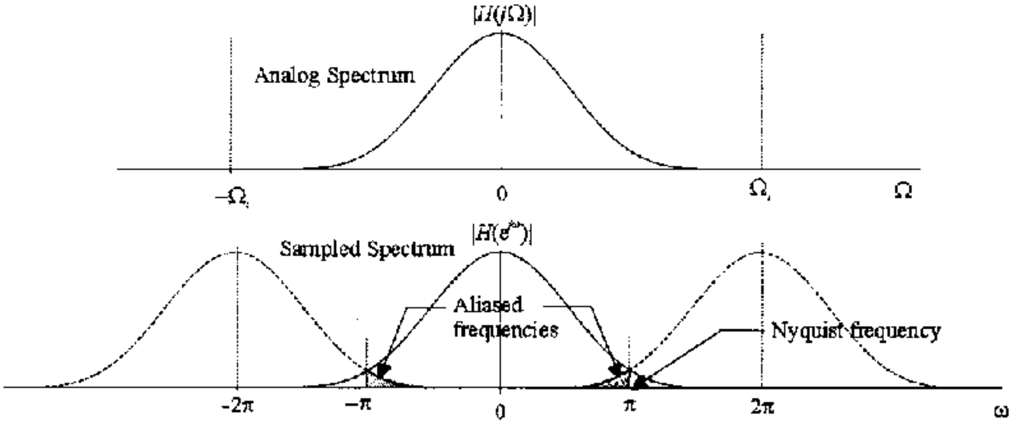


Figure 12 Spectrum of a z-transformed filter.

$$h(t) = \frac{1}{RC} e^{-t/RC} u(t)$$

For a given periodic sampling period of T_s, the resulting sampled impulse response is given by

$$h_d[k] \Leftrightarrow T_s h(kT_s)$$

or, for $k \geq 0$

$$h_d[k] = \frac{T_s}{RC} e^{-kT_s/RC} = \frac{T_s}{RC} \alpha^k$$

where $\alpha = e^{-T_s/RC}$ it follows that

$$H(z) = \frac{T_s}{RC} \frac{1}{(1 - \alpha z^{-1})} = \frac{T_s}{RC} \frac{z}{(z - \alpha)}$$

The frequency response of the impulse-invariant filter is given by

$$H(e^{j\phi}) = \frac{T_s}{RC} \left(\frac{e^{j\phi}}{e^{j\phi} - \alpha} \right)$$

which is periodic with a normalized period π.

3.6.5.2 Bilinear z-Transform

Lowpass filters have known advantages as a signal interpolator (see Chap. 3.4). In the continuous-time domain, an integrator is a standard lowpass filter model. A continuous-time integrator/interpolator is given by

$$H(s) = \frac{1}{s} \tag{48}$$

which has a common discrete-time Reimann model given by

$$y[k+1] = y[k] + \frac{T}{2}(x[k] + x[k+1]) \tag{49}$$

which has a z-transform given by

$$Y(z) = z^{-1} Y(z) + \frac{T}{2}(z^{-1} X(z) + X(z)) \tag{50}$$

which results in the relationship

$$s = \frac{2}{T_s} \frac{(z+1)}{(z-1)} \tag{51}$$

or

$$z = \frac{\frac{2}{T_s} + s}{\frac{2}{T_s} - s} \tag{52}$$

Equation (51) is called a *bilinear z-transform*. The advantage of the bilinear z-transform over the standard z-transform is that it eliminates aliasing errors introduced when an analog filter model (with are arbitrarily long nonzero frequency response) was mapped into the z-plane. The disadvantage, in some applications, is that the bilinear z-transform is not impulse invariant. As a result, the bilinear z-transform is applied to designs which are specified in terms of frequency-domain attributes and ignore time-domain qualifiers. If impulse invariance is required, the standard z-transform is used with an attendant loss of frequency-domain performance.

3.6.6 Warping

The frequency response of a classic analog filter, denoted $H_a(j\Omega)$, $\Omega \in [-\infty, \infty]$, eventually needs to be interpreted as a digital filter denoted $H(e^{j\omega})$, where $\omega \in [-\pi, \pi]$. The bilinear z-transform can map the analog frequency axis onto the digital frequency axis without introducing aliasing, or leakage as was found with a standard z-transform. To demonstrate this claim, consider evaluating Eq. (51) for a given analog frequency $s = j\Omega$. Then

$$j\Omega = \frac{2e^{j\omega} - 1}{T_s e^{j\omega} + 1} = \frac{2}{T_s} \frac{j \sin(\omega/2)}{\cos(\omega/2)} = \frac{2}{T_s} j \tan(\omega/2) \tag{53}$$

which, upon simplification reduces to

$$\Omega = \frac{2}{T_s} \tan(\omega/2) \tag{54}$$

or

$$\omega = 2 \tan^{-1}(\Omega T_s/2) \tag{55}$$

Equation (55) is graphically interpreted in Fig. 13.

Equation (55) is called the *warping equation* and Eq. (54) is referred to as the *prewarping equation*. The nonlinear mapping that exists between the analog- and discrete-frequency axes will not, in general, directly map analog to identical frequencies. While the mapping is nonlinear, the benefit is the elimination of aliasing. From Fig. 13 it can be seen that maps $\Omega \to \infty$ to the continuous-frequency axis, $\omega \to \pi f_s$ (equivalently Nyquist frequency $\omega \to \pi$) in the digital-frequency domain. Because of this, the bilinear z-transform is well suited to converting a classic analog filter into a discrete-time IIR model which preserves the shape of the magnitude frequency response of its analog parent. The design process that is invoked must however, account for these nonlinear effects and is presented

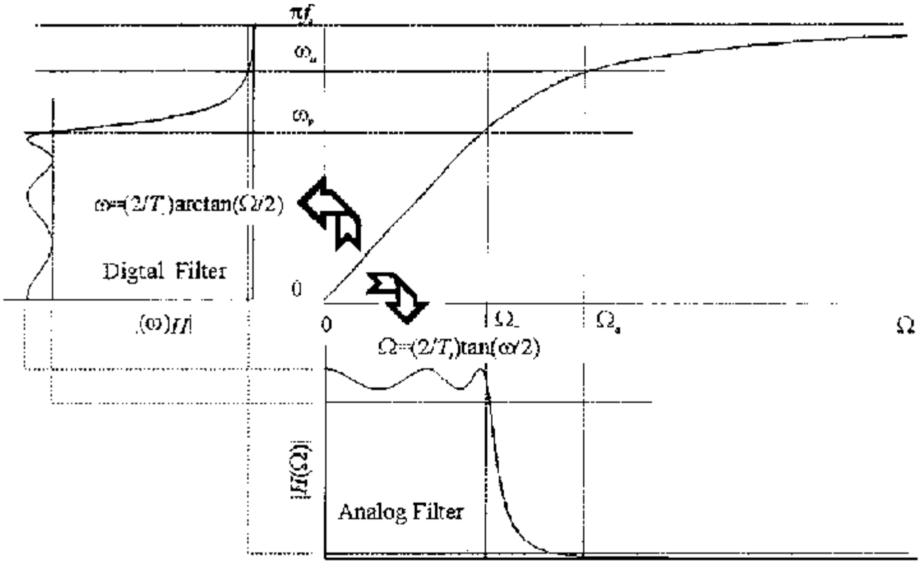


Figure 13 Relationship between the analog and digital frequency axes under the bilinear *z*-transform.

in Fig. 14. Such a process is outlined below as a step-by-step procedure.

1. Specify the desired discrete-time filter requirements and attributes.
2. Prewarp the discrete-time critical frequencies into corresponding analog frequencies and estimate analog filter order.
3. Design an analog prototype filter from the given continuous-time parameters.
4. Convert the analog prototype into an analog filter $H(s)$ using frequency-to-frequency transforms.
5. Design a digital filter $H(z)$ using a bilinear z-transform of $H(s)$ which warps the frequency axis, which has previously been prewarped.

While this method may initially seem to be complicated, it is a simple procedure which can be reduced to a digital computer program. To exemplify the procedure, several simple examples are offered.

3.6.6.1 Classical IIR Design

Classical lowpass Butterworth, Chebyshev-I, Chebyshev-II, and elliptical IIR filters can be designed which meet or exceed the following specifications:

Desired passband attenuation = 1.0 dB
Desired stopband attenuation = 40.0 dB
Sampling frequency = 10,000.0 Hz
Passband edge = 1500.0 Hz
Stopband edge = 2500.0 Hz.

The filters can be designed using MONARCH$^{\text{TM}}$ CAD tools (The Athena Group Inc.) which automatically implement the design steps shown in Fig. 13. The filter orders, required to meet or exceed the design specifications are:

Order(Butterworth) − $N = 8$
Order(Chebyshev-I and -II) − $N = 5$
Order(Elliptical) − $N = 4$

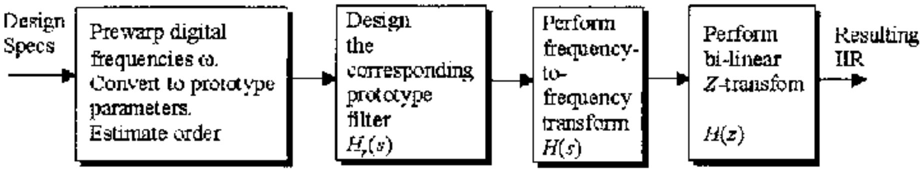

Design Specs
Prewarp digital frequencies ω. Convert to prototype parameters. Estimate order
Design the corresponding prototype filter $H_p(s)$
Perform frequency-to-frequency transform $H(s)$
Perform bi-linear Z-transfom $H(z)$
Resulting IIR

Figure 14 Design of a discrete-time IIR from an analog model using a bilinear *z*-transform.

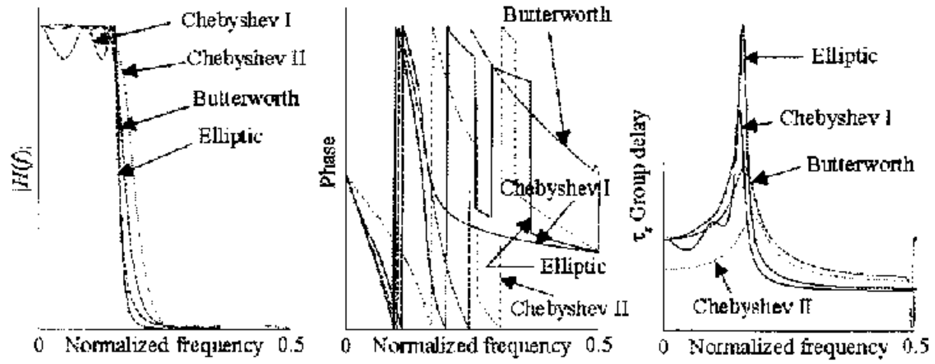


Figure 15 Comparison of magnitude and log magnitude frequency response, phase, response, and group delay of four classical IIRs.

The magnitude frequency responses of the derived filters are shown in Fig. 15.

It can be seen that the magnitude frequency of each classic IIR approximates the magnitude frequency response envelope of an ideal filter in an acceptable manner. The Cheybshev-I and elliptical introduce ripple in the passband, while the Chebyshev-II and elliptical exhibit ripple in the stopband. The Butterworth is ripple-free but requires a high-order implementation. The filters are seen to differ radically in terms of their phase and group-delay response. None of the IIR filters, however, is impulse invariant.

3.6.7 Finite Impulse Response (FIR) Filters

A *finite impulse response* (*FIR*) filter has an impulse response which consists only of a finite number of sample values. The impulse response of an Nth-order FIR is given by

$$h[k] = \{h_0, h_1, \ldots, h_{N-1}\} \tag{56}$$

The time-series response of an FIR to an arbitrary input $x[k]$, is given by the linear convolution sum

$$y[k] = \sum_{m=0}^{N-1} h_m x[k-m] \tag{57}$$

It is seen that the FIR consists of nothing more than a shift-register array of length $N-1$, N multipliers (called *tap weight multipliers*), and an accumulator. Formally, the z-transform of a filter having the impulse response described by Eq. (57) is given by

$$H(z) = \sum_{m=0}^{N-1} h_m z^{-m} \tag{58}$$

The normalized two-sided frequency response of an FIR having a transfer function $H(z)$ is $H(e^{j\omega})$, where $z = e^{j\omega}$ and $w \in [-\pi, \pi]$. The frequency response of an FIR can be expressed in magnitude–phase form as

$$H(e^{\omega}) = |H(e^{j\omega})| \angle \phi(\omega) \tag{59}$$

A system is said to have a *linear phase* respfonse if the phase response has the general form $\phi(\omega) = \alpha\omega + \beta$. Linear phase behavior can be guaranteed by any FIR whose tap weight coefficients are symmetrically distributed about the filter's midpoint, or center tap.

The most popular FIR design found in practice today is called the *equiripple method*. The equiripple design rule satisfies the minimax error criteria

$$\varepsilon_{\text{minimax}} = \text{minimize}\{\text{maximum}(\varepsilon(\omega) \mid \omega \in [0, \omega_s/2]\} \tag{60}$$

where ε is the error measured in the frequency domain measured as

$$\varepsilon(\omega) = W(\omega)|H_d(e^{j\omega}) - H(e^{j\omega})| \tag{61}$$

where $W(\omega) \geq 0$ is called the error weight. The error $\varepsilon(\omega)$ is seen to measure the weighted difference between the desired and realized filter's response at frequency ω. For an Nth-order equiripple FIR, the maximum error occurs at discrete *extremal frequencies* ω_i. The location of the maximum errors are found using the *alternation theorem* from polynomial approximation theory since the signs of the maximal errors alternate [i.e., $\varepsilon(\omega_i) = -\varepsilon(\omega_{i+1})$]. This method was popularized by Parks and McClelland who solved the alternative theorem problem iteratively using the Remez exchange algorithm. Some of the interesting properties of an equiripple FIR is that *all* the maximal errors, called extremal errors, are equal. That is, $\varepsilon_{\text{minimax}} = |\varepsilon(\omega_i)|$

for $i \in [0, N-1]$ for ω_i an extremal frequency. Since all the errors have the same absolute value and alternate in sign, the FIR is generally referrd to by its popular name, *equiripple*. This method has been used for several decades and continues to provide reliable results.

Example 2. Weighted Equiripple FIR: *The 51st-order bandpass equiripple FIR is designed to have a* $-1\,dB$ *pass band and meet the following specifications. The weights W(f) are chosen to achieve the passband attenuation requirements for* $f_s = 100\,\text{kHz}$:

Band 1: $f \in [0.0, 10]\,\text{kHz}$; *desired gain* $= 0.0$, $W(f) = 4$, *stopband.*
Band 2: $f \in [12, 38]\,\text{kHz}$; *desired gain* $= 1.0$, $W(f) = 1$, *passband.*
Band 3: $f \in [40, 50]\,\text{kHz}$; *desired gain* $= 0.0$, $W(f) = 4$, *stopband.*

The FIR is to have a passband and stopband deviation from the ideal of $\delta_p \sim -1\,\text{dB}$ *and* $\delta_p \sim -30\,\text{dB}$ *(see Fig. 16). While the passband deviation has been relaxed to an acceptable value, the stopband attenuation is approximately* $-23.5\,\text{dB}$ *to* $-30\,\text{dB}$.

The advantage of an FIR is found in its implementation simplicity and ability to achieve linear phase performance, if desired. With the advent of high-speed DSP microprocessors, implementation of relatively high-order ($N \sim 100$) FIRs are feasible. As a result, FIRs are becoming increasingly popular as part of a DSP solution.

3.6.8 Multirate Systems

One of the important functions that a digital signal processing system can serve is that of sample rate conversion. A sample-rate converter changes a system's sample rate from a value of f_{in} samples per second, to a rate of f_{out} samples per second. Systems which contain multiple sample rates are called *multirate* systems. If a time series $x[k]$ is accepted at a sample rate f_{in} and exported at a rate f_{out} such that $f_{\text{in}} > f_{\text{out}}$, then the signal is said to be *decimated* by M where M is an integer satisfying

$$M = \frac{f_{\text{out}}}{f_{\text{in}}} \tag{62}$$

A decimated time series $x_d[k] = x[Mk]$ saves only every Mth sample of the original time series. Furthermore, the effective sample rate is reduced from f_{in} to $f_{\text{dec}} = f_{\text{in}}/M$ samples per second, as shown in Fig. 17.

Decimation is routinely found in audio signal processing applications where the various subsystems of differing sample rates (e.g., 40 kHz and 44.1 kHz) must be interconnected. At other times multirate systems are used to reduce the computational requirements of a system. Suppose an algorithm requires K operations be completed per algorithmic cycle. By reducing the sample rate of a signal or system by a factor M, the arithmetic bandwidth requirements are reduced from Kf_s operations per second to Kf_s/M (i.e., M-fold decrease in bandwidth requirements). Another class of applications involves resampling a signal at a lower rate to allow it to pass through a channel of limited bandwidth. In other cases, the performance of an algorithm or transform is based on multirate system theory (e.g., wavelet transforms).

The spectral properties of a decimated signal can be examined in the transform domain. Consider the decimated time series modeled as

$$x_d[k] = \sum_{m=-\infty}^{\infty} x[k]\,\delta[m - kM] \tag{63}$$

which has a z-transform given by

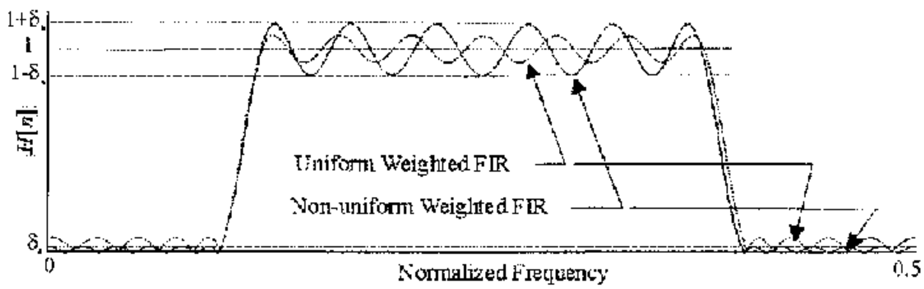


Figure 16 Magnitude frequency response for an equiripple FIR using the design weights $W = \{4, 1, 4\}$. Also shown is the design for $W = \{1, 1, 1\}$.

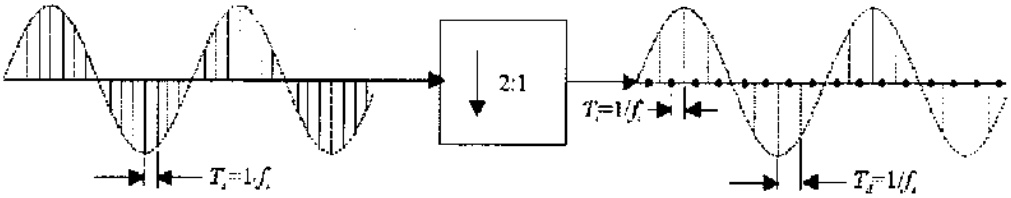


Figure 17 Decimation-by-two example.

$$X_d(z) = X(z^M) \tag{64}$$

The frequency signature of the decimated signal, relative to the undecimated parent signal, is therefore given by

$$X_d(e^{j\phi}) = X(e^{jM\phi}) \tag{65}$$

It can therefore be seen to be a frequency-scaled version of the original signal spectrum repeated on $2\pi/M$ centers when plotted on a frequency axis defined with respect to the original sampling frequency.

Example 3. Decimation: *A signal given by* $x(t) = \cos(2\pi f_s t/16)$ *is sampled at a* 1 kHz *rate to form a time series* $x[k]$. *The spectrum is given by* $X(e^{m}) = 0.5\delta(\omega - 2\pi \times 10^3/16) + 0.5\delta(\omega + 2\pi \times 10^3/16)$. *The time series and spectrum are shown in Fig. 18. What is the spectrum of the decimated-by-four version of* $x[k]$?

From Shannon's sampling theorem, the highest frequency found in $x[k]$ *is* B Hz. *Aliasing can be avoided if the decimating sampling rate exceeds* $f_d = 2B$ Hz. *This means that there is a practical upper bound to the decimation rate which is given by*

$$\frac{f_s}{M} - B > B \tag{66}$$

or

$$M < \frac{f_s}{2B} \tag{67}$$

Increasing the decimation rate beyond this maximal value will introduce aliasing errors into the decimated time series. The maximal decimation rate, however, is rarely used. Instead, a more conservative value is generally adopted which creates a guard band existing between the bandlimited spectra center on f_s/M Hz *centers.*

Example 4. Decimation Rate: *The highest frequency found in the signal* $x(t) = \cos(2\pi \times 10^3 t)$ *is* $B = 10^3$ Hz. *Suppose* $x(t)$ *is highly oversampled at a rate* $f_s = 10^5$ Hz *The minimum lower bound on the sampling rate, namely the Nyquist frequency, is* 2×10^3 Hz. *Therefore the maximum decimation rate is* $10^5/2 \times 10^3 = 50$.

The antithesis of decimation is called interpolation. In the context of multirate signal processing, interpolation simply refers to a mechanism of increasing the effective sample rate of a signal. Formally, suppose a signal $x_d[k]$ is interpolating by a factor N to form $x_i[k]$, then

$$x_i[k] = \begin{cases} x_d[k] & \text{if } n = 0 \bmod N \\ 0 & \text{otherwise} \end{cases} \tag{68}$$

That is, the act of interpolation inserts $N-1$ zeros between the samples of the original time series. This action is sometimes referred to as zero-padding, being

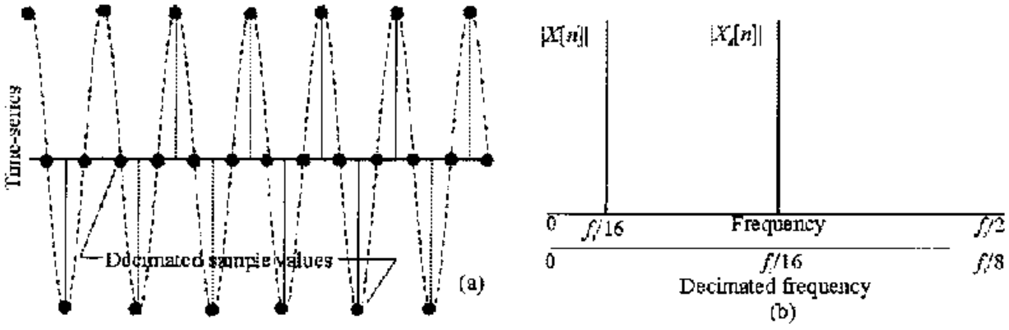


Figure 18 (a) Parent time series $x[n]$ and decimated-by-four time series $x_d[k]$; (b) magnitude frequency responses $X[n]$ and $X_d[n]$.

sampled at a rate f_{in}, which produces a new time series sampled at rate $f_{\text{out}} = Nf_{\text{in}}$.

Interpolation is often directly linked to decimation. Suppose $x_d[k]$ is a decimated-by-M version of a time series $x[k]$ which was sampled at a rate f_s. Then $x_d[k]$ contains only every Mth sample of $x[k]$ and is defined with respect to a decimated sample rate $f_d = f_s/M$. Interpolating $x_d[k]$ by N would result in a time series $x_i[k]$, where $x_i[Nk] = x_d[k]$ and 0 otherwise. The sample rate of the interpolated signal would be increased from f_d to $f_i = Nf_d = Nf_s/M$. If $N = M$, it can be seen that the output sample rate would be restored to f_s.

3.7 DSP TECHNOLOGY AND HARDWARE

The semiconductor revolution of the mid-1970s produced the tools needed to effect many high-volume real-time DSP solutions. These include medium, large, and very large integrated circuit (MSI, LSI, VLSI) devices. The ubiquitous *microprocessor*, with its increasing capabilities and decreasing costs, now provides control and arithmetic support to virtually every technical discipline. Industry has also focused on developing *application-specific* single-chip dedicated DSP units called *ASICs*. The most prominent has been the *DSP microprocessor*. There is now an abundance of DSP chips on the market which provide a full range of services.

Perhaps the most salient characteristic of a DSP chip is its multiplier. Since multipliers normally consume a large amount of chip real estate, their design has been constantly refined and redefined. The early AMI2811 had a slow $12 \times 12 = 16$ multiplier, while a later TMS320 had a $16 \times 16 =$ 32-bits 200 nsec multiplier that occupied about 40% of the silicon area. These chips include some amount of onboard RAM for data storage and ROM for fixed coefficient storage. Since the cost of these chips is very low (tens of dollars), they have opened many new areas for DSP penetration. Many factors, such as speed, cost, performance, software support and programming language, debugging and emulation tools, and availability of peripheral support chips, go into the hardware design process. The Intel 2920 chip contained onboard ADC and DAC and defined what is now called the first-generation DSP microprocessor. Since its introduction in the late 1970s, the Intel 2920 has given rise to three more generations of DSP microprocessors. The second removed the noise-sensitive ADC and DAC from the digital device and added a more powerful multiplier and additional memory. Generation three introduced floating-point. Generation four is generally considered to be multiprocessor DSP chips. DSP has traditionally focused on its primary mission of linear filtering (convolution) and spectral analysis (Fourier transforms). These operations have found a broad application in scientific instrumentation, commercial products, and defense systems. Because of the availability of low-cost high-performance DSP microprocessors and ASICs, DSP became a foundation technology during the 1980s and 90s.

DSP processors are typified by the following characteristics:

Only one or two data types supported by the processor hardware
No data cache memory
No memory management hardware
No support for hardware context management
Exposed pipelines
Predictable instruction execution timing
Limited register files with special-purpose registers
Nonorthogonal instruction sets
Enhanced memory addressing modes
Onboard fast RAM and/or ROM, and possibly DMA.

Digital signal processors are designed around a different set of assumptions than those which drive the design of general-purpose processors. First, digital signal processors generally operate on arrays of data rather than scalars. Therefore the scalar load–store architectures found in general-purpose RISCs are absent in DSP microprocessors. The economics of software development for digital signal processors is different from that for general-purpose applications. Digital signal processing problems tend to be algorithmically smaller than, for example, a word processor. In many cases, the ability to use a slower and therefore less expensive digital signal processor by expending some additional software engineering effort is economically attractive. As a consequence, real-time programming of digital signal processors is often done in assembly language rather than high-level languages.

Predicting the performance of a DSP processor in general and application-specific settings is the mission of a benchmark. A typical benchmark suite has been developed by Berkeley Design Technologies and consists of (1) real FIR, (2) complex FIR, (3) real single sample FIR, (4) LMS adaptive FIR, (5) real IIR, (6) vector dot product, (7) vector add, (8) vector maximum, (9) convolution encoder, (10) finite-state machine, and (11) radix-2 FFT.

DSP theory is also making advances that are a logical extension of the early work in algorithms. DSP algorithm development efforts typically focus on linear filtering and transforms along with creating CAE environments for DSP development efforts. DSP algorithms have also become the core of image processing and compression, multimedia, and communications. Initiatives are also found in the areas of adaptive filtering, artificial neural nets, multidimensional signal processing, system and signal identification, and time–frequency analysis.

3.8 SUMMARY

Even though the field of digital signal processing is relatively young, it has had a profound impact on how we work and recreate. DSP has become the facilitating technology in industrial automation, as well as providing a host of services that would otherwise be impossible to offer or simply unaffordable. DSP is at the core of computer vision and speech systems. It is the driving force behind data communication networks whether optical, wired, or wireless. DSP has become an important element in the fields of instrumentation and manufacturing automation. The revolution is continuing and should continue to provide higher increased levels of automation at lower costs from generation to generation.

BIBLIOGRAPHY

Antonious A. Digital Filters: Analysis and Design. New York, McGraw-Hill, 1979.

Blahut R. Fast Algorithms for Digital Signal Processing. Reading, MA: Addison-Wesley, 1985.

Bracewell R. Two Dimensional Imaging. New York: Prentice-Hall, 1995.

Brigham EO. The Fast Fourier Transform and Its Application. New York: McGraw-Hill, 1988.

Haykin S. Adaptive Filter Theory, 3rd ed. New York: Prentice-Hall, 1996.

Oppenheim AV, ed. Application of Digital Signal Processing. Englewood Cliffs, NJ: Prentice-Hall, 1978.

Proakis J, Manolakis DG. Digital Signal Processing: Principles, Algorithms, and Applications, 3rd ed. New York: Prentice-Hall, 1996.

Rabiner LR, Gold B. Theory and Applications of Digital Signal Processing. Englewood Cliffs, NJ: Prentice-Hall, 1975.

Taylor F. Digital Filter Design Handbook. New York: Marcel Dekker, 1983.

Taylor, F. and Millott, J., "Hands On Digital Signal Processing," McGraw-Hill, 1988.

Zelniker G, Taylor F. Advanced Digital Signal Processing: Theory and Applications. New York: Marcel Dekker, 1994.

Chapter 3.4

Sampled-Data Systems

Fred J. Taylor
University of Florida, Gainesville, Florida

4.1 ORIGINS OF SAMPLED-DATA SYSTEMS

The study of signals in the physical world generally focuses on three signal classes called *continuous-time* (*analog*), *discrete-time* (*sampled-data*), and *digital*. Analog signals are continuously refined in both amplitude and time. Sampled-data signals are continuously refined in amplitude but discretely resolved in time. Digital signals are discretely resolved in both amplitude and time. These signals are compared in Fig. 1 and are generally produced by different mechanisms. Analog signals are naturally found in nature and can also be produced by electronic devices. Sampled-data signals begin as analog signals and are passed through an electronic sampler. Digital signals are produced by digital electronics located somewhere in the signal stream. All have an important role to play in signal processing history and contemporary applications. Of these cases, sampled data has the narrowest application-base at this time. However, sampled data is also known to be the gateway to the study of *digital signal processing* (*DSP*), a field of great and growing importance (see Chap. 3.3). Sampled-data signal processing formally refers to the creation, modification, manipulation, and presentation of signals which are defined in terms of a set of sample values called a *time series* and denoted $\{x[k]\}$. An individual sample has the value of an analog signal $x(t)$ at the *sample instance* $t = kT_s$, namely, $x(t = kT_s) = x[k]$, where T_s is the *sample period*, and $f_s = 1/T_s$ is the *sample rate* or *sample frequency*.

The sampling theorem states that if a continuous-time (analog) signal $x(t)$, band limited to $B\ H_z$, is periodically at a rate $f_s > 2B$, the signal $x(t)$ can be *exactly* recovered (reconstructed) from its sample values $x[k]$ using the *interpolation* rule

$$x(t) = \sum_{k=-\infty}^{\infty} x[k]h(t - kT_s) \tag{1}$$

where $h(t)$ has the $\sin(x)/x$ envelope and is defined to be

$$h(t) = \frac{\sin(\pi t/T_s)}{\pi t/T_s} \tag{2}$$

The interpolation process is graphically interpreted in Fig. 2. The lower bound on the sampling frequency f_s is $f_L = 2B$, and is called the *Nyquist sample rate*. Satisfying the sampling theorem requires that the sampling frequency be strictly greater than the Nyquist sample rate or $f_s > f_L$. The frequency $f_N = f_s/2 > B$ called the *Nyquist frequency*, or *folding frequency*. This theory is both elegant and critically important to all sample data and DSP studies.

Observe that the interpolation filter $h(t)$ is both infinitely long and exists before $t = 0$ [i.e., $t \in (-\infty, \infty)$]. Thus the interpolator is both impractical from a digital implementation standpoint and *noncausal*. As such,

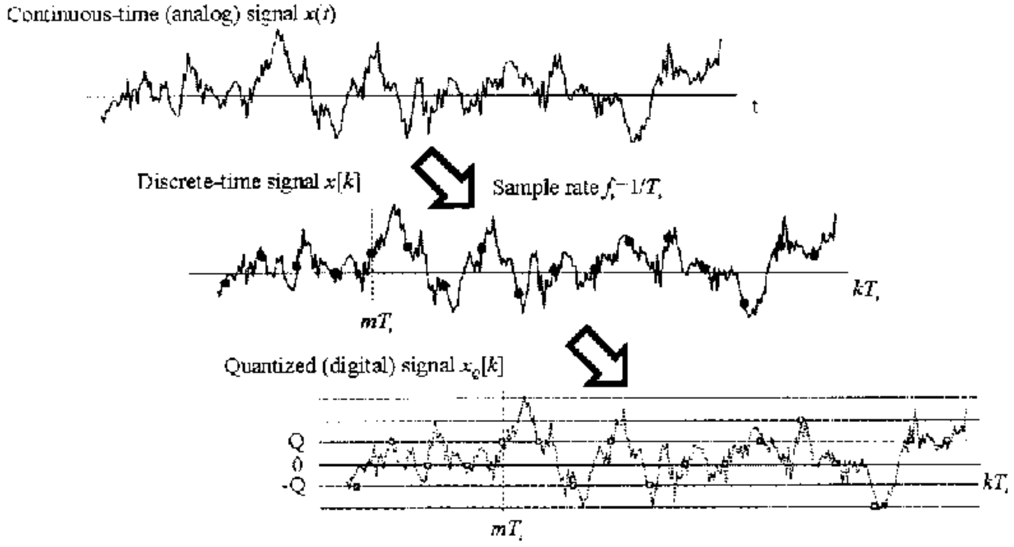

Continuous-time (analog) signal $x(t)$
t
Discrete-time signal $x[k]$
Sample rate $f_s=1/T_s$
kT_s
mT_s
Quantized (digital) signal $x_Q[k]$
Q
0
-Q
kT_s
mT_s

Figure 1 Signal hierarchy consisting of analog, discrete-time or sampled-data, and digital or quantized signal processes. Sampled-data and digital signals are quantized in time. Digital signals are also quantized in amplitude.

alternative interpolation schemes are generally used which approximate a Shannon interpolator. One of the most basic is called the *zero-order hold.* A zero-order hold circuit simply "holds" the value of $x[k]$ for a sample interval T_s beginning at $t = kT_s$. The zero-order interpolated signal is therefore a piecewise constant approximation of the true $x(t)$ as shown in Fig. 3. The quality of the zero-order hold approximation of $x(t)$ is influenced by the choice of T_s. If $x(t)$ rapidly changes in time, T_s must be extremely small in order for a piecewise constant zero-order hold signal to be in close agreement with the analog parent $x(t)$.

Another important practical interpolation procedure is the *first-order hold.* The first-order hold linearly interpolates the values of $x(t)$ between two adjacent sample values $x[k]$ and $x[k+1]$ for $t \in [kT_s$, (k+)$T_s)$.

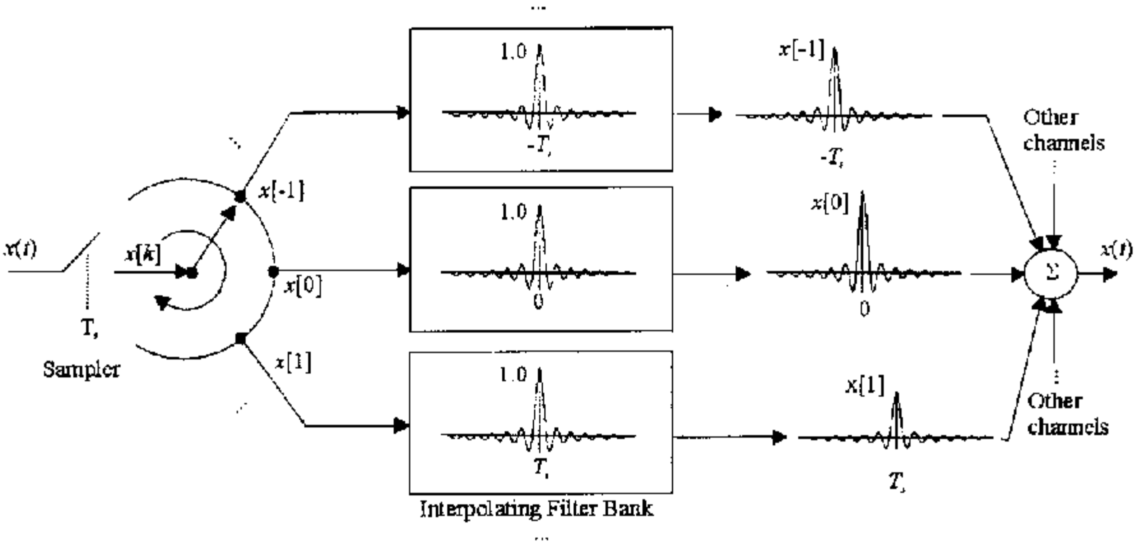


Figure 2 Shannon interpolator showing a sampler (left), Shannon interpolating filters (middle), and reconstruction after Eq. (1) (right).

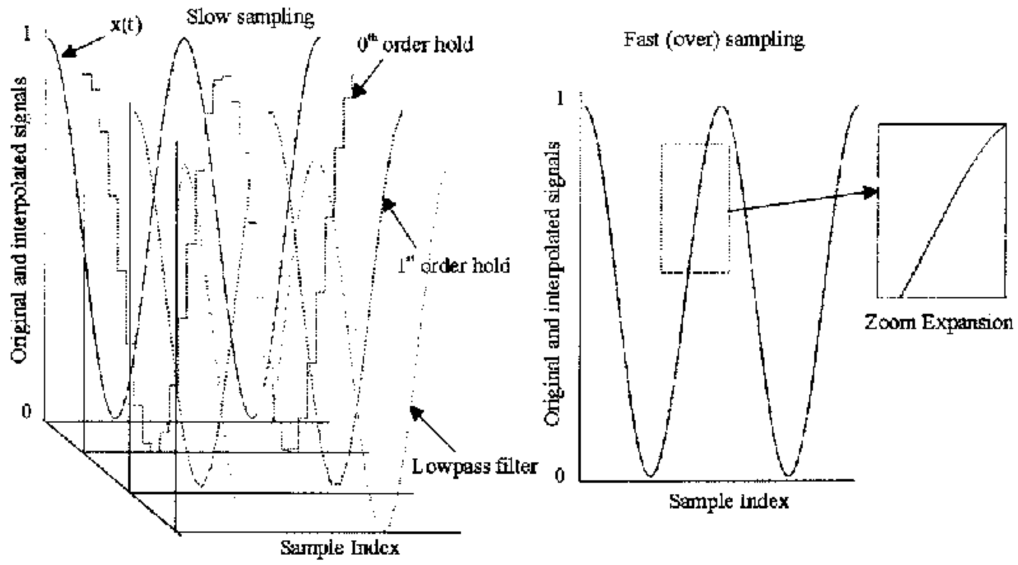


Figure 3 Zero- and first-order hold and lowpass filter interpolators. Shown on the left is the interpolation process for a sowly sampled signal with the piecewise constant envelope of the zero-order hold clearly visible. The other interpolators are seen to provide reasonably good service. On the right is an oversampled case where all interpolators work reasonably well.

The first-order hold interpolation scheme is graphically interpreted in Fig. 3. Again the quality of the interpolation is seen to be correlated to the value of T_s, but to a lesser degree than in the zero-order hold case.

Another popular interpolation method uses a lowpass filter and is called a *smoothing filter*. It can be argued from the duality theorem of Fourier transforms that the inverse Fourier transform of Eq. (2) is itself an ideal lowpass filter. A practical lowpass filter will permit only small incremental changes to take place over a sample interval and does so in a smooth manner. The smoothing filter should be matched to the frequency dynamics of the signal. If the signal contains frequency components in the stopband of the smoothing filter, the interpolator will lose its ability to reconstruct sharp edges. If the smoothing filter's bandwidth is allowed to become too large, the interpolator will become too sensitive to amplitude changes and lose its ability to interpolate.

4.2 MATHEMATICAL REPRESENTATION OF SAMPLED-DATA SIGNALS

Sampled-data or discrete-time signals can be produced by presenting a continuous-time signal $x(t)$ to an ideal sampler which is assumed to be operating above the Nyquist rate. The connection between continuous- and sampled-data signals is well known in the context of a Laplace transform. Specifically, if $x(t) \leftrightarrow X(s)$, then

$$x(t - kT_s) \xleftrightarrow{L} e^{-skT_s} X(s) \tag{3}$$

The time series $\{x[k]\} = \{x[0], x[1], \ldots\}$ would therefore have a Laplace transform given by

$$\begin{aligned} X(s) &= x[0] + x[1]e^{-2sT_s} + x[2]e^{-2sT_s} + \cdots \\ &= \sum_{k=0}^{\infty} x[k]e^{-ksT_s} \end{aligned} \tag{4}$$

It can be seen that in the transform domain the representation of a sampled-data signal is punctuated with terms the form e^{-skT_s}. For notational purposes, they have been given the shorthand representation

$$z = e^{sT_s} \qquad \text{or} \qquad z^{-1} = e^{-sT_s} \tag{5}$$

Equation (5) defines what is called the z-operator and provides the foundation for the z-transform. The complex mapping $z = e^{\sigma + j\varphi} = re^{j\varphi}$, where $r = e^{\sigma}$ and $\varphi = k2\pi + \varphi_0$, results in a contour in the z-plane given by $z = re^{j(2\pi + \varphi_0)} = re^{j\varphi_0}$. If uniqueness is required, the imaginary part of s must be restricted to a range $|\varphi_0| \le \pi$ which corresponds to bounding the normalized frequency range by plus or minus Nyquist frequency in the s-plane. For values of s outside this range, the mapping $z = e^{sT_s}$ will "wrap" around the unit circle modulo $(2\pi f_s)$ radians per second.

The *two-sided z-transform*, for a double-ended time series $\{x[k]\}$, is formally given by

$$X(z) = \sum_{k=-\infty}^{\infty} x[k]z^{-k} \tag{6}$$

if the sum converges. If the time series is defined for positive time instances only, called a *right-sided time series*, the *one-sided z-transform* applies and is given by

$$X(z) = \sum_{k=0}^{\infty} x[k]z^{-k} \tag{7}$$

which again exists only if the sum converges. The range of values of z over which the z-transform will converge is called the *region of convergence*, or *ROC*. The z-transforms of elementary functions are generally cataloged, along with their ROCs, in Table 1.

It is generally assumed that most of important signals can be represented as a mathematical combination of manipulated elementary functions. The most commonly used mapping techniques are summarized in Table 2.

In addition to the properties listed in Table 2, there are several other z-transform relationships which are of significant importance. One is the *initial-value theorem* which states

$$x[0] = \lim_{z\to\infty} (X(z)) \tag{8}$$

if $x[k]$ is causal. The second property is called the *final-value theorem* which is given by

$$x[\infty] = \lim_{z\to\infty} (z-1)\,X(z) \tag{9}$$

provided $X(z)$ has no more than one pole on the unit circle and all other poles are interior to the unit circle.

4.3 INVERSE *z*-TRANSFORM

The inverse z-transform of a given $X(z)$ is defined by

$$x[k] = Z^{-1}(X(z)) = \frac{1}{2\pi j}\oint_C X(z)z^{n-1}\,dz \tag{10}$$

where C is a restricted closed path which resides in the ROC of $X(z)$. Solving the integral equation can obviously be a very tedious process. Fortunately, algebraic methods can also be found to perform an inverse z-transform mapping. *Partial fraction expansion* is by far the most popular z-transform inversion method in contemporary use to map a given $X(z)$ into the original time series. A partial fraction expansion of $X(z)$ repre-

Table 1 z-Transform and ROCs

Time-domain	z-Transform	Region of convergence: $\vert z\vert > R$
$\delta[k]$	1	Everywhere
$\delta[k-m]$	z^{-m}	Everywhere
$u[k]$	$z/(z-1)$	1
$ku[k]$	$z/(z-1)^2$	1
$k^2u[k]$	$z(z+1)/(z-1)^3$	1
$k^3u[k]$	$z(z^2+4z+1)/(z-1)^4$	1
$\exp[akT_s)\,u[kT_s]$	$z/(z-\exp(aT_s))$	$\vert\exp(aT_s)\vert$
$kT_s\exp[akT_s]\,u[kT_s]$	$zT_s\exp(aT_s)/(z-\exp(aT_s))^2$	$\vert\exp(aT_s)\vert$
$(kT_s)^2\exp[akT_s]\,u[kT_s]$	$z(T_s)^2\exp(aT_s)(z+\exp(aT_s)/(z-\exp(aT_s))^3$	$\vert\exp(aT_s)\vert$
$a\,u[k]$	$z/(z-a)$	$\vert a\vert$
$ka\,u[k]$	$az/(z-a)^2$	$\vert a\vert$
$k^2a\,u[k[$	$az(z+a)/(z-a)^3$	$\vert a\vert$
$\sin[bkT_s]\,u[kT_s]$	$z\sin(bT_s)/(z^2-2z\cos(bT_s)+1)$	1
$\cos[bkT_s]\,u[kT_s]$	$z(z-\cos(bT_s))/(z^2-2z\cos(bT_s)+1)$	1
$\exp[akT_s]\sin[bkT_s]\,u[kT_s]$	$z\exp(aT_s\sin(bT_s)/z^2-2z\exp(aT_s)\cos(bT_s)+\exp(2aT_s))$	$\vert\exp(aT_s)\vert$
$\exp[akT_s]\cos[bkT_s]\,u[kT_s]$	$x(z-\exp(aT_s)\cos(bT_s))/(z^2-2z\exp(aT_s)\cos(bT_s)+\exp(2aT_s))$	$\vert\exp(aT_s)\vert$
$a^k\sin(bkT_s)\,u[kT_s]$	$az\sin(bT_s)/(z^2-2az\cos(bT_s)+a^2)$	$\vert a\vert$
$a^k\cos(bkT_s)\,u[kT_s]$	$z(z-a\cos(bT_s))/(z^2-2az\cos(bT_s)+a^2)$	$\vert a\vert$
$a^k, k\in[0, N-1]$	$(1-a^Nz^{-N})/(1-az^{-1})$	Everywhere

Table 2 Mapping Rules in the z-Domain

Property	Time series	z-Transform
Linearity	$x_1[k] + x_2[k]$	$X_1(z) + X_2(z)$
Real scaling	$ax[k]$	$aX(z)$
Complex scaling	$w^k x[k]$	$X(z/w)$
Delay	$x[k-L]$	$z^{-L}X(z) + \sum_{k=-L}^{-1} z^{-(L+n)}x[k]$
Time reversal	$x[-k]$	$X(1/z)$
Modulation	$e^{-ak}x[k]$	$X(e^a z)$
Rampling	$kx[k]$	$-z\dfrac{dX(z)}{dz}$
Summation	$\sum_{n=-\infty}^{k} x[n]$	$\dfrac{zX(z)}{z-1}$

sents the transform as a linear combination of terms having a known correspondence to the primitive functions found in most standard tables of z-transforms (e.g., Table 1).

The practical fraction expansion of the $X(z)$ having the rational polynomial representation

$$X(z) = \frac{N(z)}{D(z)} \tag{11}$$

is a mechanical process. The values of $N(z)$ which satisfy $N(z) = 0$ are called the *zeros* of $X(z)$. The values of $D(z)$ which satisfy $D(z) = 0$ are called the *poles* of $X(z)$. The denominator term $D(z)$ is assumed to be an Nth-order polynomial which can be expressed in product form as

$$D(z) = \prod_{i=1}^{L} (z - \lambda_l)^{n(i)} \tag{12}$$

where

$$N = \sum_{i=1}^{L} n(i) \tag{13}$$

That is, there are L distinct roots in $D(z)$ having values λ_i respectively where the integer $n(i)$ is called the *multiplicity* of the root λ_i. If $n(i) > 1$, then λ_i is said to be a repeated root and if $n(i) = 1$, λ_i is said to be distinct. If $X(z)$ is *proper* (i.e., $M \leq N$), then the *partial fraction* or *Heaviside* expansion of $X(z)$ is given by

$$X(z) = \alpha_0 + \frac{N'(z)}{D(z)} = \alpha_0 + \sum_{i=1}^{L} \sum_{j=1}^{n(i)} \frac{\alpha_{i,j} z}{(z - \lambda_i)^i} \tag{14}$$

where the coefficients $\alpha_{i,j}$ are called the *Heaviside coefficients* and $N'(z)$ is the quotient polynomial obtained by formally dividing $N(z)$ by $D(z)$. Once the Heaviside coefficients $\alpha_{i,j}$ are computed, $x[k]$ can be directly computed by weighting the inverse z-transform of $z/(z - \lambda_i)^j$ found in a standard table (see Table 1) by an amount $\alpha_{i,j}$. A local expansion of Eq. (14) would produce terms of the form

$$X(z) = \alpha_0 + \frac{z\alpha_{i,n(i)}}{(z - \lambda_i)^{n(i)}} + \cdots + \frac{z\alpha_{j,n(j)}}{(z - \lambda_j)^{n(j)}} + \cdots + \frac{z\alpha_{j,k}}{(z - \lambda_j)^k} + \cdots \tag{15}$$

The Heaviside coefficients are given by

$$\alpha_{j,n(j)} = \lim_{z \to \lambda_j} \left(\frac{(z - \lambda_i)^{n(j)} X(z)}{z} \right) \tag{16}$$

$$\alpha_{j,n(j)-1} = \lim_{z \to \lambda_j} \left(\frac{1}{2} \frac{d^2 \left\{ \dfrac{(z - \lambda_i)^{n(j)} X(z)}{z} \right\}}{dz^2} \right) \tag{17}$$

and, in general,

$$\alpha_{j,k} = \lim_{z \to \lambda_j} \left(\frac{1}{(n(j) - k)!} \frac{d^{(n(j)-k)} \left\{ \dfrac{(z - \lambda_i)^{n(j)} X(z)}{z} \right\}}{dz^{(n(j)-k)}} \right) \tag{18}$$

The process of computing Heaviside coefficients can therefore be seen to consist of a number of steps, namely

Determine α_0 in Eq. (14) along with $N'(z)$.
Factor $D(z)$ to obtain the pole locations.
Classify the poles as being distinct or repeated and if repeated, determine their multiplicity.
Use Eq. (16) through (18) to determine the Heaviside coefficients.
Substitute the Heaviside coefficients into Eq. (15).
Use standard tables of z-transforms to invert Eq. (15).

Example 1. Inverse z-transform: *To compute the inverse z-transform of*

$$X(z) = \frac{3x^3 - 5z^2 + 3z}{(z-1)^2(z-0.5)}$$

using Heaviside's method, it is required that X(z) be expanded in partial fraction form as

$$X(z) = \alpha_0 + \alpha_1 \frac{z}{(z-0.5)} + \alpha_{21}\frac{z}{(z-1)} + \alpha_{22}\frac{z}{(z-1)^2}$$

In this case, the pole at $z = 1$ has a multiplicity of 2. Using the production rules defined by Eq. (16) through (18), one obtains

$$\begin{aligned}
\alpha_0 &= \lim_{z\to 0} \frac{zX(z)}{z} = 0 \\
\alpha_1 &= \lim_{z\to 0.5} \frac{(z-0.5)X(z)}{z} = \lim_{z\to 0.5}\left(\frac{3z^3 - 5z^2 + 3z}{z(z-1)}\right) \\
&= 5 \\
\alpha_{22} &= \lim_{z\to 1} \frac{(z-1)^2 X(z)}{z} = \lim_{z\to 1}\left(\frac{3z^3 - 5z^2 + 3z}{z(z-0.5)}\right) = 2 \\
a_{21} &= \lim_{z\to 1} \frac{d}{dz}\left(\frac{(z-1)^2 X(z)}{z}\right) = \lim_{z\to 1}\left(\frac{9z^2 - 10z + 3}{z(z-0.5)}\right. \\
&\qquad \left. - \frac{(3z^3 - 5z^2 + 3z)(2z-0.5)}{(z(z-0.5))^2}\right) = -2
\end{aligned}$$

which states that the inverse z-transform of X(z) is given by $x[k] = [5(0.5)^k - 2 + 2k]\,u[k]$.

4.4 LINEAR SHIFT-INVARIANT SYSTEMS

One of the most important concepts in the study of sampled-data systems is the superposition principle. A system $\boldsymbol{S}$ has the *superposition property* if the output of $\boldsymbol{S}$ to a given input $x_i[k]$ is $y_i[k]$, denoted $y_i[k] = \boldsymbol{S}(x_i[k])$, then the output of $\boldsymbol{S}$ to $x[k]$ is $y[k]$ where

$$x[k] = \sum_{m=1}^{L} a_i x_i[k] \Rightarrow \sum_{m=1}^{L} a_i S(x_i[k]) = y[k] \tag{19}$$

A system is said to be a *linear* system if it exhibits the superposition property. If a system is not linear it is said to be *nonlinear*. A sampled-data system $\boldsymbol{S}$ is said to be *shift invariant* if a shift, or delay in the input time series, produces an identical shift or delay in the output. That is, if

$$x[k] \xrightarrow{S} y[k] \tag{20}$$

and $\boldsymbol{S}$ is shift invariant, then

$$x[k+m] \xrightarrow{S} y[k+m] \tag{21}$$

If a system is *both* linear and shift invariant, then it is said to be a *linear shift-invariant* (LSI) system. LSI systems are commonly encountered in studies of sampled-data and DSP which consider Nth-order system modeled as

$$\sum_{m=0}^{N} a_m y[k-m] = \sum_{m=0}^{M} b_m x[k-m] \tag{22}$$

If $N \geq M$, the system is said to be *proper* and if $a_0 = 1$, the system is classified as being *monic*. What is of general interest is determining the forced, or *inhomogeneous*, solution $y[k]$ of the LSI system defined in Eq. (22) to an arbitrary input $x[k]$. The input–output relationship of a causal *at-rest* (zero initial condition) LSI system to a forcing function $x[k]$ is given by

$$y[k] = \frac{1}{a_0}\left(\sum_{m=0}^{M} b_m x[k-m] - \sum_{m=1}^{N} a_m y[k-m]\right) \tag{23}$$

The solution to Eq. (23) is defined by a *convolution sum* which is specified in terms the discrete-time system's *impulse response* $h[k]$, the response of an at-rest LSI to an input $x[k] = \delta[k]$. The convolution of an arbitrary time series $x[k]$ by a system having an impulse response $h[k]$, denoted $y[k] = h[k] * x[k]$, is formally given by

$$\begin{aligned}
y[k] = h[k] * x[k] &= \sum_{m=0}^{\infty} h[k-m]\,x[m] \\
&= \sum_{m=0}^{\infty} h[m]\,x[k-m]
\end{aligned} \tag{24}$$

Computing a convolution sum, however, often presents a challenging computational problem. An alternative technique, which is based on direct z-transform methods, can generally mitigate this problem. Suppose that the input $x[k]$ and impulse response $h[k]$ of an at-rest discrete-time LSI system have z-transforms given by

$$\begin{aligned} h[k] &\overset{Z}{\longleftrightarrow} H(z) \\ x[k] &\overset{Z}{\longleftrightarrow} X(z) \end{aligned} \tag{25}$$

respectively. Then, the z-transform of Eq. (24) would result in

$$\begin{aligned} Y(z) = Z(y[k]) &= \sum_{m=0}^{\infty} h[m] \left(\sum_{p=0}^{\infty} x[p] z^{-(p+m)} \right) \\ &= \sum_{m=0}^{\infty} h[m] z^{-m} \left(\sum_{p=0}^{\infty} x[p] z^{-p} \right) = H(z)\, X(z) \end{aligned} \tag{26}$$

Therefore the z-transform of the convolution sum $y[k] = h[k] * x[k]$ is mathematically equivalent to multiplying the z-transforms of $h[k]$ and $x[k]$ in the z-domain, and then computing the inverse z-transform of the result. Equation (26) is also known by its popular name, the *convolution theorem* for z-transforms and provides a bridge between time-domain convolution and transform operations. If the regions of convergence for $X(z)$ and $H(z)$ are R_x and R_h respectively, then the region of convergence of $Y(z)$ is R_y where $R_y \supset R_x \cap R_h$. This process is graphically interpreted in Fig. 4. The attraction of the convolution theorem is that it replaces a challenging convolution sum computation with a set of simple algebraic z- and inverse z-transform calls.

4.5 TRANSFER FUNCTION

Applying the convolution theorem to the at-rest LSI model found in Eq. (22) produces

$$\sum_{m=0}^{N} a_m Y(z) z^{-m} = \sum_{m=0}^{M} b_m X(z) z^{-m} \tag{27}$$

The ratio of $Y(z)$ to $X(z)$ is formally called the *transfer function*, denoted $H(z)$, and given by

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\left(\sum\limits_{m=0}^{M} b_m z^{-m} \right)}{\left(\sum\limits_{m=0}^{N} a_m z^{-m} \right)} \tag{28}$$

The transfer function describes how the z-transform of the input signal is transformed to the z-transformed of the output signal. An LSI system which has all its poles and zeros residing interior to the unit circle is called a *minimum-phase system*. Minimum phase systems are known to have strong transient responses and are important to the study of inverse systems [i.e., $G(z) = 1/H(z)$].

Example 2. RLC Circuit: *The RLC electrical circuit is assumed to satisfy the second-order ordinary differential equation*

$$\frac{d^2 y(t)}{dt^2} + 3\frac{dy(t)}{dt} + 2y(t) = x(t)$$

which has a continuous-time system's impulse response given by

$$h(t) = (e^{-t} - e^{-2t})\, u(t)$$

For a sample period of $T_s = 1/f_s$ seconds, the discrete-time impulse response satisfies

$$h[k] = h(kT_s) = e^{-kT_s} - e^{-2kT_s} = a^k - b^k$$

where a and b are defined in the obvious manner. The z-transform of $h[k]$ *is given by*

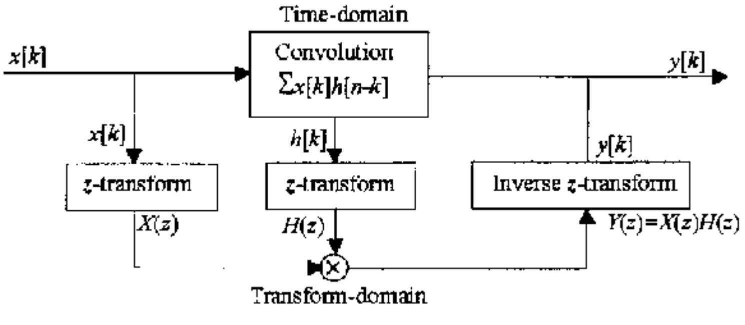


Figure 4 Convolution theorem.

$$H(z) = \frac{z}{z-a} - \frac{z}{z-b} = \frac{(a-b)z}{(z-a)(z-a)}$$

It is known that the input $x[k]$ is a unit step, then $X(z) = u(z) = z/(z-1)$. It immediately follows that

$$Y(z) = X(z)\,H(z) = \frac{(a-b)z^2}{(z-a)(z-a)(z-1)}$$

Using previously established methods of inverting a z-transform, namely partial fraction expansion, the inverse of $Y(z)$ is a time series:

$$y[k] = \left(\frac{(a-b)}{(1-a)(1-b)} + \frac{a^{k+1}}{(1-a)} + \frac{b^{k+1}}{(1-b)} \right) u[k]$$

which is also the step response of the LSI system.

4.6 STABILITY OF AN LSI SYSTEM

If, for all possible bounded initial conditions, the at-rest solution of an LSI $y[k] \to 0$ as $k \to \infty$, then the system is said to be *asymptotically stable*. If an LSI is asymptotically stable, then it is also *bounded-input–bounded-output* (*BIBO*) stable. BIBO stability simply states that the output will remain bounded provided the input is bounded. The stability of an LSI system can be determined in the transform domain. Suppose an LSI given by $H(z)$ contains N poles which are located at $z = p_i$, where p_i may be real or complex, distinct or repeated. Then, in general, the partial fraction expansion of a strictly proper $H(z)$ is given by

$$\begin{aligned} H(z) = \sum_{r=1}^{L} \sum_{m=1}^{n(r)} \left(\frac{\alpha_{r,i}}{(z-p_r)^m} \right) &= \cdots + \frac{\alpha_{r,1} z}{(z-p_r)^1} \\ + \frac{\alpha_{r,2} z}{(z-p_r)^2} + \cdots &+ \frac{\alpha_{r,n(r)} z}{(z-p_r)^{n(r)}} + \cdots \end{aligned} \tag{29}$$

where $n(r)$ is the multiplicity of the pole located at p_r, and

$$\sum_{i=1}^{L} n(r) = N \tag{30}$$

The coefrficients α's are computed using Heaviside's method. The inverse z-transform of $H(z)$ is the system's impulse response $h[k]$ which would have the general form

$$\begin{aligned} h[k] = \cdots + \alpha_{r,1}(p_r)^k + \beta_1 \alpha_{r,2} k (p_r)^k + \cdots \\ + \beta_{(n(r)-1)} \alpha_{r,n(r)} k^{(n(r)-1)} (p_r)^k + \cdots \end{aligned} \tag{31}$$

where the β_i's are constants corresponding to the numerator weights of z-transforms of the form $z/(z-a)^m$ found in Table 1 scaled by the corresponding α. Assume that $p_r = \sigma_r + j\omega_r$, then $h[k]$ converges asymptotically to zero if $|\sigma_r|^k \to 0$ as $k \to \infty$. The system is *conditionally stable* if $|\sigma_r|^k < V$ as $k \to \infty$. Otherwise the system is *unstable*. Asymptotic stability can be insured if all the poles of $H(z)$ are interior to the unit circle. This gives rise to the so-called *unit-circle criterion* for stability. Since the poles of an LSI system can be easily computed with a modern digital computer, this test is generally considered to be adequate. It should be noted that if a pole is on the unit circle (i.e., $|p_r| = 1$), it must appear with a multiplicity of 1 if the system is to remain conditionally stable. If a conditionally stable system is presented with an input signal at the frequency occupied by the conditionally stable pole, instability will result. In this case the conditionally stable system is resonant and will diverge if driven at its resonate frequency. Finally, if any pole is unstable, the entire system is unstable. If all the poles are stable, but one or more is conditionally stable, the entire system is conditionally stable. In order for the system to be asymptotically stable, all the poles must be asymptotically stable. The establishment of the stability of a nonlinear system is a completely different story and generally requires considerable mathematical sophistication to establish stability. The stability cases are summarized in Table 3. The relationship between pole location and stability case is graphically motivated in Fig. 5.

Example 3. Stability: *Three strictly proper filters are considered having a transfer function $H(z)$, where*

Table 3 Pole Stability Conditions

Stability classification	Pole multiplicity	Pole magnitude $\lvert p_r \rvert$	BIBO stable
Asymptotic	$\le N$	< 1	Yes
Conditional	$= 1$	$= 1$	No
Unstable	> 1	$= 1$	No
Unstable	$\le N$	> 1	No

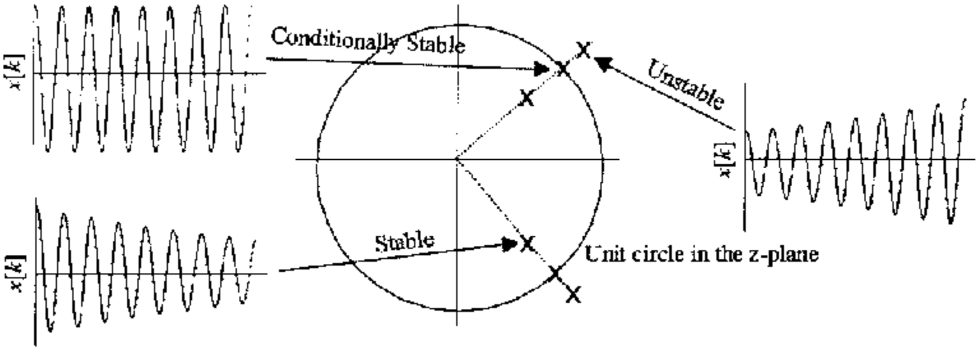


Figure 5 z-Domain relationship to the stability cases: stable (asymptotic), conditional, and unstable system behavior.

$$H(z) = K\frac{\prod_{m=1}^{4}(z-z_m)}{\prod_{m=1}^{5}(z-p_m)}$$

Three possible pole-zero distributions are shown below:

$$H_1(z) = \frac{0.002(z^4+z^3+0.25z^2+0.25z)}{(z^5-3.3z^4+5.04z^3-4.272z^2+2.002z-0.441)}$$

$$H_2(z) = \frac{0.002(z^4+z^3+0.25z^2+0.25z)}{(z^5-3.314z^4+5.086z^3-4.329z^2+2.036z-0.45)}$$

$$H_3(z) = \frac{0.002(z^4+z^3+0.25z^2+0.25z)}{(z^5-3.304z^4+4.886z^3-3.869z^2+1.572z-0.308)}$$

and their pole locations summarized in Table 4 (determined using a general-purpose computer).

The stability classification of the systems immediately follows from the study of the pole locations and, in particular, results in

$H_1(z) =$ asymptotically (BIBO) stable (all poles interior to the unit circle)

$H_1(z) =$ conditionally stable (two poles on the unit circle at $z = -0.707 \pm j0.707$)

$H_1(z) =$ unstable (three poles exterior to the unit circle at $z = -0.767 \pm j0.767$ and 1.07)

4.7 LSI SYSTEM FREQUENCY RESPONSE

If the LSI is asymptotically stable, then it can be assumed that the response to any nonzero initial condition will eventually decay to zero. This gives rise to the concept of *steady-state* analysis which states that any output present as $t \to \infty$ must be due to an external input. If the input to an LSI system is a sinusoidal time-series of the form $x[k] = Ae^{j\omega k}$, then the output of an LSI has a structure given by $y[k] = Ve^{j(\omega k+\phi)}$, which corresponds to a possible amplitude and phase change relative to the input reference. Formally, the steady-state frequency response of an LSI system having a transfer function $H(z)$ to an assumed harmonic input $x[k] = Ae^{j\omega k}$ is given by

$$|H(e^{j\omega})| = |H(z)|_{z=e^{j\omega}} \quad \text{(magnitude frequency response)} \tag{32}$$

$$\phi(e^{j\omega}) = \arg(H(e^{j\omega})) = \arctan\left(\frac{\mathrm{Im}(H(e^{j\omega}))}{\mathrm{Re}(H(e^{j\omega}))}\right) \quad \text{(phase response)} \tag{33}$$

where $\omega \in [-\pi, \pi]$ along the normalized frequency axis. The magnitude frequency response details the frequency selectivity of an LSI. The phase response establishes the amount of phase shift imparted by the discrete-time LSI. If $\phi(e^{j\omega})$ is positive, then the system is called a *lead system*. If negative, the system is called a

Table 4 Stability Summary

Filter	K	z_1	z_2	z_3	z_4	p_1	p_2	P_3	p_4	p_5
$H_1(z)$	0.002	0	-1	$j0.5$	$-j0.5$	$5+j0.5$	$5-j0.5$	$0.7+j0.7$	$0.7-j0.7$	0.9
$H_2(z)$	0.002	0	-1	$j0.5$	$-j0.5$	$5+j0.5$	$5-j0.5$	$0.707+j0.707$	$0.707-j0.707$	0.9
$H_3(z)$	0.002	0	-1	$j0.5$	$-j0.5$	$5+j0.5$	$5-j0.5$	$0.767+j0.767$	$0.767-j0.767$	1.07

lag system. Regardless, a discrete-time LSI system has a steady-state response to a sinusoidal input of frequency ω given by

$$H(e^{j\omega}) = K\frac{\prod_{m=1}^{N}(e^{j\omega} - z_m)}{\prod_{m=1}^{N}(e^{j\omega} - p_m)} = K\frac{\prod_{m=1}^{N}(\alpha_m(j\omega))}{\prod_{m=1}^{N}(\beta_m(j\omega))} \tag{34}$$

where

$$\begin{aligned}\alpha_m(j\omega) &= |\alpha_m(j\omega)|e^{j\phi_m}\\ \beta_m(j\omega) &= |\beta_m(j\omega)|e^{j\theta_m}\end{aligned} \tag{35}$$

Equation (34) can be alternatively expressed as

$$H(e^{j\omega}) = |H(e^{j\omega})| \arg(H(e^{j\omega})) \tag{36}$$

where

$$|H(e^{j\omega})| = K\frac{\prod_{m=1}^{N}|\alpha_m(j\omega)|}{\prod_{m=1}^{N}|\beta_m(j\omega)|} \tag{37}$$

and

$$\arg(H(e^{j\omega})) = \sum_{m=1}^{N}\phi_m - \sum_{m=1}^{N}\theta_m + (0 \text{ if } K > 0 \text{ and } \pi \text{ if } K < 0) \tag{38}$$

Example 4. IIR: *An eighth-order discrete-time filter is designed to meet or exceed the following specifications:*

Sampling frequency = 100 kHz.

Allowable passband deviation = 1 dB, *passband range* $f \in [0, 20]$ kHz.

Minimum stopband attenuation = 60 dB, *stopband range* $f \in [22.5, 50]$ kHz.

Using a commercial filter design package (Monarch), and eighth-order filter was designed which has a 1 dB maximum passband deviation and the minimum stopband attenuation is 69.97. The derived filter satisfied

$$H(z) = 0.00658\frac{z^8 + 1.726z^7 + 3.949z^6 + 4.936z^5 + 5.923z^4 + 4.936z^3 + 3.949z^2 + 1.726z + 1}{z^8 - 3.658z^2 + 7.495z^6 - 11.432z^5 + 11.906z^4 - 8.996z^3 + 4.845z^2 - 1.711z + 0.317}$$

The frequency response of the eighth-order filter is reported in Fig. 6. The magnitude frequency response is seen to exhibit what is considered to have a classic pass- and stopband shape. Observe also that most of the phase variability is concentrated in the pass- and transition-, and early stopband. This is verified by viewing the group delay which indicates that a delay of about 20 samples occurs at a transition band frequency.

4.8 STATE-VARIABLE REPRESENTATION OF LSI SYSTEMS

Many important LSI systems are *single-input–single-output* (SISO) systems which can be modeled as a *monic* Nth-order difference equation

$$\begin{aligned} y[k] + a_1y[k-1] + \cdots + a_Ny[k-N] &= b_0u[k]\\ + b_1u[k-1] + \cdots + b_Nu[k-N] \end{aligned} \tag{39}$$

or as the transfer function $H(z)$:

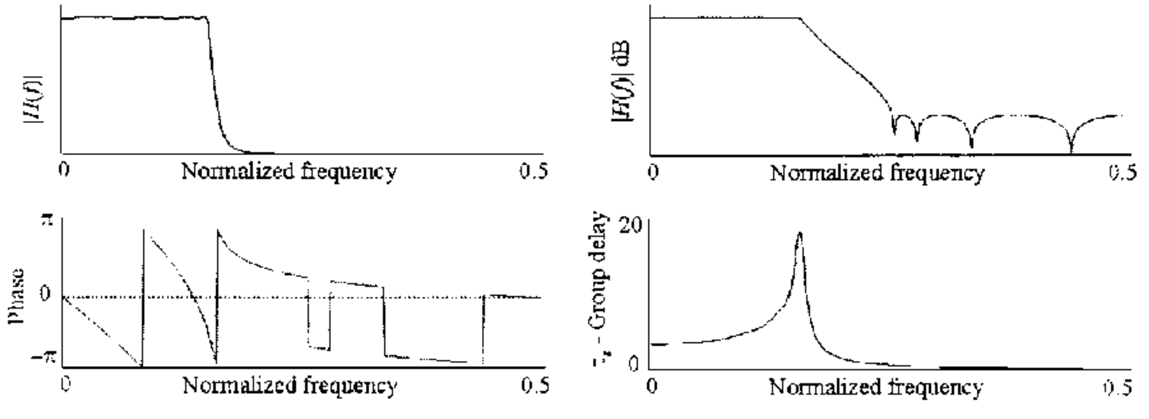


Figure 6 Response of an eight-order filter showing magnitude frequency response in linear and logarithmic (dB) units, phase response, and group delay (phase slope). (Courtesy of the Athena Group, Monarch ® software.)

$$\begin{aligned}H(z) &= \frac{b_0 + b_1 z^{-1} + \cdots + b_N z^{-N}}{1 + a_0 + a_1 z^{-1} + \cdots + a_N z^{-N}} \\ &= b_0 + \frac{(b_1 - b_0 a_1) z^{-1} + \cdots + (b_N - b_0 a_N) z^{-N}}{1 + a_0 + a_1 z^{-1} + \cdots + a_N z^{-N}} \\ &= b_0 + \frac{c_1 z^{-1} + \cdots + c_N z^{-N}}{1 + a_0 + a_1 z^{-1} + \cdots + a_N z^{-N}} \\ &= b_0 + C(z)\left(\frac{1}{D(z)}\right)\end{aligned} \tag{40}$$

The transfer function is seen to consist of three distinct subsystems called

1. A constant gain path (b_0)
2. An all feedforward system denoted $C(z)$
3. An all feedback system $D(z)$.

In general, a discrete-time system, consisting of p-inputs, r-outputs, and N-states, has a *state variable representation* given by

$$\vec{x}[k+1] = \mathbf{A}[k]\vec{x}[k] + \mathbf{B}[k]\vec{u}[k] \qquad \text{(state equation)} \tag{41}$$

$$\vec{x}[0] = x_0 \qquad \text{(initial condition)} \tag{42}$$

$$\vec{y}[k] = \mathbf{C}^T[k]\vec{x}[k] + \mathbf{D}[k]\vec{u}[k] \qquad \text{(output equation)} \tag{43}$$

where $\mathbf{A}[k]$ is an $N \times N$ matrix, $\mathbf{B}[k]$ is an $N \times P$ matrix, $\mathbf{C}[k]$ is an $N \times r$ matrix, and $\mathbf{D}[k]$ is an $R \times P$ matrix, $\vec{u}[k]$ is an arbitrary $P \times 1$ input vector, $\vec{x}[k]$ is an $N \times 1$ state vector, and $\vec{y}[k]$ is an $R \times 1$ output vector. Such a system can also be represented by the four-tuple of matrices and vectors in the form $\mathcal{S} = \{\boldsymbol{A}[k], \boldsymbol{B}[k], \boldsymbol{C}[k], \boldsymbol{D}[k]\}$. If the system is also an LSI system, then the state four-tuple is given by $\mathcal{S} = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}\}$. The state-determined system, described by Eqs. (41), (42), and (43), is graphically interpreted in Fig. 7. The *states* of an LSI system serve as information repositories and are saved in memory and/or shift registers. If an Nth-order system can be implemented with N shift registers, or N states, the system is said to be *canonic*. The states of the system reside at the shift-register locations and contain sufficient information to completely characterize both the solution and the system architecture. Architecture corresponds to the method by which the fundamental building blocks of a sampled-data system are connected (wired) together. The coefficient a_{ij} of $\boldsymbol{A}$ describes the gain of the path connecting the output of shift register j (state $x_j[k]$) with the input to shift register i (state $x_i[k+1]$). Two of the more popular architectures found in common use are the Direct II and cascade.

4.9 DIRECT II ARCHITECTURE

The system characterized by Eq. (40) can be placed into what is called a *Direct II* architectural model shown in Fig. 8. The canonic Direct II state model is defined in terms of an N-dimensional state vector given by

$$\vec{x}[k] = \begin{bmatrix} x_1[k] \\ x_2[k] \\ \vdots \\ x_N[k] \end{bmatrix} = \begin{bmatrix} x[k-N] \\ x[k-N+1] \\ \vdots \\ x[k] \end{bmatrix} \tag{44}$$

and the following state assignments:

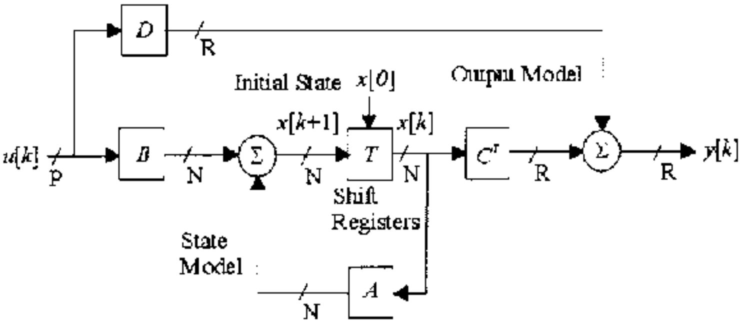


Figure 7 Discrete state-variable system model.

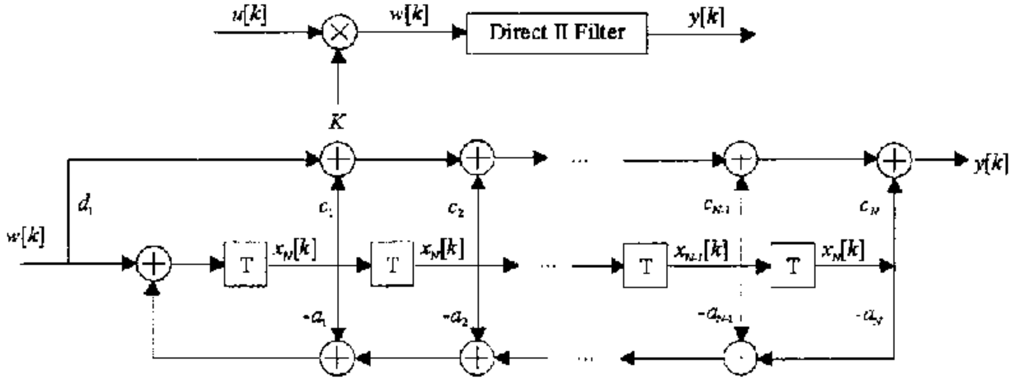


Figure 8 Direct II architecture.

$$\vec{x}[k+1] = \begin{bmatrix} x_1[k+1] \\ x_2[k+1] \\ \vdots \\ x_N[k+1] \end{bmatrix} = \begin{bmatrix} x_2[k] \\ x_3[k] \\ \vdots \\ -a_N x_1[k] - a_{N-1} x_2[k] \cdots - a_2 x_{N-1}[k] - a_1 x_N[k] - u[k] \end{bmatrix} \tag{45}$$

which results in the *state equation* for a Direct II architecture which is given by

$$\vec{x}[k+1] = A\vec{x}[k] + bu[k] \tag{46}$$

Here $\boldsymbol{A}$ is the $N \times N$ coefficient matrix

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_N & -a_{N-1} & -a_{N-2} & \cdots & -a_1 \end{bmatrix} \tag{47}$$

where, again a_{ij} defines the path gain existing between state $x_j[k]$ and $x_i[k+1]$. Continuing, b is a $N \times 1$ vector satisfying

$$b = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \tag{48}$$

It can be immediately seen that Eq. (46) and (47) define the feedback system found in the presence of the $1/D(z)$ term in Eq. (40).

The output, or *output state equation*, is given

$$y[k] = c^T x[k] + d_0 u[k] \tag{49}$$

where c is a $1 \times N$ vector and $\boldsymbol{d}$ is a scalar satisfying

$$c^T = (\, b_N - b_0 a_N \quad b_{N-1} - b_0 a_{N-1} \quad \cdots \quad b_1 - b_0 a_1 \,) \tag{50}$$

$$d_0 = b_0 \tag{51}$$

Example 5. Direct II Architecture: *An eighth-order discrete-time filter has a transfer function given by*

$$H(z) = 0.088 \frac{z^8 + 4.43z^7 + 10.76z^6 + 17.46z^5 + 20.48z^4 + 17.46z^3 + 10.76z^2 + 4.43z + 1}{z^8 + 1.10z^7 + 1.97^6 + 1.55z^5 + 1.22z^4 + +0.61z^3 + 0.24z^2 + 0.061z + 0.008}$$

A commercial CAD tool (Monarch) was used to convert the transfer function into the Direct II model shown on page 265.

4.10 CASCADE ARCHITECTURE

A cascade architecture is shown in Fig. 9 and implements the transfer function factored as

$$H(z) = K \prod_{i=1}^{Q} H_i(z) \tag{52}$$

where $H_i(z)$ is a first- or second-order subsystem defined with respect to real coefficients. Each subsystem is represented by a state-determined model $\mathcal{S}_i = (\boldsymbol{A}_i, \boldsymbol{b}_i, \boldsymbol{c}_i, \boldsymbol{d}_i)$ and

$$\sum_{i=1}^{Q} \text{order}(H_i(z)) = N \tag{53}$$

A cascade architecture, as the name implies, links the output of one subsystem to the input of its succes-

```
DIRECT II STATE-VARIABLE FILTER DESCRIPTION
        Scale Factor=0.08883285197457290
        A Matrix
        A[1, i]; i ∈ [0, 8]
         0.000000000000000            1.000000000000000
         0.000000000000000            0.000000000000000
         0.000000000000000            0.000000000000000
         0.000000000000000            0.000000000000000
        A[2, i]; i ∈ [0, 8]
         0.000000000000000            0.000000000000000
         1.000000000000000            0.000000000000000
         0.000000000000000            0.000000000000000
         0.000000000000000            0.000000000000000
        A[3, i]; i ∈ [0, 8]
         0.000000000000000            0.000000000000000
         0.000000000000000            1.000000000000000
         0.000000000000000            0.000000000000000
         0.000000000000000            0.000000000000000
        A[4, i]; i ∈ [0, 8]
         0.000000000000000            0.000000000000000
         0.000000000000000            0.000000000000000
         1.000000000000000            0.000000000000000
         0.000000000000000            0.000000000000000
        A[5, i]; i ∈ [0, 8]
         0.000000000000000            0.000000000000000
         0.000000000000000            0.000000000000000
         0.000000000000000            1.000000000000000
         0.000000000000000            0.000000000000000
        A[6, i]; i ∈ [0, 8]
         0.000000000000000            0.000000000000000
         0.000000000000000            0.000000000000000
         0.000000000000000            0.000000000000000
         1.000000000000000            0.000000000000000
        A[7, i]; i ∈ [0, 8]
         0.000000000000000            0.000000000000000
         0.000000000000000            0.000000000000000
         0.000000000000000            0.000000000000000
         0.000000000000000            1.000000000000000
        A[8, 8]; i ∈ [0, 8]
        −0.007910400932499942        −0.06099774584323624
        −0.2446494077658335          −0.616051520514172
        −1.226408547493966           −1.556364236494628
        −1.978668209561079           −1.104614299236229

        B Vector                       C′ Vector              D Scalar
         0.000000000000000            0.9920989599067499      1.000000000000000
         0.000000000000000            4.376921859721373
         0.000000000000000           10.52456679079099
         0.000000000000000           16.85365679994142
         0.000000000000000           19.17645033204060
         0.000000000000000           15.91334407549821
         0.000000000000000            8.790547988995748
         1.000000000000000            3.333305306328382
```

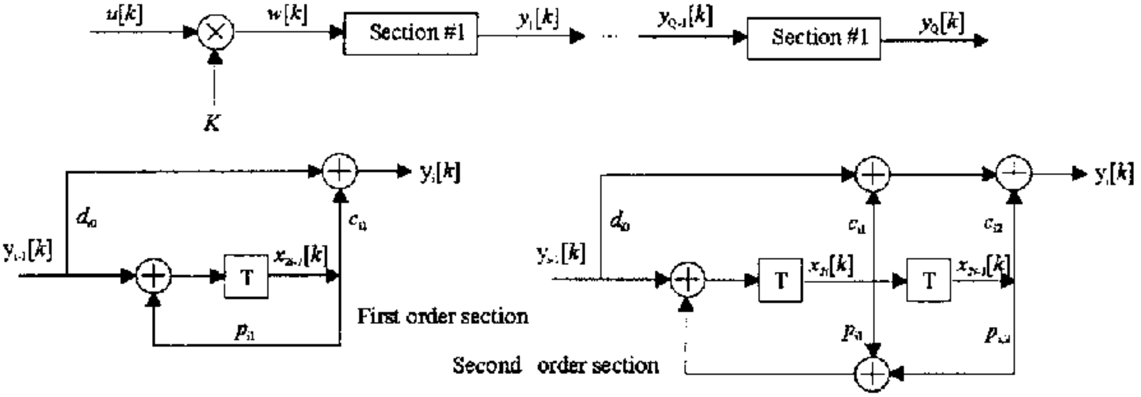


Figure 9 Cascade architecture.

sor. Specifically, $\mathcal{S}_i = (A_i, b_i, c_i, d_i)$ and $\mathcal{S}_{i=1} = (A_{i+1}, b_{i+1}, c_{i+1}, d_{i+1})$ can be chained together by mapping the $y_i[k]$ (output of $\mathcal{S}_i$) to $u_{i+1}[k]$ (input of $\mathcal{S}_{i+1}$). Following this procedure the state-variable model for a cascade system, given by $\mathcal{S} = (A, b, c, d)$ where

$$A = \begin{pmatrix} A_1 & 0 & 0 & \cdots & 0 \\ b_2 c_1^T & A_2 & 0 & \cdots & 0 \\ b_3 d_2 c_1^T & b_3 c_1^T & A_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_Q(d_{Q-1}d_{Q-2}\cdots d_2)c_1^T & b_Q(d_{Q-1}d_{Q-2}\cdots d_3)c_2^T & b_Q(d_{Q-1}d_{Q-2}\cdots d_4)c_3^T & \cdots & A_Q \end{pmatrix} \tag{54}$$

$$b = \begin{pmatrix} b_1 \\ d_1 b_2 \\ \vdots \\ (d_{Q-1}\cdots d_1)b_Q \end{pmatrix} \tag{55}$$

$$c = \begin{pmatrix} (d_q d_{Q-1}\cdots d_2)c_1 \\ (d_Q d_{Q-1}\cdots d_3)c_2 \\ \vdots \\ c_Q \end{pmatrix} \tag{56}$$

$$d = d_Q d_{Q-1}\cdots d_1 d_1 \tag{57}$$

The elements of $\boldsymbol{A}$ having indices a_{ij}, for $i+2=j$, correspond to the coupling of information from $\mathcal{S}_i$ into $\mathcal{S}_k$ where $k > i$. It can also be seen that the construction rules for a cascade design are also very straightforward. A cascade implementation of an Nth-order system can also be seen to require at most N multiplications from $\boldsymbol{A}$ (the terms a_{ij}, for $i+2=j$, are not physical multiplications), N from $\boldsymbol{b}$ and $\boldsymbol{c}$, and one from $\boldsymbol{d}$ for a total complexity measure of $M_{\text{multiplier}} \leq 3N+1$, which is larger than computed for a Direct II filter. In practice, however, many Cascade coefficients are of unit value which will often reduce the complexity of this architecture to a level similar to that of a Direct II.

Example 6. Cascade Architecture

Problem statement. *Implement the eighth-order discrete-time studied in the Direct II example. Using a commercial CAD tool (Monarch) the following Cascade architecture was synthesized. The state-variable model presented over was produced using the Cascade architecture option (see page 267). The system is reported in terms of the state model for each second-order subfilter as well as the overall system.*

4.11 SUMMARY

Sampled-data systems, per se, are of diminishing importance compared to the rapidly expanding field of *digital signal processing* or *DSP* (see Chap. 3.3). The limiting factor which has impeded the development of sampled-data systems on a commercial scale has been technological. The basic building blocks of a

```
CASCADE STATE-VARIABLE FILTER DESCRIPTION
        Scale Factor = 0.08883285197457290
        A Matrix
        A[1, i]; i ∈ [0, 8]
         0.000000000000000           1.000000000000000
         0.000000000000000           0.000000000000000
         0.000000000000000           0.000000000000000
         0.000000000000000           0.000000000000000
        A[2, i]; i ∈ [0, 8]
        −0.09123912551991054           −0.5174788167607780
         0.000000000000000           0.000000000000000
         0.000000000000000           0.000000000000000
         0.000000000000000           0.000000000000000
        A[3, i]; i ∈ [0.8]
         0.000000000000000          −0.000000000000000
         0.000000000000000           1.000000000000000
         0.000000000000000           0.000000000000000
         0.000000000000000           0.000000000000000
        A[4, i]; i ∈ [0.8]
         0.9087608744800895         −0.06330133197939697
        −0.2311930335002332         −0.3741770017630547
         0.000000000000000           0.000000000000000
         0.000000000000000           0.000000000000000
        A[5, i] : i ∈ [0.8]
         0.000000000000000          −0.000000000000000
         0.000000000000000           0.000000000000000
         0.000000000000000           1.000000000000000
         0.000000000000000           0.000000000000000
        A[6, i]; i ∈ [0.8]
         0.9087608744800895         −0.06330133197939697
         0.7688069664997668          0.3793401006819637
        −0.4720529007738288         −0.1778396572194128
         0.000000000000000           0.000000000000000
        A[7, i]; i ∈ [0, 8]
         0.000000000000000          −0.000000000000000
         0.000000000000000           0.000000000000000
         0.000000000000000           0.000000000000000
         0.000000000000000           1.000000000000000
        A[8, i]; i ∈ [0, 8]
         0.9087608744800895         −0.6330133197939697
         0.7688069664997668          0.3793401006819637
         0.5279470992261710          1.149648111678130
        −0.7944232896725211         −0.03511882349298313
        B Vector                      C′ Vector                D Scalar
         0.000000000000000           0.9087608744800895       1.000000000000000
         1.000000000000000          −0.06330133197939697
         0.000000000000000           0.7688069664997668
         1.000000000000000           0.3793401006819637
         0.000000000000000           0.5279470992261710
         1.000000000000000           1.149648111678130
         0.000000000000000           0.2055767103274789
         1.000000000000000           1.867618425947684
```

sampled-data system would include samplers, multipliers, adders, and delays. Of this list, analog delays are by far the msot difficult to implement in hardware. Digital systems, however, are designed using ADCs, multipliers, adders, and delays. Delays in a digital technology are nothing more than clocked shift registers of digital memory. These devices are inexpensive and highly accurate. As a result, systems which are candidates for sampled-data implementation are, in a contemporary setting, implemented using DSP techniques and technology.

BIBLIOGRAPHY

Antonious A. Digital Filters: Analysis and Design. New York: McGraw-Hill, 1979.

Blahut R. Fast Algorithms for Digital Signal Processing. Reading, MA: Addison-Wesley, 1985.

Brigham EO. The Fast Fourier Transform and Its Application. New York: McGraw-Hill, 1988.

Oppenheim AV, ed. Application of Digital Signal Processing. Englewood Cliffs, NJ: Prentice-Hall, 1978.

Oppenheim AV, Schafer R. Digital Signal Processing. Englewood Cliffs, NJ: Prentice-Hall, 1975.

Proakis J, Manolakis DG. Introduction to Digital Signal Processing. New York: Macmillan, 1988.

Rabiner LR, Gold B. Theory and Applications of Digital Signal Processing. Englewood Cliffs, NJ: Prentice-Hall, 1975.

Taylor F. Digital Filter Design Handbook. New York: Marcel Dekker, 1983.

Zelniker G, Taylor F. Advanced Digital Signal Processing: Theory and Applications. New York: Marcel Dekker, 1994.

Chapter 4.1

Regression

Richard Brook
Off Campus Ltd., Palmerston North, New Zealand
Denny Meyer
Massey University–Albany, Palmerston North, New Zealand

1.1 FITTING A MODEL TO DATA

1.1.1 What is Regression?

1.1.1.1 Historical Note

Regression is, arguably, the most commonly used technique in applied statistics. It can be used with data that are collected in a very structured way, such as sample surveys or experiments, but it can also be applied to observational data. This flexibility is its strength but also its weakness, if used in an unthinking manner.

The history of the method can be traced to Sir Francis Galton who published in 1885 a paper with the title, "Regression toward mediocrity in hereditary stature." In essence, he measured the heights of parents and found the median height of each mother–father pair and compared these medians with the height of their adult offspring. He concluded that those with very tall parents were generally taller than average but were not as tall as the median height of their parents; those with short parents tended to be below average height but were not as short as the median height of their parents. Female offspring were combined with males by multiplying female heights by a factor of 1.08.

Regression can be used to explain relationships or to predict outcomes. In Galton's data, the median height of parents is the explanatory or predictor variable, which we denote by X, while the response or predicted variable is the height of the offspring, denoted by Y. While the individual value of Y cannot be forecast exactly, the average value can be for a given value of the explanatory variable, X.

1.1.1.2 Brief Overview

Uppermost in the minds of the authors of this chapter is the desire to relate some basic theory to the application and practice of regression. In Sec 1.1, we set out some terminology and basic theory. Section 1.2 examines statistics and graphs to explore how well the regression model fits the data. Section 1.3 concentrates on variables and how to select a small but effective model. Section 1.4 looks to individual data points and seeks out peculiar observations.

We will attempt to relate the discussion to some data sets which are shown in Sec 1.5. Note that data may have many different forms and the questions asked of the data will vary considerably from one application to another. The variety of types of data is evident from the description of some of these data sets.

Example 1. Pairs (Triplets, etc.) of Variables (Sec. 1.5.1): *The Y-variable in this example is the heat developed in mixing the components of certain cements which have varying amounts of four X-variables or chemicals in the mixture. There is no information about how the various amounts of the X-variables have been chosen. All variables are continuous variables.*

Example 2. Grouping Variables (Sec. 1.5.2): *Qualitative variables are introduced to indicate groups allocated to different safety programs. These qualitative variables differ from other variables in that they only take the values of 0 or 1.*

Example 3. A Designed Experiment (Sec. 1.5.3): *In this example, the values of the X-variables have been set in advance as the design of the study is structured as a three-factor composite experimental design. The X-variables form a pattern chosen to ensure that they are uncorrelated.*

1.1.1.3 What Is a Statistical Model?

A statistical model is an abstraction from the actual data and refers to all possible values of Y in the population and the relationship between Y and the corresponding X in the model. In practice, we only have sample values, y and x, so that we can only check to ascertain whether the model is a reasonable fit to these data values.

In some area of science, there are laws such as the relationship $e = mc^2$ in which it is assumed that the model is an exact relationship. In other words, this law is a deterministic model in which there is no error. In statistical models, we assume that the model is stochastic, by which we mean that there is an error term, e, so that the model can be written as

$$Y = f(X = x) + e$$

In a regression model, $f(.)$ indicates a linear function of the X-terms. The error term is assumed to be random with a mean of zero and a variance which is constant, that is, it does not depend on the value taken by the X-term. It may reflect error in the measurement of the Y-variable or by variables or conditions not defined in the model. The X-variable, on the other hand, is assumed to be measured without error.

In Galton's data on heights of parents and offspring, the error term may be due to measurement error in obtaining the heights or the natural variation that is likely to occur in the physical attributes of offspring compared with their parents.

There is a saying that "No model is correct but some are useful." In other words, no model will exactly capture all the peculiarities of a data set but some models will fit better than others.

1.1.2 How to Fit a Model

1.1.2.1 Least-Squares Method

We consider Example 1, but concentrate on the effect of the first variable, x_1, which is tricalcium aluminate, on the response variable, which is the heat generated. The plot of heat on tricalcium aluminate, with the least-squares regression line, is shown in Fig. 1. The least-squares line is shown by the solid line and can be written as

$$\hat{y} = f(X = x_1) = a + bx_1 = 81.5 + 1.87x_1 \qquad (1)$$

where $\hat{y}$ is the predicted value of y for the given value x_1 of the variable X_1.

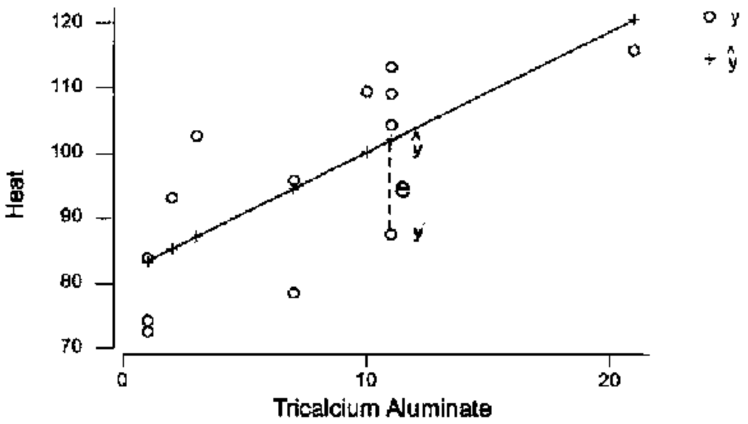


Figure 1 Plot of heat, y, on tricalcium aluminate, x_1.

All the points represented by (x_1, y) do not fall on the line but are scattered about it. The vertical distance between each observation, y, and its respective predicted value, $\hat{y}$, is called the residual, which we denote by e. The residual is positive if the observed value of y falls above the line and negative if below it. Notice in Sec. 1.5.1 that for the fourth row in the table, the fitted value is 102.04 and the residual (shown by e in Fig. 1) is -14.44, which corresponds to one of the four points below the regression line, namely the point $(x_1, y) = (11, 87.6)$.

At each of the x_1 values in the data set we assume that the population values of Y can be written as a linear model, by which we mean that the model is linear in the parameters. For convenience, we drop the subscript in the following discussion.

$$Y = \alpha + \beta x + \varepsilon \tag{2}$$

More correctly, Y should be written as $Y \mid x$, which is read as "Y given $X = x$."

Notice that a model, in this case a regression model, is a hypothetical device which explains relationships in the population for all possible values of Y for given values of X. The error (or deviation) term, ε, is assumed to have for each point in the sample a population mean of zero and a constant variance of σ^2 so that for $X =$ a particular value x, Y has the following distribution:

$Y \mid x$ is distributed with mean $(\alpha + \beta x)$ and variance σ^2

It is also assumed that for any two points in the sample, i and j, the deviations ε_i and ε_j are uncorrelated.

The method of least squares uses the sample of n ($= 13$ here) values of x and y to find the least-squares estimates, a and b, of the population parameters α and β by minimizing the deviations. More specifically, we seek to minimize the sum of squares of e, which we denote by S^2, which can be written as

$$S^2 = \sum e^2 = \sum [y - f(x)]^2 = \sum [y - (a + bx)]^2 \tag{3}$$

The symbol $\sum$ indicates the summation over the $n = 13$ points in the sample.

1.1.2.2 Normal Equations

The values of the coefficients a and b which minimize S^2 can be found by solving the following, which are called normal equations. We do not prove this statement but the reader may refer to a textbook on regression, such as Brook and Arnold [1].

$$\begin{aligned} &\sum [y - (a + bx)] = 0 \quad \text{or} \quad na + b\sum x = \sum y \\ &\sum x[y - (a + bx)] = 0 \quad \text{or} \\ &a\sum x + b\sum x^2 = \sum xy \end{aligned} \tag{4}$$

By simple arithmetic, the solutions of these normal equations are

$$\begin{aligned} a &= \bar{y} - b\bar{x} \\ b &= \left[\sum (x - \bar{x})(y - \bar{y})\right] \Big/ \sum (x - x)^2 \end{aligned} \tag{5}$$

Note:

1. The mean of y is $\sum y/n$, or $\bar{y}$. Likewise the mean of x is $\bar{x}$.
2. b can be written as S_{xy}/S_{xx}, which can be called the sum of cross-products of x and y divided by the sum of squares of x.
3. From Sec. 1.5.1, we see that the mean of x is 7.5 and of y is 95.4.

The normal equations become

$$\begin{aligned} 13a + 97b &= 1240.5 \\ 97a + 1139b &= 10{,}032 \end{aligned} \tag{6}$$

Simple arithmetic gives the solutions as $a = 81.5$ and $b = 1.87$.

1.1.3 Simple Transformations

1.1.3.1 Scaling

The size of the coefficients in a fitted model will depend on the scales of the variables, predicted and predictor. In the cement example, the X variables are measured in grams. Clearly, if these variables were changed to kilograms, the values of the X would be divided by 1000 and, consequently, the sizes of the least squares coefficients would be multiplied by 1000. In this example, the coefficients would be large and it would be clumsy to use such a transformation.

In some examples, it is not clear what scales should be used. To measure the consumption of petrol (gas), it is usual to quote the number of miles per gallon, but for those countries which use the metric system, it is the inverse which is often quoted, namely the number of liters per 100 km travelled.

1.1.3.2 Centering of Data

In some situations, it may be an advantage to change x to its deviation from its mean, that is, $x - \bar{x}$. The fitted equation becomes

$$\hat{y} = a + b(x - \bar{x})$$

but these values of x and b may differ from Eq. (1). Notice that the sum of the $(x - \bar{x})$ terms is zero as

$$\sum(x - \bar{x}) = \sum x - \sum \bar{x} = n\bar{x} - n\bar{x} = 0$$

The normal equations become, following Eq. (4),

$$\begin{aligned} na + 0 &= \sum y \\ 0 + b\sum(x - \bar{x})^2 &= \sum(x - \bar{x})y \end{aligned} \tag{7}$$

Thus,

$$a = \sum y/n = \bar{y}$$

which differs somewhat from Eq. (5), but

$$b = \left[\sum(x - \bar{x})y\right] \Big/ \sum(x - \bar{x})^2$$

which can be shown to be the same as in Eq. (5).

The fitted line is

$$\hat{y} = 95.42 + 1.87(x - \bar{x})$$

If the y variable is also centered and the two centered variables are denoted by y and x, the fitted line is

$$y = 1.87x$$

The important point of this section is that the inclusion of a constant term in the model leads to the same coefficient of the X term as transforming X to be centered about its mean. In practice, we do not need to perform this transformation of centering as the inclusion of a constant term in the model leads to the same estimated coefficient for the X variable.

1.1.4 Correlations

Readers will be familiar with the correlation coefficient between two variables. In particular the correlation between y and x is given by

$$r_{xy} = S_{xy}/\sqrt{(S_{xx}S_{yy})} \tag{8}$$

There is a duality in this formula in that interchanging x and y would not change the value of r. The relationship between correlation and regression is that the coefficient b in the simple regression line above can be written as

$$b = r\sqrt{S_{yy}/S_{xx}} \tag{9}$$

In regression, the duality of x and y does not hold. A regression line of y on x will differ from a regression line of x and y.

1.1.5 Vectors

1.1.5.1 Vector Notation

The data for the cement example (Sec. 1.5) appear as equal-length columns. This is typical of data sets in regression analysis. Each column could be considered as a column vector with 13 components. We focus on the three variables $\mathbf{y}$ (heat generated), $\hat{\mathbf{y}}$ (FITS1 = predicted values of $\mathbf{y}$), and $\mathbf{e}$ (RESI1 = residuals).

Notice that we represent a vector by bold types: $\mathbf{y}$, $\hat{\mathbf{y}}$, and $\mathbf{e}$.

The vectors simplify the columns of data to two aspects, the lengths and directions of the vectors and, hence, the angles between them. The length of a vector can be found by the inner, or scalar, product. The reader will recall that the inner product of $\mathbf{y}$ is represented as $\mathbf{y} \cdot \mathbf{y}$ or $\mathbf{y}^{\mathrm{T}}\mathbf{y}$, which is simply the sum of the squares of the individual elements.

Of more interest is the inner product of $\hat{\mathbf{y}}$ with $\mathbf{e}$, which can be shown to be zero. These two vectors are said to be orthogonal or "at right angles" as indicated in Fig. 2.

We will not go into many details about the geometry of the vectors, but it is usual to talk of $\hat{\mathbf{y}}$ being the projection of $\mathbf{y}$ in the direction of $\mathbf{x}$. Similarly, $\mathbf{e}$ is the projection of $\mathbf{y}$ in a direction orthogonal to $\mathbf{x}$, orthogonal being a generalization to many dimensions of "at right angles to," which becomes clear when the angle ϕ is considered.

Notice that $\mathbf{e}$ and $\hat{\mathbf{y}}$ are "at right angles" or "orthogonal." It can be shown that a necessary and sufficient condition for this to be true is that $\mathbf{e}^{\mathrm{T}}\hat{\mathbf{y}} = 0$.

In vector terms, the predicted value of $\mathbf{y}$ is

$$\hat{\mathbf{y}} = a\mathbf{1} + \mathbf{bx}$$

and the fitted model is

$$\mathbf{y} = a\mathbf{1} + \mathbf{bx} + \mathbf{e} \tag{10}$$

Writing the constant term as a column vector of '1's pave the way for the introduction of matrices in Sec. 1.1.7.

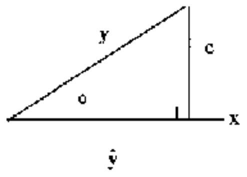


Figure 2 Relationship between $\mathbf{y}$, $\hat{\mathbf{y}}$ and $\mathbf{e}$.

1.1.5.2 Vectors—Centering and Correlations

In this section, we write the vector terms in such a way that the components are deviations from the mean; we have

$$\hat{\mathbf{y}} = b\mathbf{x}$$

The sums of squares of $\mathbf{y}$, $\hat{\mathbf{y}}$, and $\mathbf{e}$ are

$$\begin{aligned}\mathbf{y}^{\mathrm{T}}\mathbf{y} = {} & S_{yy} = (78.5 - 95.42)^2 + (74.3 - 95.42)^2 + \cdots \\ & + (109.4 - 95.42)^2 = 2715.8\end{aligned}$$

$$\hat{\mathbf{y}}^{\mathrm{T}}\hat{\mathbf{y}} = S_{\hat{y}\hat{y}} = 1450.1 \qquad \mathbf{e}^{\mathrm{T}}\mathbf{e} = S_{ee} = 1265.7$$

As we would expect from a right-angled triangle and Pythagoras' theorem,

$$\mathbf{y}^{\mathrm{T}}\mathbf{y} = \hat{\mathbf{y}}^{\mathrm{T}}\hat{\mathbf{y}} + \mathbf{e}^{\mathrm{T}}\mathbf{e}$$

We discuss this further in Sec. 1.2.1.5 on ANOVA, the analysis of variance.

The length of the vector $\mathbf{y}$, written as $|\mathbf{y}|$, is the square root of $(\mathbf{y}^{\mathrm{T}}\mathbf{y}) = 52.11$. Similarly the lengths of $\hat{\mathbf{y}}$ and $\mathbf{e}$ are 38.08 and 35.57, respectively.

The inner product of $\mathbf{y}$ with the vector of fitted values, $\hat{\mathbf{y}}$, is

$$\mathbf{y}^{\mathrm{T}}\hat{\mathbf{y}} = \sum y_i \hat{y}_i = 1450.08$$

The angle ϕ in Fig. 2 has a cosine given by

$$\cos\phi = \mathbf{y}^{\mathrm{T}}\hat{\mathbf{y}}/(|\mathbf{y}||\hat{\mathbf{y}}|) = \sqrt{(1450.1/2715.8)} = 0.73 \tag{11}$$

As $\mathbf{y}$ and $\mathbf{x}$ are centered, the correlation coefficient of y on x can be shown to be $\cos\phi$.

1.1.6 Residuals and Fits

We return to the actual values of the X and Y variables, not the centered values as above. Figure 2 provides more insight into the normal equations, as the least-squares solution to the normal equation occurs when the vector of residuals is orthogonal to the vector of predicted values. Notice that $\hat{\mathbf{y}}^{\mathrm{T}}\mathbf{e} = 0$ can be expanded to

$$(a\mathbf{1} + b\mathbf{x})^{\mathrm{T}}\mathbf{e} = a\mathbf{1}^{\mathrm{T}}\mathbf{e} + b\mathbf{x}^{\mathrm{T}}\mathbf{e} = 0 \tag{12}$$

This condition will be true if each of the two parts are equal to zero, which leads to the normal equations, Eq. (4), above.

Notice that the last column of Sec. 1.5.1 confirms that the sum of the residuals is zero. It can be shown that the corollary of this is that the sum of the observed y is the same as the sum of the fitted y values; if the sums are equal the means are equal and Section 1.5.1 shows that they are both 95.4.

The second normal equation in Eq. (4) could be checked by multiplying the components of the two columns marked x_1 and RESI1 and then adding the result.

In Fig. 1.3, we would expect the residuals to approximately fall into a horizontal band on either side of the zero line. If the data satisfy the assumptions, we would expect that there would not be any systematic trend in the residuals. At times, our eyes may deceive us into thinking there is such a trend when in fact there is not one. We pick this topic up again later.

1.1.7 Adding a Variable

1.1.7.1 Two-Predictor Model

We consider the effect of adding the second term to the model:

$$Y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

The fitted regression equation becomes

$$y = b_0 x_0 + b_1 x_1 + b_2 x_2 + e$$

To distinguish between the variables, subscripts have been reintroduced. The constant term has been written as $b_0 x_0$ and without loss of generality, $x_0 = 1$.

The normal equations follow a similar pattern to those indicated by Eq. (4), namely,

$$\begin{aligned}\sum[b_0 + b_1 x_1 + b_2 x_2] &= \sum y \\ \sum x_1[b_0 + b_1 x_1 + b_2 x_2] &= \sum x_1 y \\ \sum x_2[b_0 + b_1 x_1 + b_2 x_2] &= \sum x_2 y\end{aligned} \tag{13}$$

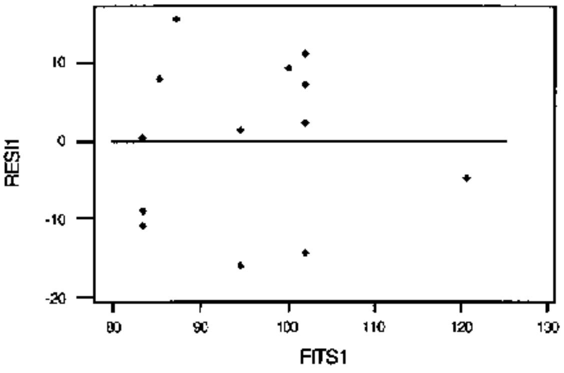


Figure 3 Plot of residuals against fitted values for y on x_1.

These yield

$$\begin{aligned}\mathbf{13b_0} + \mathbf{97b_1} + 626b_2 &= \mathbf{1240.5}\\ \mathbf{97b_0} + \mathbf{1139b_1} + 4922b_2 &= \mathbf{10{,}032}\\ 626b_0 + 4922b_1 + 33{,}050b_2 &= 62{,}027.8\end{aligned} \tag{14}$$

Note that the entries in bold type are the same as those in the normal equations of the model with one predictor variable. It is clear that the solutions for b_0 and b_1 will differ from those of a and b in the normal equations, Eq. (6). It can be shown that the solutions are: $b_0 = 52.6$, $b_1 = 1.47$, and $b_2 = 0.622$.

Note:

1. By adding the second prediction variable x_2, the coefficient for the constant term has changed from $a = 81.5$ to $b_0 = 52.6$. Likewise the coefficient for x has changed from 1.87 to 1.47. The structure of the normal equations give some indication why this is so.
2. The coefficients would not change in value if the variables were orthogonal to each other. For example, if x_0 was orthogonal to x_2, $\sum x_0x_2$ would be zero. This would occur if x_2 was in the form of deviation from its mean. Likewise, if x_1 and x_2 were orthogonal, $\sum x_1x_2$ would be zero.
3. What is the meaning of the coefficients, for example b_1? From the fitted regression equation, one is tempted to say that "b_1 is the increase in y when x_1 increases by 1." From 2, we have to add to this, the words "in the presence of the other variables in the model." Hence, if you change the variables, the meaning of b_1 also changes.

When other variables are added to the model, the formulas for the coefficients become very clumsy and it is much easier to extend the notation of vectors to that of matrices. Matrices provide a clear, generic approach to the problem.

1.1.7.2 Vectors and Matrices

As an illustration, we use the cement data in which there are four predictor variables. The model is

$$y = \beta_0x_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \varepsilon$$

The fitted regression equation can be written in vector notation,

$$\mathbf{y} = b_0\mathbf{x}_0 + b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + b_3\mathbf{x}_3 + b_4\mathbf{x}_4 + \mathbf{e} \tag{15}$$

The data are displayed in Sec. 1.5.1. Notice that each column vector has $n = 13$ entries and there are $k = 5$ vectors. As blocks of five vectors, the predictors can be written as an $n \times k = 13 \times 5$ matrix, $\mathbf{X}$.

The fitted regression equation is

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} \tag{16}$$

It can be shown that the normal equations are

$$\mathrm{X^TX}b = \mathrm{X^Ty} \tag{17}$$

Expanded in vector terms,

$$\begin{aligned}\mathbf{x}_0^T\mathbf{x}_0b_0 + \mathbf{x}_0^T\mathbf{x}_1b_1 + \cdots + \mathbf{x}_0^T\mathbf{x}_4b_4 &= \mathbf{x}_0^T\mathbf{y}\\ \mathbf{x}_1^T\mathbf{x}_0b_0 + \mathbf{x}_1^T\mathbf{x}_1b_1 + \cdots + \mathbf{x}_1^T\mathbf{x}_4b_4 &= \mathbf{x}_1^T\mathbf{y}\\ \mathbf{x}_4^T\mathbf{x}_0b_0 + \mathbf{x}_4^T\mathbf{x}_1b_1 + \cdots + \mathbf{x}_4^T\mathbf{x}_4b_4 &= \mathbf{x}_4^T\mathbf{y}\end{aligned}$$

These yield the normal equations

$$13b_0 + 97b_1 + 626b_2 + 153b_3 + 39064b_4 = 1240.5$$

$$97b_0 + 1130b_1 + 4922b_2 + 769b_3 + 2620b_4 = 10{,}032$$

$$626b_0 + 4922b_1 + 33050b_2 + 7201b_3 + 15739b_4 = 62{,}027.8$$

$$153b_0 + 769b_1 + 7201b_2 + 2293b_3 + 4628b_4 = 13{,}981.5$$

$$39{,}064b_0 + 2620b_1 + 15{,}739b_2 + 4628b_3 + 15{,}062b_4 = 34{,}733.3$$

Notice the symmetry in the coefficients of the b_i.

The matrix solution is

$$\begin{aligned}\mathrm{b} &= (\mathrm{X^TX})^{-1}\mathrm{X^TY}\\ \mathrm{b^T} &= (62.4, 1.55, 0.510, 0.102, -0.144)\end{aligned} \tag{18}$$

With the solution to the normal equations written as above, it is easy to see that the least-squares estimates of the parameters are weighted means of all the y values in the data. The estimates can be written as

$$b_i = \sum w_iy_i$$

where the weights w_i are functions of the x values.

The regression coefficients reflect the strengths and weaknesses of means. The strengths are that each point in the data set contributes to each estimate but the weaknesses are that one or two unusual values in the data set can have a disproportionate effect on the resulting estimates.

1.1.7.3 The Projection Matrix, P

From the matrix solution, the fitted regression equation becomes

$$\hat{y} = \mathrm{xb} = \mathrm{x}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{y} \qquad \text{or} \qquad \mathrm{Py} \tag{19}$$

$\mathrm{P} = \mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T$ is called the projection matrix and it has some nice properties, namely

1. $\mathrm{P}^T = \mathrm{P}$ that is, it is symmetrical.
2. $\mathrm{P}^T\mathrm{P} = \mathrm{P}$ that is, it is idempotent.
3. The residual vector $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (I - P)\mathbf{y}$.
   I is the identity matrix with diagonal elements being 1 and the off-diagonal elements being 0.
4. From the triangle diagram, **e** is orthogonal to $\hat{\mathbf{y}}$, which is easy to see as

   $$\mathbf{e}^T\hat{\mathbf{y}} = \mathbf{y}^T(\mathrm{I} - \mathrm{P}^T)\mathrm{P}\mathbf{y} = \mathbf{y}^\mathrm{T}(\mathrm{P} - \mathrm{P}^\mathrm{T}\mathrm{P})\mathbf{y} = 0$$

5. P is the projection matrix onto X and $\hat{\mathbf{y}}$ is the projection of **y** onto X.
6. I − P is the projection matrix orthogonal to X and the residual, 1, is the projection of y onto a direction orthogonal to X.
   The vector diagram of Fig. 2 becomes Fig. 4.

1.1.8 Normality

1.1.8.1 Assumptions about the Models

In the discussion so far, we have seen some of the relationships and estimates which result from the least-squares method which are dependent on assumptions about the error, or deviation, term in the model. We now add a further restriction to these assumptions, namely that the error term, e, is distributed normally. This allows us to find the distribution of the residuals and find confidence intervals for certain estimates and carry out hypothesis tests on them.

The addition of the assumption of normality adds to the concept of correlation as a zero correlation coefficient between two variables will mean that they are statistically independent.

1.1.8.2 Distributions of Statistics

The variance of the constant term is

$$\mathrm{Var}\, b_0 = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

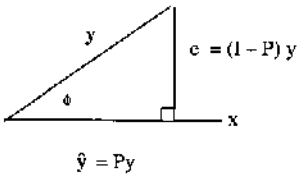


Figure 4 Projections of **y** in terms of **P**.

and the variance of the coefficient of the x variable is

$$\mathrm{Var}\, b_1 = \sigma^2/S_{xx} \tag{20}$$

We are usually more interested in the coefficient of the x term. The confidence interval (CI) for this coefficient (β_1) is given by

$$\mathrm{CI} = b_1 \pm t_{n-2}\sqrt{s^2/S_{xx}} \tag{21}$$

1.1.8.3 Confidence Interval for the Mean

The 95% confidence interval for the predicted value, $\hat{y}$, when $x = x_0$ is given by

$$\hat{y}_0 \pm t_{n-2}s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \tag{22}$$

Note that the width of the confidence interval is smallest when the chosen x_0 is close to the mean, $\bar{x}$, but the width diverges the further the x_0 is from the mean. A more important point is the danger of extrapolating outside of the range of values of X as the model may not be appropriate outside these limits.

This confidence interval is illustrated in Fig. 5 using the cement data.

1.1.8.4 Prediction Interval for a Future Value

At times one wants to forecast the value of y for a given single future value x_0 of x. This prediction interval for a future single point is widier than the confidence interval of the mean as the variance of single value of y around the mean is σ^2. In fact, the “1”

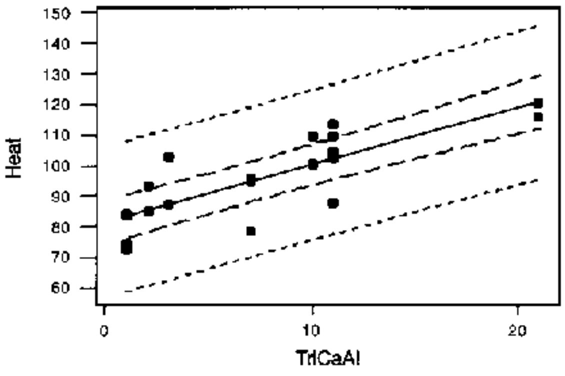


Figure 5 Confidence and prediction intervals.

under the square root symbol may dominate the other terms. The formula is given by

$$\hat{y}_0 \pm t_{n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{\mathrm{xx}}}} \qquad (23)$$

1.1.9 Conclusions

Regression is a widely used and flexible tool, applicable to many situations.

The method of least squares is the most commonly used in regression.

The resulting estimates are weighted means of the response variable at each data point. Means may not be resistant to extreme values of either X or y.

The normal, gaussian, distribution is closely linked to least squares, which facilitates the use of the standard statistical methods of confidence intervals and hypothesis tests.

In fitting a model to data, an important result of the least-squares approach is that the vector of fitted or predicted values is orthogonal to the vector of residuals. With the added assumptions of normality, the residuals are statistically independent of the fitted values.

The data appear as columns which can be considered as vectors. Groups of X vectors can be manipulated as a matrix. A projection matrix is a useful tool in understanding the relationships between the observed values of y, the predicted y and the residuals.

1.2 GOODNESS OF FIT OF THE MODEL

1.2.1 Regression Printout from MINITAB

1.2.1.1 Regression with One or More Predictor Variables

In this section, comments are made on the printout from a MINITAB program on the cement data using the heat evolved as y and the number of grams of tricalcium aluminate as x. This is extended to two or more variables.

1.2.1.2 Regression Equation

```
The regression equation is
y = 81.5 + 1.87 x 1
```

In keeping with the terminology we are using in this chapter, the y above should be $\hat{y}$. Alternatively, if a residual term e is added to the equation, we have termed this "the fitted regression equation." With one predictor variable, the fitted equation will represent a line.

We have noted in Sec. 1.1.7.1 that the estimated coefficients will vary depending on the other variables in the model. With the first two variables in the model, the fitted regression equation represents a plane and the least-squares solution is

$$y = 52.6 + 1.47x_1 + 0.662x_2$$

In vector terms, it is clear that $\mathbf{x}_1$ is not orthogonal to $\mathbf{x}_2$.

1.2.1.3 Distribution of the Coefficients

```
Predictor        Coef      StDev       T       P
Constant       81.479      4.927   16.54   0.000
x1             1.8687     0.5264    3.55   0.005
```

The formula for the standard deviation (also called the standard error by some authors) of the constant term and for the x_1 term is given in Sec. 1.1.8.1.

The T is the t-statistic = (estimator − hypothesized parameter)/standard deviation. The hypothesized parameter is its value under the null hypothesis, which is zero in this situation. The degrees of freedom are the same as those for the error or residual term. One measure of the goodness of fit of the model is whether the values of the estimated coefficients, and hence the values of the respective t-statistics, could have arisen by chance and these are indicated by the p-values.

The p-value is the probability of obtaining a more extreme t-value by chance. As the p-values here are small, we conclude that small t-value is due to the presence of x_1 in the model. In other words, as the probabilities are small (< 0.05 which is the common level used), both the constant and b_1 are significant at the 5% level.

1.2.1.4 *R*-Squared and Standard Error

```
S = 10.73   R-Sq = 53.4%    R-Sq(adj) = 49.2%
```

$S = 10.73$ is the standard error of the residual term. We would prefer to use lower case, s, as it is an estimate of the S in the S^2 of Eq. (3).

R–Sq (short for *R*-squared) is the coefficient of determination, R^2, which indicates the proportion of the variation of Y explained by the regression equation:

$$R^2 = S_{\hat{y}\hat{y}}/S_{yy} \quad \text{and recall that} \quad S_{yy} = \sum(y - \bar{y})^2$$

It is shown that R is the correlation coefficient between $\hat{y}$ and y provided that the x and y terms have been centered.

In terms of the projection matrices,

$$\mathrm{R}^2 = \frac{\sum \hat{\mathrm{y}}_i^2}{\sum \mathrm{y}^2} = \frac{\mathrm{y}^T\mathrm{Py}}{\mathrm{y}^T\mathrm{y}} \tag{24}$$

R^2 lies between 0, if the regression equation does not explain any of the variation of Y, and 1 if the regression equation explains all of the variation. Some authors and programs such as MINITAB write R^2 as a percentage between 0 and 100%. In this case, R^2 is only about 50%, which does not indicate a good fit. After all, this means that 50% of the variation of y is unaccounted for.

As more variables are added to the model, the value of R^2 will increase as shown in the following table. The variables x_1, x_2, x_3, and x_4 were sequentially added to the model. Some authors and computer programs consider the increase in R^2, denoted by ΔR^2. In this example, x_2 adds a considerable amount to R^2 but the next two variables add very little. In fact, x_4 appears not to add any prediction power to the model but this would suggest that the vector x_4 is orthogonal to the others. It is more likely that some rounding error has occurred.

Number of predictor variables	1	2	3	4
R^2	53.4	97.9	98.2	98.2
R^2 (adjusted)	49.2	97.4	97.6	97.4
Increase in R^2, ΔR^2		44.5	0.3	0.0

One peculiarity of R^2 is that it will, by chance, give a value between 0 and 100% even if the X variable is a column of random numbers. To adjust for the random effect of the k variables in the model, the R^2, as a proportion, is reduced by $k/(n-1)$ and then adjusted to fall between 0 and 1 to give the adjusted R^2. It could be multiplied by 100 to become a percent:

$$\text{Adjusted } R^2 = [R^2 - k/(n-1)][n-1]/[n-k-1] \tag{25}$$

1.2.1.5 Analysis of Variance

```
Analysis of Variance
Source          DF      SS        MS       F    P
Regression       1    1450.1    1450.1  12.600.005
Residual
  Error         11    1265.7     115.1
Total           12    2715.8
```

The SS (sums of squares) can best be understood by referring to Fig. 4 (Sect. 1.7.3) which showed the relationship between the three vectors, $\mathbf{y}$, $\hat{\mathbf{y}}$, and $\mathbf{e}$ provided that the Y- and X-variables are centered around their means. By Pythagoras' theorem,

$$\begin{aligned}\text{Sums of squares of } \mathbf{y} &= \text{Sums of squares of } \hat{\mathbf{y}}\\ &\quad + \text{Sums of squares of } \mathbf{e}\end{aligned}$$

That is,

$$\begin{aligned}&\text{Sums of squares of total} =\\ &\text{Sums of squares for regression} +\\ &\text{Sums of squares for residual}\end{aligned} \tag{26}$$

The ANOVA table is set up to test the hypothesis that the parameter $\beta = 0$. If there are more than one predictor variable, the hypothesis would be,

$$\text{H: } \beta_1 = \beta_2 = \beta_3 = \cdots \beta_\kappa = 0$$

If this is the case, it can be shown that the mean, or expected, value of y, ŷ, and e will all be zero. An unbiased estimated of the variance of y, σ^2, could be obtained from the mean squares of each of the three rows of the table by dividing the sums of squares by their degrees of freedom. From Fig. 4, we are now well aware that the vector of fitted values is orthogonal to the vector of residuals and, hence, we use the first two rows as their mean squares are independent and their ratio follows a distribution called the F-statistic. The degrees of freedom of the F-test will be 1 and 11 in this example.

The p-value of 0.005 is the probability that by chance the F-statistic will be more extreme than the value of 12.6. This confirms that the predictor variable, $x_1 =$ tricalcium aluminate, predicts a significant amount of the heat generated when the cement is mixed.

What are the effects of adding variables to the model? These can be demonstrated by the cement data. The regression sum of squares monotonically increase as variables are added to the model; the residual sum of squares monotonically decrease; residual mean squares reduce to a minimum and then increase.

One method of selecting a best-fit model is to select the one with the minimum residual mean squares.

Number of predictor variables	1	2	3	4
Regression sum of squares	1450	2658	2668	2668
Residual sum of squares	1266	58	48	48
Residual mean squares = s^2	115	5.8	5.4	6.0

1.2.1.6 Unusual Observations

```
Unusual Observations
Obs   x1        y      Fit    StDev  Residual        St
                                Fit               Resid
 10 21.0   115.90   120.72     7.72     -4.82   -0.65 X
```

Individual data points may be unusual because the y-values are unusually large or small which would be measured according to whether they fall within a 95% confidence interval. Alternatively, specific x-values may differ from the others and have an unduly large effect on the regression equation and its coefficients. More will be said on this in Section 1.4.

1.2.2 Power Transformations

Two variables, x and y, may be closely related but the relationship may not be linear. Ideally, theoretical clues would be present which point to a particular relationship such as an exponential growth model which is common in biology. Without such clues, we could firstly examine a scatter plot of y against x.

Sometimes we may recognize a mathematical model, which fits the data well. Otherwise, we try to choose a simple transformation such as raising the variable to a power p as in Table 1. A power of 1 leaves the variable unchanged as raw data. As we proceed up or down the table from 1, the strength of the transformation increases; as we move up the table the transformation stretches larger values relatively more than smaller ones. Although the exponential does not fit in very well, we have included it as it is the inverse of the logarithmic transformation. Other fractional powers could be used but they may be difficult to interpret.

Table 1 Common Power Transformations

p	Name	Effect
—	Exponential	Stretches
3	Cube	Large
2	Square	Values
1	"Raw"	
0.5	Square root	Shrinks
0	Logarithmic	Large
−0.5	Reciprocal/root	Values
−1	Reciprocal	

It would be feasible to transform either y or x, and, indeed, a transformation of y would be equivalent to the inverse transformation of x. For example, squaring y would have similar effects to taking the square root of x. If there are two or more predictor variables, it may be advisable to transform these in different ways rather than y, for if y is transformed to be linearly related to one predictor variable it may then not be linearly related to another.

In general, hwoever, it is usual to transform the y, rather than the x, variable as this transformation may lead to a better-fitting model and also to a better distribution of the response variable and the residuals.

Number of predictor variables	1	2	3	4
R^2	53.4	97.9	98.2	98.2
R^2 (adjusted)	49.2	97.9	98.2	98.2
Increase in R^2, ΔR^2	—	44.5	0.3	0.0

1.2.3 Resistant Regression

The traditional approach to regression is via the least-squares method, which has close relationships with means and the normal distribution. This is a powerful approach that is widely used. It does have problems in that the fitted regression line can be greatly affected by a few unusually large or small y-values.

Another approach, which is resistant to extreme values, is based on medians rather than means, as medians are not affected so much by strange values. The method is shown in Fig. 6 (data from Sec. 1.5.4). The x-values are divided, as closely as possible, into three groups according to size. In this simple example, there are only nine points so that each group consists of three points. For the lower third, the median value of x is found and the median value of y giving the point (2, 12); this is repeated for the middle and upper third. The middle and upper median points are (5, 25) and (8, 95). The resistant line is found by joining the lower point, (2, 12), to the upper point, (8, 95) and is shown in Fig. 6 as a solid line. To check whether a curve would be more appropriate than a line, the three pairs of medians are linked by the dashed lines (- - -). If the slopes of the dashed lines differ from the slope of the resistant line, it would suggest that a curve should be used or the response variable, y, should be transformed. A rule of

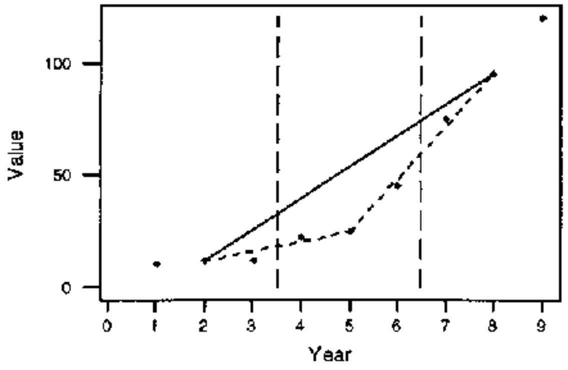


Figure 6 Resistant line of value of stamp by year.

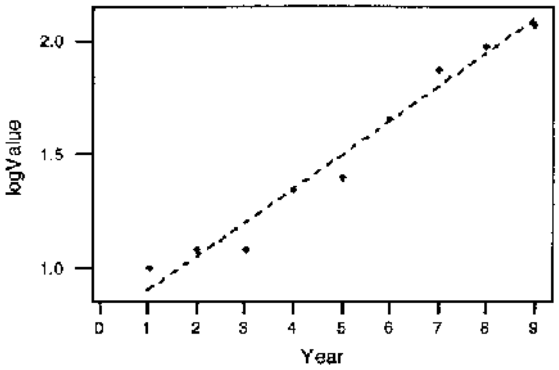


Figure 7 Plot of logarithm of stamp value by year.

thumb is that if the slope of the dashed line joining the middle and upper medians is more than two times or less than 0.5 times the slope of the dashed line joining the lower and the middle medians, a transformation is required. Here the upper slope is 70/3, whereas at the lower end the slope is only 13/3, suggesting a transformation is needed.

Notice that the two dashed lines take on the appearance of an arrow head pointing down in this case which suggests that a transformation should be tried which shrinks the large y-values, namely a p-value less than 1. For example, a logarithm transformation may be appropriate, as shown in Fig. 7.

1.2.4 More on Residuals

Is the model using the logarithm of the value of the stamp data in Sect. 1.5.4, better than using the raw value? Some clues are provided by a comparison of the plots of the residuals from these models. The residuals could be plotted against predicted values or, as shown in Fig. 8, against the observation order. Figure 8 shows the strong pattern which remains in the residuals of the raw data and the curvature in the pattern indicates that a quadratic term could be added to the model.

In Fig. 9, the transformed data, logarithm of the value, have removed the strong quadratic trend. In each residual plot, there remain certain trends in the residuals. If one considers the signs (+ or −) of the residuals, the raw data gives a pattern of $++-----+++$, whereas the logarithm of y yields the pattern $++---+++-$. These patterns indicate a correlation between successive residuals which is termed autocorrelation. The strength of this autocorrelation seems to suggest that a basic assumption has

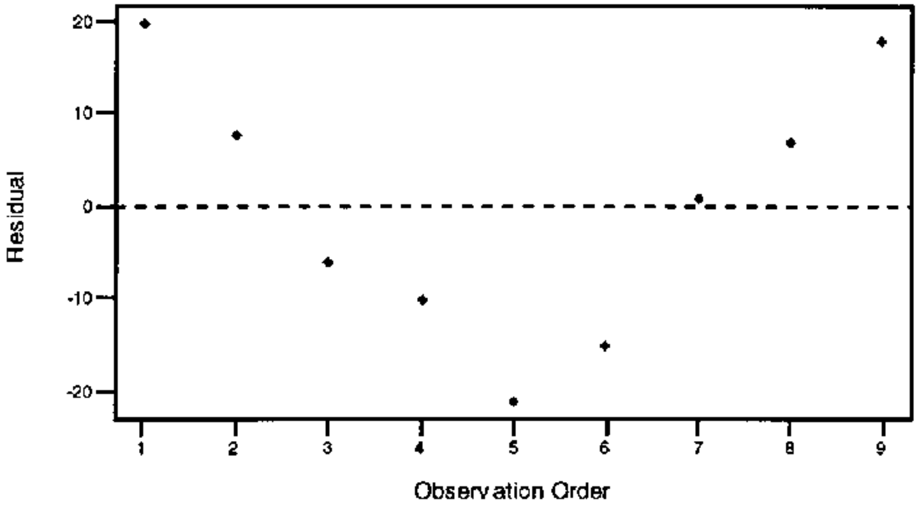


Figure 8 Residuals versus the order of the data (response is value).

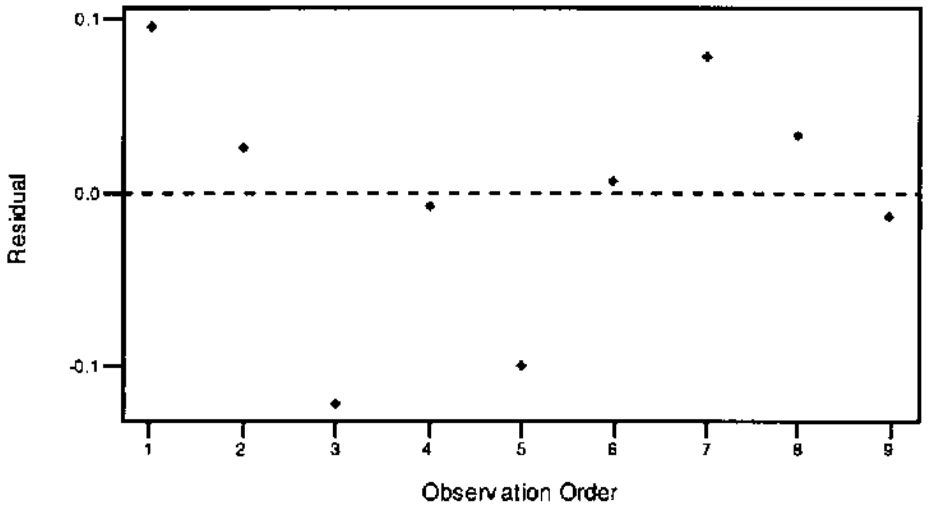


Figure 9 Residuals versus the order of the data (response is log value).

been violated, namely that successive deviations, e_i and e_{i+1}, are correlated. This autocorrelation is tested by the Durbin–Watson statistic, d, where

$$d = \sum_{t=2}^{n}(e_t - e_{t-1})^2 \bigg/ \sum_{t=1}^{n} e_t^2$$

If successive residuals are positively correlated with each other, the test statistic will be low. The hypothesis of uncorrelated errors is rejected for small values of this test statistic. If the test statistic d is below a lower critical value, d_L, then the hypothesis of independence of errors is rejected. In other words, we conclude that the errors are autocorrelated. Values greater than the upper critical value, d_U, allow us to accept the hypothesis of independence. Values of d between d_L and d_U are interpreted as "inconclusive."

With this stamp data set, the length of the series is very short with only nine data points so that strong probability statements are not appropriate. Tables of the d statistic do not usually include information on d_L and d_U for $n = 9$. However, for other sample sizes, n, the critical values for one predictor variable and a significance level of 5%, are:

n	d_L	d_U
15	1.08	1.36
20	1.20	1.41
25	1.29	1.45

With the raw stamp data, $d = 0.56$ and for the transformed data, $d = 1.64$. At least, we can conclude that the logarithm transformation has improved the autocorrelation of the residuals.

It has been mentioned before that the residuals should roughly fall in a horizontal band. An example where this is not so is shown in Fig. 10. The variance of the residuals seems to be increasing in absolute value as the fitted value increases, which suggests that the assumption of constant variance may be violated. To reduce this effect, one should consider a power transformation, such as a logarithm.

1.2.5 Conclusions

For a model to be a good fit to a data set, the F-test from the ANOVA should be significant. Some, if not all, of the estimated coefficients in the model will also be significant.

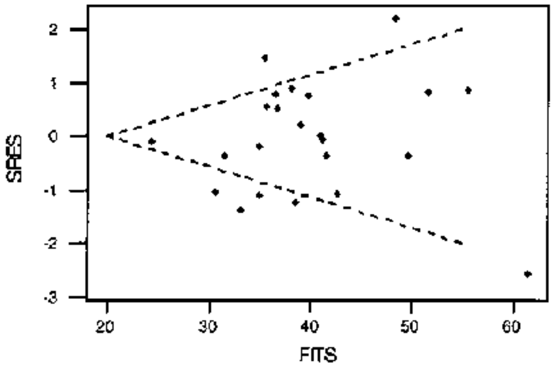


Figure 10 Funnel plot of standardized residuals.

R^2 indicates the proportion of the variation of the response variable which is explained by the model. This should be more than 50%, but one should take into account that R^2 depends on the number of predictor variables, and this is corrected for to some extent by the adjusted R^2.

The residual mean square is an estimate of the variance of the deviations in the model and a small value indicates a good fit. Unfortunately, this statistic depends on the scale of the response variable.

An examination of the residuals indicates if the model does not fit well. The residuals can be plotted against variables in the model or yet to be entered but they should always be plotted against the fitted, predicted variable.

Residuals also should be examined to check for autocorrelations, constant variance by observation number, or other variables.

To provide a better check on the goodness of fit of a model, other variables to enter the model should be considered. In other words, one checks the goodness of fit by looking for a model which provides a better fit. A discussion on this follows in the next section.

1.3 WHAT VARIABLES SHOULD BE INCLUDED IN THE MODEL?

1.3.1 Introduction

When a model can be formed by including some or all of the predictor variables, there is a problem in deciding how many variables to include. A small model with only a few predictor variables has the advantage that it is easy to understand the relationships between the variables. Furthermore, a small (parsimonious) model will usually yield predictions which are less influenced by peculiarities in the sample and, hence, are more reliable.

Another important decision which must be made is whether to use the original predictor variables or to transform them in some way. For example, the cost of laying rectangular concrete slabs largely depends on the length, breadth, and height of such slabs. Instead of using these three variables as predictors of the cost it would be much more sensible to use a single predictor which is the product of length, breadth, and height, that is, volume. This transformation has two advantages. It reduces the number of predictor variables and it introduces a predictor variable (volume) which has obvious physical implications for cost.

We shall start our discussion on variable selection by considering uncorrelated predictor variables. Such variables are often created in experimental design situations.

1.3.2 Uncorrelated Predictor Variables

In a regression with uncorrelated predictor variables the coefficients for each predictor variable in the regression equation are not affected by the presence or absence of other predictor variables. In this case it is very easy to determine which predictor variables should be included in the model. This is done by fitting a full regression model and then excluding any predictor variables which do not make a significant contribution to the regression. We shall illustrate this idea using the following example.

In Sec. 1.5.3, a three-factor central composite design is used to determine the effect of three predictor variables on the precipitation (y) of a chemical (stoichiometric dihydrate of calcium hydrogen orthophosphate). The three predictor variables were the mole ratio of ammonia to calcium chloride in the calcium chloride solution, time in minutes (t), and the starting pH of the base solution. In order to simplify the design of the experiment the predictor variables were defined as scaled versions of the original variables as indicated in Sec. 1.5.3. In addition the experiment was designed in such a way as to ensure that the predictor variables were uncorrelated, as indicated below.

```
Correlations (Pearson)

             x1         x2
x2        0.000
          1.000

x3        0.000      0.000
          1.000      1.000
```

A full regression model has been fitted to these data in Table 2.

Only the coefficients for the constant, x_1 and x_3, are significant. Because the predictor variables are uncorrelated this means that the coefficients do not change when the nonsignificant predictor (x_2) is dropped from the model as shown in Table 3. (Note the marginal increase in the residual standard deviation (s) and the marginal reduction in R^2 when the nonsignificant predictor is dropped from the regression equation.)

If the predictor variables had not been uncorrelated this would not have been true and, in order to obtain an appropriate smaller, more parsimonious model,

Table 2 Full Regression Model

```
Regression Analysis

The regression equation is
y = 71.9 + 5.49 x1 - 0.71 x2 + 10.2 x3

Predictor      Coef    StDev        T        P
Constant     71.895    1.710    42.05    0.000
x1            5.488    2.069     2.65    0.017
x2           -0.711    2.069    -0.34    0.736
x3           10.177    2.069     4.92    0.000

S = 7.646   R-Sq = 66.2%   R-Sq(adj) = 59.9%
```

special methods would have been required. Some of these methods are discussed in the next sections. Note that the fit of the above model can be much improved by including the square of x_3 as an additional predictor. This fact becomes obvious when the residuals for the model fitted in Table 3 are plotted against x_3 as shown in Fig. 11.

1.3.3 Testing Hypotheses

Suppose that we want to compare a full regression model with some reduced model. Assume that the reduced model estimates p coefficients and the full model estimates q coefficients. Consider the error sum of squares and the associated degrees of freedom for these two models.

Model	Error sum of squares	Error degrees of freedom	Mean square error
Reduced	*SSE* (reduced)	$n - p$	*SSE* (reduced)/ $(n - p)$
Full	*SSE* (full)	$n - q = n_E$	*SSE* (full)/ $(n - q)$

Table 3 Smaller Model

```
Regression Analysis

The regression equation is
y = 71.9 + 5.49 x1 + 10.2 x3

Predictor      Coef    StDev        T        P
Constant     71.895    1.665    43.19    0.000
x1            5.488    2.015     2.72    0.014
x3           10.177    2.015     5.05    0.000

S = 7.445   R-Sq = 66.0%   R-Sq(adj) = 62.0%
```

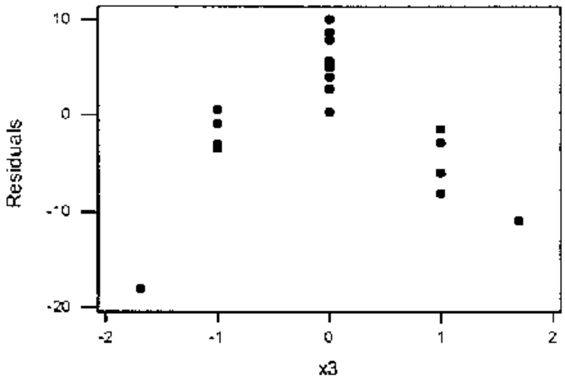


Figure 11 Relationship between residuals and x_3.

The following F-test can be used to determine whether the full model is significantly better than the reduced model:

$$F_{q-p,n-q} = \frac{[SSE(\text{Reduced}) - SSE(\text{Full})]/(q-p)}{MSE(\text{Full})}$$

If the p-value associated with this F-value is small (< 0.05) it means that the full model is significantly better than the reduced model.

1.3.4 Variable Selection—All Possible Regressions

One method which is commonly used to obtain more parsimonious models is to run all possible regressions with $k = 1, 2, 3, \ldots$ predictors and to choose the model which gives the most reliable fit. Define p equal to the number of regression coefficients including the constant (i.e., $p = k + 1$). Mallows [2] suggests using the following C_p statistic to compare the various models.

$$C_p = \frac{SSE(p)}{s^2} - (n - 2p)$$

The use of this statistic is illustrated in Example 1, which has been taken from Draper and Smith [3]. If the true model has been fitted then

$$E(C_p) = p = k + 1$$

The model with C_p closest to p therefore suggests the best model. This statistic penalizes models for which the number of coefficients including the constant is large in relation to the sample size (n), and it penalizes models for which the error sum of squares [$SSE(p)$] is

Table 4 Full Regression Model for the Cement Data

```
Regression Analysis

The regression equation is
y = 94.6 + 1.25 x1 + 0.172 x2 - 0.237 x3 -
0.460 x4

Predictor      Coef    StDev        T        P
Constant      94.57    30.40     3.11    0.014
x1           1.2508   0.2634     4.75    0.000
x2           0.1722   0.3418     0.50    0.628
x3          -0.2372   0.2773    -0.86    0.417
x4          -0.4599   0.3148    -1.46    0.182

S = 2.344   R-Sq = 98.4%   R-Sq(adj) = 97.6%
```

large in relation to the error variance (s^2) for the full model.

For the example of Sec. 1.5.1, the amount of heat evolved in cement is to be modeled in terms of the amount of four ingredients in the mix. In this example the predictor variables are not uncorrelated. This means that we cannot simply drop the nonsignificant predictor variables from the full model fitted in Table 4 in order to determine a more parsimonious model.

In the output shown in Table 5 the two best one-predictor, two-predictor, and three-predictor models have been compared with the full model. The model with C_p closest to p is obtained for the three-predictor model involving x_1, x_2, and x_4, suggesting that this is the best regression model for these data.

This model has been fitted in Table 8 producing a residual standard deviation and an R^2 marginally smaller than for the full model. All the coefficients are significant at the 5% level in this three-predictor model, confirming that, in this case, it would have been a mistake to simply drop all those predictor variables which were insigificant in the full model.

1.3.5 Variable Selection—Sequential Methods

Another approach for the selection of predictor variables involves the sequential addition and/or removal of predictor variables when there is a significant change in the regression sums of squares (SSR). When predictors are added sequentially this method is called forward selection. When the predictor variables are removed sequentially this method is called backward elimination. When both these methods are applied simultaneously this method is called sequential regression.

1.3.5.1 Forward Selection

The forward selection method starts with no predictor variables in the model and proceeds adding variables one at a time until a stopping criterion is reached. For each predictor variable an F-statistic is calculated in order to evaluate the significance of a one-predictor model. This F-statistic is the ratio of the regression sums of squares (SSR) to the mean square error (MSE) for the one-predictor model. The MSE is obtained by dividing the SSE by the error degrees of freedom. In the following equation $n_R = 1$ represents the regression degrees of freedom and n_E represents the error degrees of freedom:

$$F = \frac{SSR(\text{one-predictor})/n_R}{SSE(\text{one-predictor})/n_E} = \frac{SSR(\text{one-predictor})}{MSE(\text{one-predictor})}$$

The predictor variable with the highest F-value enters the regression, provided that this F-value is sufficiently large. Next, F-statistics are calculated for all remaining predictors, in order to evaluate the significance of an additional predictor in the regression. This F-statistic is the difference in the error sums of squares for a two-predictor and the best one-predictor model, divided by the mean square error for a two-predictor model. This means that the two-predictor model is the

Table 5 Best Subsets Regression

```
Response is y

                                      R-Sq
#Vars(k)       p        R-Sq          (adj)         Cp          S       x1x2x3x4
1              2        67.5          64.5       151.9     8.9639              X
1              2        66.6          63.6       156.0     9.0771          X
2              3        97.9          97.4         3.5     2.4063        X X
2              3        97.2          96.7         6.6     2.7343        X      X
3              4        98.3          97.8         3.3     2.2447        X    X X
3              4        98.2          97.6         3.7     2.3087        X X    X
4              5        98.4          97.6         5.0     2.3440        X X  X X
```

Table 6 Best Regression

```
Regression Analysis

The regression equation is
y = 71.6 + 1.45 x1 + 0.416 x2 - 0.237 x4

Predictor      Coef    StDev        T        P
Constant      71.65    14.14     5.07    0.000
x1           1.4519   0.1170    12.41    0.000
x2           0.4161   0.1856     2.24    0.052
x4          -0.2365   0.1733    -1.37    0.205

S = 2.309   R-Sq = 98.2%   R-Sq(adj) = 97.6%
```

full model and the one-predictor model is the reduced model. Therefore $q - p = 1$ in the formula

$$F = \frac{[SSE(\text{Reduced}) - SSE(\text{Full})]/(q-p)}{MSE(\text{Full})}$$

The predictor with the highest F-value is added to the regression provided that its F-value is sufficiently large. This process continues until all the predictors are included or until none of the remaining predictors would make a significant contribution in explaining the variation in the response variable.

Table 7 applies this method to the cement data of Sec. 1.5.1 (Example 1). In this example F-values in excess of 4 are required in order for the predictor to be added to the model. It is common to use t-values beyond $t = \pm 2$ to denote statistical significance at a 5% level for n reasonable large, say ≥ 15. This translates to $F = 4$ because $F(1, n) = t^2(n)$. The T-values in Table 7 are the square roots of the F-values defined above.

Table 7 Forward Selection for the Cement Data

```
Stepwise Regression

  F-to-Enter:  0.00  F-to-Remove:  0.00

  Response is y on 4 predictors, with N = 13

     Step        1        2        3          4
 Constant   117.57   103.10   109.78      94.57

x4          -0.738   -0.614   -0.617     -0.460
T-Value      -4.77   -12.62   -15.44      -1.46

x1                     1.44     1.15       1.25
T-Value               10.40     6.97       4.75

x3                             -0.35      -0.24
T-Value                        -2.42      -0.86

x2                                         0.17
T-Value                                    0.50

S             8.96     2.73     2.24       2.34
R-Sq         67.45    97.25    98.33      98.38
```

The last column provides the recommended four-predictor model. The coefficients in this model are not all significant. This means that a model with four predictors is not appropriate. If the F to enter had been set at 4 then the variable X_2 would not have entered the regression and the analysis would have stopped at Step 3. Consider the three-predictor model under Step 3; all the coefficients are significant, indicating that this model is appropriate. That is,

$$y = 109.8 + 1.15x_1 - 0.35x_3 - 0.617x_4$$

Note that x_4 entered the model first, followed by x_1 and then x_3. Note that this method produces the same regression model as the best subsets regression. This will not always be the case. Note that, as expected, R^2 increases as successive variables are added to the model. The coefficients for predictors included in the model (e.g., x_4) change at each step because these predictor variables are not uncorrelated.

1.3.5.2 Backward Elimination

The backward elimination method starts with all the (k) predictor variables in the model and proceeds to remove variables one at a time until a stopping criterion is reached. For each predictor variable an F-statistic is calculated in order to evaluate its contribution to the model. This F-statistic measures the effect of removing the predictor variable from the model. It is the difference in the error sums of squares for a k- and a $(k-1)$-predictor model divided by the mean square error for the k-predictor model

$$F = \frac{SSE(k-1) - SSE(k)}{MSE(k)}$$

The predictor variable with the lowest F-value makes the least contribution to the model. It leaves the regression provided that this F-value is sufficiently small. Next, F-statistics are calculated for all predictors left in the regression, in order to determine if another predictor should be eliminated from the regression. This F-statistic is the difference in the error sums of squares for a $(k-1)$-predictor and a $(k-2)$-predictor model divided by the mean square error for a $(k-1)$-predictor model.

$$F = \frac{SSE(k-2) - SSE(k-1)}{MSE(k-1)}$$

The predictor with the lowest F-value is again eliminated from the regression provided that its F-value is sufficiently small. This process continues until all the predictors are excluded or until all the predictors remaining in the regression make a significant contribution in explaining the variation in the response variable.

Table 8 applies this method to the cement data of Sec. 1.5.1. In this example predictors with F-values of less than 4 are dropped from the regression one by one. This method also produces the same regression model as the best subsets regression:

$$y = 109.8 + 1.15x_1 - 0.35x_3 - 0.617x_4$$

The last column provides the recommended three-predictor model. Notice that, in step 2, x_2 is dropped from the model because its T-value of 0.50 is the smallest T-value for step 1 and its F-value $[F = t^2 = (0.50)^2 = 0.25]$ is less than 4.

Backward elimination is recommended in preference to forward selection when dealing with pairs of predictor variables which together measure a gap. For example, if one predictor variable measures the inside diameter of a nut and another predictor variable is the diameter of the corresponding bolt, then it is the difference between these two measurements that may be of interest rather than their individual measurements. Forward selection may fail to enter either of these variables individually, hence failing to include their difference in the model. Backward elimination does not do this and is therefore recommended in this situation.

Table 8 Backward Elimination for the Cement Data

```
Stepwise Regression

 F-to-enter: 99999.00  F-to-Remove: 4.00
 Response is y on 4 predictors, with N = 13

     Step         1         2
 Constant     94.57    109.78

x1             1.25      1.15
T-Value        4.75      6.97

x2             0.17
T-Value        0.50

x3            -0.24     -0.35
T-Value       -0.86     -2.42

x4           -0.460    -0.617
T-Value       -1.46    -15.44

S              2.34      2.24
R-Sq          98.38     98.33
```

1.3.5.3 Hierarchical Regression

The above model uses a model consisting of only linear terms to fit the data. In this case the fit is so good ($R^2 =$ 98.33%) that there is no need to consider anything but linear terms. However, in other situations polynomial models are required in order to achieve a reasonable fit. In polynomial models, powers (e.g., squares and cubes) of the existing variables are added to the model. In addition, interaction terms consisting of cross-products may be considered.

In fitting such polynomial models it is usual to add the higher-order terms one at a time in layers—hence the name of hierarchical regression. Layers of higher order terms are added until there is no significant increase in R^2 as measured by the F-statistics defined in the last section. Note, however, that the stepwise methods should not be used to fit these models because they might produce a model consisting of, say, only linear and cubic terms. To omit the quadratic terms in this way is not appropriate.

Multicollinearity (i.e., high correlations between predictor variables) is a problem associated with hierarchical regression. As discussed in Sec. 1.4.4 multicollinearity causes computational and statistical difficulties in regression. In hierarchical regression the higher-order variables (e.g., x^3) are generally highly correlated with their low-order counterparts (e.g., x). This problem can often be overcome by subtracting the mean from each predictor variable. This results in regression terms of the form

$$(x - \bar{x}), (x - \bar{x})^2, (x - \bar{x})^3, \ldots$$

instead of

$$x, x^2, x^3, \ldots$$

Hierarchical regression is illustrated below for the data of Sec. 1.5.3. For these data the means of the predictor variables have already been set to zero, so multicollinearity should not pose too much of a problem. The R^2 for the linear model shown in Table 4 explained only 66.0% of the total variation in the response variable. Hopefully a model with higher-order terms will do better. Table 9 shows the second-order model including squares and an interaction term. Table 10 shows the third-order model.

R^2 for the second-order model is 94.9%, while R^2 for the third-order model is 97.1%. In order to test whether there is a signfiicant improvement when going from a second- to a third-order model the following F-statistic is used. In this equation SSE denotes error sum of squares and MSE denotes mean square error.

Table 9 Second-Order Model for Experimental Design Data

```
Regression Analysis

The regression equation is
y = 76.4 + 5.49 x1 + 10.2 x3 + 0.643 x1^2 - 7.22 x3^2 - 1.46 x1x3

Predictor             Coef      St Dev          T          P
Constant            76.389       1.099      69.52      0.000
x1                  5.4880      0.8588       6.39      0.000
x3                 10.1773      0.8588      11.85      0.000
x1^2                0.6429      0.8319       0.77      0.452
x3^2               -7.2236      0.8319      -8.68      0.000
x1x3                -1.463       1.122      -1.30      0.213

S = 3.174   R-Sq = 94.9%    R-Sq(adj) = 93.1%

Analysis of Variance

Source                  DF          SS         MS          F        P
Regression               5     2627.09     525.42      52.16    0.000
Error                   14      141.02      10.07
Total                   19     2768.11
```

Model	Error degrees of freedom	Error sums of squares	Mean square error
2nd order	$p = 14$	*SSE*(2nd order)	
3rd order	$q = 12$	*SSE*(3rd order)	*MSE*(3rd order)

Using the same *F*-test for comparing a reduced model with a full model,

$$F_{q-p,n-q} = \frac{[SSE(\text{2nd order}) - SSE(\text{3rd order})]/(q-p)}{MSE(\text{3rd order})}$$

$$= \frac{(141.02 - 81.2)/(14 - 12)}{6.77}$$

$$= 4.418$$

The degrees of freedom associated with this *F*-value are 2 in the numerator and 12 in the denominator. An *F*-value of 3.89 is significant at the 5% level, whereas an *F*-value of 6.93 is significant at the 1% level. This means that the improvement produced by fitting a third-order model as opposed to a second-

Table 10 Third-Order Model for Experimental Design Data

```
Regression Analysis

The regression equation is
y = 76.4 + 3.22 x1 +6.78 x3 + 0.643 x1^2 - 7.22 x3^2 - 1.46 x1x3 + 1.29 x1^3
         + 193 x3^3

Predictor             Coef       StDev          T          P
Constant           76.3886      0.9006      84.82      0.000
x1                   3.224       1.543       2.09      0.059
x3                   6.778       1.543       4.39      0.001
x1^2                0.6429      0.6818       0.94      0.364
x3^2               -7.2236      0.6818     -10.59      0.000
x1x3               -1.4625      0.9197      -1.59      0.138
x1^3                1.2881      0.7185       1.65      0.125
x3^3                1.9340      0.7185       2.47      0.029

S= 2.601    R-Sq = 97.1%    R-Sq(adj) = 95.4%

Analysis of Variance

Source                  DF          SS         MS          F        P
Regression               7     2686.91      83.84      56.73    0.000
Error                   12       81.20       6.77
Total                   19     2768.11
```

order model is barely significant. Indeed, it seems that all that is required is a linear model in terms of x_1 and a quadratic model in terms of x_3.

1.3.6 Indicator (Grouping) Variables

In this section we show how qualitative variables can be introduced into a regression in order to test for differences between groups. This is done by defining the qualitative (group) variable in terms of dummy variables. Dummy variables are binary variables in that they can take only the values of 0 and 1. The number of dummy variables required in order to define a single qualitative variable is always one less than the number of groups. As illustrated below, if there are only two groups then a single dummy variable is required to define the groups, but if there are three groups then two dummy variables are required. For a three-group qualitative variable DUM1 = 1 for Group 1 only, DUM2 = 1 for Group 2 only and, for Group 3, DUM1 = DUM2 = 0. Note that a third dummy variable, set equal to 1 for Group 3 and 0 otherwise, is therefore redundant.

Two-groups		
Qualitative variable	Dummy variable	
Group 1	0	
Group 2	1	
Three-groups		
Qualitive variable	First dummy variable (DUM1)	Second dummy variable (DUM2)
Group 1	1	0
Group 2	0	1
Group 3	0	0

By introducing such dummy variables into a regression, with additional terms defined as the product of the dummy variables and predictor variables, it is possible to fit separate regressions for each group. In addition it is possible to test whether the coefficients for the groups differ signficiantly. We shall illustrate this method using the following example.

In a study of the effect of company size (measured using the number of employees) on the number of work hours lost due to work-related accidents, the presence or absence of a safety program is thought to be important. The data in Sec. 1.5.2 are illustrated in Fig. 12. The two groups have been coded using the dummy variable "Dum." For companies with a safety program Dum = 1 and for companies without a safety program Dum = 0. The term DumEmp is defined as the product of Dum and the predictor variable "Employees."

Three regressions have been fitted to these data in Table 11. The first regression suggests that the term Dum is redundant, so this term is eliminated in the second regression. The second regression suggests that the constant term is redundant, so this term is eliminated in the third regression.

The final equation contains only significant terms, but note that R^2 is not given for this model. This is due to the fact that the interpretation of the R^2 for a regression model which does not have a constant term is problematic. In addition, the standard formulas given in previous sections are no longer valid. Because of these problems it may be easier to retain the constant in the regression model even when it is not significant. Note, however, that the residual standard deviation, s, is smaller for the third regression, indicating an improvement in the model.

$$\begin{aligned}\text{Hours} = {} & 0.0195454\ \text{Employees} \\ & - 0.0096209\ \text{DumEmp}\end{aligned}$$

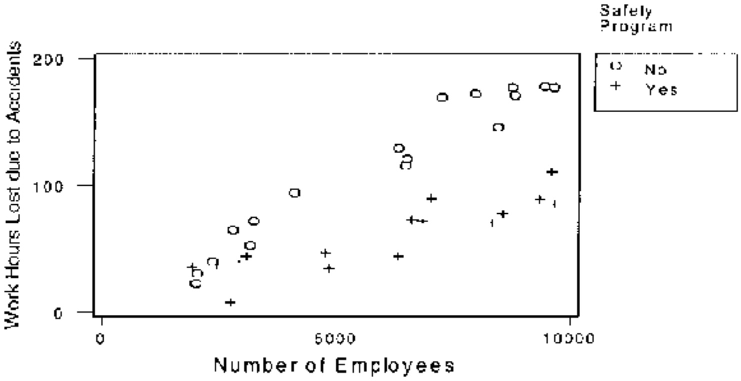


Figure 12 Hours lost by the number of employees.

Table 11 Regressions for Safety Program Data

```
First Regression Analysis

The regression equation is
Hours = -1.21 + 0.0197 Employees + 8.9 Dum - 0.0109 DumEmp

Predictor              Coef        StDev            T            P
Constant             -1.213        7.593        -0.16        0.874
Employees          0.019716     0.001179        16.72        0.000
Dum                    8.90        11.13         0.80        0.430
DumEmp            -0.010874     0.001721        -6.32        0.000

S = 13.28   R-Sq = 94.1%    R-Sq(adj) = 93.5%

Second Regression Analysis

The regression equation is
Hours = 2.93 + 0.0191 Employees - 0.00962 DumEmp

Predictor              Coef        StDev            T            P
Constant              2.934        5.517         0.53        0.599
Employees         0.0191327    0.0009214        20.76        0.000
DumEmp           -0.0096211    0.0007112       -13.53        0.000

S = 13.30   R-Sq = 94.0%    R-Sq(adj) = 93.6%

Third Regression Analysis

The regression equation is
Hours = 0.0195 Employees - 0.00962 DumEmp

Predictor              Coef        StDev            T            P
Noconstant
Employees         0.0195454    0.0004914        39.78        0.000
DumEmp           -0.0096209    0.0007030       -13.69        0.000

S = 13.05
```

Now, for companies without a safety program DumEmp equals zero, so

$$\text{Hours} = 0.0195454 \text{ Employees}$$

For companies with a safety program DumEmp, the product $1 \times$ Employees = Employees, so that

$$\begin{aligned}\text{Hours} &= 0.0195454 \text{ Employees} \\ &\quad - 0.0096209 \text{ Employees} \\ &= 0.0291663 \text{ Employees}\end{aligned}$$

There is a significant difference between the slopes of these two lines because the DumEmp coefficient is significantly different from zero. Residual analysis confirms that two lines with different slope but a common intercept of zero is a suitable model for these data. It is clear that the benefit of a safety program in terms of reducing the number of hours lost due to accidents does depend on company size. There is a larger benefit

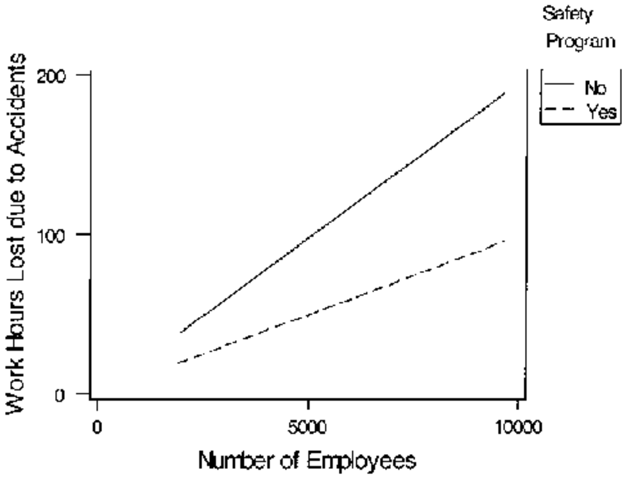


Figure 13 Fitted model of safety program data.

for large companies than for smaller companies. Figure 13 shows the fitted model for these data.

1.3.7 Conclusion

This section has dealt with the selection of predictor variables in regression. The main points are as follows:

Small, parsimonious models are easier to understand and they predict more reliably.

Predictor variables which are meaningful predictors of the response should be used.

When predictor variables are uncorrelated the full model shows which predictors can be dropped.

When predictor variables are correlated then other methods must be used.

Best subsets regression defines the optimum model as the equation for which $C_p \approx p$.

Forward selection enters predictor variables one at a time in order of importance.

Backward elimination removes predictor variables one at a time in order of least importance.

Backward elimination is recommended when the difference between two predictors may be an important predictor, whereas the individual predictors are unimportant.

$(k-1)$ dummy (zero–one) variables are required to define k groups.

Test for significant differences between the relationship for the groups by testing the significance of the dummy variable coefficients and the coefficients of the predictor–dummy-variable products.

1.4 PECULIARITIES OF OBSERVATIONS

1.4.1 Introduction

In Sec. 1.3 we considered the relationship between the response variable y and the k-predictor variables $X = (x_1, x_2, \ldots, x_k)$. This is a relationship between column vectors of data. In this section we turn our attention to rows or individual data points, $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$. Both predictor and response variable values must be considered when deciding whether a point is unusual. Points with peculiar x-values are termed sensitive or high leverage points. Points with unusual y-values are termed outliers.

1.4.2 Sensitive or High Leverage Points

The unusual x-values of high leverage points means that these points have much influence on the regression equation. In addition the variance for y-predictions at these points are large. We shall illustrate these concepts using a simple example:

x	1	2	3	4	10
y	10	8	13	11	?

Let us compare the regression lines when, for $x = 10$, the y-value is 14 or 28. As shown in Fig. 14 the regression line is greatly influenced by this y-value. This is because the x-value for this point is so unusual. The line fitted to the point (10, 14) has a much smaller slope than the line fitted to the point (10, 28). This means that for $x = 10$ the value of y has a "high leverage" on the fitted line.

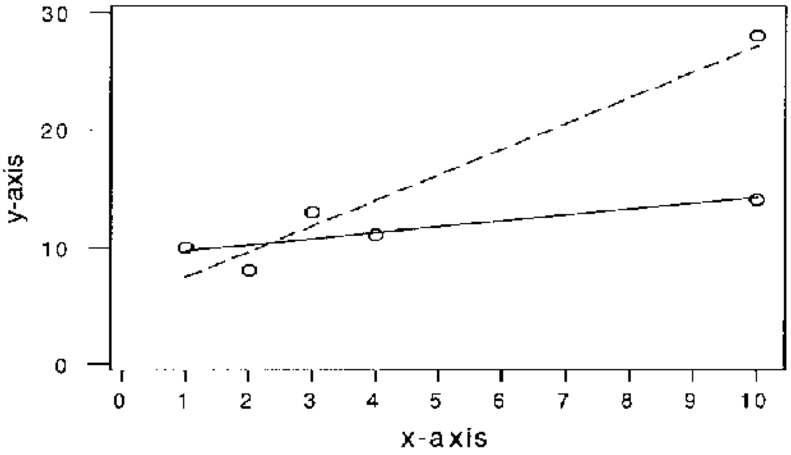


Figure 14 Effect of a high leverage point at $x = 10$.

In algebraic terms $\hat{\mathbf{y}} = P\mathbf{y}$, where $P = \{p_{ij}\} = X(X^T X)^{-1} X^T$. For this example,

$$\mathrm{P} = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 4 \\ 10 \end{bmatrix} (130)^{-1} [1 \quad 2 \quad 3 \quad 4 \quad 10]$$

$$= (130)^{-1} \begin{bmatrix} 1 & 2 & 3 & 4 & 10 \\ 2 & 4 & 6 & 8 & 20 \\ 3 & 6 & 9 & 12 & 30 \\ 4 & 8 & 12 & 14 & 40 \\ 10 & 20 & 30 & 40 & 100 \end{bmatrix}$$

and

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \hat{y}_4 \\ \hat{y}_5 \end{bmatrix} = (130)^{-1} \begin{bmatrix} 1 & 2 & 3 & 4 & 10 \\ 2 & 4 & 6 & 8 & 20 \\ 3 & 6 & 9 & 12 & 30 \\ 4 & 8 & 12 & 16 & 40 \\ 10 & 20 & 30 & 40 & 100 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix}$$

Clearly the influence of y_5 on all the predicted values is greater than all the other y-values due to its unusual x-value.

The covariance matrix for these predictions equals $\mathrm{P}\sigma^2$. As a result, the variance for the predicted value of y_5 is $p_{55}\sigma^2 = (100/130)\sigma^2$, larger than that for the other y-values. Large values in the matrix P alert us to unusual x-values. In particular, points with the diagonal values of P, namely P_{ii}, more than twice the average value of all the diagonal P values, can be regarded as high leverage points. that is, any point i with

$$p_{ii} > \frac{2\sum_{j=1}^{n} p_{jj}}{n}$$

can be regarded as a high leverage point. In the above example any point with p_{ii} greater than 0.4 (that is (2(130/130)/5) can be regarded as a high leverage point. This means that only the fifth point, with $p_{ii} = 100/130$, is a high leverage point.

1.4.3 Outliers

Any point with a large difference between its observed and predicted y-values is probably an outlier. This means that it is the model as well as the observed y-value which determine whether a point is an outlier. Therefore, if more or different predictors are included in a model then it is likely that different points will show up as being potential outliers.

Outliers are often difficult to detect because, in particular, they may be obscured by the presence of high leverage points. Generally speaking, an outlier has a high absolute value for its "Studentized" residual. Hoaglin and Welsch [4] suggest several alternative methods for detecting outliers. Many of these methods consider what happens to estimated coefficients and residuals when a suspected outlier is deleted. Instead of discussing these methods we will use a principal-component method for detecting outliers in the next section.

What should one do with outliers? If the sample size is large and it is suspected that the outlier is an error of measurement then an outlier can be deleted. However, it is important to consider such points very carefully because an outlier may suggest conditions under which the model is not valid. The most infamous outlier in recent times has been the ozone hole over the south pole. For several years this outlier was regarded as an error of measurement rather than a reality.

1.4.4 Eigenvalues and Principal Components

As mentioned above, principal-component analysis can be used to identify outliers. We shall illustrate this use of principal-component analysis using the cement data. In addition we shall show how principal-component analysis can be used to determine when there is collinearity or multicollinearity. This condition occurs when the correlations between the predictors are so large that the $X^T X$ matrix becomes nearly singular with a determinant close to zero. In this situation the variances associated with the estimated coefficients are so large that their interpretation becomes meaningless.

In principal-component analysis the predictor variables are transformed to principal components. The eigenvalues measure the variances of each of these principal components, with the first component having the largest eigenvalue and the last eigenvector having the smallest eigenvalue. In Fig. 15, x_1 and x_2 represent the original predictors, while w_1 and w_2 represent the principal components. The variance associated with w_1 is obviously maximized, while the variance associated with w_2 is minimized.

In Fig. 15 there is a strong correlation between x_1 and x_2. As a result the eigenvalue (variance) associated with w_2 is close to zero. Unusual values for w_2 fall far from the w_1 axis. Such values can be used to identify outliers. When there are more than two variables this

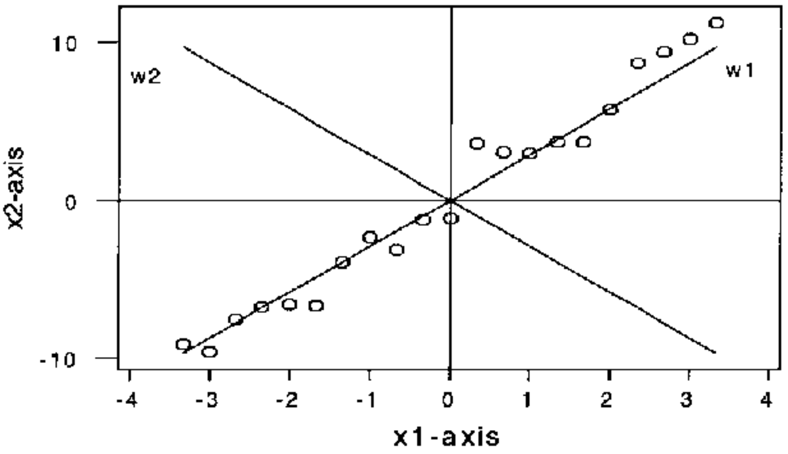


Figure 15 Original x-axis with w-principal-component axis.

rule can still be used to identify outliers. Any unusual principal-component value identifies an outlier when it is associated with an unimportant (low-variance) eigenvalue.

When there is a strong muilticollinearity, at least one of the eigenvalues is close to zero. In this situation the determinant of $X^T X$ (the product of the eigenvalues) is close to zero, causing the variance of the regression coefficients $[\sigma^2(X^T X)^{-1}]$ to be very big. According to Hoaglin and Welsch [4], if the ratio of the largest eigenvalue to the smallest eigenvalue exceeds 30 when a principal-component analysis is performed on a correlation matrix, then it means that multicollinearity is making the regression unreliable. In this situation, predictors should be combined or discarded.

The principal-component analysis for the predictors of the heat evolved in cement appears in Table 12. The largest eigenvalue is 2.3246, while the smallest eigenvalue is 0.0077. This suggests a ratio $2.3246/0.0077 = 302$. This means that there is too much multicollinearity among these predictors for a full four-predictor model. Predictors must be combined or discarded in order to ensure that the estimated coefficient variances are reasonably small, otherwise the regression will be unreliable.

Table 13 shows the principal-component scores for these data. The last principal component identifies no outliers, but the third component identifies the fourth observation as an outlier. Observation 4 has a relatively high value for w_3. The loadings for w_3 (high on x_1 and x_3) show in what sense this observation is an outlier. It has a relatively high score for $(x_1 + x_3)$.

1.4.5 Ridge Regression and Prior Information

We have seen that high correlations among the predictor variables tend to increase the variance of the estimated regression coefficients, making these estimates and any predictions unreliable. We have suggested

Table 12 Principal-Component Analysis: Predictors for Cement Data

```
Principal-Component Analysis
Eigenanalysis of the Correlation Matrix

Eigenvalue          2.3246        1.3836        0.2841        0.0077
Proportion           0.581         0.346         0.071         0.002
Cumulative           0.581         0.927         0.998         1.000


Variable               PC1           PC2           PC3           PC4
x1                   0.437         0.551         0.693         0.158
x2                   0.569        -0.411        -0.189         0.687
x3                  -0.427        -0.568         0.675         0.200
x4                  -0.550         0.453        -0.169         0.681
```

Table 13 Principal-Component Scores

y	$x_1 + x_3$	w_1	w_2	w_3	w_4	Observation
78.5	13	−1.40199	1.92387	−0.76496	0.028472	1
74.3	16	−2.06499	0.28164	−0.49613	−0.049005	2
104.3	19	1.17609	0.24906	−0.04723	−0.104609	3
87.6	29	−1.28180	0.76837	**1.01948**	0.196814	4
95.9	13	0.43662	0.50652	−0.80836	0.078454	5
109.2	20	1.00818	0.24246	0.04829	−0.036712	6
102.7	20	1.00079	−2.05957	−0.09935	0.049964	7
72.5	23	−2.18793	−0.59801	0.28506	−0.070893	8
93.1	20	−0.28684	−1.35846	−0.06863	−0.046095	9
115.9	25	1.65441	1.93414	0.76620	−0.112222	10
83.8	24	−1.59558	−1.19351	0.38017	−0.049449	11
113.3	20	1.73927	−0.31864	0.01558	0.042293	12
109.4	18	1.80377	−0.37788	−0.23013	0.0729894	13

that predictors should be combined or discarded in this situation. Another solution to the multicollinearity problem is to use a ridge regression estimator proposed by Hoerl and Kennard [5]. They suggest increasing the problematic determinant of the $X^T X$ matrix by adding a constant, k, to all diagonal terms. This obviously increases the determinant, producing smaller variances for the coefficients.

In general the value of k should be less than 0.2. Plots of the estimated coefficients against k, known as ridge traces, are useful for deciding the optimum value for k. The best value of k is the lowest value of k for which the estimated coefficients are reasonably stable. An optimum value for k of approximately 0.08 is suggested by the ridge trace shown in Fig. 16.

A similar approach is used to incorporate prior information into a regression analysis. Say that the prior information about the β coefficients can be described by the model

$$r = R\beta + \delta$$

where δ is distributed $N(0, \lambda^2 I)$. Then the original model and this additional model can be combined into a single model

$$\begin{bmatrix} y \\ r \end{bmatrix} = \begin{bmatrix} X \\ R \end{bmatrix} \beta + \begin{bmatrix} \varepsilon \\ \delta \end{bmatrix}$$

with the weighted least-squares estimator for β given by

$$b = \frac{X^T y/\sigma^2 + R^T r/\lambda^2}{(X^T X/\sigma^2 + R^T R/\lambda^2)}$$

If $\mathrm{R}^T \mathrm{r} = 0$ and $\mathrm{R}^T \mathrm{R} = \mathrm{I}$ and $k = (\sigma/\lambda)^2$, then the ridge regression estimate is obviously obtained as a special case of this estimator.

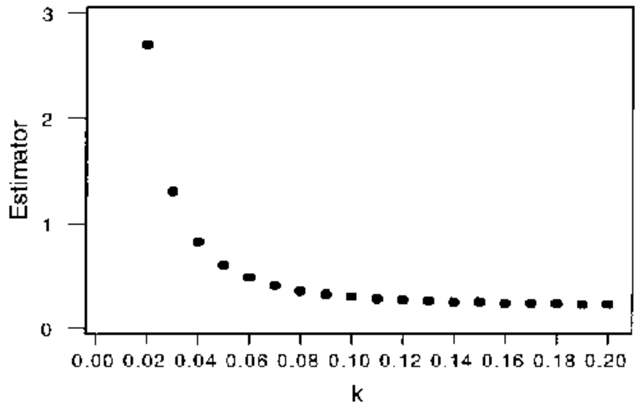


Figure 16 A ridge trace.

1.4.6 Weighted Least Squares

One of the assumptions of least-squares regression is that the variance remains constant for all values of the predictor variables. If it is thought that this assumption is not valid, it is possible to modify the least-squares method by giving different weights to different observations.

Say that the response y_i is the mean for a sample of n_i individuals, then the variance for y_i is σ^2/n_i. In order to force constant variances for each response value one must multiply through by $\sqrt{n_i}$. That is, instead of fitting the model

$$y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

the following model should be fitted in order to obtain constant variances:

$$\sqrt{n_i}y_i = \sqrt{n_i}\beta_0 + \sqrt{n_i}\beta_1 X_1 + \sqrt{n_i}\varepsilon_i$$

This model produces weighted coefficient estimates of the form

$$b = (X^T WX)^{-1} X^T Wy$$

where W is a diagonal matrix consisting of the sample sizes (n_i) for each mean observation.

When the variance for the ith mean is defined as s_i^2/n_i, with different variances, then the weighted estimates are similarly defined, but $\mathrm{W} = \mathrm{diag}(n_i/s_i^2)$.

As an illustration, the relationship between book value and resale value is to be determined for a company's plant and equipment. The following data have been collected. A log transformation of book value is required in order to linearize these data.

Book value ($)	Midpoint	Number of items (n_i)	Mean resale value	Resale standard deviation (s_i)
< 1000	500	71	184.9	21.24
1000 +	1500	250	231.8	16.64
2000 +	2500	210	254.8	15.3
3000 +	3500	59	272.5	11.08
4000 +	4500	44	271.5	12.93
5000 +	5500	34	278.3	10.03
6000 +	6500	12	284.9	7.24
7000 +	7500	9	281.8	11.05
8000 +	8500	8	285.0	13.3
9000 +	9500	7	272.3	17.73

Applying the above formulae to these data the following regression equation is obtained:

$$\text{Mean resale value} = 218 + 36.8 \log(\text{book value})$$

The slope of this equation differs quite markedly from 32.8, the slope that is obtained when no weighting is applied.

1.4.7 Conclusion

This section has dealt with the identification and handling of peculiar observations.

High leverage points are points which have unusual predictor variable values. These points have a strong influence on the coefficient estimates so deserve special care and checking.

Outliers are points for which the observed responses are strange in relation to the model.

Outliers can be deleted only when there are grounds for doing so (e.g., very unusual known causes).

Outliers can be detected by identifying high standardized residuals or by identifying points which change the results (e.g., the slope of a line) markedly when they are ignored.

Unusual principal-component scores for the unimportant predictor principal components are associated with outliers.

Serious multicollinearity is present when the ratio of the largest eigenvalue to the smallest eigenvalue exceeds 30.

Serious multicollinearity causes unreliable results and, in particular, high estimated coefficient variances.

Multicollinearity can be overcome by combining or removing predictors or by performing a ridge regression.

A weighted regression is necessary whenever the response variances differ for each observation. The weights are defined as the inverse of these variances.

1.5 DATA SETS

1.5.1 Cement Data

The amount of heat evolved in cement is to be modeled in terms of the amount of four ingredients in the mix. The heat evolved (y) is measured in calories per gram of cement and the four ingredient amounts are (see, table at the top of pg. 294)

x_0 = constant term

x_1 = grams of tricalcium aluminate

x_2 = grams of tricalcium silicate

x_3 = grams of tetracalcium alumino ferrite

x_4 = grams of dicalcium silicate

1.5.2 Safety Program Data

In a study of the effect of company size (measured using the number of employees) on the number of work hours lost due to work-related accidents, the presence or absence of a safety program is thought to be important. The two groups have been coded using the dummy variable "Dum." For companies with a safety

		Matrix X					Regression for y in terms of x_1		Regression for y in terms of x_1 and x_2	
	y	x_0	x_1	x_2	x_3	x_4	FITS1	RES1	FITS2	RES12
	78.5	1	7	26	6	60	94.561	−16.0606	80.074	−1.57400
	74.3	1	1	29	15	52	83.348	−9.0481	73.251	1.04908
	104.3	1	11	56	8	20	102.036	2.2644	105.815	−1.51474
	87.6	1	11	31	18	47	102.036	−14.4356	89.258	−1.65848
	95.9	1	7	52	6	33	94.561	1.3394	97.293	−1.29251
	109.2	1	11	55	9	22	102.036	7.1644	105.152	4.04751
	102.7	1	3	71	17	6	87.086	15.6144	104.002	−1.30205
	72.5	1	1	31	22	44	83.348	−10.8481	74.575	−2.07542
	93.1	1	2	54	18	22	85.217	7.8832	91.275	1.82451
	115.9	1	21	47	4	26	120.723	−4.8230	114.538	1.36246
	83.8	1	1	40	23	34	83.348	0.4519	80.536	3.26433
	113.3	1	11	66	9	12	102.036	11.2644	112.437	0.86276
	109.4	1	10	68	8	12	100.167	9.2332	112.293	−2.89344
Mean	95.4	1	7.5	48.2	11.8	30	95.4	0	95.4	0

program Dum = 1 and for companies without a safety program Dum = 0. The term DumEmp is defined as the product of Dum and the predictor variable "Employees".

1.5.3 Experimental Design Data Set

A three-factor central composite design was used to determine the effect of three predictor variables on the precipitation (y) of a chemical (stoichiometric dihydrate of calcium hydrogen orthophosphate). The three predictor variables were the mole ratio of ammonia to calcium chloride in the calcium chloride solution, time in minutes (t), and the starting pH of the base solution. In order to simplify the experimental design the predictor variables were defined as scaled versions of the original variables, as indicated below.

Employees	Dum	Hours lost	DumEmp	Employees	Dum	Hours lost	DumEmp
6490	0	121	0	3077	1	44	3077
7244	0	169	0	6600	1	73	6600
7943	0	172	0	2732	1	8	2732
6478	0	116	0	7014	1	90	7014
3138	0	53	0	8321	1	71	8321
8747	0	177	0	2422	1	37	2422
2020	0	31	0	9581	1	111	9581
4090	0	94	0	9326	1	89	9326
3230	0	72	0	6818	1	72	6818
8786	0	171	0	4831	1	35	4831
1986	0	23	0	9630	1	86	9630
9653	0	177	0	2905	1	40	2905
9429	0	178	0	6308	1	44	6308
2782	0	65	0	1908	1	36	1908
8444	0	146	0	8542	1	78	8542
6316	0	129	0	4750	1	47	4750
2363	0	40	0				

x_1	x_2	x_3	y
−1	−1	−1	52.8
1	−1	−1	67.9
−1	1	−1	55.4
1	1	−1	64.2
−1	−1	1	75.1
1	−1	1	81.6
−1	1	1	73.8
1	1	1	79.5
−1.6818	0	0	68.1
1.6818	0	0	91.2
0	−1.6818	0	80.6
0	1.6818	0	77.5
0	0	−1.6818	36.8
0	0	1.6818	78.0
0	0	0	74.6
0	0	0	75.9
0	0	0	76.9
0	0	0	72.3
0	0	0	75.9
0	0	0	79.8

1.5.4 Value of Postage Stamp

The value of an Australian stamp (1963 £2 saffron colored) in pounds sterling is shown over a 9-year period.

Year	1972	1973	1974	1975	1976
Coded year	1	2	3	4	5
Value of stamp (£)	10	12	12	22	25
Year	1977	1978	1979	1980	
Coded year	6	7	8	9	
Value of stamp (£)	45	75	95	120	

REFERENCES

1. RJ Brook, GC Arnold. Applied Regression Analysis and Experimental Design. New York: Marcel Dekker, 1985.
2. CP Mallows. Some comments on C_p. Technometrics 15: 661–675, 1973.
3. NR Draper, H Smith. Applied Regression Analysis. New York: John Wiley, 1981.
4. DC Hoaglin, RE Welsch. The hat matrix in regression and ANOVA. Am Statist 32: 17–22, 1978.
5. AE Hoerl, RW Kennard. Ridge regression. Biased estimation for non-orthogonal problems. Technometrics 12: 55–67, 1970.

Chapter 4.2

A Brief Introduction to Linear and Dynamic Programming

Richard B. Darst
Colorado State University, Fort Collins, Colorado

2.1 LINEAR PROGRAMMING

2.1.1 Introduction

Linear programming is a method for dealing with an exceptionally diverse class of situations. A situation is usually presented in the form of a moderately amorphous mass of information by a group of one or more people who wish to do something. To deal with a situation effectively, you need to understand the situation, so you know what the group wants to do. Then you need to analyze the situation, so you can tell the group how to do what it wants to do. Typically the goal is to make decisions to optimize an objective while satisfying constraints. To understand a situation you need to formulate an accurate conceptual model for the situation; generally this encompasses stating an appropriate set of decisions that must be made, together with the objective to be optimized and the constraints which must be satisfied. The corresponding analysis is typically a process to determine which set of values for the decisions will optimize the objective while satisfying the constraints; it often involves formulating an appropriate mathematical model for the situation, solving the model and interpreting the implications of the model for the situation.

Linear programming (LP), is an extremely effective method for formulating both the conceptual and mathematical models. Good LP software is available to solve the problem if you can formulate a valid LP model of reasonable size. Linear programming encompasses several activities. It includes recognizing whether or not an LP model is appropriate for a specific situation. An LP model is both a conceptual and a mathematical model for the situation. If an LP model is appropriate, LP includes formulating an accurate LP model (also called simply an LP), solving the LP, and applying the solution to the LP to the situation.

Formulating an LP model for a situation involves three basic steps:

1. Specify the decision variables.
2. Specify the constraints.
3. Specify the objective function.

Quantifying an appropriate set of decisions is the crucial first step to understanding a situation and to formulating a valid model for the situation; it addresses the question: What decisions need to be made in order to optimize the objective? The first step in formulating an LP model for a situation helps you understand the situation by requiring you to specify an appropriate set of decision variables for the model. After the decision variables are specified, one tries to specify the constraints and objective function as linear functions of the decision variables. Either you get an LP model or you discover where essential nonlinearities appear to be.

Linear programming has its theoretical side too, with algebraic and geometrical components, and various algorithms for solving LP models. However, the common procedures are variations of the simplex method discovered by George Dantzig. The simplex method uses elementary linear algebra to exploit the geometrical and algebraic structure of an LP; it is elegant and efficient. Learning the rudiments of the simplex method will enable you to utilize output from LP software packages more effectively. In addition to giving you a solution to your LP, or telling you that there is no solution, the software may provide you with additional information that can be very helpful if you know how to use it. Theoretical concepts are useful in LP modeling. While theory will not be discussed in this chapter, some theoretical concepts will appear during our discussions of the examples which are presented in the sequel.

Sometimes you will discover that the problem has a special structure which permits you to formulate a different type of model for the problem, one that exploits the special structure more efficiently. Transportation problems have a special structure which has been studied from several points of view. Transportation problems provide a good bridge between LP and network models. Network models are associated with a diverse class of methods. After using examples to introduce several types of situations to which LP can be applied, we will apply dynamic programming to some situations to illustrate how special structure and conditions can be exploited.

Programming has artistic and technical aspects, and like learning to paint or play a game, one learns to program by doing it. Reading can help you get started, but you need to participate to become good at it. As you participate you will be able to model an increasing diverse collection of situations effectively. An effective model is valid, it can be solved using software that is available to you, and it will provide clear usable output with which to make your decisions.

Finding an optimal solution to a complicated problem may well depend on recognizing whether LP or some other special structure can be used to model your situation accurately. Often, the first part of a solution is to determine whether there is a valid LP model for your situation. If you decide that a valid LP model can be constructed, then either find special structure that will let you to use a significantly faster algorithm or construct an efficient LP model. Otherwise, look for special structure in the situation which will permit you to use some known modeling strategy.

Examples of types of situations which can be modeled using LP will be discussed below. The examples begin with some information being provided, followed by an LP model and a brief discussion of the situation. Often there are several valid LP models for a situation. If you model the same situation on different days, even your decision variables may be quite different because you may be emphasizing different aspects of the situation. A different focus may well result in a model with different constraints and a different objective function. You may wish to formulate and solve your own models for these situations; do not be surprised if what you get differs from what you see here.

The basic ingredients of linear programming are introduced in the first three examples.

2.1.2 Examples 1–3

Example 1 A Production Problem

Information Provided. An organization has an abundance of two types of crude oil, called light crude and dark crude. It also has a refinery in which it can process light crude for 25 dollars per barrel and dark crude for 17 dollars per barrel. Processing yields fuel oil, gasoline, and jet fuel as indicated in the table below.

	Input	
Output	Light crude	Dark crude
Fuel oil	0.21	0.55
Gasoline	0.5	0.3
Jet fuel	0.25	0.1

The organization requries 3 million barrels of fuel oil, 7 million barrels of gasoline, and 5 million barrels of jet fuel.

Understanding the Situation: Interpreting the Information Provided. Part of programming is interpreting what people are trying to tell you and feeding back your impressions to those people until you believe you understand what they are trying to tell you. For example, the entry 0.21 in this table seems to imply that each unit of light crude oil which you process produces 0.21 units of fuel oil. If the objective is to process enough oil to meet the listed demands at minimum processing cost, then we wish to determine how much light and dark crude should be processed to meet the requirements with minimal processing cost. To begin recall the three steps in formulating an LP model:

1. Specify the decision variables.
2. Specify the constraints.
3. Specify the objective function.

Decision Variables. We must decide how much light crude oil to process and how much dark crude oil to process. A set of decision variables for Example 1 follows.

Let $L =$ the number of million barrels of light crude to process.
Let $D =$ the number of million barrels of dark crude to process.

The values of decision variables are numbers. *I cannot overemphasize the importance of doing Step 1.* Specifying the decision variables as numerical valued variables forces you to focus on exactly what decisions you need to make.

The next two steps are to specify a set of *constraints to model the requirements* and to specify an appropriate *objective function to model the cost*, which we wish to minimize. The form and order of these specifications depend on our choice of decision variables and our thought processes at the time when we are making the model. In this example we consider the constraints first.

Constraints. I interpreted the table to imply that each million barrels of light crude processed produces 0.21 million barrels of fuel oil, etc. Thus, processing L million barrels of light crude and D million barrels of dark crude produces $0.21L + 0.55D$ barrels of fuel oil. Since 3 million barrels are required, we have the constraint $0.21L + 0.55D = 3$; this constraint is a linear equation in L and D. Similar equations for gasoline and jet fuel apply. Putting these equations together gives us the linear system

$$\begin{aligned} 0.21L + 0.55D &= 3 \qquad \text{(fuel oil)} \\ 0.50L + 0.30D &= 7 \qquad \text{(gasoline)} \\ 0.25L + 0.10D &= 5 \qquad \text{(jet fuel)} \end{aligned}$$

Necessity of Inequality Constraints. Unfortunately, because processing either type of crude oil produces at least twice as much gasoline as jet fuel, there is no way to produce 5 million barrels of jet fuel without producing at least 10 million barrels of gasoline. Thus, there is no feasible solution to this linear system. Consequently, we are forced to formulate our constraints as a system of linear inequalities:

$$\begin{aligned} 0.21L + 0.55D &\geq 3 \\ 0.50L + 0.30D &\geq 7 \\ 0.25L + 0.10D &\geq 5 \end{aligned}$$

Standard Form for Constraints. I would prefer to have the constraints in the form of a system of linear equations because I know how to solve systems of linear equations, so we put them in that form by introducing surplus variables F, G, and J to denote the number of millions of barrels of surplus fuel oil, gasoline, and jet fuel produced. The preceding system of linear inequalities is replaced by the following system of linear equations:

$$\begin{aligned} 0.21L + 0.55D &= 3 + F \\ 0.50L + 0.30D &= 7 + G \\ 0.25L + 0.10D &= 5 + J \end{aligned}$$

which can be rewritten in *standard form*:

$$\begin{aligned} 0.21L + 0.55D - F &= 3 \\ 0.50L + 0.30D - G &= 7 \\ 0.25L + 0.10D - J &= 5 \end{aligned}$$

Objective Function. The cost of processing L million barrels of light crude and D million barrels of dark crude is $25L + 17D$ millions of dollars. The objective function for this problem is the linear function of L and D given by the formula

$$25L + 17D$$

We put the pieces together in a standard form LP model below.

Standard Form LP Model

$$\begin{aligned} \text{Minimize} \quad & 25L + 17D \\ \text{subject to} \quad & 0.21L + 0.55D - F = 5 \\ & 0.50L + 0.30D - G = 7 \\ & 0.25L + 0.10D - J = 5 \end{aligned}$$

A standard form LP has the form: minimize a linear function of the decision variables subject to a system of linear equations being satisfied by the decision variables, plus *implicit constraints* which require that the decision variables be *nonnegative*. This is the form that is processed by the simplex method.

Note: Unless it is specifically stated otherwise, the values of all decision variables are nonnegative numbers.

Canonical Form. Some authors use the name canonical LP, or canonical form LP, to denote an LP which we have called a standard form LP. Sometimes standard form or canonical form refers to a max LP with equality constraints, so be aware of these variations in terminology when you read about or discuss LP.

After a comment about vectors and matrices, elementary concepts from linear algebra, we will use them to write the model for Example 1 in a very compact form which displays its structure.

Comments about Vectors and Matrices

A vector is simply a list of numbers. This simple concept is extremely useful both conceptually and computationally. Imagine a cash register drawer with eight compartments, labelled 1 to 8, for pennies, nickels, dimes, quarters, dollar bills, 5 dollar bills, 10 dollar bills, and 20 dollar bills, respectively. Suppose that I count the number of pieces of money in each compartment and write those numbers in a list. If that list is $(3, 5, 2, 4, 11, 4, 1, 6)$, then how much money is in the drawer? If you trust my counting and your answer is \$162.48, then you understand how vectors work.

We can either write the list as a row of numbers: a row vector, denoted [...], or as a column of numbers: a column vector. Column vectors take more space on a page, so we generally use a row format using parentheses (...), to denote column vectors; for example, the preceding vector, representing a list of the numbers of the various items in the drawer, really looks like

$$\begin{bmatrix} 3 \\ 5 \\ 2 \\ 4 \\ 11 \\ 4 \\ 1 \\ 6 \end{bmatrix}$$

A matrix is a rectangular array of numbers. We can think of a matrix as a row of column vectors or as a column of row vectors when it is convenient to do so.

We will illustrate how one can use vectors and matrices to formulate LPs in a very compact form by discussing Example 1 from a vector–matrix perspective below.

Vector–Matrix Formulation of Example 1. The unit output vectors

$$\begin{bmatrix} 0.21 \\ 0.5 \\ 0.25 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0.55 \\ 0.3 \\ 0.1 \end{bmatrix}$$

provide lists of the number of units of fuel oil, gasoline, and jet fuel produced by processing one unit of light and dark crude oil, respectively.

Let A denote the 3×2 matrix whose columns are the unit output vectors of light and dark crude oil, let b denote the column vector of requirements, let c denote the row vector of unit processing costs, and let x denote the column vector of decision variables:

$$A = \begin{bmatrix} 0.21 & 0.55 \\ 0.5 & 0.3 \\ 0.25 & 0.1 \end{bmatrix} \qquad b = \begin{bmatrix} 3 \\ 7 \\ 5 \end{bmatrix} \qquad c = [25, 17]$$

$$x = \begin{bmatrix} L \\ D \end{bmatrix}$$

Then the LP model for Example 1 can be written as follows:

$$\begin{array}{ll} \text{Minimize} & cx \\ \text{subject to} & Ax \geq b, x \geq 0 \end{array}$$

Example 2 A Generic Diet Problem

Information Provided. A large institution wishes to formulate a diet to meet a given set of nutritional requirements at minimal cost. Suppose that n foods are available and m nutritional components of the foods are to be considered (each of n and m denotes an arbitrary but fixed positive integer determined by the individual situation). Label the foods $f_1, \ldots, f_n$ and label the nutrients $N_1, \ldots, N_m$. The costs and nutritional characteristics of the foods, and the nutritional requirements of the diet, are provided in the following format. *A* unit of f_j costs c_j money units and contains a_{ij} units of N_i; at least b_i units of N_i are required for the diet. Use summation notation to formulate a compact LP model for the diet problem.

Summation Notation: a Compact Way to Write the Sum of n Quantities. Given n quantities $Q_1, Q_2, \ldots, Q_i, \ldots, Q_n$ which can be added, their sum $Q_1 + Q_2 + \cdots + Q_i + \cdots + Q_n$ is denoted by $\sum_{i=1}^{n} Q_i$. The index i is a symbol which represents a generic integer between 1 and n; instead of i, one can use j, or any other symbol except n, which represents the number of terms to be added.

Solution. The n decisions are how many units of each of the foods to put in the diet. The constraints are the m nutritional requirements that must be satisfied, and the objective is to find a minimal cost combination of available foods which meets the nutritional requirements. Let x denote the column vector $(x_1, \ldots, x_i, \ldots, x_n)$, let b denote the column vector $(b_1, \ldots, b_i, \ldots, b_m)$, let c denote the row vector $[c_1, \ldots, c_i, \ldots, c_n]$, and let A denote the $m \times n$ matrix which has a_{ij} in its ith row and jth column. Then the LP model for Example 2 has the form

$$\begin{array}{ll} \text{Minimize} & cx \\ \text{subject to} & Ax \geq b, x \geq 0 \end{array}$$

A numerical example follows.

Example 2a "What's for Lunch?" Suppose seven foods: green beans, soybeans, cottage cheese, Twinkies, vegetable juice, spaghetti and veggie supreme cheese pizza, are available. Suppose that units of these foods cost 0.35, 0.2, 0.28, 0.2, 0.15, 0.99, and 2.49 dollars, respectively. Suppose that four properties of each food: calories, protein, carbohydrates, and vitamins, are required to be considered, and the following properties matrix A and requirements vector b are provided:

$$A = \begin{bmatrix} 1 & 3 & 3 & 6 & 1 & 14 & 16 \\ 0 & 1 & 4 & 0 & 0 & 8 & 8 \\ 1 & 2 & 3 & 6 & 0 & 14 & 14 \\ 6 & 2 & 2 & 0 & 4 & 7 & 13 \end{bmatrix}$$

$$b = \begin{bmatrix} 16 \\ 8 \\ 14 \\ 13 \end{bmatrix}$$

$$c = [0.35, 0.2, 0.28, 0.2, 0.15, 0.99, 2.49]$$

The four numbers listed in a column of the matrix represent the numbers of units of the four properties in one unit of the food corresponding to that column; for instance, column three corresponds to cottage cheese, so each unit of cottage cheese contains three units of calories, four units of protein, three units of carbohydrates, and two units of vitamins.

A unit of pizza has been designed to meet the dietary requirements exactly. A meal composed of one unit of green beans, two units of soybeans, 3/2 units of cottage cheese, and 3/4 units of Twinkie also fits the minimal requirements of the diet exactly; this meal costs \$1.32, while pizza costs \$2.49. Notice that a meal composed of one unit of spaghetti and two units of vegetable juice will also meet the requirements, at a cost of \$1.29 (this meal provides two extra units of vitamins). Which of these three meals would you choose?

The transportation problem is another classical example; a standard form LP model for it is developed in Example 3.

Example 3 A Transportation Problem

The goal of a transportation problem is to ship quantities of a material from a set of supply points to a set of demand points at minimal cost, A model for a transportation problem consists of supplying *five lists*: a list of supply points, a list of demand points, a list of the numbers of units available at each supply point, a list of the numbers of units desired at each demand point, and a list of the costs of shipping a unit of material from each supply point to each demand point, and *an a priori constraint: total supply is equal to total demand.* These five lists will be displayed in a transportation tableau. The tableau for an example which deals with shipping truckloads of blueberries from orchards to warehouses follows.

	Warehouses				
Orchards	California	Arizona	Colorado	New Mexico	Supplies
Washington	460	550	650	720	100
Oregon	350	450	560	620	170
Michigan	990	920	500	540	120
Demands	175	100	80	35	390

Total supply = total demand = 390. There are 12 decisions to be made; we must decide how many truckloads of blueberries to ship from each of the three supply points to each of the four demand points. We replace the unit shipping costs with a list of names for the decision variables below.

	Warehouses				
Orchards	California	Arizona	Colorado	New Mexico	Supplies
Washington	$X1$	$X2$	$X3$	$X4$	100
Oregon	$X5$	$X6$	$X7$	$X8$	170
Michigan	$X9$	$X10$	$X11$	$X12$	120
Demands	175	100	80	35	

There are three supply constraints and four demand constraints:

$$\begin{aligned}
&X1 + X2 + X3 + X4 = 100 && \text{(Washington)} \\
&X5 + X6 + X7 + X8 = 170 && \text{(Oregon)} \\
&X9 + X10 + X11 + X12 = 120 && \text{(Michigan)} \\
&X1 + X5 + X9 = 175 && \text{(California)} \\
&X2 + X6 + X10 = 100 && \text{(Arizona)} \\
&X3 + X7 + X11 = 80 && \text{(Colorado)} \\
&X4 + X8 + X12 = 35 && \text{(New Mexico)}
\end{aligned}$$

Put

$x = (X1, \ldots, X12)$

$b = (100, 170, 120, 175, 100, 80, 35)$

$c = [460, 550, 650, 720, 350, 450, 560, 620, 990, 920, 500, 540]$

and

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Then the blueberry example has the standard form LP model displayed below.

$$\begin{array}{ll} \text{Minimize} & \text{cx} \\ \text{Subject to} & Ax = \text{b}, x \geq 0 \end{array}$$

2.1.3 Duality

An LP has a specific algebraic form. There is a corresponding algebraic form, called the dual LP. The original problem is called the primal problem, and the primal together with the dual is called a primal–dual pair; the primal is the first LP in the pair and the dual LP is the second LP in the pair. When an LP is a model of a "real-world" situation, very often there is a different (dual) perspective of the situation which is modeled by the dual LP. Knowing the existence and form of the dual LP provides a vantage point from which to look for a dual interpretation of the situation; examples are provided below by exhibiting dual linear programs for Examples 1 and 3.

2.1.3.1 Dual Problem for Example 1

A related problem, called the dual problem, will be introduced by considering Example 1 from a different perspective below.

A group has ample supplies of fuel oil, gasoline, and jet fuel which they would like to sell to the organization. The group wishes to maximize income from selling 3 million barrels of fuel oil, 7 million barrels of gasoline and 5 million barrels of jet fuel to the organization. The group's decisions are to determine the numbers of millions of dollars (prices), pF, pG, and pJ, to charge for a million barrels of fuel oil, gasoline, and jet fuel, respectively. For algebraic reasons we write their decision variable vector in row form $y = [pF, pG, pJ]$. Then the group's objective can be expressed algebraically in the form: maximize yb, where A and b are defined in Example 1. For the group to be competitive, the price for the output of processing a million barrels of light crude can be no more than the processing cost; this constraint has the algebraic form $0.21pF + 0.5pG + 0.25pJ \leq 25$. The corresponding constraint for dark crude is $0.55pF + 0.3pG + 0.1pJ \leq 17$. These constraints can be written in vector–matrix form as $yA \leq c, y \geq 0$. Thus, the dual can be written

$$\begin{array}{ll} \text{Maximize} & yb \\ \text{Subject to} & yA \leq c, y \geq 0 \end{array}$$

2.1.3.2 Comment on Standard form

By introducing slack variables, surplus variables, and other appropriate modifications, any LP can be put in standard form. For instance, Example 1 showed how to write $\geq$ constraints in standard form by introducing surplus variables. The dual constraints introduced above can be written in standard form by introducing nonnegative *slack variables*, sL and sD:

$$0.21pF + 0.5pG + 0.25pJ + sL = 25$$
$$0.55pF + 0.3pG + 0.1pJ + sD = 17$$

The objective "Maximize yb" is equivalent to "Minimize $y(-b)$."

2.1.3.3 Dual Problem for Example 3

A wholesale produce corporation wishes to buy the blueberries at the orchards and deliver the required amounts at the demand points. Their objective is to maximize net income from this transaction. Their decisions are prices to charge per truckload, but the prices are not all ≥ 0 because buying blueberries at the orchards represents negative income. Referring to Example 3, put $y = [y1, y2, y3, y4, y5, y6, y7]$, where $y1, y2$, and $y3$ are ≤ 0 and the others are ≥ 0. The net income to be maximized is yb. The constraints are that the buying price at an orchard plus the selling price at a state is no more than the cost of shipping a truckload from the orchard to the state; these 12 constraints are written algebraically as $yA \leq c$. Thus, the dual problem can be written

$$\begin{array}{ll} \text{Maximize} & yb \\ \text{Subject to} & yA \leq c \end{array}$$

We do not have the constraints $y \geq 0$ in the dual of a standard form LP.

2.1.4 Absolute Values and Integer Variables

Before presenting more examples to illustrate a variety of situations which can be modeled using LP, we mention two useful concepts, absolute values and integer variables.

2.1.4.1 Absolute Values

The absolute value of a number is the magnitude of the difference between the number and zero. The absolute value, $|X|$, of a number X can be expressed as the solution of an LP:

$$|X| = \text{minimum } P$$
$$\text{Subject to} \qquad -P \leq X \leq P, \qquad P \geq 0.$$

or

$$|X| = \text{minimum } P + N$$
$$\text{Subject to} \qquad P - N = X, \qquad P \geq 0, N \geq 0$$

2.1.4.2 Integer Variables

Realistic solutions to situations must often be integer valued. An LP model does not require integer values for solutions; however, in many cases either the LP solution turns out to be integer valued, or one can use the LP solution to guess an integer valued solution that is good enough. The definition of "good enough" depends the tools, time, and creativity available to apply to the problem. Problems that were intractable for a PC a few years ago can be solved easily now. Computer hardware and software that is available for a reasonable price has been getting faster and better at a rapid pace. Some problems that had to be modeled carefully in order to be solvable with available tools a few years ago can be solved using a sloppy model now. However, good modeling habits are worth cultivating because they enable you to deal effectively with a larger set of situations at any point in time.

There are two kinds of integer-value variables, *INT variables* and *GIN variables*, which are often available in LP software. Each INT variable may essentially double the computational complexity of a LP model and each GIN variable may increase it at least that much. Consequently, they should be used carefully. Nevertheless, they can be very useful.

INT variables (sometimes called 0–1-valued integer variables) are variables whose values are either 0 or 1. An INT variable can be used as a switch to switch between two possible situations. The INT variable insures that at most one of the possibilities occurs in any one possible solution.

For example, given a positive number M, the absolute value of X in a range from $-M$ to M can also be expressed using an INT variable, Z, as follows:

$$|X| = P + N \quad P \leq MZ, \quad N \leq M(1 - Z)$$
$$P \geq 0 \quad N \geq 0$$

The inequalities involving Z require that at least one of P and N be equal to zero. In some situations it is useful to consider "big M" constants, denoted by M; "big M's" are constants which are big enough to impose no new constraint on a problem in which they are used. We illustrate this with the following example.

Constraints of the form

$$a|X| + y \leq, =, \text{ or } \geq c$$

where each of X and y is a linear term and each of a and c is a constant, can be put in LP format. If $a = 0$, we have an LP constraint; otherwise, there are two cases to consider, $a > 0$ and $a < 0$. In both cases, we divide the constraint by a. If $a > 0$ we get $|X| + (1/a)y \leq, =,$ or $\geq c/a$. Putting $Y = (1/a)y$ and $C = c/a$, the constraint becomes

$$|X| + Y \leq, =, \text{ or } \geq C$$

If $a < 0$, the inequalities are reversed. Consequently, we have three possibilities to consider. The $\leq$ case is the easiest. The constraint $|X| + Y \leq C$ is equivalent to the following two LP constraints

$$u + v + Y \leq C \tag{1}$$

and

$$X = u - v \tag{2}$$

To model the other two cases, $|X| + Y =$ or $\geq C$, in LP format, in addition to (2) and the appropriate modification of (1) above, we use an INT variable Z and a "big M" in the following additional constraints:

$$u \leq MZ \tag{3}$$

and

$$v \leq M(1 - Z) \tag{4}$$

These latter two constraints that at least one of u and v be zero, consequently $|X| = u + v$.

General integer (GIN) variables are variables whose values are integers. INT and GIN variables will be used in some of the examples.

Example 4 A Primal-Dual Pair

Information Provided. The Volkswagen Company produces three products: the bug, the superbug, and the van. The profit from each bug, superbug, and van is \$1000, \$1500, and \$2000, respectively. It takes 15, 18, and 20 labor-hours to produce an engine; 15, 19, and 30 labor-hours to produce a body; and 10, 20, and 25 minutes to assemble a bug, superbug, and van. The engine works has 10,000 labor-hours available, the body works has 15,000 labor-hours available, and the assembly line has 168 hours available each week. Plan weekly production to maximize profit.

Solution

Let $B =$ number of bugs produced per week.
Let $S =$ number of superbugs produced per week.
Let $V =$ number of vans produced per week.

An LP model and solution for Example 4 follow.

```
Maximize 1000 B + 1500 S + 2000 V
Subject to
    2) 15 B + 18 S + 20 V <= 10000
    3) 15 B + 19 S + 30 V <= 15000
    4) 10 B + 20 S + 25 V <= 10080

OBJECTIVE FUNCTION VALUE 861714.300

  VARIABLE   VALUE
    B 276.571400
    S  .000000
    V 292.571400

  ROW  SLACK OR SURPLUS
    2)  .000000
    3) 2074.286000
    4)  .000000
```

The solution says that all the engine works and assembly time is used, but about 2074 labor-hours of body works time is not used.

The values of the decision variables in the LP solution are not all integers. However, the feasible, integer valued solution $B = 276$, $S = 1$, and $V = 292$ has objective function value equal to 861,500. Thus, this solution is optimal because its value, 861,500, is the largest multiple of 500 which is $\leq 861{,}714.3$, the LP optimum.

Now we introduce a dual problem for Example 4: demographics change with time and BMW Corp. decides to increase production. BMW decides to approach Volkswagen about leasing their plant. They want to determine the smallest offer that Volkswagen might accept. A model for their situation follows.

Let $E =$ no. of dollars to offer per hour for engine works time.
Let $W =$ no. of dollars to offer per hour for body works time.
Let $A =$ no. of dollars to offer per hour for assembly time.

```
Minimize  10000 E + 15000 W + 10080 A
Subject to
  2) 15 E + 15 W + 10 A >= 1000
  3) 18 E + 19 W + 20 A >= 1500
  4) 20 E + 30 W + 25 A >= 2000
```

The constraints for this dual problem say that the number of dollars offered to pay to rent must be enough to make Volkswagen's rental income at least as much as the profit it would get by using its resources to manufacture vehicles instead of renting them to BMW. It turns out that the optimal objective function values for the primal and its dual are always equal. A solution follows.

```
VARIABLE   VALUE
   E   28.571430
   W    .000000
   A   57.142860
```

Example 5 Percentage Constraints

Information Provided. A farmer requires that each of his cows receives between 16,000 and 18,000 calories, at least 2 kg of protein, and at least 3 g of vitamins per day. Three kinds of feed are available; the following table lists their relevant characteristics per kilogram.

Feed	Cost	Calories	Kg. of protein	Grams of vitamins
1	\$0.8	3600	0.25	0.7
2	\$0.6	2000	0.35	0.4
3	\$0.2	1600	0.15	0.25

The farmer also requires that the mix of feeds contain at least 20% (by weight) feed 1 and at most 50% (by weight) feed 3. The farmer wishes to formulate a diet which meets his requirements at minimum cost.

Solution

Let $A =$ no. of kilograms of feed 1 to put in the diet.
Let $B =$ no. of kilograms of feed 2 to put in the diet.
Let $C =$ no. of kilograms of feed 3 to put in the diet.

The first four constraints correspsond to the calorie, protein and vitamin requirements of the diet. The last two correspond to the percentage requirements: $A \geq 0.2$ $(A + B + C)$ and $C \leq 0.5(A + B + C)$. A model appears below, followed by a solution. The constraints have been adjusted to make the coefficients integers.

```
Minimize  8 A + 6 B + 2 C
Subject to
   2) 36 A + 20 B + 16 C >= 160
   3) 36 A + 20 B + 16 C <= 180
   4) 25 A + 35 B + 15 C >= 200
   5) 70 A + 40 B + 25 C >= 300
   6) 8 A - 2 B - 2 C >= 0
   7) 5 A + 5 B - 5 C >= 0

OBJECTIVE FUNCTION VALUE 38.3132600

VARIABLE  VALUE
   A  1.686747
   B  2.831325
   C  3.915662
```

Example 6 A Processing Problem

Information Provided. Fruit can be dried in a dryer according to the following table. The dryer can hold 1 m^3 of fruit.

	Relative volume		
Drying hours	Grapes	Apricots	Plums
1	0.30	0.46	0.53
2	0.23	0.44	0.51
3	0.16	0.34	0.47
4	0.13	0.31	0.42

Formulate an LP to estimate the minimum time in which 20 m^3 of grapes, 10 m^3 of apricots, and 5 m^3 of plums can be dried to a volume of no more than 10 m^3.

Solution. We begin by making some simplifying assumptions. The dryer will be filled with one kind of fruit and operated for 1, 2, 3, or 4 hours, then the dried fruit will be removed from the dryer and the dryer will be refilled. In accord with these assumptions, we have the following decision variables:

Let $GI =$ no. of cubic meters of grapes to dry for I hours.

Let $AI =$ no. of cubic meters of apricots to dry for I hours.

Let $PI =$ no. of cubic meters of plums to dry for I hours.

Time spent filling the dryer and removing dried fruit is assumed to be independent of the type of fruit being dried and the number of hours the fruit is dried. Then these factors do not need to be explicitly incorporated into the model.

```
Minimize G1 + 2 G2 +3 G3 + 4 G4 + A1 + 2 A2
+3 A3 + 4 A4 + P1 + 2 P2 + 3 P3 + 4 P4
Subject to
   2) G1 + G2 + G3 + G4 = 20
   3) A1 + A2 + A3 + A4 = 10
   4) P1 + P2 + P3 + P4 = 5
   5) .3 G1 + .23 G2 + .16 G3 + .13 G4 + .46
      A1 + .44 A2 + .34 A3 + .31 A4 + .53 P1
      + .51 P2 + .47 P3 + .42 P4 ≤ 10

OBJECTIVE FUNCTION VALUE 82.4999900

VARIABLE  VALUE
   G1      .000000
   G2      .000000
   G3    20.000000
   G4      .000000
   A1     6.250004
   A2      .000000
   A3     3.749996
   A4      .000000
   P1     5.000000
   P2      .000000
   P3      .000000
   P4      .000000
```

$A1$ and $A3$ are not integer valued. However, if we change $A1$ to 6 and $A3$ to 4, then we get an integer-valued feasible solution with objective function value of 83. The objective function value of any integer valued solution will be an integer. Since 83 is the smallest integer which is ≥ 82.5, we know that changing $A1$ to 6 and $A3$ to 4 provides use with an optimal integer valued solution.

Example 7 A Packaging Problem

Information Provided. The Brite-Lite Company receives an order for 78 floor lamps, 198 dresser lamps, and 214 table lamps from Condoski Corp. Brite-Lite ships orders in two types of containers. The first costs $15 and can hold two floor lamps and two table lamps or two floor lamps and two table lamps and four dresser lamps. The second type costs $25 and can hold three floor lamps and eight table lamps or eight table lamps and 12 dresser lamps. Minimize the cost of a set of containers to hold the order.

Solution

Let $CIJ =$ no. of containers of type I to pack with mix $J,\ I = 1,\ 2,\ J = 1,\ 2$

```
Minimize  15 C11 + 15 C12 + 25 C21 + 25 C22
Subject to
   2) 2 C11+ 2 C12 + 3 C21 ≥ 78
   3) 4 C11 + 2 C12 + 8 C21 + 8 C22 ≥ 214
   4) 4 C12 + 12 C22 ≥ 198

OBJECTIVE FUNCTION VALUE 852.500000

VARIABLE  VALUE
   C11      .000000
   C12    21.000000
   C21    12.000000
   C22     9.500000
```

This solution is not integer valued; however, it puts six dresser lamps in the half carton of type 2 packed with mix 2 and no other listed packing option will allow six dresser lamps in one carton. Consequently, I could put $C22 = 10$, increase the cost to 865, and claim that I now have the optimal solution. But claiming does not necessarily make it optimal. Is there a better solution? Using GIN variables, I found a better solution. The best integer-valued solution is $C11 = 4$, $C12 = 20$, $C21 = 10$ and $C22 = 10$, with objective function value 860; 5 may seem to be a trivial number of dollars, but if we are shipping different products and we are talking about hundreds of thousands of dollars, the difference might be significant to you.

Example 8 LP Can Model Diminishing Return Thresholds

Information Provided. The Model-Kit Company makes two types of kits. They have 215 engine assemblies, 525 axle assemblies, 440 balsa blocks, and 560 color packets in stock. Their earthmover kit contains two engine assemblies, three axle assemblies, four balsa blocks, and two color kits; its profit is $15. Their racing kit contains one engine assembly, two axle assemblies, two balsa blocks, and three color kits; its profit is $9.50. Sales have been slow and Sears offers to buy all that Model-Kit can supply using components in stock if Model-Kit will sell all earthmover kits over 60 at a $5 discount and all racing kits over 50 at a $3 discount. Determine which mix of model kits to sell to Sears to maximize profit using components in stock.

Solution. The numbers 60 and 50 are thresholds for earthmover and racing kits. *As production increases across the threshold, the return per unit diminishes.*

Let E and R denote the number of earthmover and racing kits to sell to Sears. Let $E1$ denote the number of earthmover kits to be sold at the regular price and let $E2$ denote the number to be sold for $5 less: $E2 = E - E1$. Let $R1$ and $R2$ play similar roles for racing kits.

```
Maximize  15 E1 + 10 E2 + 9.5 R1 + 6.5 R2
Subject to
  2) 2 E + R <= 215
  3) 3 E + 2 R <= 525
  4) 4 E + 2 R <= 440
  5) 2 E + 3 R <= 560
  6) E - E1 - E2 = 0
  7) R - R1 - R2 = 0
  8) E1 <= 60
  9) R1 <= 50
```

Because E1 contributes more profit per unit than E2, the model will automatically increase E1 to 60 before E2 becomes greater than zero. A similar remark applies to racing kits.

```
OBJECTIVE FUNCTION VALUE 1667.50000

VARIABLE  VALUE
  E    60
  R    95
  E1   60
  E2    0
  R1   50
  R2   45
```

Example 9 Multiple Solutions

Information Provided. A heating and air-conditioning manufacturer wishes to rent warehouse storage space. Space is available on a monthly, quarterly, semiannual, or annual basis. The following tables provide lists of space required in thousands of square feet and cost of space in dollars per thousand square feet per lease period. Quarterly lease periods begin January 1, April 1, July 1, and October 1. Half-year lease periods begin February 1 and August 1. Formulate an appropriate LP model to lease space at minimal cost for a calendar year.

Month	1	2	3	4	5	6	7	8	9	10	11	12
Space reqd.	10	15	23	32	43	52	50	56	40	25	15	10

Leasing period	Dollars per 1000 ft^2
Month	300
Quarter	500
Half-year	700
Year	1000

Solution

Let MI = no. of thousands of square feet leased for month I, $1 <= I <= 12$.

Let QI = no. of thousands of square feet leased for quarter I, $1 <= I <= 4$.

Let HI = no. of thousands of square feet leased for half year I, $I = 1, 2$.

Let Y = no. of thousands of square feet leased for the year.

```
Minimize 10 Y + 7 H1 + 7 H2 + 5 Q1 + 5 Q2 +
5 Q3 + 5 Q4 + 3 M1 + 3 M2 + 3 M3 + 3 M4 +
3 M5 + 3 M6 + 3 M7 + 3 M8 + 3 M9 + 3 M10 +
3 M11 + 3 M12
Subject to
    2) Y + Q1 + M1 ≥ 10
    3) Y + H1 + Q1 + M2 ≥ 15
    4) Y + H1 + Q1 + M3 ≥ 23
    5) Y + H1 + Q2 + M4 ≥ 32
    6) Y + H1 + Q2 + M5 ≥ 43
```

```
 7) Y + H1 + Q2 + M6 ≥ 52
 8) Y + H1 + Q3 + M7 ≥ 50
 9) Y + H2 + Q3 + M8 ≥ 56
10) Y + H2 + Q3 + M9 ≥ 40
11) Y + H2 + Q4 + M10 ≥ 25
12) Y + H2 + Q4 + M11 ≥ 15
13) Y + H2 + Q4 + M12 ≥ 10
```

I have included a solution in the form of output from a software package so we can discuss it. Reduced costs and dual prices are discussed in books on linear programming. If you are going to use LP with some frequency it is worth learning about the algebra and geometry of LP. If you simply want to know whether an LP may have multiple solutions, a first clue is a decision variable with value equal to zero and reduced cost equal to zero: that does not occur below. A second clue is a slack or surplus equal to zero with corresponding dual price equal to zero: that does occur below.

```
OBJECTIVE FUNCTION VALUE 510.000000

VARIABLE    VALUE    REDUCED COST
    Y    25.000000     .000000
   H1    10.000000     .000000
   H2      .000000    4.000000
   Q1      .000000    5.000000
   Q2     8.000000     .000000
   Q3    15.000000     .000000
   Q4      .000000    5.000000
   M1      .000000    3.000000
   M2      .000000    3.000000
   M3      .000000    3.000000
   M4      .000000    3.000000
   M5      .000000    1.000000
   M6     9.000000     .000000
   M7      .000000    1.000000
   M8    16.000000     .000000
   M9      .000000    3.000000
  M10      .000000    3.000000
  M11      .000000    3.000000
  M12      .000000    3.000000

ROW SLACK OR SURPLUS DUAL PRICES
 2)   15.000000         .000000
 3)   20.000000         .000000
 4)   12.000000         .000000
 5)   11.000000         .000000
 6)     .000000       -2.000000
 7)     .000000       -3.000000
 8)     .000000       -2.000000
 9)     .000000       -3.000000
10)     .000000         .000000
11)     .000000         .000000
12)   10.000000         .000000
13)   15.000000         .000000
```

There are many ways to look for multiple optimal solutions. I put the objective function equal to C (for cost) and added this as a constraint. I was going to add the constraint $C = 510$ and use a new objective function to search for another solution, but look what the first change caused to happen.

```
Minimize 10 Y + 7 H1 + 7 H2 + 5 Q1 + 5 Q2 +
  5 Q3 + 5 Q4 + 3 M1 + 3 M2 + 3 M3 + 3 M4 +
  3 M5 + 3 M6 + 3 M7 + 3 M8 + 3 M9 + 3 M10 +
  3 M11 + 3 M12
Subject to
    2) Y + Q1 + M1 >= 10
    3) Y + H1 + Q1 + M2 >= 15
    4) Y + H1 + Q2 + M3 >= 23
    5) Y + H1 + Q2 + M4 >= 32
    6) Y + H1 + Q2 + M5 >= 43
    7) Y + H1 + Q2 + M6 >= 52
    8) Y + H1 + Q3 + M7 >= 50
    9) Y + H2 + Q3 + M8 >= 56
   10) Y + H2 + Q3 + M9 >= 40
   11) Y + H2 + Q4 + M10 >= 25
   12) Y + H2 + Q4 + M11 >= 15
   13) Y + H2 + Q4 + M12 >= 10
   14) 10 Y + 7 H1 + 7 H2 + 5 Q1 + 5 Q2 +
       5 Q3 + 5 Q4 + 3 M1 + 3 M2 + 3 M3 +
       3 M4 + 3 M5 + 3 M6 + 3 M7 + 3 M8 +
       3 M9 + 3 M10 + 3 M11 + 3 M12 - C = 0

OBJECTIVE FUNCTION VALUE 510.00000

VARIABLE    VALUE     REDUCED COST
    Y    43.000000     .000000
   H1      .000000     .000000
   H2      .000000    4.000000
   Q1      .000000    5.000000
   Q2      .000000     .000000
   Q3     7.000000     .000000
   Q4      .000000    5.000000
   M1      .000000    3.000000
   M2      .000000    3.000000
   M3      .000000    3.000000
   M4      .000000    3.000000
   M5      .000000    1.000000
   M6     9.000000    3.000000
   M7      .000000    1.000000
   M8     6.000000     .000000
   M9      .000000    3.000000
  M10      .000000    3.000000
  M11      .000000    3.000000
  M12      .000000    3.000000
    C   510.000000     .000000
```

Not only do we have a different solution, but now we have variables, $H1$ and $Q2$ with value equal to zero and reduced cost equal to zero. I changed the objective function to Maximize $Q2$ and added the constraint: 15) $C = 510$. The following solution appeared.

```
OBJECTIVE FUNCTION VALUE 18.000000

VARIABLE     VALUE     REDUCED COST
   Q2    18.000000     .000000
    Y    25.000000     .000000
   Q1      .000000    2.500000
   M1      .000000    1.500000
   H1      .000000    1.000000
   M2      .000000    1.500000
   M3      .000000    1.500000
   M4      .000000    1.500000
   M5      .000000    1.500000
   M6     9.000000     .000000
   Q3    25.000000     .000000
   M7      .000000     .500000
   H2      .000000    1.000000
```

```
      M8      6.000000      .000000
      M9       .000000     1.500000
      Q4       .000000     1.500000
     M10       .000000      .500000
     M11       .000000     1.500000
     M12       .000000     1.500000
       C    510.000000      .000000
```

The point of this discussion is to make you aware of the possibility of multiple optimal solutions and indicate how you can begin to look for them.

Example 10 A Loading Problem

Information Provided. Quick Chemical Corporation supplies three types of chemicals to Traction Tire Company. The chemicals are shipped on pallets loaded with one type of chemical. Type 1, 2, and 3 pallets weigh 2000, 2500, and 3000 lb, respectively. Currently, Quick has 50 type 1 pallets, 100 type 2, and 30 type 3 pallets in stock. Quick ships pallets in a truck which can hold 11,000 lb. Quick requires that its truck be fully loaded. How many truckloads can Quick deliver using its current stock?

Solution. There are four ways to load the truck.

Let $L1$ = no. of trucks loaded with one type 1 pallet and three type 3 pallets.
Let $L2$ = no. of trucks loaded with four type 1 pallets and one type 3 pallet.
Let $L3$ = no. of trucks loaded with three type 1 pallets and two type 2 pallets.
Let $L4$ = no. of trucks loaded with two type 2 pallets and 2 type 3 pallets.

```
Maximize L1 + L2 + L3 + L4
Subject to
  2) L1 + 4 L2 + 3 L3 <= 50
  3) 2 L3 + 2 L4 <= 100
  4) 3 L1 + L2 + 2 L4 <= 30

OBJECTIVE FUNCTION VALUE 31.6666700

VARIABLE     VALUE
   L1        .000000
   L2        .000000
   L3      16.666670
   L4      15.000000
```

This solution is not integer valued, so we change the value of $L3$ to 16 to get an integer-valued solution and claim that this solution is optimal: the value of the best LP solution is less than 32, so the value of any integer solution will be at most 31; we have an integer solution with value equal to 31.

Example 11 A Shipping Problem

Information Provided. A company has a large order to ship. There are four types of products to be shipped. The company transports the products from its factory to a dock. Then the orders are transported from the dock to customers by boat. The boat leaves the dock each Sunday and returns to the dock the following Friday. The company uses two trucks, say A and B, to transport products to the dock. The trucks can hold the following quantities of the four types of products:

	Truck capacity			
Truck	1	2	3	4
A	25	20	15	10
B	42	37	34	31

It takes one hour for a truck to load at the factory, deliver to the dock, unload at the dock, and return to the factory (independent of the type of product being transported). However, the per truckload cost (in US dollars) of transporting the types of products varies as follows:

	Truck delivery costs			
Truck	1	2	3	4
A	18	16	12	10
B	26	26	18	16

Types of products cannot be mixed on a truckload. Each of the trucks can be run up to 40 hr per week. The order to be shipped is for 7000 of type 1, 6500 type 2, 7500 of type 3, and 8000 of type 4 products. Determine how to ship the order as quickly and cheaply as possible.

Solution

Let AI = no. of truckloads of type I to transport to the dock on truck A, $I \leq 4$.
Let BI = no. of truckloads of type I to transport to the dock on truck B, $I \leq 4$.

```
Minimize T
Subject to
  2) 25 A1 + 42 B1 >= 7000
  3) 20 A2 + 37 B2 >= 6500
  4) 15 A3 + 34 B3 >= 7500
  5) 10 A4 + 31 B4 >= 8000
  6) A1 + A2 + A3 + A4 - T <= 0
  7) B1 + B2 + B3 + B4 - T <= 0
```

```
OBJECTIVE FUNCTION VALUE 522.985100

VARIABLE     VALUE
   A1    280.000000
   A2    242.985100
   A3       .000000
   A4       .000000
   B1       .000000
   B2     44.332410
   B3    220.588200
   B4    258.064500
    T    522.985100
```

Since it takes 14 weeks to process the order, we add the constraint $T <= 560$ and change the objective function to get an LP model which will find the minimal cost of transporting the order to the dock in 14 weeks below.

```
Minimize 18 A1 + 16 A2 + 12 A3 + 10 A4 +
26 B1 + 26 B2 + 18 B3 + 16 B4
Subject to
  2) 25 A1 + 42 B1 >= 7000
  3) 20 A2 + 37 B2 >= 6500
  4) 15 A3 + 34 B3 >= 7500
  5) 10 A4 + 31 B4 >= 8000
  6) A1 + A2 + A3 + A4 - T <= 0
  7) B1 + B2 + B3 + B4 - T <= 0
  8) T <= 560

OBJECTIVE FUNCTION VALUE 17994.7100

VARIABLE     VALUE
   A1    143.336600
   A2    325.000000
   A3       .000000
   A4       .000000
   B1     81.347260
   B2       .000000
   B3    220.588200
   B4    258.064500
    T    560.000000
```

We cannot send parts of a truck to the dock; so we don't have a solution. Seven years ago my computer was much slower and my software did not have a GIN capability, so I wrote: "But we do have some information which we can use to get an approximate solution. The slack in constraint 6) tells me that we can increase the number of loads which we send to the dock on truck *A*. Also, we can round one of the *BI*s up if we round the other two down. Rounding *B*3 down will require sending truck *A* twice; but, we can decrease *B*1 to 81 and increase *A*1 to 144, and we can decrease *B*4 to 258 and increase *A*4 to 1. Consequently, I will choose the approximate solution

```
VARIABLE     VALUE
   A1         144
   A2         325
   A3           0
   A4           1
   B1          81
   B2           0
   B3         221
   B4         258
   T          560
```

with cost equal to 18,014. Can one find a better solution?" My current software and computer asserts that the answer is NO.

Exercise: Using Overtime to Fill a Rush Order. Suppose that the trucks could be run a sixth 8-hr day each week at a cost per hour $4 higher than those listed in the table. How much would transportation costs increase if the company were to use some overtime to transport the order to the dock in 12 weeks?

Let $CI =$ no. of overtime truckloads of type I to transport to the dock on truck A.
Let $DI =$ no. of overtime truckloads of type I to transport to the dock on truck B.

```
Minimize 18 A1 + 16 A2 + 12 A3 + 10 A4 +
26 B1 + 26 B2 + 18 B3 + 16 B4 + 22 C1 +
20 C2 + 16 C3 + 14 C4 + 30 D1 + 30 D2 +
22 D3 + 20 D4
Subject to
   2) 25 A1 + 42 B1 + 25 C1 + 42 D1 >= 7000
   3) 20 A2 + 37 B2 + 20 C2 + 37 D2 >= 6500
   4) 15 A3 + 34 B3 + 15 C3 + 34 D3 >= 7500
   5) 10 A4 + 31 B4 + 10 C4 + 31 D4 >= 8000
   6) A1 + A2 + A3 + A4 - T <= 0
   7) B1 + B2 + B3 + B4 - T <= 0
   8) T <= 480
   9) C1 + C2 + C3 + C4 <= 96
  10) D1 + D2 + D3 + D4 <= 96

OBJECTIVE VALUE

1) 18310.8700

VARIABLE     VALUE     REDUCED COST
   A1     116.456700      .000000
   A2     325.000000      .000000
   A3        .000000     2.188234
   A4        .000000     3.470967
   B1       1.347229      .000000
   B2        .000000      .640001
   B3     220.588200      .000000
   B4     258.064500      .000000
   T      480.000000      .000000
   C1        .000000     4.000000
   C2        .000000     4.000000
   C3        .000000     6.188234
   C4        .000000     7.470967
   D1      96.000000      .000000
   D2        .000000      .640001
   D3        .000000      .000000
   D4        .000000      .000000
```

Variables *D*3 and *D*4 have value equal to zero and reduced cost zero; thus, there may be multiple solutions to the LP model. This suggests that the optimal integer-valued solution may look quite different from the optimal LP solution. What would be your integer-valued solution?

Example 12 Using INT Variables to Model Increasing Return Thresholds

Information Provided. Speedway Toy Company manufactures tricycles and wagons, which require 17 and 8 min to shape and 14 and 6 min to finish. Tricycles require one large wheel and two small wheels; wagons require four small wheels. Speedway has decided to allocate up to 520 hr of shaping time and up to 400 hr of finishing time to these two products during the next month. Speedway buys wheels from Rollright Corporation, which sells large wheels for \$5 and small wheels for \$1. Speedways has been selling its entire output of these products to Toy World, Inc. at a profit of \$11 and \$5 per unit. The holidays approach; Rollright offers to sell small wheels in excess of 6000 for \$0.75 each, and Toy World is willing to pay a \$2 bonus for tricycles in excess of 1200. Speedway wants to plan production of tricycles and wagons for the next month.

Solution

Let T = no. of tricycles to produce.
Let W = no. of wagons to produce.
Let $S1$ = no. of small wheels to purchase at \$1 each.
Let $S2$ = no. of small wheels to purchase at \$0.75 each.
Let $T2$ = no. of tricycles to produce in excess of 1200.
Let $T1 = T - T2$.

We have thresholds of 6000 and 1200 for small wheels and tricycles. However, in contrast with Example 8, *as we cross these thresholds the return per unit increases*. An LP model for this situation like the one used for Example 8 would prefer to use $S2$ and $T2$ rather than $S1$ and $T1$, so we will use two INT variables, X and Y, to force $S2$ and $T2$ to have value equal to zero until $S1$ and $T1$ reach 6000 and 1200.

```
Maximize 11 T1 + 13 T2 + 5 W + .25 S2
Subject to
  2) 17 T + 8 W ≤ 31200
  3) 14 T + 6 W ≤ 24000
  4) T1 + T2 - T = 0
  5) 2 T + 4 W - S1 - S2 = 0
  6) 6000 X - S1 ≤ 0
  7) S2 - 14000 X ≤ 0
  8) 1200 Y - T1 ≤ 0
  9) T2 - 800 Y ≤ 0
```

Since no more than 2000 tricycles can be produced and no more than 4000 wagons can be produced, 14,000 is an upper bound for $S2$ in constraint 7 and 800 is an upper bound for $T2$ in constraint 9. If $S1$ is less than 6000, then constraint 6 requires that X be equal to zero, which forces $S2$ to be equal to zero according to constraint 7. If $S1$ is equal to 6000, then X can be equal to 1, which permits $S2$ to be any nonnegative number no greater than 14000. Similarly, constraints 8 and 9 reqire that Y be zero and $T2$ be zero if $T1 < 1200$ and permit $T2$ to be any number between zero and 800 if $T1 = 1200$.

Example 13 INT Variables

Information Provided. A paper-recycling machine can produce toilet paper, writing pads, and paper towels, which sell for 18, 29, and 25 cents and consume 0.5, 0.22, and 0.85 kg of newspaper and 0.2, 0.4, and 0.22 min. Each day 10 hr and 1500 kg of newspaper are available; at least 1000 rolls of toilet paper and 400 rolls of paper towels are required. If any writing pads are manufactured, then at least 500 must be made; moreover, the government will pay a bonus of \$20 if at least 1200 rolls of toilet paper are produced. The objective is to determine which mix of products will maximize daily income.

Solution

Let T = no. of rolls of toilet paper to produce in a 10 hr day.
Let W = no. of writing pads to produce in a 10 hr day.
Let P = no. of rolls of paper towels to produce in a 10 hr day.
Let Y be an INT variable which is equal to 1 when writing pads are produced, and let Z be an INT variable which is equal to 1 if at least 1200 rolls of toilet paper are produced.

The model appears after some comments which explain how the INT variables work in the model.

Constraints 6 and 7 take care of writing pads. Constraint 7 forces Y to be one if $W > 0$; I chose to use the number 3000 in constraint 7 because it is clear from constraint 3 that $W \leq 3000$ whenever W is feasible; 1500 or any larger number would work equally well. Constraint 6 forces W to be ≥ 500 if $Y > 0$. Constraint 8, $T \geq 1200Z$, permits Z to be one if $T \geq 1200$; the presence of the term $20Z$ in the objective function causes the value of Z to pop up to 1 as soon as the value of T reaches 1200.

```
Maximize 20 Z + .18 T + .29 W + .25 P
Subject to
  2) .5 T + .22 W + .85 P <= 1500
  3) .2 T + .4 W + .22 P <= 600
  4) T >= 1000
  5) P >= 400
  6) -500 Y + W >= 0
  7) -3000 Y + W <= 0
  8) -1200 Z + T >= 0

OBJECTIVE FUNCTION VALUE 562.818200

VARIABLE    VALUE
   Y         1.000000
   Z         1.000000
   T      1200.000000
   W       500.000000
   P       727.272700
```

Example 14 GIN Variables

Information Provided. A product is assembled from three parts that can be manufactured on two types of machines, *A* and *B*. Each assembly uses one part 1, two part 2, and one part 3. There are three type *A* machines and five type *B* machines. Each machine can run up to 16 hr per day. A machine can process at most one type of part during the day. The number of parts manufactured by each type of machine per hour is summarized below.

Part	No. of parts per hour	
	Machine *A*	Machine *B*
1	12	6
2	15	12
3	—	25

Management wants a daily schedule for the machines that will produce parts for the maximal number of assemblies.

Solution

Let *AI* = no. of type *A* machines on which to make part *I*, *I* <= 2.

Let *BI* = no. of type *B* machines on which to make part *I*, *I* <= 3.

```
Maximize N
Subject to
  2) 192 A1 + 96 B1 - N >= 0
  3) 192 B2 - 2 N + 240 A2 >= 0
  4) -N + 400 B3 >= 0
  5) A1 + A2 = 3
  6) B1 + B2 + B3 = 5

GIN A1, B1, B2

OBJECTIVE FUNCTION VALUE 400

VARIABLE    VALUE
   A1      2.000000
   B1      1.000000
   B2      3.000000
   N     400.000000
   A2      1.000000
   B3      1.000000
```

Example 15 Use GIN Variables Sparingly

Information Provided. The management of the Campus Cafe decided to use LP to determine the minimal numbers of cooks and waitresses needed to meet the following requirements.

There are six 4 hr shifts for cooks each day. Cooks work two consecutive shifts. The minimum number of cooks needed on a shift is tabulated below; the numbers do not change from day to day. Cooks are paid the same hourly salary, independent of shifts worked.

Shift	Minimum number of cooks
1: Midnight to 4 AM	6
2: 4 AM to 8 AM	12
3: 8 AM to noon	16
4: Noon to 4 PM	12
5: 4 PM to 8 PM	14
6: 8 PM to midnight	8

Waitresses are hired to work one 6 hr shift per day. They are paid for 6 hr per day for five consecutive days, then they are off for two days. They are all paid the same amount per day of work. The following schedule of minimal waitress hours needed per day follows.

Day	Minimum number of waitress hours
Sunday	400
Monday	400
Tuesday	400
Wednesday	400
Thursday	400
Friday	200
Saturday	None

Up to 50% of the waitresses who are off on a given day can be hired as overtime employees that day. If they are hired for a day of overtime, they are expected to be available to work up to 6 hr that day and they are paid 150% of a regular day's pay.

Solution. Scheduling cooks and waitresses are effectively separate problems. Cooks will be scheduled first. Since cooks are all paid the same it suffices to minimize the total number of cooks needed each day.

Cooks. Let *NI* = number of cooks to begin working at start of shift *I*, *I* <= 6, and let *N* denote the total number of cooks needed each day.

```
Minimize N
Subject to
  2) N - N1 - N2 - N3 - N4 - N5 - N6 =0
  3) N1 + N6 >= 6
```

```
   4) N1 + N2 >= 12
   5) N2 + N3 >= 16
   6) N3 + N4 >= 12
   7) N4 + N5 >= 14
   8) N5 + N6 >= 8

OBJECTIVE FUNCTION VALUE 36.000000

VARIABLE    VALUE
   N         36
   N1         0
   N2        12
   N3         4
   N4        12
   N5         2
   N6         6
```

The LP solution turned out to have integer values, which was fine. There are many alternate solutions to the problem.

Waitresses. To determine the number of regular-time waitresses needed, we will determine the number of waitresses that will begin their work week on each day of the week, and we will determine the number of waitresses to work overtime on each day of the week.

Starting with Sunday corresponding to day 1, let *WI* denote the number of waitresses to begin their regular work week on day *I*, let *DI* denote the total number of regular time waitresses working on day *I*, and let *OI* denote the number of waitresses to work overtime on day *I*, $I \leq 7$.

```
Minimize 5 W + 1.5 O
Subject to
   2) W - W1 - W2 - W3 - W4 - W5 - W6 - W7 =
      0
   3) - W1 - W4 - W5 - W6 - W7 + D1 = 0
   4) - W2 + W4 - D1 + D2 = 0
   5) - W3 + W5 - D2 + D3 = 0
   6) - W4 + W6 - D3 + D4 = 0
   7) - W5 + W7 - D4 + D5 = 0
   8) W1 - W6 - D5 + D6 = 0
   9) W2 - W7 - D6 + D7 = 0
  10) 6 D1 + 6 O1 >= 402
  11) 6 D2 + 6 O2 >= 402
  12) 6 D3 + 6 O3 >= 402
  13) 6 D4 + 6 O4 >= 402
  14) 6 D5 + 6 O5 >= 402
  15) 6 D6 + 6 O6 >= 204
  16) 6 D7 + 6 O7 >= 0
  17) O - O1 - O2 - O3 - O4 - O5 -O6 - O7 =
      0
  18) W2 + W3 - 2 O1 >= 0
  19) W3 + W4 - 2 O2 >= 0
  20) W4 + W5 - 2 O3 >= 0
  21) W5 + W6 - 2 O4 >= 0
  22) W6 + W7 - 2 O5 >= 0
  23) W1 + W7 - 2 O6 >= 0
  24) W1 + W2 - 2 O7 >= 0
```

Constraints 4 to 9 are recursive formulations which relate the *WI*s to the *DI*s. The numbers on the right side of the inequalities in constraints 10 to 15 are the smallest multiples of 6 which are greater than or equal to the numbers listed in the table for the corresponding day. Because waitresses are available to work 6 hr per day, making these changes does not change the solution, but does encourage the solution to be interger valued, which we need. Before making these changes, the solution was a mess. After making them, it was still not integer valued, however the optimal value of the LP was 386.1, which implies that an integer solution with objective function value 386.5 is optimal. I added the constraint "GIN *W*" and got the following integer solution.

```
NEW INTEGER SOLUTION OF 386.5

VARIABLE     VALUE
   W          68
   O          31
   W1         64
   W2          2
   W3          0
   W4          1
   W5          0
   W6          1
   W7          0
   D1         66
   D2         67
   D3         67
   D4         67
   D5         67
   D6          4
   D7          2
   O1          1
   O2          0
   O3          0
   O4          0
   O5          0
   O6         30
   O7          0
```

Example 16 Allocating Limited Resources in a Transportation Problem

Information Provided. A power authority operates four power plants, located near Denver, Fort Collins, Pueblo, and Salt Lake City. It can purchase coal from three suppliers, located in Colorado, Utah and Wyoming. It must have at least 400, 80, 120, and 560 boxcars of coal per month to keep the plants operating. It would like to have 1000, 230, 430, and 1400 boxcars per month at the power plants. The suppliers can provide 800, 600, and 1000 boxcars per month. The following table lists costs (including shipping) for a boxcar of coal delivered from a supplier to a power plant.

	Denver	Fort Collins	Pueblo	Salt Lake
Colorado	403	441	458	430
Utah	530	510	550	350
Wyoming	360	340	380	410

Solution. The total supply is 2400 boxcars and the total demand is 3030 boxcars, so total demand cannot be met. Consequently, we split the demand for each city into two demands, a minimal demand to keep the power plant running and an optional demand which represents the number of boxcars above minimal demand that the city would like to receive. The total minimal demand is 1160 boxcars. We will use available supply to meet the minimal demands. Consequently, there are 1240 boxcars available to meet some optimal demand. There are 630 boxcars of optimal demand that cannot be met. We will introduce a dummy supply of 630 boxcars to represent unmet demand, and we will introduce a big M cost of shipping from the dummy supply to the minimal demands to insure that minimal demands are met from real supplies. The costs from the dummy supply to the optimal demand will be put equal to zero, since we are not really shipping anything. A transportation tableau follows.

	Den	Den	Fort	Fort	Peub	Peub	SL	SL	
	min	opt	min	opt	min	opt	min	opt	Supply
Colorado	403	403	441	441	458	458	430	430	800
Utah	530	530	510	510	550	550	350	350	600
Wyoming	360	360	340	340	380	380	410	410	1000
Dummy	M	0	M	0	M	0	M	0	660
Demand	400	600	80	150	120	310	560	840	3060

Example 17 “Everyone Wants More Money”

Information Provided. Four research centers are applying for grants from four government agencies. The centers need 20, 30, 40, and 50 million dollars next year to keep operating. They would each like an infinite amount of money. The agencies have 40, 30, 30, and 80 million dollars available for research grants to the four centers next year. The government has decided to keep the centers operating and has prepared the following table of estimated benefit per million dollars granted.

	Center benefit			
Agency	1	2	3	4
1	2	5	7	6
2	6	3	3	6
3	8	9	6	4
4	5	7	4	10

Solution. This situation can also be modeled as a transportation problem. It can be modeled as a “maximize” (benefit) transportation problem, or as a “minimize” transportation problem after changing the benefits to costs by multiplying them by -1, and then adding 10 to each of these negative numbers to get nonnegative numbers. We present a transportation tableau for the minimize formulation below. The four research centers require a total of 140 million dollars to keep operating, this number is the sum of the minimal needs; enough money to meet minimal needs will be shipped from the agencies to the research centers. In addition to this 140 million, there are 40 million dollars available to meet optional demand (wants) by the research centers. Consequently, even though the research centers might prefer more, the most that any one can get is 40 million dollars; thus, the optional demands are set at 40 and a dummy supply of 120 is incorporated to balance total supply with total demand. We use 20 as a big $-M$ to keep from shipping artificial money to meet minimal demands. The minimal and optional demands of research center I are denoted by MI and OI.

	$M1$	$O1$	$M2$	$O2$	$M3$	$O3$	$M4$	$O4$	Supplies
1	8	8	5	5	3	3	4	4	40
2	4	4	7	7	7	7	4	4	30
3	2	2	1	1	4	4	6	6	30
4	5	5	3	3	6	6	0	0	80
Dummy	20	0	20	0	20	0	20	0	120
Demands	20	40	30	40	40	40	50	40	—

A solution in the form of an allocation table follows.

	$M1$	$O1$	$M2$	$O2$	$M3$	$O3$	$M4$	$O4$	Supplies
1	—	—	—	—	40	—	—	—	40
2	20	10	—	—	—	—	—	—	30
3	—	—	30	—	—	—	—	—	30
4	—	—	—	—	—	—	50	30	80
Dummy	—	30	—	40	—	40	—	10	120

Example 18 A Multiperiod Problem

Information Provided. A company produces two products, which we denote by *P* and *Q*. During the next four months the company wishes to produce the following numbers of products *P* and *Q*.

	No. of products			
Product	Month 1	Month 2	Month 3	Month 4
P	4000	5000	6000	4000
Q	6000	6000	7000	6000

The company has decided to install new machines on which to produce product *Q*. These machines will become available at the end of month 2. The company wishes to plan production during a four-month changeover period. Maximum monthly production of product Q is 6000 during months 1 and 2, and 7000 during months 3 and 4; maximum total monthly production is 11,000 during months 1 and 2, and 12,000 during months 3 and 4. Manufacturing costs of product Q are $15 per unit less during months 3 and 4. Also, during the changeover period, units of product Q can be delivered late at a penalty of $10 per unit per month. Monthly inventory costs are $10 per unit per month for product *P* and $8 per unit per month for product Q. Formulate an LP to minimize production, penalty, and inventory costs during the four-month changeover period.

Solution. Let *PI* and *QI* denote the numbers of units of products *P* and *Q* produced during month *I*, $1 \leq I \leq 4$. Let *XI* denote the number of units of product *Q* in inventory at the end of month *I*, and let *YI* denote the number of units of product *Q* backlogged at the end of month *I* (i.e., not able to be delivered by the end of month *I*).

Constraints:

```
1)  P1 >= 4                (Production for month 1)
2)  P1 + P2 >= 9           (Production for months 1 and 2)
3)  P1 + P2 + P3 >= 15     (Production for months 1-3)
4)  P1 + P2 + P3 + P4 = 19 (Production for months 1-4)
5)  Q1, Q2 <= 6
6)  Q3, Q4 <= 7
7)  P1 + Q1 <= 11
8)  P2 + Q2 <= 11
9)  P3 + Q3 <= 12
10) P4 + Q4 <= 12
11) Q1 - 6 = X1 - Y1
12) Q1 - 6 + Q2 - 6 = X2 - Y2
13) Q1 + Q2 - 12 + Q3 - 7 = X3 - Y3
14) Q1 + Q2 + Q3 + Q4 = 25
```

Constraints 5–10 are production capacity constraints. Constraints 11–13 model the effects of production of *Q* during months 1–3, and constraint 14 is the production requirement on *Q* for months 1–4: backlogging is only permitted during the changeover period, and inventory at the end of month 4 would incur an unnecessary cost.

Constraints 12 and 13 can be written in a *recursive formulation* which is useful to know about; we do that now. Using constraint 11 we replace $Q1 - 6$ in constraint 12 by $X1 - Y1$ to get

```
12)* X1 - Y1 + Q2 - 6 = X2 - Y2
```

Next we replace $Q1 + Q2 - 12$ in constraint 13 by $X2 - Y2$ to get

```
13)* X2 - Y2 + Q3 - 7 = X3 - Y3
```

Recursive formulation of constraints can save typing and clarify relationships between periods (cf. constraints 4 to 9 in Example 15).

Objective function. The objective (in thousands of dollars) is to

$$\text{Minimize } \{10(3P1 + 2P2 + P3) - 4 - 9 - 15\} + \{8(X1 + X2 + X3)\} + \{10(Y1 + Y2 + Y3)\} - \{15(Q3 + Q4)\}$$

The first block in the objective function represents inventory cost for product *P*; the number −28 can be deleted because it is a constant. Likewise, production cost for product *P* is a constant which is ignored in the objective function. The next two blocks represent inventory and backlogging costs for product *Q*. The last block represents production savings due to manufacturing product *Q* during periods 3 and 4. Because both *XI* and *YI* incur a positive cost, at most one of each pair will be positive.

Example 19 Another Multiperiod Problem

Information Provided. A cheese company is sailing along smoothly when an opportunity to double its sales arises, and the company decides to "go for it." The company produces two types of cheese, Cheddar and Swiss. Fifty experienced production workers have been producing 10,000 lb of Cheddar and 6000 lb of Swiss per week. It takes one worker-hour to produce 10 lb of Cheddar and one worker-hour to produce 6 lb of Swiss. A workweek is 40 hr. Management has decided to double production by putting on a second shift over an eight-week period. The weekly demands (in thousands of pounds) during the eight-week period are tabulated below.

Week	1	2	3	4	5	6	7	8
Cheddar	10	10	12	12	16	16	20	20
Swiss	6	7.2	8.4	10.8	10.8	12	12	12

An experienced worker can train up to three new employees in a two-week training period during which all involved contribute nothing to production; nevertheless, each trainee receives full salary during the training period. One hundred experienced (trained) workers are needed by the end of week 8. Trained workers are willing to work overtime at time and a half during the changeover period. Experienced workers earn $360 per week. Orders can be backlogged during the transition period at a cost of $0.50 per pound for Cheddar and $0.60 per pound for Swiss for each week that shipment is delayed. All back orders must be filled by the end of week 8.

Solution. We will start by setting up an LP model.

Let $CI =$ no. of workers to make Cheddar during week I, $1 \leq I \leq 8$.
Let $SI =$ no. of workers to make Swiss during week I.
Let $QI =$ no. of workers to work overtime during week I.
Let $TI =$ no. of workers to begin training new workers during week I.
Let $NI =$ no. of new workers to begin training during week I.
Let $AI =$ no. of thousands of pounds of Cheddar to backlog at end of week I.
Let $BI =$ no. of thousands of pounds of Swiss to backlog at end of week I.
(also see Page 316)

This solution is not satisfactory because we cannot use parts of experienced workers to train parts of new employees. What would you do next?

Example 20 Sometimes It Pays to Permit Idle Time

Information Provided. ATV corporation has predicted delivery requirements of 3000, 6000, 5000, and 2000 units in the next four months. Current regular time workforce is at the 4000 units per month level. At the moment there are 500 units in inventory. At the end of the four months the company would like its inventory to be 500 units and its regular time workforce to be at the 3000 units per month level. Regular time workforce has a variable cost of $100 per unit. Overtime can be hired in any period at a cost of $140 per unit. Regular time workforce size can be increased from one month to the next at a cost of $300 per unit of change in capacity, it can be decreased at a cost of $80 per unit. There is a charge of $5 per unit for inventory at the end of each month.

Solution
Let $RK =$ no. of units of regular time workforce in period K.
Let $OK =$ no. of units of overtime workforce in period K.
Let $IK =$ no. of units of inventory at end of period K.
Let $HK =$ no. of units of regular time workforce hired at beginning of period K.
Let $FK =$ no. of units of regular time workforce fired at beginning of period K.

Two models are presented below; the first requires that all production capacity be utilized to produce units of the product, while the second model permits some of the workforce to be idle. See page 317.

Example 21 An Assignment Problem

This is the last example in the linear programming section of this chapter. The variables are actually INT variables; however, the LP solution to the model turns out to be integer valued, so the LP solution solves the INT-variable problem. We will also present a dynamic-programming model for this situation in Example 25, the concluding example in the next section.

Information Provided. A corporation has decided to introduce three new products; they intend to produce 2000, 1200 and 1600 units of products 1, 2, and 3 weekly. The corporation has five locations where the products could be produced at the following costs per unit.

	Unit production costs				
Product	1	2	3	4	5
1	90	82	92	84	86
2	62	58	64	56	58
3	76	70	80	—	—

Each location has ample production capacity; however, they have decided to produce each product at only one location, and to produce no more than one product at any location: three locations get a product and two locations get no product. Product 3 can only be produced at locations 1, 2, or 3.

Solution. We will formulate an LP model to assign products to plants. In actuality, the variables are INT variables, but we will run the LP model without requir-

```
Minimize 540 O1 + 540 O2 + 540 O3 + 540 O4 + 540 O5 + 540 O6 + 540 O7 + 540 O8 + 500 A1 + 500
A2 + 500 A3 + 500 A4 + 500 A5 + 500 A6 + 500 A7 + 600 B1 + 600 B2 + 600 B3 + 600 B4 + 600 B5
+ 600 B6 + 600 B7 + 2880 N1 + 2520 N2 + 2160 N3 + 1800 N4 + 1440 N5 + 1080 N6 + 720 N7
Subject to
 2) A1 + .4 C1 = 10
 3) A2 + .4 C1 + .4 C2 = 20
 4) A3 + .4 C1 + .4 C2 + .4 C3 = 32
 5) A4 + .4 C1 + .4 C2 + .4 C3 + .4 C4 = 44
 6) .4 C1 + .4 C2 + .4 C3 + .4 C4 + .4 C5 + A5 = 60
 7) .4 C1 + .4 C2 + .4 C3 + .4 C4 + .4 C5 + .4 C6 + A6 = 76
 8) .4 C1 + .4 C2 + .4 C3 + .4 C4 + .4 C5 + .4 C6 + .4 C7 + A7 = 96
 9) .4 C1 + .4 C2 + .4 C3 + .4 C4 + .4 C5 + .4 C6 + .4 C7 + .4 C8 = 116
10) .24 S1 + B1 = 6
11) .24 S1 + .24 S2 + B2 = 13.2
12) .24 S1 + .24 S2 + .24 S3 + B3 = 21.6
13) .24 S1 + .24 S2 + .24 S3 + .24 S4 + B4 = 32.4
14) .24 S1 + .24 S2 + .24 S3 + .24 S4 + 24 S5 + B5 = 43.2
15) .24 S1 + .24 S2 + .24 S3 + .24 S4 + .24 S5 + .24 S6 + B6 = 55.2
16) .24 S1 + .24 S2 + .24 S3 + .24 S4 + .24 S5 + .24 S6 + B7 + .24 S7 = 67.2
17) .24 S1 + .24 S2 + .24 S3 + .24 S4 + .24 S5 + .24 S6 + .24 S7 + .24 S8 = 79.2
18) -O1 + C1 + S1 + T1 <= 50
19) -O2 + C2 + S2 + T1 + T2 <= 50
20) -O3 + C3 + S3 + T2 + T3 - N1 <= 50
21) -O4 + C4 + S4 + T3 + T4 - N1 - N2 <= 50
22) -O5 + C5 + S5 + T4 + T5 - N1 - N2 - N3 <= 50
23) -O6 + C6 + S6 + T5 + T6 - N1 - N2 - N3 - N4 <= 50
24) -O7 + C7 + S7 + T6 + T7 - N1 - N2 - N3 - N4 - N5 <= 50
25) -O8 + C8 + S8 + T7 - N1 - N2 - N3 - N4 - N5 - N6 <= 50
26) N1 + N2 + N3 + N4 + N5 + N6 + N7 = 50
27) -3 T1 + N1 <= 0
28) -3 T2 + N2 <= 0
29) -3 T3 + N3 <= 0
30) -3 T4 + N4 <= 0
31) -3 T5 + N5 <= 0
32) -3 T6 + N6 <= 0
33) -3 T7 + N7 <= 0

OBJECTIVE FUNCTION VALUE 131118.800

VARIABLE    VALUE    REDUCED COST             VARIABLE    VALUE    REDUCED COST
   C1     25.000000      .000000                A5        .000000    500.000000
   C2     25.000000      .000000                A6        .000000    478.906300
   C3     30.000000      .000000                A7        .000000    942.968800
   C4     30.000000      .000000
   C5     40.000000      .000000                B1        .000000    600.000000
   C6     40.000000      .000000                B2        .000000     37.500000
   C7     50.000000      .000000                B3        .000000    600.000000
   C8     50.000000      .000000                B4        .000000    459.375000
                                                B5        .000000    600.000000
   S1     25.000000      .000000                B6        .000000    564.843800
   S2     30.000000      .000000                B7        .000000   1338.218000
   S3     35.000000      .000000
   S4     45.000000      .000000                T1       6.979167       .000000
   S5     45.000000      .000000                T2       2.812501       .000000
   S6     50.000000      .000000                T3       3.124996       .000000
   S7     50.000000      .000000                T4       1.250005       .000000
   S8     50.000000      .000000                T5       2.499993       .000000
                                                T6        .000008       .000000
   O1      6.979167      .000000                T7        .000000    708.750000
   O2     14.791670      .000000
   O3       .000000   135.000000                N1      20.937500       .000000
   O4       .000000   135.000000                N2       8.437503       .000000
   O5       .000000   168.750000                N3       9.374987       .000000
   O6       .000000   168.750000                N4       3.750015       .000000
   O7       .000000   177.187500                N5       7.499979       .000000
   O8       .000029      .000000                N6        .000015       .000000
                                                N7        .000000       .000000
   A1       .000000   500.000000
   A2       .000000   162.500000
   A3       .000000   500.000000
   A4       .000000   415.625000
```

Model 1. All available production capacity be utilized to produce units of the product:

```
Minimize 100 R + 140 O + 300 H + 80 F + 5 I
Subject to
 2) -R + R1 + R2 + R3 + R4 = 0
 3) -O + O1 + O2 + O3 + O4 = 0
 4) -H + H1 + H2 + H3 + H4 + H5 = 0
 5) -F + F1 + F2 + F3 + F4 + F5 = 0
 6) -I + I1 + I2 + I3 = 0
 7) I0 = 500
 8) I4 = 500
 9) R0 = 4000
10) R1 + O1 - I1 + I0 = 3000
11) R2 + O2 + I1 - I2 = 6000
12) R3 + O3 + I2 - I3 = 5000
13) R4 + O4 + I3 - I4 = 2000
14) R1 - H1 + F1 - R0 = 0
15) -R1 + R2 - H2 + F2 = 0
16) -R2 + R3 - H3 + F3 = 0
17) -R3 + R4 - H4 + F4 = 0
18) R4 + H5 - F5 = 3000

OBJECTIVE FUNCTION VALUE 1937500.00

VARIABLE       VALUE
   R      14500.000000
   O       1500.000000
   H        500.000000
   F       1500.000000
   I       1500.000000
   R1      4000.000000
   R2      4000.000000
   R3      4000.000000
   R4      2500.000000
   O1          .000000
   O2       500.000000
   O3      1000.000000
   O4          .000000
   H1          .000000
   H2          .000000
   H3          .000000
   H4          .000000
   F1          .000000
   F2          .000000
   F3          .000000
   F4      1500.000000
   I1      1500.000000
   I2          .000000
   I3          .000000
   I0       500.000000
   I4       500.000000
   R0      4000.000000
   H5       500.000000
   F5          .000000
```

Model 2. Surplus production capacity variables SK are used to incorporate the possibility of idle workforce:

```
Minimize 100 R + 140 O + 300 H + 80 F + 5 I
Subject to
 2) -R + R1 + R2 + R3 + R4 = 0
 3) -O + O1 + O2 + O3 + O4 = 0
 4) -H + H1 + H2 + H3 + H4 + H5 = 0
 5) -F + F1 + F2 + F3 + F4 + F5 = 0
 6) -I + I1 + I2 + I3 = 0
 7) I0 = 500
 8) I4 = 500
 9) R0 = 4000
10) R1 + O1 - I1 + I0 - S1 = 3000
11) R2 + O2 + I1 - I2 - S2 = 6000
12) R3 + O3 + I2 - I3 - S3 = 5000
13) R4 + O4 + I3 - I4 - S4 = 2000
14) R1 - H1 + F1 - R0 = 0
15) -R1 + R2 - H2 + F2 = 0
16) -R2 + R3 - H3 + F3 = 0
17) -R3 + R4 - H4 + F4 = 0
18) R4 + H5 - F5 = 3000

OBJECTIVE FUNCTION VALUE 1797500.00

VARIABLE       VALUE
   R      15000.000000
   O       1500.000000
   H           .000000
   F       1000.000000
   I       1500.000000
   R1      4000.000000
   R2      4000.000000
   R3      4000.000000
   R4      3000.000000
   O1          .000000
   O2       500.000000
   O3      1000.000000
   O4          .000000
   H1          .000000
   H2          .000000
   H3          .000000
   H4          .000000
   F1          .000000
   F2          .000000
   F3          .000000
   F4      1000.000000
   I1      1500.000000
   I2          .000000
   I3          .000000
   I0       500.000000
   I4       500.000000
   R0      4000.000000
   H5          .000000
   F5          .000000
   S1          .000000
   S2          .000000
   S3          .000000
   S4       500.000000
```

ing that the variables be INT variables and see what happens.

Let

$$AIJ = 1$$

denote that product I is assigned to plant J.

The cost of assigning production of a product to a plant is determined because the numbers of items to be produced are given parameters of the problem.

```
Minimize 1800 A11 + 1640 A12 + 1840 A13 +
1680 A14 + 1720 A15 + 744 A21 + 696 A22 +
768 A23 + 672 A24 + 696 A25 + 1216 A31 +
1120 A32 + 1280 A33
Subject to
 2) A11 + A12 + A13 + A14 + A15 = 1
 3) A21 + A22 + A23 + A24 + A25 = 1
 4) A31 + A32 + A33 = 1
 5) A11 + A21 + A31 <= 1
 6) A12 + A22 + A32 <= 1
 7) A13 + A23 + A33 <= 1
 8) A14 + A14 <= 1
 9) A15 + A25 <= 1

OBJECTIVE FUNCTION VALUE 3496

VARIABLE   VALUE
   A11       0
   A12       0
   A13       0
   A14       1
   A15       0
   A21       0
   A22       0
   A23       0
   A24       0
   A25       1
   A31       0
   A32       1
   A33       0
```

Since the solution to the LP is an INT variable solution, we know that it is an optimal solution to the assignment problem that we wished to solve.

2.2 DYNAMIC PROGRAMMING

2.2.1 Introduction

Like linear programming, dynamic programming (DP) can be an effective way to model situations which have an appropriate structure. This section begins by introducing three characteristics of problems which fit into a (backward) dynamic programming format and outlining the steps involved in formulating a DP model. Next an example is used to explain the terminology (stages, states, etc.) of DP. Then three more examples are used to illustrate some types of situations which fit into a dynamic programming format, and show you how to use DP to model these situations.

2.2.2 Three Characteristics Necessary for a (Backward) DP Model

1. The problem can be organized into a finite sequence of stages, with a decision to be made at each stage. A set of initial states enters a stage. A decision transforms each initial state to a terminal state; the terminal state leaves the stage and becomes an initial state for the next stage. Suppose there are n stages for a problem. Then a solution to the problem amounts to making a sequence of n decisions, one for each stage. This sequence of decisions is called an optimal policy or strategy for the problem.
2. A key characteristic of problems which can be modeled effectively by dynamic programming is the Markov property or principle of optimality: given any initial state at any stage, an optimal policy for the successive stages does not depend on the previous stages (i.e., an optimal policy for the future does not depend on the past).
3. The optimal decisions for the final stage are determined (known in advance).

When these three properties are satisfied, the problem can be solved (dynamically) by moving backward stage by stage, making an optimal decision at each stage.

2.2.3 Formulating a DP Model

This includes four steps:

1. Specifying a sequence of stages.
2. Noting that the final stage policies are determined.
3. Checking that the Markov property is satisfied.
4. Formatting each stage; this involves
   a. Specifying initial states
   b. Specifying how initial states are transformed
   c. Specifying terminal states
   d. Specifying a decision rule that determines the optimal terminal state (or states) corresponding to each initial state at each stage.

2.2.4 Examples 22–25

Example 22 Travel Plans

We are back in the 1800s and wish to travel by stagecoach for the least possible cost from San Francisco to New York. There are four time zones to cross and we have several alternative possible stagecoach rides across each time zone. Each stagecoach ride has a published cost. The possible stagecoach rides go across a time zone, from an initial location to a terminal location. There are ten locations, labeled 1 to 10; the costs of the stagecoach rides are tabulated below. San Francisco is location 1 and New York is location 10. We need to determine a minimal cost path (trip), composed of four rides across the time zones, from location 1 to location 10. We can visualize the locations as points on a map, or nodes, where the paths intersect. States correspond to locations: points or nodes. Stages correspond to time zones. At Stage k an optimal path from each initial state for Stage k to New York is determined. The four stages are displayed on the following table.

	Cost to destination (state)								
Origins		Stage 1			Stage 2		Stage 3		Stage 4
	2	3	4	5	6	7	8	9	10
1	5	4	6						
2				7	5				
3				5	6	4			
4					3	5			
5							5	3	
6							4	6	
7							4	2	
8									4
9									3

State 10 is the only terminal state for Stage 4. The initial states for Stage 4 are 8 and 9, and the optimal trips from states 8 and 9 to state 10 are determined; they cost 4 and 3, respectively. The Markov property is satisfied because if an optimal trip from state 1 to state 10 goes through state i, then the part of the trip from state i to state 10 is an optimal trip from state i to state 10. Thus, the first three steps to formulating a DP model are done, and we can focus on the stages. We can visualize a stage as a literal stage on which a path enters on the left at an initial state, proceeds across the stage, and exits at a terminal state on the right. In this example Stage 4 corresponds to the Eastern time zone, states 8 and 9 are on the left and state 10 is on the right, and the path pays a known fare to ride across Stage 4. Thus, optimal paths from states 8 and 9 to New York are known. Looking at Stage 3, the initial states are states 5, 6, and 7 and the terminal states are states 8 and 9. There are two possible paths from each initial state to state 10; one path goes through state 8 and the other goes through state 9. The costs of these paths are tabulated below, together with the optimal terminal state for each initial state and the cost of an optimal path from each initial state to State 10.

Stage 3

	Cost from initial state to state 10 via		Optimal terminal state	Cost of optimal path to state 10
Initial states	State 8	State 9		
5	$5+4$	$3+3$	9	6
6	$4+4$	$6+3$	8	8
7	$4+4$	$2+3$	9	5

At this point we know optimal paths, and their costs, from states 5, 6, 7, 8, and 9 to state 10. After filling in the corresponding table for Stage 2 below, we will also know optimal paths, and their costs, from states 2, 3, and 4 to state 10. Then the table for Stage 1 will determine an optimal path from state 1 to state 10.

Stage 2. The entries in the table for Stage 3 were unambiguous because there was only one path from each terminal state to state 10. At Stage 2 we consider only optimal paths from terminal states to state 10. *Compare the various possibilities, choose the best, and discard the rest*. This step can lead to a great reduction in computation time (compared to other types of models) for a valid DP model of a situation. A wide variety of notation and terminology is used to describe this process in the literature (stage transition functions, optimal stage transition states, etc.). We avoid introducing a bunch of notation for this example by displaying this step for Stages 2 and 1 in tables below.

Initial states	Cost from initial states at Stage 2 to state 10 via			Optimal terminal state	Cost of optimal path to state 10
	State 5	State 6	State 7		
2	$7+6$	$5+8$		5 or 6	13
3	$5+6$	$6+8$	$4+5$	7	9
4		$3+8$	$5+5$	7	10

Stage 1

Initial state	Cost from state 1 to state 10 via			Optimal terminal state	Cost of optimal path to state 10
	State 2	State 3	State 4		
1	$5+13$	$4+9$	$6+10$	3	13

We can read the unique optimal policy, stage by stage, from the tables: go from State 1 to State 3 to State 7 to State 10 for a total cost of 13.

Example 23 A Mixed Assignment Problem

Assign two medical teams and one dental team to three countries according to the following benefit table.

	Benefit of assigning team to country		
	1	2	3
1 Dental	25	40	30
1 Medical	40	20	50
2 Medical	70	60	60
1 Dental + 1 medical	70	65	70
1 Dental + 2 medical	110	100	105

We will consider three stages; at stage k we will decide which teams to send to country k, $1 \leq k \leq 3$. At stage k the initial states will be the numbers of dental and medical teams which have not been assigned to a preceding country: initial states represent teams available to send to country k and subsequent countries; terminal states represent teams availabe to send to subsequent countries. Let (x, y) denote the case where x dental teams and y medical teams are available to be assigned. The set of initial states for all three stages is $\{(0,0), (1,0), (0,1), (0,2), (1,1), (1,2)\}$. Because assigning more teams increases the benefit, it suffices to consider only the initial state $(1,2)$ at Stage 1 and only the terminal state $(0,0)$ at Stage 3. The Stage 3 benefits corresponding to the initial states are listed in the last column of the table of benefits, they are determined. The Markov property is satisfied, so it remains to format Stages 1 and 2.

We order the states as follows: $(u, v) \leq (x, y)$ if, and only if, $u \leq x$ and $v \leq y$. At both stages, the terminal states corresponding to an initial state (x, y) are the states (u, v) which are $\leq (x, y)$.

Initial states appear in the first column and terminal states are in the first row in the tables for Stages 2 below and 1 on next page.

The optimal benefit is 130; the optimal policy is to send one medical team to country 1, one dental team to country 2 and one medical team to country 3.

Example 24 Problem Solving

Three teams are trying to solve a problem. The probability of their being able to solve the problem is estiamted to be 0.6, 0.4, and 0.2. Two additional people become available to be assigned to the teams; the estimated probability of success with additional team members is tabulated below. Assign the two additional people to the three teams to maximize the estimated probability that at least one team solves the problem.

Additional members	Probability of success		
	Team 1	Team 2	Team 3
0	0.6	0.4	0.2
1	0.8	0.6	0.5
2	0.85	0.8	0.7

We will decide the number of people to be assigned to team k at Stage k. Maximizing the probability that at least one team succeeds is equivalent to minimizing the probability that all three teams fail. The probability that all teams in a group of teams fail is the product of the individual probabilities of failure; thus, the transition functions for this example are multiplicative, not additive. Dynamic programming can deal with nonlinear functions, but less software

Stage 2

Initial state	Benefit for terminal state						Maximum benefit	Optimal terminal state
	(0,0)	(1,0)	(0,1)	(0,2)	(1,1)	(1,2)		
(0,0)	0						0	(0,0)
(1,0)	40	30					40	(0,0)
(0,1)	20		50				50	(0,1)
(0,2)	60		20 + 50	60			70	(0,1)
(1,1)	65	20 + 30	40 + 50		70		90	(0,1)
(1,2)	100	60 + 30	65 + 50	40 + 60	20 + 70	105	115	(0,1)

Stage 1

Initial state	Benefit for terminal state (0,0)	(1,0)	(0,1)	(0,2)	(1,1)	(1,2)	Maximum benefit	Optimal terminal state
(1,2)	110	110	120	95	130	115	130	(1,1)

is available and what is available tends to be quite limited in scope, or less user friendly than LP or other network software. The Markov property is satisfied by this model.

Focus on Stage 3 first. The initial states are 0, 1, and 2 (people to assign to team three). The probability of success table implies that all available people will be assigned. The probabilities of failure for Team 3 follow.

Stage 3 Assigning Members to Team 3

Initial state	Probability of team 3 failing, terminal state 0
0	0.8
1	0.5
2	0.3

Stage 2 Assigning Members to Team 2

Initial state	Probability of teams 2 and 3 failing for terminal state 0	1	2	Optimal terminal state	Probability of failure
0	0.48			0	0.48
1	0.32	0.3		1	0.3
2	0.16	0.2	0.18	0	0.16

Stage 1 Assigning Members to Team 1

Initial state	Probability of all three teams failing for terminal state 0	1	2	Optimal terminal state	Probability of failure
2	0.072	0.06	0.064	1	0.06

Consequently, the optimal probability of success is 0.94; it occurs when one person is assigned to team 1 and one person is assigned to team 3.

Example 25 A Dynamic Programming Solution for Example 21

In Example 21 we assigned products to plants; here we will assign plants to products: a plant will be assigned to product k at Stage k. The following table gives costs (in hundreds of dollars) of assigning a plant to a product.

Plant	Cost of assigning plant to product 1	2	3
1	1800	744	1216
2	1640	696	1120
3	1840	768	1280
4	1680	672	
5	1720	696	

Stage 3. At Stage 3 we assign a plant to product 3. Since plants 4 and 5 cannot be assigned to product 3, the initial states are chosen to be the one-element sets: $\{1\}, \{2\}$, and $\{3\}$, of plants containing, 1, 2, and 3, respectively. One of these plants is assigned to product 3. The following table lists the cost of assigning plants 1, 2, and 3 to product 3.

Plant	Cost
1	1216
2	1120
3	1280

Stage 2. At Stage 2 we assign a plant to product 2 and send a plant on to product 3. Initial states are two-element sets of plants different from $\{4, 5\}$, and terminal states corresponding to an initial state are one-element subsets of the initial state which contain one of the numbers 1, 2, or 3.

Stage 1. At stage 1 the initial states are the (ten) three-element sets of plants, and the terminal states are the (nine) two-element sets of plants different from $\{4, 5\}$. To illustrate the process, one case is computed.

Stage 2

	Cost for terminal state				
Initial state	{1}	{2}	{3}	Optimal cost	Optimal terminal state
{1, 2}	696 + 1216	744 + 1120		1864	{2}
{1, 3}	768 + 1216		744 + 1280	1984	{1}
{1, 4}	672 + 1216			1888	{1}
{1, 5}	696 + 1216			1912	{1}
{2, 3}		768 + 1120	696 + 1280	1888	{2}
{2, 4}		672 + 1120		1792	{2}
{2, 5}		696 + 1120		1816	{2}
{3, 4}			672 + 1280	1952	{3}
{3, 5}			696 + 1280	1976	{3}

Stage 1

Initial	Cost for terminal state									Optimal	Optimal
state	{1, 2}	{1, 3}	{1, 4}	{1, 5}	{2, 3}	{2, 4}	{2, 5}	{3, 4}	{3, 5}	cost	terminal state
{1, 2, 3}											
{1, 2, 4}											
{1, 2, 5}											
{1, 3, 4}											
{1, 3, 5}											
{1, 4, 5}											
{2, 3, 4}											
{2, 3, 5}											
{2, 4, 5}						1720 + 1792	1680 + 1816			3496	{2, 5}
{3, 4, 5}											

Chapter 4.3

Simulation and Analysis of Manufacturing Systems

Benita M. Beamon
University of Washington, Seattle, Washington

3.1 INTRODUCTION

An engineer working in a chemical processing company is asked to analyze the problem of reported shortages in the supply of portable stainless steel tanks that are used for chemical processing by a number of different departments. After careful quantitative analysis, the engineer concludes that there is actually an *excess* in the number of available tanks; in fact, there is a four-tank overage. Based on the engineer's recommendations, management orders *five more tanks*! This example illustrates a common source of frustration faced by many engineers and operations analysts: the frustration of solving assigned problems, and then finding the results of the analysis completely discarded, largely as a result of management nervousness and internal organizational pressures.

Although this example paints a somewhat bleak picture for problem solvers in today's business environment, there is still a need for systems analysis. The purpose of this chapter is to provide a basic introduction to simulation as a tool for analyzing manufacturing systems. More specifically, this chapter will provide a step-by-step introduction to building, verifying, and validating simulation models, as an aid in developing and convincingly communicating quantitative solutions to real-world problems.

3.1.1 What Is Simulation?

Simulation may be defined as ‘‘the imitation of the operation of a real-world process or system over time’’ [1]. This imitation may exist in the form of a hand calculation or a computer model, but in either case, it uses historical or hypothesized system information to make inferences about particular operational aspects of the process or system. More specifically, simulation allows for indirect quantitative analysis of: (1) proposed systems that have not yet been built and/or (2) existing systems that, for reasons of practicality or expense, preclude direct analysis. Thus, simulation is a modeling tool that allows for the analysis of systems without actual modification or disruption of the actual system. The general structure of a simulation model, and categorizations of common model inputs and outputs are shown in Fig. 1.

3.1.2 When Is Simulation Appropriate?

Simulation is a commonly used tool for studying various aspects of existing and proposed systems. Some of the conditions under which simulation is appropriate are given below.

Simulation may be used when an analytical model of the system of interest is too complex; thus,

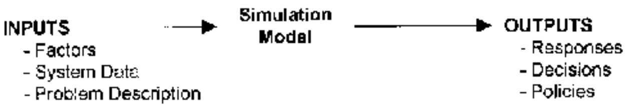


Figure 1 General structure of a simulation model.

simulation allows for modeling of very complicated systems.

During the process of building the simulation, the modeler may gain important insights regarding the behavior of the system, particularly in understanding to what extent particular system characteristics have an effect on specific variables or performance measures of interest.

Simulation models can be important as a communication tool, as their results are often more readily understandable than other quantitative tools.

Simulation is often used to verify or validate analytical models and solutions.

Simulation allows for the study of existing systems, without tampering with or destroying the system; it also allows for the study of proposed systems, without actually building them.

3.1.3 Advantages and Disadvantages of Simulation

As with any tool, simulation possesses a number of inherent advantages and disadvantages. Some of the advantages are implicit in Sec. 3.1.2 above, in the discussion of the appropriateness of simulation, such as its ability to: (1) perform what-if analyses, (2) analyze complex systems, and (3) aid in the understanding of relationships among system characteristics, variables, and performance measures. Some additional advantages are: (4) repeatability of solutions, (5) ease of application, and (6) generally less restrictiveness than analytical methods in that simulation models require fewer simplifying assumptions than do analytical models.

There are also a number of disadvantages associated with simulation models. These disadvantages are: (1) building simulation models may be time consuming and expensive, (2) simulation models may be applied incorrectly, when analytical methods may be more appropriate, and (3) simulation modeling is a skill that requires specific training [1]. Although these disadvantages may appear to be significant, their effects are mitigated by: (1) the fact that there are a number of simulators available on the market that do not require a great deal of special training, and (2) most real-world problems are far too complex to be studied with analytical methods [1].

3.1.4 Simulation Examples

Simulation may be applied to a number of different systems in various environments, including manufacturing, transportation, and public service. The following is a list of potential simulation applications:

Designing and analyzing the number of material-handling vehicles required for a particular manufacturing system

Designing transportation facilities (e.g., freeways, subways, and airports)

Determining optimal inventory levels and management policies for a warehouse

Determining the optimal time-varying number of servers required at a fast-food restaurant

Determining the number of loading docks and forklift drivers to service incoming trucks at a warehouse

Determining the optimal number of parallel-processing machines in a flexible manufacturing system (FMS) cell.

There are countless additional examples beyond what has been presented here, and of course, the list of simulation examples is constantly expanding.

3.1.5 Developing a Simulation Project—an Overview

There are a number of steps that must be completed when developing a simulation project. Each of these steps is illustrated in Fig. 2. Each of these development steps will be addressed in detail throughout the remainder of this chapter.

3.2 PROBLEM DEFINITION AND OBJECTIVES

The first steps in solving any complex problem are to: (1) define the problem, (2) define the objectives of the analysis, (3) define the system, and (4) select the solution method.

3.2.1 Problem Definition

Certainly, without an actual problem, there would be no need to utilize simulation or any other analytical

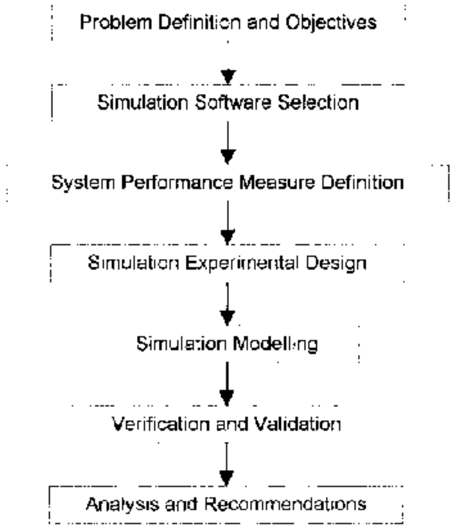

Problem Definition and Objectives
Simulation Software Selection
System Performance Measure Definition
Simulation Experimental Design
Simulation Modelling
Verification and Validation
Analysis and Recommendations

Figure 2 Simulation project overview.

tool. The problem definition is not as trivial as it sounds. In fact, it is perhaps the single most important step in completing a successful simulation study. The problem of interest usually falls into one of the two following categories: (1) “what-if” analyses (referring to a proposed system, or proposed changes to an existing system) or (2) comparison of competing (existing and/or proposed) system alternatives.

3.2.2 Analytical Objectives

Once the problem has been defined, the natural next step is to determine what is to be gained from the analysis. That is, given the problem identified above in Sec. 3.2.1, what answers could be developed that would enable decision makers to solve the problem? Often, it is practical during this phase to actually construct the questions that should be answered by the analysis, e.g., “How would the addition of two parallel cutting machines affect our throughput?,” or “Which of these four system configurations would yield the most substantial reduction in customer service time?” It is important during this phase to make certain that, if answered, the questions developed here would allow decision makers to solve the problems posed during the problem definition phase (Sec. 3.2.1 above).

3.2.3 System Definition

A system may be defined as a collection of objects that interact for the purpose of achieving one or more objectives. One fact that holds for all systems is that systems are affected by changes occurring beyond their defined boundaries. The set of factors that exist outside of the system, but nevertheless have an effect on the system, constitutes the system environment. Since systems are affected by these external forces, defining the system, requires defining the boundaries of the system (or equivalently, specifying the exact collection of objects to be modeled). That is, during this phase the modeler must decide what will be included in the system and what will be excluded. This choice is directly influenced by the questions posed during the identification of the analytical objectives (Sec. 3.2.2, above). For example, if one is interested in determining the throughput effect of increasing the parallel machines for one station in a flexible manufacturing system, there may be no need to model the operation of the loading dock.

Usually, the definition of the system and its boundaries reduces to a decision regarding external system factors. If these factors completely control the system, then there would be nothing to be gained by studying the defined system [2]. However, if the factors only partially control the defined system, then there are three basic ways to treat each of these factors within the model [1]: (1) include them in the system definition, (2) ignore them, or (3) include the factors as system inputs. There is an important distinction here between (1) and (3). If the factors are included in the system definition, then they are a fixed, unchanging constituent of the model; if they are included only as system inputs, then they are specified as constant values or functions and can therefore be modified relatively easily. This would allow the modeler to vary these inputs and study the effects of these variations, if desired. The relationship between the system and its external factors is illustrated in Fig. 3. The system comprises various personnel, machines, processes, and the organization itself. Each of these components, therefore, is included within the system boundary. External factors that influence the system exist outside of the system boundary, and thus comprise the system environment. These factors include: union rules, financial factors, marketing activities, and raw materials supplies (and suppliers). The relationship between the system and the system environment is illustrated in the following example.

Example 1. System Boundaries and the System Environment [1]: *In the case of [1] factory system, the factors controlling the arrival of orders may be considered to be outside the influence of the factory*

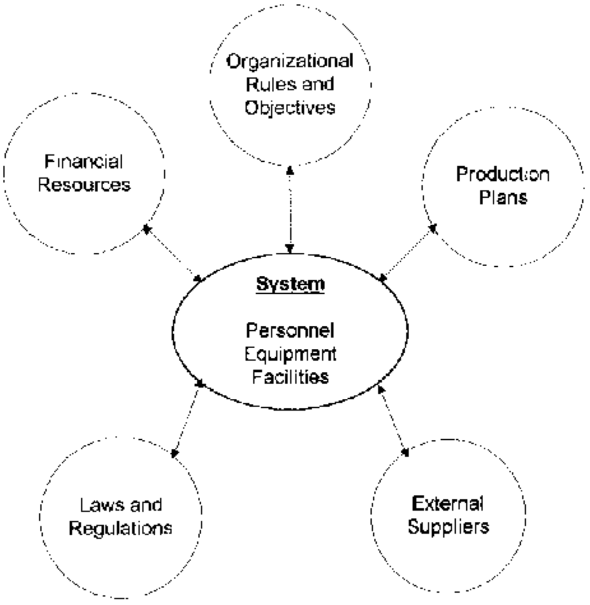


Figure 3 External factors and system boundaries.

and therefore part of the system environment. However, if the effect of supply on demand is to be considered, there will be a relationship between factory output and arrival of orders, and this relationship must be considered an activity of the system. Similarly, in the case of a bank system, there may be a limit on the maximum interest rate that can be paid. For the study of a single bank, this would be regarded as a constraint imposed by the environment. In a study of the effects of monetary laws on the banking industry, however, the setting of the limit would be an activity of the system.

3.2.4 Selection of Solution Method

One of the purposes of this chapter is to aid in the determination of whether or not simulation is an appropriate analytical tool for a given application. Determining whether or not simulation is the correct approach to a given problem requires a two-part evaluation [3]: (1) appropriateness versus other alternative modeling methods (see Sec. 3.1.2 above), and (2) benefits versus costs.

That is, building a simulation model requires a significant: (1) monetary investment (if it must be purchased), and (2) time investment (which varies, depending on the simulation tool used, and the project complexity). Therefore, the cost (in time *and* money) versus the potential savings must be evaluated. For example, if one is interested in analyzing and improving the efficiency of a corner lemonade stand, given the monetary and time requirements of building a simulation model, it would probably be wise to consider an alternative method.

3.3 SELECTION OF SIMULATION SOFTWARE

Once it is determined that simulation is the best approach to solving a particular problem, the next step is to choose the simulation software. When choosing a simulation tool, a number of factors must be considered. The first consideration involves the class of the simulation tool to be used.

3.3.1 Simulation Tool Classes

There are three basic classes (or levels) of simulation modeling tools available: (1) simulators, (2) multipurpose simulation packages, and (3) general-purpose programming languages. Generally, simulators are designed primarily for applications to manufacturing systems, and their high-level building blocks eliminate the need for programming, but may include simplifying and inflexible assumptions. Multipurpose simulation packages are powerful simulation tools that require a greater degree of programming, but still require far less time than simulating in general-purpose computer languages. General-purpose computer languages represent the lowest-level simulation tool, since most functions and all coding must be specifically input by the modeler. Table 1 describes each of these classes, gives examples within each class, and compares each class on the bases of modeling flexibility, ease of use (including interfaces and learning requirements), and time requirements for model building. Readers interested in obtaining more detailed information regarding simulators are referred to Hlupic and Paul [4]; a detailed comparison of simulators and simulation packages is given in Banks [5].

3.3.2 Final Selection

Although the tools available for simulation analysis are constantly changing, selection of the appropriate simulation tool will primarily depend on four factors [1]: (1) input features, (2) processing features, (3) output features, and (4) technical support. Input features refer to the ease and sophistication of data input, interfaces with programming languages, and flexibility.

Table 1 Simulation Tool Class Specifications

Tool Class	Examples	Flexibility	Ease of use	Time requirements
Simulators	WITNESS, ProModel, XCELL	Low	Easy	Low
Simulation packages	SLAM II, SIMAN, GPSS	Moderate–high	Moderate	Moderate
Programming languages	Fortran, Pascal, C	High	Advanced	High

Processing features pertain mainly to running speed and model size limitations, and output features refer to the flexibility and limitations of output reports, graphs, and statistical analysis. Finally, the quality and accessibility of a technical support staff must also be considered.

Specifically, when selecting a simulation tool, the modeler must consider the following factors:

1. *Expense*. There is typically a tradeoff between the power and feature quality of simulation tools: generally speaking, the greater the power and feature quality, the greater the expense. Therefore, given existing budgetary constraints, the modeler must be careful in selecting the package that allows for the most comprehensive set of desirable features for the price.
2. *Flexibility*. The modeler must consider the system(s) for which the tool will be used to simulate. For example, if the tool will be used to simulate fairly simple manufacturing systems, a simulator may be appropriate; however, if many different kinds of systems will be simulated and/or if the systems under study are complex, a simulation language may be a better choice.
3. *Manufacturer reputation/stability*. There are a number of software companies in the simulation business, and this number is always changing. Therefore, when making a selection, an important consideration is the stability of the manufacturer. Since simulation software usually requires a substantial investment (in time and money), it is important that the manufacturer will continue to produce upgrades and provide technical support well into the future.

3.4 DEFINING SYSTEM PERFORMANCE MEASURES

Defining system performance measures is a critical step in the development of a simulation project. In fact, it is during this step that the actual objectives of the analysis are formally addressed. Unfortunately, this is a step that is commonly ignored, or postponed until the last minute. It is important, however, to define the performance measure(s) before building the model, since the measure(s) used are likely to have a significant effect on the model itself.

A *performance measure* may be defined as a metric for quantifying efficiency and/or effectiveness [6]. The *effectiveness* of a system describes to what extent the system performs the required tasks, whereas *efficiency* describes how economically these tasks are performed. Thus, it is possible for an effective system to be inefficient; it is also possible for an efficient system to be ineffective.

Performance measures may be categorized on the bases of: quality, time, cost, resource utilization, operating efficiency, and flexibility measures. Choosing adequate and relevant performance measures is critical in accurately analzying any system. This is because the performance measures are directly associated with the objectives of the study. For example, if an objective of a particular simulation study is to determine how many bank tellers should staff a particular bank, a possible performance measurement system for this study might include server utilization, customer queue length and total cost. For a general case, the following guidelines are suggested for selecting performance measures [7]: (1) methods of calculating the performance criteria must be clearly defined, (2) objective performance criteria are preferred to subjective performance criteria, and (3) ratio performance criteria are preferred to absolute numbers, since ratios provide a comparison of two or more factors. Moreover, the two most important considerations when defining system performance measures are that: (1) the measure(s) are consistent with the overall operational goals of the system (and the organization) and (2) the measures selected sufficiently describe how well the system operates. For example, if a modeler is modeling a production system, and chooses to use total cost as a performance measure, it is clear that this measure does not sufficiently describe the operation of this system. This is because other factors, such as resource utilization and throughput are also of interest, and minimizing cost may correspond to an unacceptable reduction in utilization and throughput performance. Thus, it is important to consider all relevant measures when defining the performance measures of interest, so that the description of system performance is accurate and inclusive.

3.5 DESIGNING A SIMULATION EXPERIMENT

3.5.1 Warm-Up Period

3.5.1.1 Definition and Objectives

A warm-up period for a simulation run is a prespecified length of simulation time during which data are not collected. The purpose of the warm-up period is to reduce bias in the statistical estimates by eliminating the data during the initial (sometimes called transient) period of the simulation. The objective in doing so is to collect data after the simulation has reached a steady state. It is important to note that in certain cases, a warm-up period is not appropriate, particularly when the system of interest begins empty, as in a teller station at a bank. In this case, the system begins when the bank opens (with no customers); thus, for this simulation model, it would make sense to collect statistics from the beginning, without a warm-up period.

3.5.1.2 Determining the Length of the Warm-Up Period

Generally, the most commonly used method for determining the length of the warm-up period is by experimentation. More specifically, this method takes some particular performance measure [e.g., throughput, work-in-progress (WIP)], and plots this value versus simulated time. The time at which the performance curve appears to stabilize should represent the end of the warm-up period. This concept is illustrated in Fig. 4. Notice that for the example shown in Fig. 4, the appropriate length of the warm-up period is T^*. This is because the system is in a transient period between times zero and T^*, and does not reach steady-state until time T^*.

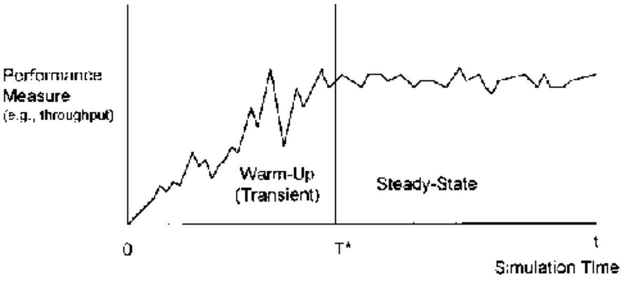


Figure 4 Determining the length of the warm-up period.

3.5.2 Run Length and Replications

After the decision has been made regarding the length of the warm-up period, the next major decisions involve the run length and the number of replications. When making these decisions, it is important to note that there is a tradeoff between the run length and the number of replications. In general, using a small number of long runs yields better estimates of statistical mean,, since the initial bias (introduced at the beginning of each run) occurs fewer times, but because there are fewer runs (replications), the variance of that mean increases [2]. The opposite situation occurs when using a large number of shorter runs: although the large number of runs will reduce the variance, the shorter run lengths are more sensitive to the initial bias, which will adversely affect the estimate of the mean. This discussion yields the general conclusion that large numbers of long runs is probably best; the problem is that, due to a number of situational constraints, this is not always possible.

There are a number of ways to control the running time of the simulation. Two of the most common methods are: specifying a simulation end time and specifying a system-based stop condition. By specifying the ending time of the simulation, the number of data points that are collected is undetermined (i.e., a random variable). By specifying a stop condition (e.g., stop the simulation when the number of data points collected equals 1000), the simulation time is now the random variable. The selection of the type of running time control depends on the specific system of interest.

Example 2. Manufacturing Job Shop: *Suppose a modeler decides to use as a stop condition the number of completed jobs for a job shop operating on a shortest processing time (SPT) discipline. In this case, at the termination time of the simulation, the jobs that are still remaining in the system may not be representative, since each machine within the system is processing jobs with the smallest processing time first. Thus, when the simulation has terminated, it is reasonable to expect that a high percentage of the jobs that are still remaining in the system will have a much higher than average processing time. This can significantly bias the results. In this case, the modeler may decide to use a different stop condition, by introducing a fixed number of parts to be processed, and then specifying that the simulation terminates when all of the parts have been processed.*

3.5.3 Factorial Design

3.5.3.1 Introduction

The final step in designing a simulation experiment is to determine what experiments need to be performed. In fact, this is a crucial step in the development of the simulation project because it is at this point that a plan is established that will allow the modeler to accomplish the objectives of the simulation study. When designing the simulation experiments, there are two methods that may be used: (1) comparing a number of alternatives and (2) searching for an optimal solution [3]. Comparing alternatives involves evaluating a number of different system options with respect to some particular performance criteria, and then selecting the alternative that yields the best performance. When searching for an optimal solution, the modeler varies a number of predetermined experimental factors until some particular target or optimum value of some particular performance criteria is achieved [3]. In this case, the objective is to estimate the overall effects of each of the various experimental factors on the overall performance of the system. This method of experimentation is commonly referred to as factorial design. In this section, factorial design of simulation experiments is introduced. For an in-depth classical treatment of experimental design, the reader is referred to Box [8].

3.5.3.2 Objectives and Terminology

Generally, experimental design establishes a set of principles that govern the method by which an experiment is carried out. The purpose of the design is to determine the effects of system variables on a particular response variable (the response variable is usually some measure of performance). An experimental factor is an input variable; a level is a possible value of a particular factor. Thus, a given factor may have a number of different levels. For example, if a modeler wishes to study the effects of three different part types (Type A, Type B, and Type C) on processing time, the factor would be the part type and the levels would be A, B, and C. A factorial design, then, is a set of experiments that comprise the various levels of experimental factors. Table 2 illustrates the structure of a factorial design with two factors, with three levels for each factor.

It is important to note that the number of experiments required for a full factorial design required follows the equation

$$N = L^f$$

where

$N =$ Number of experiments
$L =$ Number of levels
$f =$ Number of experimental factors

Therefore, for the two-factor factorial design given in Table 2, the number of experiments is $N = 3^2 = 9$, as shown.

Example 3. Vehicle Fleet Size and Vehicle Speed Specification: *Suppose a modeler wishes to study the effects of the number of vehicles and vehicle speeds (two factors) on the throughput of an automated guided vehicle (AGV) system. Suppose also that the modeler decides to study four levels for each factor. These levels are specified as follows:*

1. *Factor 1: number of vehicles. Levels: 1, 2, 3, 4.*
2. *Factor 2: vehicle speed. Levels (in feet per minute): 200, 300, 400, 500.*

Table 2 Two-Factor Factorial Design (Three Levels Per Factor)

Experiment no.	Level of factor 1	Level of factor 2
1	1	1
2	1	2
3	1	3
4	2	1
5	2	2
6	2	3
7	3	1
8	3	2
9	3	3

Table 3 Full Factorial Design: Vehicle Specification Problem

Experiment no.	Level of factor 1 (no. of vehicles)	Level of factor 2 (vehicle speed)
1	1	200
2	1	300
3	1	400
4	1	500
5	2	200
6	2	300
7	2	400
8	2	500
9	3	200
10	3	300
11	3	400
12	3	500
13	4	200
14	4	300
15	4	400
16	4	500

Note that the total number of experiments is $N = 4^2 = 16$. The full factorial design for this case is given in Table 3. In this case, the modeler would run 16 experiments, varying the levels of the two factors as dictated in Table 3.

3.6 SIMULATION MODELING

3.6.1 Simulation Model Structure

After the simulation experiment is designed, and before the simulation model is verified and validated, the following steps are completed within the simulation modeling phase: (1) collecting system data, (2) identifying assumptions, (3) coding the simulation model, and (4) running the simulation model. These steps are illustrated in Fig. 5.

3.6.2 Data Collection

During the data collection phase, the first task is to identify the data needed for the simulation analaysis. Usually, the data comprise one or more of the following: (1) the physical characteristics of an existing or proposed system, (2) historical process data, or (3) current data, obtained from existing databases or samples collected by the modeler, expressly for the purposes of the simulation.

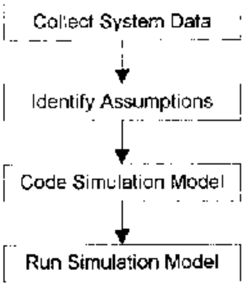


Figure 5 Simulation modeling structure.

It is important to note that actual data may not be collected during this phase. Sometimes, instead of collecting and using actual data, hypotheses are made regarding certain input variables. For example, instead of actually collecting data for the customer arrival rates (which could be expensive, time consuming, or intrusive to the system), an educated guess may be used for the frequency of those arrivals. Additionally, sometimes the actual data may not be used directly, but may instead be used to estimate specific input probability distributions (standard or empirical) and their parameters. Finally, the data may not be used for the purposes of input at all; sometimes, the data are used to validate the simulation model. It is important to note that the output or results of the simulation can only be as good as the input or data used. If the data are poorly estimated or hypothesized, they may significantly reduce the reliability of the simulation results.

3.6.3 Modeling Assumptions

By definition, models include assumptions. Simulation models are no different; indeed, it is virtually impossible to model every detail of the simulation model. Therefore, specific assumptions must be made to make the simulation model tractable. For simulation modeling in most industrial applications, there are a number of governing general assumptions [9]:

The system changes as specific (discrete) points in time.
There is a process network through which one or more entities flow.
There are activities that engage the various entities.
There are conditions that change the states of the entities, and consequently, the system.

Beyond these assumptions, there will be specific assumptions for a given simulation model, e.g., the pickup time for an automated guided vehicle is negligible, the parts arrive to a particular queue according to an exponential distribution with parameter 3 min, etc.

3.6.4 Simulation Modeling

3.6.4.1 Modeling Terminology

The terminology in Table 4 describes the various components of a simulation model. An example of each term is also provided.

Example 4. Serial Production Line: *A simulation has been developed to study the performance of a serial production line. The production line comprises three processing machines (M1, M2, and M3) and three buffers (B1, B2, and B3). Notice that B1 is the incoming buffer for the system. The machines process two part types (P1 and P2), and each part must be processed on all machines in sequence (i.e., each part must be processed first on M1, then on M2, and finally on M3 before exiting the system). The system is depicted in Fig. 6.*

In this example, the entities *are: the three machines, the three buffers, and the collection of parts. Examples of* attributes *are: machine speeds, buffer capacities, machine failure rates, and part types. Some* activities *could be: the time a part is processed on M2 and the time M1 is down. Some* state variables *for this example might be: the number of parts in each buffer, the number of parts processed, and the number of busy machines. Therefore, the state at some particular time t might be (B1, B2, B3, no. of parts processed, busy machines) = (16, 11, 5, 38, 1), which means that at time t there are 16 parts in buffer 1, 11 parts in buffer 2, five parts in buffer 3, 38 completed parts, and one busy machine. Particular* events *for this system could be: a part finishes processing on M1, a new (incoming) part arrives to the system, and M3 breaks down.*

3.6.4.2 Model Inputs

When developing a simulation model, the modeler must specify the particular model inputs. In Example 4 above, the inputs that must be specified are: (1) an arrival discipline (or probability distribution of arrivals) for incoming parts, (2) a constant value or distribution for machine processing speeds or service times (for each machine and for each part type), (3) buffer capacities, (4) breakdown and repair frequencies or breakdown and repair probability distributions (if applicable), and (5) a simulation stopping condition (number of parts processed, run time, etc.). The specifications from these inputs are generally obtained from available data and hypotheses regarding the behavior of specific system elements.

3.7 VERIFICATION AND VALIDATION

3.7.1 Objectives and Definitions

It is impossible to overstate the importance of verification and validation in a simulation experiment. It is during these steps that the modeler confirms that the simulation model is working properly, and that the simulation model is an accurate representation of the real system. Verification and validation are treated separately here, but it is important to realize that the two processes overlap significantly. The placement of

Table 4 Simulation Terminology

Term	Definition	Examples
Entity	Any object within the system that might be of interest to the modeler	Customers, parts, machines
Attribute	A particular property of an entity	Customer priority, part routing, machine speed
Activity	A time period of a particular length	Service time, processing time
State	The collection of (state) variables that are necessary to fully describe the system at a particular time	Queue lengths, number of busy machines, total number of parts processed
Event	Any instantaneous occurrence that changes the state variables (and therefore, the state) of the system	A customer finishes service, a new part arrives to be processed, a machine breaks down

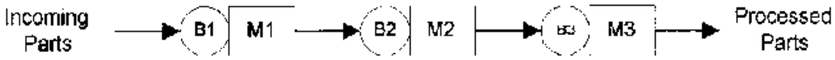


Figure 6 Serial production line.

the verification and validation steps within the modeling process is illustrated in Fig. 7.

3.7.2 Verification

Verification refers to the process of determining whether or not the simulation model is performing as intended, and that it is an accurate representation of the conceptual model [10]. There are many ways in which modelers accomplish the verification step. A number of the most common techniques are given below [10]:

1. Write and debug the computer program in logical submodules, testing each of the model elements separately.
2. Vary the input parameters and confirm that the output is reasonable.
3. Run the model under simplifying conditions for which the output can be easily computed.
4. If possible, observe an animation of the simulation output.

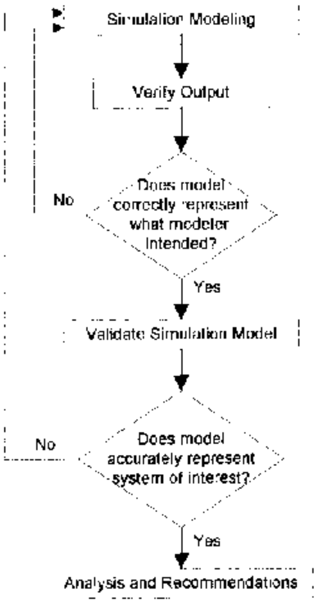

Simulation Modeling
Verify Output
Does model correctly represent what modeler intended?
No
Yes
Validate Simulation Model
Does model accurately represent system of interest?
No
Yes
Analysis and Recommendations

Figure 7 Verification and validation.

3.7.3 Validation

The *validation* process confirms that the simulation model is an approximate representation of the actual system of interest [10]. There are two levels of validation: (1) face validity, and (2) comparison with the real system. Face validity is often considered the first test for model validation. Face validity is a quality given to a model that appears to be reasonable to an individual who is familiar with the actual system of interest [3]. Face validity is often confirmed by group demonstrations and animations (also see step 4 of verification, Sec. 3.7.2 above). When comparing the simulation model to the actual system, analysis of historical data may prove useful. Additionally, given the same inputs, simulation outputs should compare favorably to the real system outputs.

3.8 OUTPUT ANALYSIS, RECOMMENDATIONS, AND DOCUMENTATION

The final step in developing a simulation project (prior to implementation) is analyzing the output and making recommendations based on that analysis. The sequence of this final step within the project development process is illustrated below in Fig. 8.

There are two basic types of analysis a modeler may perform on simulation output: (1) measures of location (or point estimates) and (2) measures of variability (or spread). For a given simulation study, it is generally advisable to use both types of measures in order to achieve confidence in the results.

3.8.1 Output Analysis

There is a countless number of methods available that may be used to analyze a data set. This section presents an introduction to simple analytical techniques that may be used to analyze simulation output.

3.8.1.1 Measures of Location

Measures of location refer to statistical values that sumarize a data set into a single value. The most common measures of location are the mean, the median,

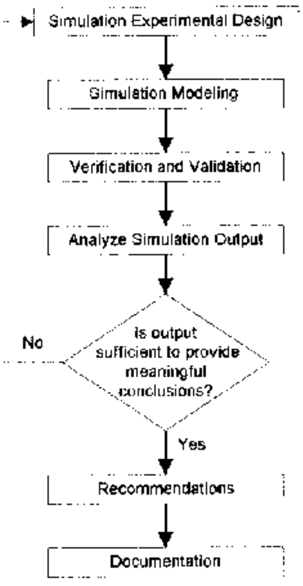

Simulation Experimental Design
Simulation Modeling
Verification and Validation
Analyze Simulation Output
Is output sufficient to provide meaningful conclusions?
No
Yes
Recommendations
Documentation

Figure 8 Output analysis and recommendations.

and the mode. Each of these measures may be calculated on the entire statistical population of interest, or they may be calculated on a sample (a subset of the population), in which case they are referred to as the sample mean, the sample median, and the sample mode.

The Mean. The *sample mean* $\bar{x}$ of a set of n numbers $x_1, x_2, \dots, x_n$ is defined as

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example 5. Rework Station Arrivals: *Suppose a particular simulation model outputs the number of parts that arrive to a rework station during an 8 hr shift. Suppose also that the simulation is run six times, and that the results of each run are as follows:* $\{4, 8, 9, 10, 36, 4\}$. *Therefore, the mean number of parts reworked over an 8 hr shift for the six simulation runs is*

$$\bar{x} = \frac{4 + 8 + 9 + 10 + 36 + 4}{6} = \frac{71}{6} = 11.8 \text{ parts}$$

The Median. It has often been said that the median, as a measure of centrality, is preferable to the mean, since the median is more insensitive to statistical outliers. On the other hand, the mean also directly excludes a number of the data values in the calculation. The *sample median*, $\tilde{x}$, of a set of n observations is calculated by first ordering the observations in magnitude from smallest to largest and then setting $\tilde{x}$ equal to the middle value, if n is odd, or equal to the mean of the two middle values, if n is even. Thus, for the simulation output introduced above (in Example 5), the median is $(8 + 9)/2 = 8.5$. Notice that if there had only been output available for five simulation runs, so that the output values were $\{4, 8, 9, 10, 36, 4\}$, the median would be 9.

The Mode. The mode of a set of numbers is the number that appears with highest frequency. So, the mode of the same simulation output given above in Example 5 is 4, since 4 occurs twice; all other values occur only once.

3.8.1.2 Measures of Variability

No single measure of location can completely describe a data set. Although the measures of location, discussed above, provide a description of where the "middle value" lies in a particular data set, they say nothing directly about how dispersed (or how widely varied) the numbers are for a particular set. It is because of this limitation that we need measures of variation (or equivalently, ways to measure the dispersion of the observations within the data set).

The sample variance, the sample standard deviation, and the range are, perhaps, the most widely used of all measures of variation. The sample variance, s^2, is given by

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

and the sample standard deviation, s, is given by $s = \sqrt[+]{s^2}$. The range of a set of numbers is simply the maximum value minus the minimum value. Unlike the range, neither s^2 nor s have simple intuitive meanings. That is, if one were asked the question "What is the meaning of the sample variance?," there would be no nice, intuitive answer. What one can say, however, is that if the observations in a data set A are widely dispersed (i.e., Set A has many data points that lie far away from its mean), and the observations of another data set B are generally close to its mean, we would expect the sample variance and sample standard deviation for the data set A to be much larger than the sample variance and sample standard deviation for the data set B, as shown in the histograms in Fig. 9. The following example illustrates the importance of using point estimates *and* measures of variability in a simulation study.

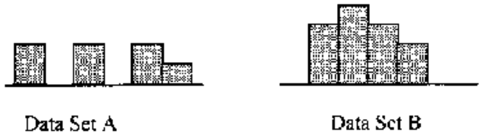


Figure 9 Histograms for two data sets.

Example 6. Manufacturing System Design: *Suppose an engineer is faced with the task of designing a manufacturing line to produce a new product. Suppose also that management has identified that this line should occupy a minimum space, since space in this facility is limited. Consistent with this request, the engineer realizes that minimizing the work-in-progress (WIP) inventory will be a critical step in space minimization. The engineer decides to test three alternative system designs, varying machine types and machine configurations. The mean and variance of the total WIP levels for each system, calculated over an 8 hr shift, are given below. These results are based on 1000 replications per system (3000 total replications).*

In this case, System C yields the lowest (i.e., most desirable) mean WIP level (391 units), but has a much higher variance (65) than the system with the second-lowest mean WIP level (System B, 403 units). These results illustrate that the mean is not the only measure of importance when analyzing simulation results. In this example, even though System C yields the lowest WIP levels, the low variance yielded by System B (with a small sacrifice in WIP), may make System B the better choice. This is because having more predictable *performance levels (i.e., lower variability) is often preferable to a less predictable, although better, mean performance.*

3.8.2 Sensitivity Analysis

Example 6 introduces the need to assess the sensitivity of the model and/or system to changes in input data and variables. The purposes of sensitivity analysis are: (1) to determine *how* the model output reacts to these input changes, and (2) to determine the *extent* to which the model changes when the input is changed (and ultimately, what range of input data results in the same solution and/or conclusions).

Table 5 Simulation Results: Manufacturing System Design Problem

	WIP level	
System	Mean no. of units	Variance
A	553	79
B	403	48
C	391	66

When choosing among system alternatives, systems that are more robust (that is, less sensitive to changes) are generally preferred over less robust systems, when all other measures are approximately equal. This is true for a similar reason as was given in Example 6, namely, that most real-world systems are subject to minor changes, sometimes on a daily basis. In this case, the modeler would like to be relatively certain that the system will continue to perform well in the face of these changes.

3.8.3 Recommendations

After the simulation analysis has been executed, the final step in the simulation project is to make recommendations. Although there are no algorithms or formulae to follow, there are a few guidelines that should be followed when making recommendations:

1. The simulation model is a decision-support tool, and must be treated as such. That is, the purpose of the simulation model is *not* to actually make decisions and/or recommendations; the purpose of the model is to provide results that can be used by a decision maker to make decisions and/or recommendations. Therefore, the model output, and the analysis of that output must be considered in the context of the organization, and the recommendations made must be consistent with the goals of the organization.
2. It is extremely important that the results of the simulation study, including all pertinent assumptions and that all data sources are clearly explained. The results of the study are useless if they are not communicated effectively; thus the responsibility rests with the modeler to make sure that the results are clear and understood by all involved.

3.8.4 Documentation

Although listed as the final step in the development of a simulation project, documentation is an important ongoing step. Every step in the project development process, should be carefully documented, paying special attention to data and assumptions used, model construction (including well-commented simulation code), verification, and validation. The documentation should be maintained (changed, when appropriate), and complete, so that the results could be replicated by a reasonably qualified individual. The final documentation for the entire project should also include the simulation results, analysis, and final recommendations.

3.9 SUMMARY

This chapter provided a basic introduction to the principles and methods used in developing a simulation project. In particular, this chapter defined simulation, and introduced the general objectives of a simulation study, and identified some advantages and disadvantages of simulation as an analytical tool. Additionally, this chapter discussed the following steps in simulation analysis: (1) defining the problem and model objectives, (2) selecting simulation software, (3) defining system performance measures, (4) designing the simulation experiment, (5) simulation modeling, (6) verification and validation, and (7) output analysis, recommendations, and documentation. The objective of this chapter was to provide a practical approach to building, verifying, validating, and communicating simulation solutions to real-world problems.

REFERENCES

1. J Banks, JS Carson, BL Nelson. Discrete-Event System Simulation. 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1996.
2. AAB. Pritsker. Introduction to Simulation and SLAM II. 4th ed. New York: John Wiley, 1995.
3. S Robinson. Successful Simulation. London: McGraw-Hill (UK) International, 1994.
4. V Hlupic, RJ Paul. A critical evaluation of four manufacturing simulators. Int. J Prod Res 33(10): 2757–2766, 1995.
5. J Banks. Interpreting simulation software checklists. OR/MS Today (June): 74–78, 1996.
6. A Neely, M Gregory, K Platts. Performance measurement system design: a literature review and research agenda. Int J Operat Prod Manag 15(4): 80–116, 1995.
7. S Globerson. Issues in developing a performance criteria system for an organization. Int J Prod Res 23(4): 693–646, 1985.
8. GEP Box, WG Hunter, JS Hunter. Statistics for Experimenters. New York: John Wiley, 1978.
9. AAB Pritsker. Modeling for simulation analysis. In: G Salvendy, ed. The Handbook of Industrial Engineers. 2nd ed. New York: John Wiley, 1992.
10. AM Law, WD Kelton. Simulation Modeling and Analysis. 2nd ed. New York: McGraw-Hill, 1991.

Chapter 4.4

Petri Nets

Frank S. Cheng
Central Michigan University, Mount Pleasant, Michigan

4.1 INTRODUCTION

Carl A. Petri invented a net-theoretical approach to model and analyze communication systems in his dissertation in 1962 [1]. Since then, Petri nets have been proven to be a very useful approach for modeling, control, simulation, and performance analysis of discrete-event systems (DESs) [2–4]. This is because Petri nets have the ability to capture the precedence relations and interactions among concurrent and asynchronous events. In addition, a strong mathematical foundation exists for describing these nets, which allows a qualitative analysis of system properties.

This chapter introduces Petri net basics and Petri-net-based modeling techniques. As the background knowledge to Petri nets, the event-driven characteristics and the state evolution of a DES are described in Sect. 4.2. The basis of establishing a formal way to model the logic behavior of a DES is illustrated by state automata in Sect. 4.3. Sects 4.4, 4.5, and 4.6 give a formal introduction to Petri nets basics, properties, and subclasses, respectively. Graph- and matrix-based techniques for analyzing Petri net properties are presented in Sect. 4.7. Sect. 4.8 covers the methods of developing timed Petri net models for system performance analysis. To aid the construction and the analysis of large Petri net models, Sect. 4.9 provides Petri net model synthesis techniques that lead to modeling modules and procedures. Finally, extensions to an original Petri net are described in Sect. 4.10.

4.2 DISCRETE-EVENT SYSTEMS

A DES is characterized by the progression of discrete events. Specifically,

1. The event may be identified with a specific action taken, or viewed as a spontaneous occurrence dictated by nature.
2. The states can only be changed at discrete points in time, which physically correspond to occurrences of these asynchronously generated discrete events.

A *sample path* is often used to describe the state evolution of a DES as shown in Fig. 1. From the sample path, it is easy to see that there are two ways to determine the inputs of a DES:

1. The inputs are deterministically specified by an event sequence $\{e_1, e_2, e_3, \ldots\}$, without considering any time information about the occurrence of these events.
2. The inputs are deterministically specified as a timed event sequence $\{(e_1, t_1), (e_2, t_2), \ldots\}$.

4.3 STATE AUTOMATA

The theory of languages and automata is the basis of establishing a formal way to model the logical behavior of a DES [3]. Assume $E = \{e_1, e_2, e_3\}$ is a finite

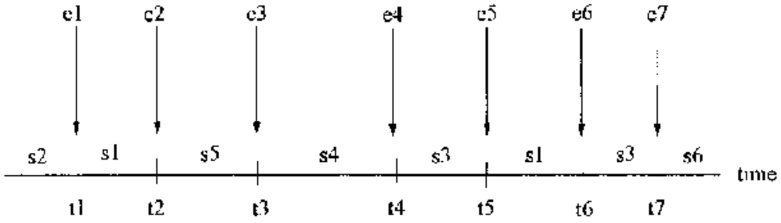


Figure 1 An example of the sample path in a DES.

event set that describes three different events of a given DES. A sequence of events taken out from E forms a string that could be thought of as the "words" in a language L. For example, let $L = \{e_1e_3, e_1e_2e_2\}$ be a language that contains two strings, e_1e_3 and $e_1e_2e_2$. Each string describes a specified sequence of events that are taken out from the event set $E = \{e_1, e_2, e_3\}$.

Such a language could be automatically generated by a "device" called an *automaton*. The characteristics of an automaton include:

1. It is intended to observe strings of events formed according to some event set E.
2. It is endowed with a state set S, and some $s_0 \in S$ is designated as the initial state.
3. It obeys a set of rules of the form: if the state is s and event e is observed, then the state must change to $s' = f(s, e)$ where s' denotes the next state, and f is a state transition function, $f\colon S \times E \to S$.

Definition 1. An *automaton* is a five-tuple

$$(E, S, f, s_0, F)$$

where

E is a finite event set.
S is a finite state set.
f is a state transition function, $f\colon S \times E \to S$.
s_0 is an initial state, $s_0 \in S$.
F is a set of final states, $F \subseteq E$.

If the state space of an automaton is constrained to be finite, then it is called a "finite state" automaton.

A *state transition diagram* is often used to represent an automaton as shown in Fig. 2. In a state transition diagram, circles denote states and arcs denote events. An arc that connects two circles represents a state transition as a result of an event taking place.

Decidability is an attractive feature of a finite-state automaton model. For example, by looking at a state transition diagram, it is easy to find whether or not a particular event sequence is recognized by an automaton model.

The limitations imposed by an automaton model include that:

1. The generation of an automaton model becomes more and more difficult as the modeled system becomes complex. This is because the total states of the system must be known.
2. An automaton model must be modified as the state information changes. This is because each automaton model represents only a fixed number of states.
3. The combination of multiple systems or asynchronous processes of a modeled system quickly increases the complexity of the model and hides some of the intuitive structure involved in this combination.

4.4 ORIGINAL PETRI NET BASICS

Like state automata, a Petri net is a "device" which manipulates events according to certain rules. Unlike

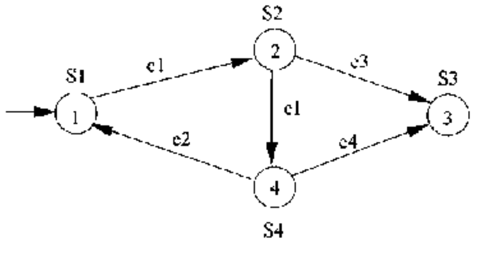


Figure 2 A state transition diagram of a finite-state automaton.

state automata, a Petri net includes explicit conditions under which an event can be enabled. Using this feature of a Petri net, one may represent a very general DES whose operation depends on complex control schemes [3].

4.4.1 Petri Net Notation and Definitions

A Petri net consists of *transitions*, *places*, *arcs*, and *tokens*. Generally, transitions represent the events driving a DES, and are normally denoted by the set $T = \{t_j\}$ for $j = 1, 2, ..., m$. Places describe the conditions related to the occurrence of events, and are normally denoted by the set $P = \{p_i\}$ for $i = 1, 2, ..., n$. Some places are viewed as the *input* places to a transition and denoted by the set $I(t_j)$; they are associated with conditions required for this transition to occur. Other places are viewed as the *output* places of a transition and denoted by the set $O(t_j)$; they are associated with conditions that are affected by the occurrence of this transition. A typical arc is of the form (p_i, t_j) or (t_j, p_i) that represents the connection of the input place p_i to transition t_j, or the output place p_i to transition t_j. If the input place p_i of transition t_j is also the output place of the same transition, then the arc forms a *self-loop* for transition t_j.

Definition 2. A *pure* Petri net has no self-loops.

If a Petri net is not pure, one can transform it into a pure Petri net by adding appropriate dummy places and transitions.

Multiple arcs are allowed to connect place p_i and transition t_j, or, equivalently, a weight can be assigned to a single arc representing the number of arcs. For example, if $w(p_i, t_j) = k$, it means either there are k arcs from p_i to t_j or a single arc is accompanied by its weight k.

A *token* in a Petri net, represents something we "put in a place" essentially to indicate the fact that the condition described by that place is satisfied. Tokens in a Petri net form a marking.

Definition 3. A *marking* of a Petri net is defined as a function $m : P \rightarrow \{0, 1, 2, ...\}$. Actually, a marking is expressed by a column vector $m = [m(p)_1), m(p_2), ..., m(p_n)]^T$, where n is the number of places in a Petri net, and $m(p_i) \in \{0, 1, 2, ...\}$ is the ith entry of this vector that indicates the (nonnegative integer) number of tokens in place p_i.

Formally, a Petri net is defined as the following:

Definition 4. A marked *Petri net* is a five-tuple

$$(P, T, A, w, m_0)$$

where:

P is a finite set of *places*.
T is a finite set of *transitions*.
A is a set of *arcs,* a subset of the set $(P \times T) \cup (T \times P)$.
w is a *weight function*, $w: A \rightarrow \{1, 2, 3, ...\}$.
m_0 is an *initial marking*.

A *Petri net graph* is often used to explicitly list the places, transitions, arcs, arc weights, and tokens of a Petri net as shown in Fig. 3.

Example 1. The Petri net in Fig. 3 is specified by

$$\begin{aligned}
&P = \{p_1, p_2, p_3,\}\\
&T = \{t_1, t_2, t_3\}\\
&A = \{(p_1, t_1), (p_2, t_2), (p_2, t_3), (p_3, t_3),\\
&\qquad (t_1, p_2), (t_2, p_3), (t_2, p_1), (t_3, p_3)\}\\
&w(p_1, t_1) = 1, w(p_2, t_2) = 1, w(p_2, t_3) = 1,\\
&\qquad w(p_3, t_3) = 1\\
&w(t_1, p_2) = 1, w(t_1, p_3) = 1, w(t_2, p_1) = 1,\\
&\qquad w(t_3, p_3) = 1
\end{aligned}$$

The given initial marking is

$$m_0 = [m(p_1),\ m(p_2),\ m(p_3)]^T = [1, 0, 0]^T$$

Definition 5. The *state* of a Petri net is its marking $m = [m(p_1), m(p_2), ..., m(p_n)]^T$. Theoretically, the num-

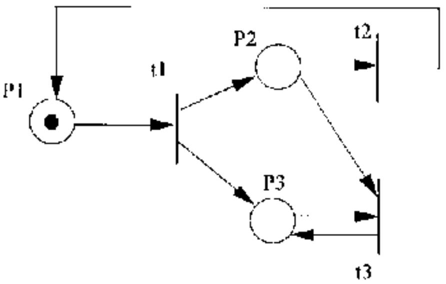


Figure 3 A Petri net described by a Petri net graph.

ber of tokens assigned to a place is an arbitrary nonnegative integer, not necessarily bounded. This causes the fact that the number of markings of a Petri net may be, in general, infinite. Thus, the state space S of a marked Petri net with n places is defined by all n-dimensional vectors whose entities are nonnegative integer markings, that is $S = \{0, 1, 2, ...\}^n$.

4.4.2 State Transition Function of a Petri Net

A Petri net can be used to model a dynamic DES, because it allows tokens to flow through the net as transitions become enabled. To enable a transition, tokens must present in each input place of the transition.

Definition 6. If a transition $t_j \in T$ in a Petri net is said to be *enabled*, the following condition must be true:

$$m(p_i) \geq w(p_i, t_j) \qquad \text{for all } p_i \in I(t_j) \tag{1}$$

An enabled transition can *fire*. The firing of a transition causes a change of the marking in a Petri net. This is defined as the *state transition function* of a Petri net.

Definition 7. The *state transition function* $f(m, t_j)$ of a Petri net (P, T, A, w, m_0) is defined for an enabled transition $t_j \in T$ as $m' = f(m, t_j)$, where

$$\begin{aligned} m'(p)_i) = m(p_i) + w(t_j, p_i) - w(p_i, t_j) \\ \text{for } i = 1, 2, ..., n \end{aligned} \tag{2}$$

It is interesting to note that:

1. The state transition function is defined only for transitions that satisfy Eq. (1). In this sense, an "enabled transition" in a Petri net is equivalent to the idea of a "feasible event" in state automata.
2. The state transition of a Petri net is based on the structure defined by Eq. (2). Thus, the state transition function is not arbitrary.

According to Eq. (2), the number of tokens in a marked Petri net need not to be conserved. This implies that tokens may be completely lost, or grown to infinity after several firings of transitions.

Example 2. To illustrate the process of firing transitions and changing the state of a Petri net, consider the Petri net in Fig. 3. Since transition t_1 requires a single token in place p_1 to enable itself and place p_1 initially has one token [i.e., $m(p_1) = 1 = w(p_1, t_1)$], thus Eq. (1) is satisfied for t_1. When t_1 fires, one token is removed from place p_1 and placed in places p_2 and p_3. Using Eq. (2),

$$\begin{aligned} &m'(p_i) = m(p_i) + w(t_j, p_i) - w(p_i, t_j) \\ &\text{for } i = 1, 2, 3 \text{ and } j = 1 \end{aligned}$$

the new marking m_1 is calculated as follows:

$$\begin{aligned} m_1(p_1) &= m_0(p_1) - w(p_1, t_1) + w(t_1, p_1) \\ &= 1 - 1 + 0 = 0 \\ m_1(p_2) &= m_0(p_2) - w(p_2, t_2) + w(t_1, p_2) \\ &= 0 - 0 + 1 = 1 \\ m_1(p_3) &= m_0(p_3) - w(p_3, t_1) + w(t_1, p_3) \\ &= 0 - 0 + 1 = 1 \end{aligned}$$

Therefore,

$$m_1 = [m(p_1), m(p_2), m(p_3)]^T = [0, 1, 1]^T$$

4.4.3 State Equation of a Petri Net

Given a Petri net that has n places and m transitions, a vector state equation can be generated from Eq. (2) and its *incidence matrix* [2,5].

Definition 8. The *incidence matrix* of a Petri net, $A = [a_{ij}]$, is an $(n \times m)$ matrix whose (i, j) entry is of the form

$$a_{ij} = w(t_j, p_i) - w(p_i, t_j) \tag{3}$$

where

$w(t_j, p_i)$ = the number of arcs (or a single arc accompanied by its weight) from transition j to output place i.

$w(p_i, t_j)$ = the number of arcs (or a single arc accompanied by its weight) from input place i to transition j.

[Note: the definition of a_{ij} also matches the weight difference that appears in Eq. (2) in updating $m(p_i)$].

Definition 9. The *firing vector* u_k is an $(m \times 1)$ vector of the form

$$u_k = [0, 0, 0, 0, ..., 1, 0, ..., 0]^T \tag{4}$$

where the number 1 that appears in the jth position, $j = 1, 2, 3, ..., m$, indicates the fact that jth transition is currently firing.

Let m_k denote the marking of the Petri net after its kth execution, with $k \geq 0$. Using Eq. (2) and its incidence matrix A, a vector state equation can be written as

$$m_{k+1} = m_k + Au_k \tag{5}$$

The ith expression of Eq. (5) is precisely Eq. (2). Therefore, the transition function $f(m, t_j)$ in Def. 7, now becomes $m_k + Au_k$, where the argument t_j in this function represents the jth nonzero entry in u_k.

Equation (5) is a simple linear-algebraic equation. It may be used to model the dynamic behavior of a concurrent system and aid in determining the properties of the model. Since the number of tokens in a place is nonnegative, the legal execution (firing of enabled transitions) of a Petri net guarantees that:

$$m_k + Au_k \geq 0 \qquad \text{for all } k \geq 0 \tag{6}$$

Using Eq. (5), it is possible to test if a given marking m_f is reachable from an initial marking m_0. Suppose that m_f is reachable from m_0 by the successive firing of certain transitions. Then,

$$\begin{aligned} m_1 &= m_0 + Au_0 \\ m_2 &= m_1 + Au_1 \\ &\vdots \\ m_f &= m_{f-1} + Au_{f-1} \end{aligned}$$

Combining the above equations,

$$m_f = m_0 + Au \tag{7}$$

where u is an $(m \times 1)$ column vector of nonnegative integers that sums all u_i, $i = 1, 2, ..., f - 1$. (Note: u is also called the *firing count vector*.)

The ith entry of u denotes the number of times that t_i must fire to bring the system from m_0 to m_f.

Example 3. Let us reconsider the Petri net in Ex. 2 and obtain state transitions using Eq. (5) and Eq. (7). We first write down the incidence matrix A.

Using Eq. (3) and the results from Ex. 1, then

$$\begin{aligned}
a_{11} &= w(t_1, p_1) - w(p_1, t_1) = 0 - 1 = -1 \\
a_{12} &= w(t_2, p_1) - w(p_1, t_2) = 1 - 0 = 1 \\
a_{13} &= w(t_3, p_1) - w(p_1, t_3) = 0 - 0 = 0 \\
a_{21} &= w(t_1, p_2) - w(p_2, t_1) = 1 - 0 = 1 \\
a_{22} &= w(t_2, p_2) - w(p_2, t_2) = 0 - 1 = -1 \\
a_{23} &= w(t_3, p_2) - w(p_2, t_3) = 0 - 1 = -1 \\
a_{31} &= w(t_1, p_3) - w(p_3, t_1) = 1 - 0 = 1 \\
a_{32} &= w(t_2, p_3) - w(p_3, t_2) = 0 - 0 = 0 \\
a_{33} &= w(t_3, p_3) - w(p_3, t_3) = 1 - 1 = 0
\end{aligned}$$

Thus, the incidence matrix is

$$A = \left[a_{ij}\right] = \begin{bmatrix} -1 & 1 & 0 \\ 1 & -1 & -1 \\ 1 & 0 & 0 \end{bmatrix}$$

Using Eq. (5), the reachable markings can be determined as

$$\begin{aligned}
m_1 &= m_0 + Au_0 \\
&= [1\ 0\ 0]^T + A[1\ 0\ 0]^T = [1\ 0\ 0]^T + [-1\ 1\ 1]^T \\
&= [0\ 1\ 1]^T \\
m_2 &= m_1 + Au_1 \\
&= [0\ 1\ 1]^T + A[0\ 1\ 0]^T = [0\ 1\ 1]^T + [1\ -1\ 0]^T \\
&= [1\ 0\ 1]^T \\
m_3 &= m_1 + Au_2 \\
&= [0\ 1\ 1]^T + A[0\ 0\ 1]^T = [0\ 1\ 1]^T + [0\ -1\ 0]^T \\
&= [0\ 0\ 1]^T
\end{aligned}$$

The reachable marking m_2 and m_3 can also be determined by Eq. (7). In the case of m_2, we have

$$\begin{aligned}
u &= [1, 1, 0]^T \\
m_2 &= m_0 + Au = [1\ 0\ 0]^T + A\,[1\ 1\ 0]^T = [1\ 0\ 1]^T
\end{aligned}$$

Similarly, in the case of m_3, we have

$$\begin{aligned}
u &= [1, 0, 1]^T \\
m_3 &= m_0 + Au = [1\ 0\ 0]^T + A\,[1\ 0\ 1]^T = [0\ 0\ 1]^T
\end{aligned}$$

4.5 PROPERTIES OF ORDINARY PETRI NETS

Petri nets have two types of properties. Behavioral properties of a Petri net are dependent on the initial marking, and structural properties of a Petri net are independent of the initial marking.

4.5.1 Behavioral Properties of a Petri Net

4.5.1.1 Reachability

Given a Petri net, it is interesting to know the sequence of firing transitions which lead the initial marking m_0 to a given marking m_r.

Definition 10. A marking m_r can be reached from m_0 if there exists a *firing sequence* s_r that will yield m_r. In fact, s_r is the same as the firing count vector u defined in Eq. (7).

A *reachability set* is the set that contains all possible markings reachable from m_0. A state reachability problem for a Petri net is to ensure some desirable states must be reachable and some undesirable states must be avoided. Hence, finding the reachable firing sequence (i.e., s_r or u) becomes an important issue in the analysis of Petri net models.

4.5.1.2 Boundedness

Definition 11. A *k-bounded* Petri net with respect to an initial marking m_0 has at most k tokens for all markings in the reachability set. A Petri net is *safe* if it is k-bounded and $k = 1$.

Unbounded growth in markings leads to some form of instability of a Petri net model. A boundedness problem of a Petri net is to check if the net is bounded, and determine a bound. If the boundedness property is not satisfied, the Petri net model must be altered so as to ensure the boundedness.

4.5.1.3 Liveness

A *deadlock* occurs in a Petri net when a transition or a set of transitions cannot fire. Transitions which are not involved in a deadlock are called *live*.

Definition 12. A *live* Petri net with respect to a given initial marking m_0 always has some firing sequence such that any transition can eventually fire from any state reached from m_0.

The *liveness* defined in Def. 12 is a strong property that guarantees the absence of deadlocks in a Petri net model regardless of the firing sequence. Practically, it is often infeasible to check the liveness defined in Def. 12 for many systems modeled using Petri nets. As a result, the degree of the friability of transitions in a Petri net is defined in terms of levels of liveness.

Given an initial marking m_0, a transition in a Petri net may be:

Level 0: it never fires (dead transition).
Level 1: it fires at least once for some firing sequence from m_0.
Level 2: it fires k times for some given positive integer k.
Level 3: it fires infinitely for some infinite firing sequence.
Level 4: it fires at *level* 1 for every possible state reached from m_0.

A Petri net is i level live if every transition is live at level i. Deadlock avoidance in Petri net is to find potential transition sequences that can lead system to some form of deadlock.

4.5.1.4 Reversibility

Definition 13. A *reversible* Petri net is the net that the initial marking m_0 can be reached from all reachable markings.

A Petri net with the property of *reversibility* implies that the modeled system (or Petri net) can return to an admissible state (or marking) in a finite number of steps (or transition firings).

4.5.1.5 Conservation

Definition 14. A *strict conservative* Petri net with respect to a given initial marking m_0, remains a constant number of tokens in the net, that is, $\Sigma\, m'(p_i) = \Sigma\, m_0(p_i)$.

The *strict conservation* defined in Def. 14 is a strong property, because it requires that all reachable markings from m_0 must have the same number of tokens. Practically, one only requires that the weighted sum of tokens for all reachable markings be constant.

Definition 15. A *conservative* Petri net with a given initial marking m_0 has:

1. a $(1 \times n)$ vector $x = [x_1, x_2, ..., x_n]$;
2. a condition: $x[m(p_1), m(p_2), ..., m(p_n)]^T =$ Constant, for all reachable markings.

4.5.2 Structural Properties of a Petri Net

Because the structural properties of a Petri net is independent of the initial marking, the determination of structural properties of a Petri net is closely related to the topological structure of the net. The following are the definitions of several structural properties [6,7]:

Definition 16. A *structural bounded* Petri net is bounded for any initial marking.

Definition 17. A *structural live* Petri net is live for any initial marking.

Definition 18. A *structural conservative* Petri net has an vector x for any initial marking m_0 and a reachable marking m, such that $x_i \neq 0$ for all $i = 1, 2, ..., n$, and $x^T m = x^T m_0$.

Definition 19. A *structural consistent* Petri net has a marking m, and a firing sequence s_c (called cyclic firing sequence) that is expressed by a firing vector v with nonzero element in it such that s_c brings the net back to marking m by firing each transition at least once.

Definition 20. A *partial consistent* Petri net has a firing vector v that has some zero elements. This implies that some transitions do not occur in v when s_c brings the net from m marking back to itself.

Definition 21. A *complete controllable* Petri net has any marking that is reachable from any initial marking. (This is a strong property.)

Definition 22. A *repetitive* Petri net has a finite marking m_0 and a firing sequence s such that the elements of the associated firing vector v are infinity. If v contains only some of the transitions, the Petri net is *partially repetitive*.

4.6 SUBCLASSES OF PETRI NETS

Two basic subclasses of an original Petri nets are called *marked graph* and *state machine* as shown in Fig. 4. Each class has a certain modeling power and its own properties.

4.6.1 Marked Graph

A marked graph can be used to model the situations that exhibit the characteristics of concurrence and synchronization.

Definition 23. If each place of a Petri net has exactly one input and one output transition, the Petri net is called a *marked graph.*

The analysis of a marked graph uses the concepts of *directed path* and *directed circuit* defined in a Petri net graph.

Definition 24. A *directed path* is the path that starts from one node (place or transition) and ends on another one.

Definition 25. A *directed circuit* is a directed path from one node back to itself. This also implies that a directed circuit has no places that (1) have no input transitions (*source places*); and (2) have no output transitions (*sink places*).

Definition 26. An *elementary circuit* has no node that appears more than once on a directed circuit, other than the starting node.

Example 4. For the marked graph in Fig. 4a, there are total four elementary circuits. They are: circuit 1: $p_2t_1, p_3t_2, p_5t_5, p_7t_4$; circuit 2: p_2t_1, p_4t_3, p_6t_4; circuit 3: $p_1t_1, p_3t_2, p_5t_5, p_7t_4$; and circuit 4: p_1t_1, p_4t_3, p_6t_4.

The liveness and boundedness properties of a marked graph with respect to an initial marking m_0 can be determined through its directed circuit.

Boundedness. Because each place on a directed circuit of a marked graph has only one input transition and one output transition, firing transitions

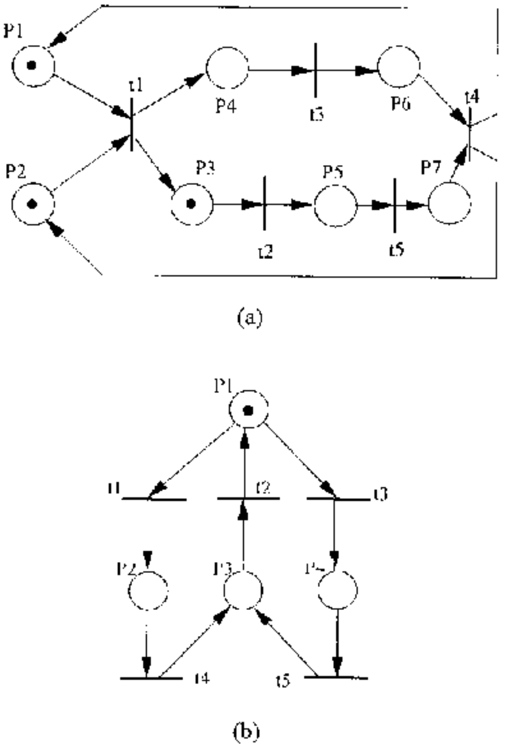


Figure 4 (a) A marked graph; (b) a state machine.

along the directed circuit does not change the number of tokens on the directed circuit.

Liveness. A live marked graph must have the facts that: (1) the directed circuits cover all transitions; and (2) the initial marking m_0 places at least one token on each directed circuit.

A safe marked graph has the facts that: (1) it is a live Petri net; (2) the directed circuits cover every place; and (3) the initial marking m_0 places exactly one token on each directed circuit.

4.6.2 State Machine

Definition 27. If each transition of a Petri net has exactly one input place and one output place, the Petri net is called a *state machine*.

The unique modeling feature of a state machine is its ability to represent conflicts and decision-making structures. However, a state machine has no ability to model the situation of synchronization or concurrence as a marked graph does. This is because: (1) each transition in a state machine can only have one input place and one output place; (2) the same place may be an input place to more than one transition. Thus, the enabled transition can be disabled by the firing of another transition.

The liveness and boundedness properties of a state machine can be determined by using the concept of *strongly connected Petri net*.

Definition 28. A *strongly connected Petri net* has a *directed path* from every node to every other node in the net.

Liveness. A live state machine with respect to initial marking m_0 must: (1) be strongly connected; and (2) have at least one token placed in m_0.

Boundedness. A safe state machine with respect to initial marking m_0 must: (1) be a live Petri net; and (2) have exactly one token placed in m_0.

Conservation. A state machine is strictly conservative. This property implies that the facts that: (1) the state space of a state machine is finite; (2) any finite state automata can be modeled with a state machine [4,6].

4.6.3 Features of Petri Nets

When modeling a given DES using Petri nets, one only uses the local state and does not need to know the state of the rest of the system. This feature means that a Petri net model is more easily expended. For example, when a component is added to the modeled system, the corresponding Petri net model only expends locally, which is achieved by simply adding a few places and/or transitions that represent the coupling effects between the added component and the original modeled system. Moreover, by looking at a Petri net graph, one can conveniently see the individual components, discern the level of their interaction, and ultimately decompose a system into logical distinct modules. Therefore, a Petri net has the ability to *decompose* or *modularize* a complex system.

It has also been proven the fact that a finite state automaton can always be represented by a Petri net. This scheme supports the claim that Petri nets are indeed very general modeling tools. However, it is not always easy for one to deal with the decidability of a Petri net model. This reflects a natural tradeoff between the decidability and richness of a Petri net model [3].

4.7 ORIGINAL PETRI NET ANALYSIS

4.7.1 Reachability Tree and Coverability Tree

The techniques of reachability and coverability form the basis of the graph-based methods for analyzing Petri net properties. The key step of using these technologies is based on constructing a *reachability* tree or a coverability tree where the nodes are Petri net markings and the arcs represent transitions as shown in Fig. 5. Generally, the following standard notations are used.

Definition 29. *Root node* is the first node of the tree, corresponding to a given initial marking.

Definition 30. *Terminal node* is any node from which no transition can fire.

Definition 31. *Duplicate node* is a node which is identical to a node already in the tree.

Let $m = [m(p_1), m(p_2), ..., m(p_n)]^T$ and $m' = [m'(p_1), m'(p_2), ..., m'(p_n)]^T$ be two any nodes in the tree.

Definition 32. *Node dominance* describes the case that for any two nodes, m' and m, in the tree, one may say m' dominates m, denoted by $m' >_d m$, if the following two conditions hold:

1. $m'(p_i) \geq m(p_i)$ for all $i = 1, ..., n$;
2. $m'(p_i) > m(p_i)^-$ for at least some $i = 1, ..., n$.

Definition 33. *The symbol* ω represents a node dominance relationship in the tree, that is, if $m' >_d m$, then for all i such that $m'(p_i) >_d m(p_i)$ the value of $m(p_i)$ is replaced by the symbol ω. Note that adding (or subtracting) tokens to (or from) a place which is already marked by ω does not have any effect, that is, $\omega \pm k = \omega$ for any $k = 0, 1, 2, ...$.

If the symbol ω appears in a reachability tree, the tree is called *coverability tree*. A coverability tree contains every reachable marking from m_0 that is either explicitly represented by a node, or covered by a node through the ω notation. Hence, the name "coverability tree" is more appropriate than "reachability tree".

Adopting the above standard notation and procedures, the coverability tree can be constructed using algorithms [2,3]. It is not difficult to see that a coverability tree is essentially a state transition diagram. Thus, by looking at the coverability tree of a Petri net model, one can examine all possible markings and transition sequences.

Example 5. To construct the reachability tree and coverability tree for the Petri net as shown in Fig. 3, one needs to find all possible markings. The process starts with initial marking $[1, 0, 0\,]^T$ which makes transition t_I enabled. The firing of t_1 leads to a new marking $[0, 1, 1]^T$. Now either t_2 or t_3 can fire creating two branches, with corresponding next markings $[1, 0, 1]^T$ and $[0, 0, 1]^T$, respectively. Since no transition can fire along the right branch, $[0, 0, 1]^T$ is a terminal node. The left branch allows transition t_1 to fire, with a new marking $[0, 1, 2\,]^T$. Once again, either t_2 or t_3 can fire, and the corresponding next markings are $[1, 0, 2\,]^T$ and $[0, 0, 2\,]^T$, respectively. The reachability tree is shown in Fig. 5a. Because of the fact: $[1, 0, 2\,]^T >_d [1, 0, 1\,]^T$, the value of $m(p_3)$ is replaced by the symbol ω in the coverability tree as shown in Fig. 5b.

4.7.2 Determining the Behavioral Properties Using Coverability Tree

Boundedness. A bounded Petri net must have no symbol ω appeared in its coverability tree. In this case, the state space of the modeled DES is finite. The largest value of $m(p_i)$ for any marking in the tree specifies a bound at place p_i.

Safeness. If the value of markings in a coverability tree only contains 0 and 1, the associated Petri net is safe.

Coverability. (1) By inspecting the coverability tree, it is always possible to determine whether marking m' covers marking m by checking the following condition:

$$m'(p_i) \geq m(p_i) \qquad \text{for all } i = 1, 2, ..., n.$$

(2) If m' contains ω in one or more places, it means this path must include a loop. Depending on the particular values of $m'(p_i)$, $i = 1, ..., n$, it is pos-

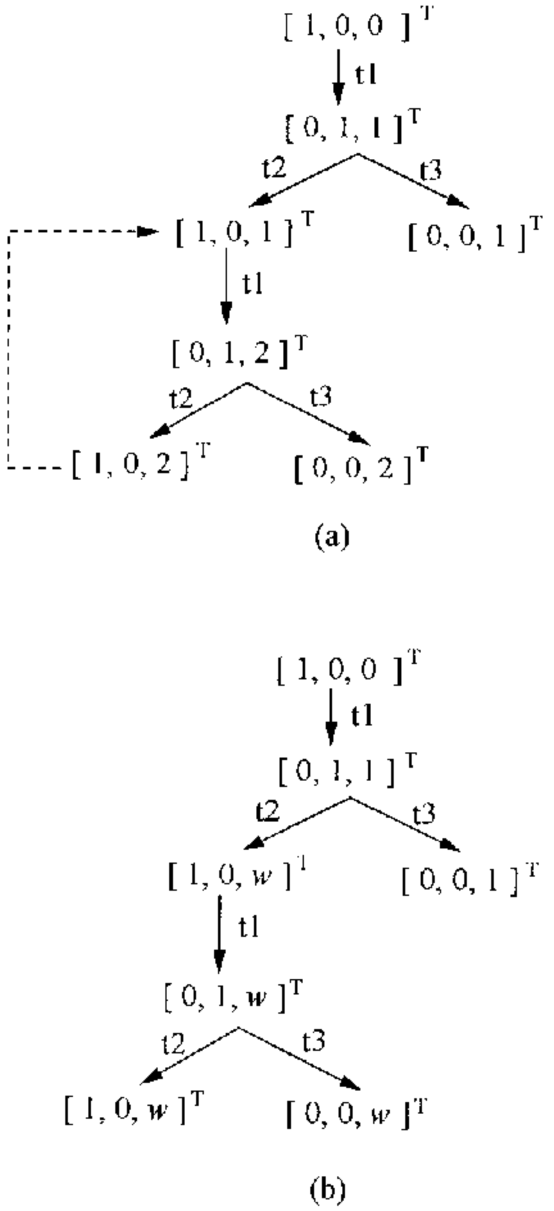


Figure 5 The associated reachability tree (a) and the coverability tree (b) of the Petri net in Fig. 3.

sible to determine the number of loops involved in this path until marking m is covered.

(3) It is also possible to identify dead transitions by checking for coverability.

Consistency. A consistent Petri net with respect to an initial marking m_0 must have a directed circuit (not necessary elementary circuit) in the coverability tree such that the directed circuit contains all the transitions at least once. A partial consistent Petri net has a directed circuit in the coverability tree such that the directed circuit contains only some of the transitions.

4.7.3 Structural Analysis Using State Equation

4.7.3.1 Controllability

One application of using Eq. (7) is to study the *complete controllability* of a Petri net model. A Petri net is *completely controllable* if any marking is reachable from any initial marking m_0.

Theorem 1. If a Petri net with n places and m transitions is completely controllable, the rank of its incidence matrix A must equal to n. (Note: Theorem 1 does not assume $m = n$.)

Since Theorem 1 is rarely satisfied in most Petri net applications, the reachability problem becomes more important. For the case of reachability, the rank of incidence matrix A is r, where $r < n$.

The following equation that is derived from Eq. (7) could be used for studying the reachability of a Petri net.

$$\Delta m = m_f - m_0 = Au = \begin{bmatrix} A & A & \ldots & A \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ \cdot \\ \cdot \\ \cdot \\ u_{f-1} \end{bmatrix} \qquad (8)$$

where Δm is the change of the marking between the initial marking and the final marking.

Given any Δm, one must solve Eq. (8) for $u_0, u_1, ..., u_{f-1}$ in order to determine if the Petri net model can reach any other marking from the initial marking.

4.7.3.2 Concepts of P-Invariants and T-Invariants

P-Invariants. In general, one would like to find a weighting vector that makes all reachable markings of a Petri net be constant. This vector is called *P-invariant*.

Definition 34. If an $(n \times 1)$ vector x is an nonnegative integer that satisfies

$$x^T A = 0 \tag{9}$$

then this vector is called a *P-invariant*.

A given Petri net may have a group of independent P-invariants as a result of Eq. (9). If every place in a Petri net belongs to some P-invariant, the Petri net is covered by P-invariants.

Multiplying a P-invariant to Eq. (7), then

$$x^T m = x^T m_0 \tag{10}$$

Equation (10) implies that the total number of initial tokens in m_0, weighted by the P-invariant, is constant.

T-Invariants

Definition 35. If an $(m \times 1)$ vector y is a nonnegative integer vector that satisfies

$$Ay = 0 \tag{11}$$

then y is called a *T-invariant*.

If the firing count vector u in Eq. (7) is identical to a T-invariant, that is, $y = u$, then

$$m_f = m_0 \tag{12}$$

This implies that the final marking is equal to the initial marking through a firing sequence defined by a T-invariant.

4.7.3.3 Obtaining Invariants of a Petri Net

For a Petri net with n places and m transitions, it can be shown that there are $(n - r)$ minimal P-invariants and $(m - r)$ minimal T-invariants, where $r = \text{rank}(A)$, $r < n$, and $r < m$.

For a Petri net, the incidence matrix A with rank(A) $= r < n$ can be partitioned by rearranging the columns of A so that A_{12} has r independent columns of A, i.e., A_{12} is a nonsingular square matrix of dimension r. As a result,

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

where A_{11} is $r \times (m - r)$, A_{12} is $r \times r$, A_{21} is $(n - r) \times (m - r)$, and A_{22} is $(n - r) \times r$.

Similarly, vector y is partitioned as

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

where y_1 is $(m - r) \times 1$ and y_2 is $r \times 1$.

Applying the partitioned version of A and y to Eq. (11), then

$$A_{11}y_1 + A_{12}y_2 = 0 \tag{13}$$

$$A_{21}y_1 + A_{22}y_2 = 0 \tag{14}$$

Since A_{12} is invertable (it is nonsingular), Eq. (13) can be rewritten as

$$y_2 = -A_{12}^{-1}A_{11}y_1$$

Substituting y_2 into Eq. (14) yields

$$(A_{21} - A_{22}A_{12}^{-1}A_{11})y_1 = 0 \tag{15}$$

It is possible to solve Eq. (15) for y_1. The solution indicates that there are $(m - r)$ minimal T-invariants.

It can be verified [2] that

$$B_p = [-A_{22}A_{12}^{-1}, I] \tag{16}$$

is a solution for P-Invariants. This solution indicates that B_p has $(n - r)$ rows and so there are $(n - r)$ minimal P-Invariants. This solution also leads to the following necessary condition for reachability.

Theorem 2. In a Petri net, the final marking m_f can be reached from the initial marking m_0 through a firing sequence if $B_p \Delta m = 0$.

The converse of Theorem 2 provides the following sufficient condition for nonreachability [8].

Theorem 3. In a Petri net, m_f can not be reached from m_0 if Δm is a linear combination of the row of B_p, i.e., $\Delta m = B_p^T K$, where K is a nonzero $(n - r) \times 1$ column vector.

It is important to note that when solving $Ay = 0$ (or $x^T A = 0$) over the field of rational numbers, ignoring the nonnegative integer constraint does not affect the application of Theorem 1 through Theorem 3 [2].

Example 6. Given a Petri net with incidence matrix A, where

$$A = \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 0 & 0 \end{bmatrix}$$

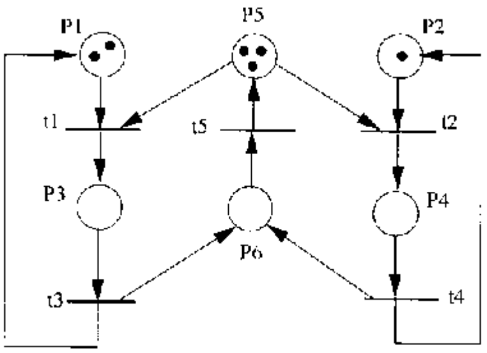


Figure 6 A Petri net model.

Determine if $m_f = (3\ 0\ 1)^T$ is reachable from initial marking $m_0 = (0\ 3\ 1)^T$. Solution:

1. Check the rank of A: $r = \text{rank}(A) = 1$.
2. Partition A as $A_{11} = 1$, $A_{12} = -1$, $A_{21} = (-10)T, and A22 = (10)T$.
3. Use Eq. (16) to determine B_p.

$$B_p = \left[-\begin{pmatrix} 1 \\ 0 \end{pmatrix} \; -1, \; \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right] = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

4. Check $B_p \Delta m$, where $\Delta m = m_f - m_0 = (3\ -3\ 0)^T$.

$$B_p = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ -3 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

According to Theorem 1, m_f can be reachable from m_0.

The solution to Eq. (9) may become more complicated when one includes the constraint that the elements of x must be nonnegative integers. In this case, an efficient algorithm [9] for computing the invariant of Petri net is proposed. The basic idea of the method can be illustrated by getting the minimal set of P-invariants through the following example.

Example 7. The incidence matrix of the Petri net in Fig. 6 is

$$A = \begin{bmatrix} -1 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ -1 & -1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & -1 \end{bmatrix}$$

A modified version of incidence matrix A is formed as

$$[A : I] = \left[\begin{array}{ccccc|cccccc} -1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ -1 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right]$$

We will add rows to eliminate a nonzero element in each row of A. Specifically, (1) adding the third row to the first row; (2) the fourth row to the second row; (3) adding the fourth, fifth, and sixth rows to the third row.

$$\left[\begin{array}{ccccc|cccccc} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ -1 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right] \Rightarrow$$

$$\left[\begin{array}{ccccc|cccccc} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ -1 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right] \Rightarrow$$

$$\left[\begin{array}{ccccc|cccccc} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ -1 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right]$$

The three P-invariants are $x_1 = (1\,0\,1\,0\,0\,0)^T$, $x_2 = (0\,1\,0\,1\,0\,0)^T$, and $x_3 = (0\,0\,1\,1\,1\,1)^T$. They are actually the first three rows of the modified identity matrix. This is because their associated three rows in the final version of modified A have all zero elements.

4.7.3.4 Invariant Analysis of a Pure Petri Net

The following example illustrates how Petri net invariants can aid the analysis of a pure Petri net.

Example 8. Consider the Petri net in Fig. 6 as a model that describes two concurrent processes (i.e., process 1 and process 2), each of which needs a dedicated resource to do the operations in the process. The tokens at p_1 and p_2, models the initial availability of each dedicated resource. Both processes also share a common resource to aid their own process. This shared resource is modeled as a token initially at p_5. The initial marking of the model is $m_0 = (2\ 1\ 0\ 0\ 3\ 0)^T$.

Applying three P-invariant results obtained from Example 7 to Eq. (10), then

$$
\begin{aligned}
&m(p_1) + m(p_3) = 2 \\
&m(p_2) + m(p_4) = 1 \\
&m(p_3) + m(p_4) + m(p_5) + m(p_6) = 3
\end{aligned}
$$

The first equation implies that the total number of resources for process 1 is 2. The second equation implies that the total number of resources for process 2 is 1. The last equation implies that the three shared resources are in a mutual exclusion situation serving for either process 1 or process 2.

The invariant analysis of a pure Petri net also includes that:

1. A structural bounded Petri net must have an $(n \times 1)$ vector x of positive integer such that $x^T A \leq 0$.
2. A conservative Petri net must have an $(n \times 1)$ vector x of positive integers such that $x^T A = 0$.
3. A repetitive Petri net must have an $(m \times 1)$ vector y of positive integers such that $Ay \geq 0$. It is partially repetitive if y contains some zero elements.
4. A consistent Petri net must have an $(m \times 1)$ vector y of positive integers such that $Ay = 0$. It is partially consistent if y contains some zero elements.

4.8 TIMED PETRI NET MODELS FOR PERFORMANCE ANALYSIS

Timed Petri net (TPN) models have been developed for studying the temporal relationships and constraints of a DES. They also form the basis of system performance analysis that includes the calculation of process cycles, resource utilization, operation throughput rate, and others. There are two approaches for modeling the time information associated with the occurrence of events in a Petri net. A *timed place Petri net* (TPPN) associates time information with places. A *timed transition Petri net* (TTPN) associates time information with transitions. Timed Petri nets can be further classified according to the time parameters assigned in the net. If all time parameters in a net are deterministic, the net is called *deterministic timed Petri net* (DTPN). If all time parameters in a net are exponentially distributed random variables, the net is called a *stochastic timed Petri net* (STPN or SPN).

4.8.1 Deterministic Timed Petri Nets

4.8.1.1 TPPN Approach

In a TPPN, every place is assigned a deterministic time parameter that indicates how long an entered token remains unavailable in that place before it can enable a transition. It is possible that during the unavailable period of a token in a place another token may arrive in the place. Only available tokens in a marking can enable a transition. The firing of a transition is carried out with zero time delay as it is in ordinary Petri nets.

The matrix-based methods can be used for the analysis of a TPPN. From Eq. (7), the marking at instant time t can be expressed as

$$m(t) = m(t_0) + Au(t)$$

where:

$m(t_0)$ is the initial marking at instant t_0 ($t_0 < t$).
$u(t)$ represents the firing vector at instant t.

Considering $\Delta t = t - t_0 \neq 0$, and $\Delta m(t) = m(t) - m(t_0) = Au(t)$, then

$$\Delta m(t)/\Delta t = Au(t)/\Delta t = Af \tag{17}$$

where:

$\Delta m(t)/\Delta t$ represents the average change in tokens over period Δt.
f represents the average firing frequency of the transitions, called *current vector*.

A direct application of Eq. (17) is to study the periodic behavior of a *consistent* Petri net. According to Definitions 19 and 20, a consistent Petri net has some firing sequence that returns a marking m back to itself. In the case of a consistent Petri net modeled as a TPPN, it has $m(t) - m(t_0) = 0$. Substituting this fact into Eq. (17), then

$$Af = 0, \ f > 0 \tag{18}$$

Actually, solving Eq. (18) for a TPPN is equivalent to determining the T-invariants with the exception of the time scaling factor. The performance of the model such as throughput and resource utilization can be determined using the current vectors that are associated with the places representing the resource.

Through P-invariants a relationship can also be established, relating the initial marking m_0, the deterministic time delays in timed places, and the firing frequencies of the transitions [2]:

$$x^T m_0 = x^T D A^+ f \tag{19}$$

where:

x is a P-invariant.
D contains the deterministic time delays for timed places, and $D = \operatorname{diag}\{d_i\}$ for $i = 1, 2, ..., n$.
A^+ is the output part of the incidence matrix A.

In Eq. (19), $A^+ f$ indicates the average frequency of token arrivals and $DA^+ f$ indicates the average number of tokens due to the delay restrictions.

If Eqs. (18) and (19) are satisfied for all P-invariants, the TPPN model functions at its the maximum rate. The applications of this approach can be found in Sifakis [10].

Example 9. Figure 6 shows a TPPN model, where

$$D = \{2, 1, 2, 3, 6, 4\}$$

$$DA^+ f = \begin{bmatrix} d_1 & & & & & \\ & d_2 & & & & \\ & & d_3 & & & \\ & & & d_4 & & \\ & & & & d_5 & \\ & & & & & d_6 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \end{bmatrix}$$

$$= \begin{bmatrix} d_1 f_1 & d_2 f_4 & d_3 f_3 & d_4 f_2 & d_5 f_5 & d_6(f_4 + f_3) \end{bmatrix}^T$$

Given the initial marking m_0 as $[K_1 \ K_2 \ 0 \ 0 \ \ K_5 \ 0]^T$, where:

K_1 represents the number of tokens (dedicated resource) at p_1.
K_2 represents the number of tokens (dedicated resource) at p_2.
K_5 represents the number of tokens (shared resource) at p_5.

Applying the P-invariants results obtained in Example 7 to Eq. (19), then

$$K_1 = (d_1 + d_3) f_1$$
$$K_2 = (d_2 + d_4) f_2$$
$$K_5 = (d_5 + d_6)(f_2 + f_1) + d_3 f_1 + d_4 f_2$$

These three equations provide a relation that relates the initial markings, the time delays, and the firing frequencies. Let $K_1 = 2$, $K_2 = 1$, then

$$f_1 = 1/2, f_2 = 1/4$$
$$K_5 = (6 + 4)(f_2 + f_1) + 2f_1 + 3f_2 = 9.25$$

For a maximum rate, the minimum number of shared resource (K_5) required in the model must be 10.

4.8.1.2 TTPN Approach

In a TTPN framework [11], deterministic time parameters are assigned to transitions. A timed transition is enabled by removing appropriate number of tokens from each input place. The enabled transition fires after a specified amount of time by releasing the tokens to its output places. A direct application of TTPN is the computation of cycle time in a marked graph [2].

Definition 36. The cycle time C_i of transition t_i is defined as

$$C_i = \lim_{n_i \to \infty} S_i(n_i)/n_i, \tag{20}$$

where $S_i(n_i)$ is the time at which transition t_i initiates its n_ith execution.

Theorem 4. A marked graph has the same cycle time for all transitions.

Theorem 5. The minimum cycle time (maximum performance) C_m of a marked graph is

$$C_m = \max_k \{T_k / K_k\} \text{ for } k = 1, ..., c \tag{21}$$

where:

T_k is the sum of transition delays in a circuit k.
K_k is the sum of the tokens in a circuit k.
c is the number of circuits in the model.

The processing rate (or throughput) can be easily determined from the cycle time by

$$\lambda = 1/C_m = \min\{K_k / T_k, k = 1, 2, ..., c\} \tag{22}$$

Example 10. Let us determine the minimum cycle time for the marked graph in Fig. 4, where $\{d_i\} = \{5, 20, 4, 3, 6\}$. Using the elementary circuit results in Example 4 and apply Eq. (21) to each elementary circuit, then

Circuit 1: $T_1/K_1 = (5 + 20 + 3 + 6) / 2 = 17.$
Circuit 2: $T_2/K_2 = (5 + 4 + 3) / 1 = 12.$
Circuit 3: $T_3/K_3 = (5 + 20 + 3+ 6) / 2 = 17.$
Circuit 4: $T_4/K_4 = (5 + 4 + 3) / 1 = 12.$

Therefore, the minimum cycle time $C_m = \max\{17, 12, 17, 12\} = 17$.

4.8.2 SPN and GSPN

Stochastic Petri nets (SPNs) have been developed [12,13] to model the nondeterministic behavior of a DES.

Definition 37. A continuous-time stochastic Petri net, SPN, is defined as a TTPN with a set of stochastic timed transitions. The firing time of transition t_i is an exponentially distributed variable with firing rate $\lambda_i > 0$.

Generalized stochastic Petri nets (GSPNs) [14] incorporate both stochastic timed transitions and immediate transitions. The immediate transitions fire in zero time. Additional modeling capabilities are introduced to GSPNs without destroying the equivalence with Markov chains. They are inhibitor arcs, priority functions, and random switches. An inhibitor arc in the net prevents a transition from firing when certain conditions are true. A priority function specifies a rule for the marking in which both timed and immediate transitions are enabled. The random switch, as a discrete probability distribution, resolves conflicts between two or more immediate transitions.

The generation of a Markov chain can be greatly simplified through SPN and GSPNs approaches. Specifically, one needs to:

1. Model the system with a SPN or GSPN.
2. Check the liveness and boundedness of the model by examining the underlying Petri net model with either reachability tree or invariants analysis. The liveness and boundedness properties are related to the existence of the steady-state probabilities distribution of the equivalent Markov chain [13].
3. Obtain the equivalent Markov chain from the reachability tree.
4. Solve the equivalent Markov chain by a set of linear algebraic equations, i.e., the steady-state probabilities.

The steady-state probabilities obtained from the Markov chain could be used to compute (1) the expected number of tokens in a place; (2) the probability that a place is not empty; (3) the probability that a transition is enabled; and (4) performance measures such as average production rate, average in-process inventory, and average resource utilization.

It is interesting to note that the solution of a GSPN may be obtained with less effort than what is required to solve the corresponding SPN, especially if many immediate transitions are involved [14].

4.9 PETRI NET MODEL SYNTHESIS TECHNIQUES AND PROCEDURES

Modeling a practical DES with Petri nets can be done using stepwise refinement technologies [2]. In this approach, Petri net modeling starts with a simple, coarsely detailed model that can be easily verified as a live, bounded, and reversible one. Then, the transitions and places of the initial model are replaced by special subnets step by step to capture more details about the system. Each successive refinement will guarantee the preservation of the desired properties of the initial model. This process is to be repeated until the required modeling detail is achieved. Through this approach, the computational difficulties of checking a large Petri net model for liveness, boundedness, and reversibility are avoided.

4.9.1 Initial Petri Net Model

Figure 7 shows an initial Petri net model that contains $n + k$ places and two transitions, where:

Places $p_1, p_2, \ldots, p_n$ are *n operation places* that represent n concurrently working subsystems.
Places $p_{n+1}, p_{n+2}, \ldots, p_{n+k}$ are *k resource places*.
Transition t_1 represents the beginning of the working of the system.
Transition t_2, represents the end of the working of the system.

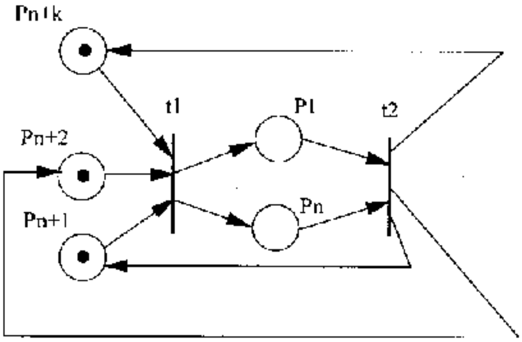


Figure 7 A safe, live, and reversible Petri net.

One can verify that this initial model is live, reversible, and safe with respect to the initial marking $m_0 = (0, ..., 0, 1, ..., 1)$, n zeros and k ones, where:

n zeros in initial marking means that all operation places are initially not marked;.

k ones in initial marking means that all resources are initially available.

Through the firing of transition t_1 the resources are occupied by the operation places. Then, the resources are released to the resource places again after the firing of the transition t_2 indicating the completion of one operation cycle.

4.9.2 Stepwise Refinement of Operation Places

An operation place in the initial Petri net can be replaced by some special design modules without changing the properties of the original model [2]. The basic modules include *sequence Petri net, parallel Petri net, choice Petri net*, and *mutual exclusion Petri net*, each of which describes a specific type of operations of the system.

4.9.2.1 Sequence Petri Net Module

Example 11. The Petri net with an initial marking of zero as shown in Fig. 8a is a sequence Petri net that has $n+1$ series places connected by n transitions ($n > 0$) for the case of $n = 2$. A sequence Petri net represents a series of successive operations.

4.9.2.2 Parallel Petri Net Module

Example 12. The Petri net with an initial marking of zero as shown in Fig. 8b is a parallel Petri net that consists of $n+2$ places and two transitions that connect n parallel places ($n > 1$) for the case of $n = 2$.

4.9.2.3 Choice Petri Net Module

Example 13. The Petri net with an initial marking of zero as shown in Fig. 9a is a choice Petri net that consists of $2n$ transitions and $(n+2)$ places with n parallel paths ($n > 1$) for the case of $n = 2$. A choice Petri net represents n choices for a successive operation.

4.9.2.4 Mutual Exclusion Petri Net Module

Example 14. The Petri net as shown in Fig. 9b is a mutual exclusion Petri net that consists of six transitions and nine places with a resource place and has its initial marking value one at the resource place and zero at all other places.

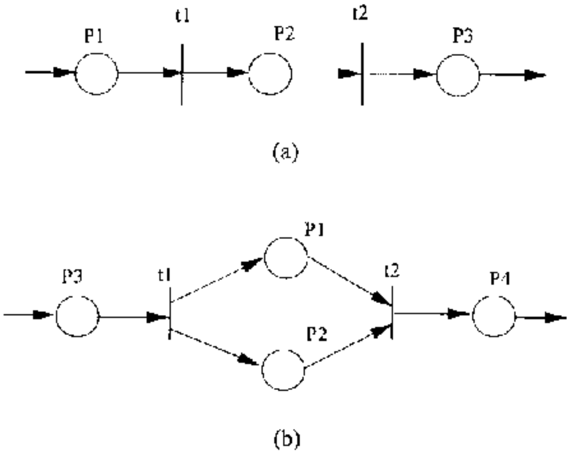
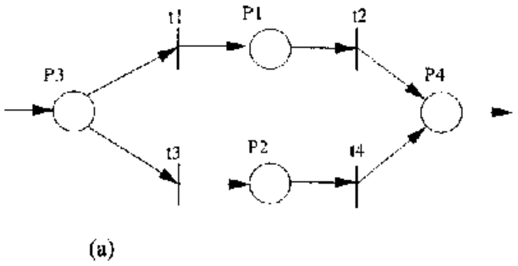


Figure 8 An example of (a) a sequence Petri net and (b) a parallel Petri net.



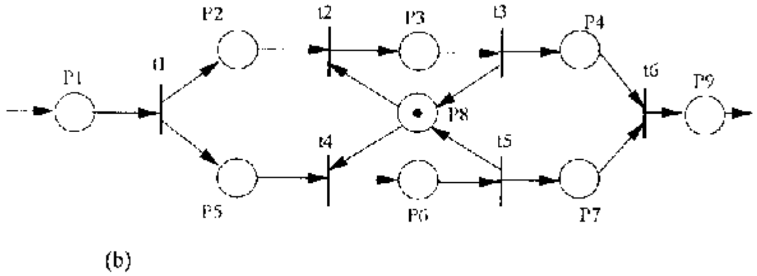


Figure 9 An example of (a) choice Petri net and (b) a mutual exclusion Petri net.

4.9.3 Modeling Dedicated Resources

Example 15. Given a subnet as shown in Fig. 10a that is a live, reversible, and safe Petri net with respect to an initial marking $m_{0,}$ one may add a dedicated resource (i.e., tokens at place p_d) to the subnet as shown in Fig. 10b. It has been verified [2] that the new Petri net is also safe, live, and reversible with respect to new initial marking M_0.

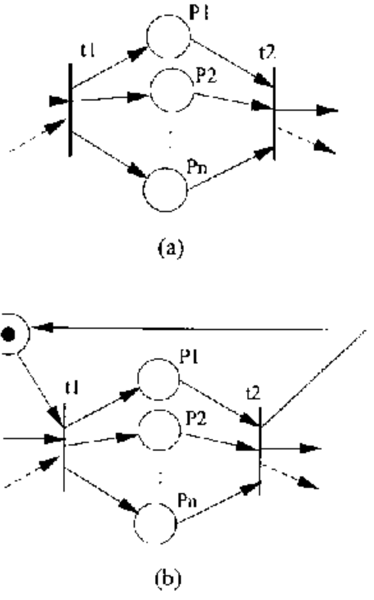


Figure 10 Petri net (a) is augmented into Petri net (b).

4.9.4 Stepwise Refinement of Transitions

Refinements of transitions in a Petri net use the concepts of a block Petri net, an associated Petri net of a block Petri net, and a well-formed block Petri net [15] as shown in Fig. 11.

Definition 38. *A block Petri net* is a Petri net that starts always from one initial transition, t_{in}, and ends with one final transition, t_f.

Definition 39. *An associated Petri net*, $\overset{\wedge}{PN}$, of a block Petri net is obtained by adding an idle place p_0 to the block Petri net such that (1) t_{in} is the only output transition of p_0; (2) t_f is the only input transition to p_0; (3) the initial marking of the associated Petri net is $\hat{m}_0$ and $\hat{m}_0(p_0) = 1$.

Definition 40. A well-formed Petri net block must be a live associated Petri net $\overset{\wedge}{PN}$ with $m_0 = \hat{m}_0(p) = \hat{m}_0(p_0) = 1$.

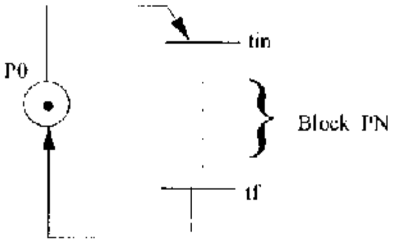


Figure 11 An associated Petri net $\hat{PN}$ of a block Petri net PN.

A transition in a Petri net can be replaced by a well-formed block Petri net. The substitution does not change the properties of the original Petri net such as boundedness, safeness, and liveness.

4.9.5 Modeling Coupled Process Using Shared Resources

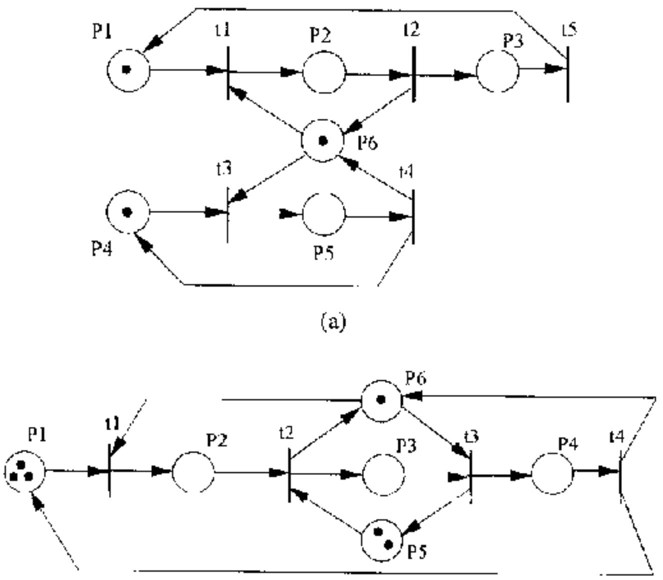
For resource places defined in a Petri net model, the need to model resource sharing has led to two essential concepts: parallel and sequential mutual exclusions (PME and SME) [16].

4.9.5.1 Parallel Mutual Exclusion (PME)

Example 16. A PME as shown in Fig. 12a has a shared resource place p_6 and two pairs of transitions, (t_1, t_2) and (t_3, t_4), each of which has a dedicated resource place p_1 and p_4, respectively. The token at p_6 models a single shared resource, and the pair of transitions models an independent process that requires that shared resource. This is a 2-PME model. Generally, a k-PME consists of k independent processes that share a resource. Let the transition pairs be denoted by $\{(t_{a1}, t_{b1}), (t_{a2}, t_{b2}), ..., (t_{ak}, t_{bk})\}$. A live, bounded, and reversible k-PME requires that:

1. t_{ai}, the dedicated resource place, and t_{bi} must be in any elementary path.
2. t_{ai}, the shared resource place, and t_{bi} must be in any elementary circuit.
3. A set of operation places should be on one elementary path between t$_{ai}$ and t_{bi}.

Condition (1) explains that the dedicated resource is always consumed by t_{ai} and released by t_{bi}. Condition (2) explains how the shared resource is consumed and released by t_{ai} and t_{bi}. Condition (3) guarantees that



(a)

(b)

Figure 12 (a) A PME; (b) an SME.

the concurrent groups of operation processes are successive.

4.9.5.2 Sequential Mutual Exclusion (SME)

One typical example of a SME is shown in Fig. 12b. It has a shared resource place p_6 and a group of sets of transition pairs (t_1, t_2) and (t_3, t_4). The token initially marked at p_6 models a single shared resource, and the groups of transitions model the processes that need the shared resource sequentially. This implies that there is a sequential relationship and a mutual dependency between the firing of a transition in one group and the firing of a transition in another group. The properties of an SME such as liveness are related to a concept called *token capacity*.

Definition 41. The maximum number of firings of t_i from the initial marking without firing t_j is the token capacity $c(t_i, t_j)$ of an SME.

The value of $c(t_i, t_j)$ depends on the structure and the initial marking of an SME. It has been shown [16] that when the initial marking (tokens) on dedicated resource places is less than or equal to $c(t_i, t_j)$, the net with and without the shared resource exhibits the same properties. For example, in Fig. 12b, p_1 is a dedicated resource place and the initial marking of the net is (3 0 0 0 2 1). It is easy to see that t_2 can only fire at most two times before t_3 must be fired to release two lost tokens at p_5. Otherwise, no processes can continue. Thus, the token capacity of the net is 2. As long as $1 \leq m_0(p_1) \leq 2$, the net is live, bounded, and reversible.

4.9.6 Petri Net Synthesis Technique Procedure [17]

1. Start an initial Petri net model that is live, bounded, and reversible. This model should be a macrolevel model that captures important system interactions in terms of major activities, choices, and precedence relations. All places are either operation places, fixed resource places, or variable resource places.
2. Use stepwise refinement to decompose the operation places using basic design modules until all the operations cannot be divided or until one reaches a point where additional detail is not needed. At each stage, add the dedicated resource places before proceeding with additional decomposition.
3. Add shared resources using bottom-up approach. At this stage, the Petri net model will be merged to form the final net. The place where the resource is shared by k parallel processes is specified so that it forms a k-PME. The place where the resource is shared by several sequentially related processes is added such that the added place and its related transitions form an SME.

4.10 EXTENSIONS TO ORIGINAL PETRI NETS

Based on the original Petri nets concept, researchers have developed different kinds of extended Petri nets (EPNs) for different purposes. The key step for developing EPNs is the developments of the theory that supports the extensions defined in the nets. As an example, timed Petri nets are well-developed EPNs that are used for system performance analysis. Similarly, to aid the modeling of the flow of control, resources, parts, and information through complex systems such as CIM and FMS, multiple classes of places, arcs, and tokens are introduced to the original Petri net to form new EPNs. With these extensions, system modeling can proceed through different levels of detail while preserving structural properties and avoiding deadlocks.

4.10.1 Multiple Places

Five types of places that are commonly used in EPNs are developed to model five common classes of conditions that may arise in a real system. They are status place, simple place, action place, subnet place, and switch place as shown in Fig. 13. Each place may also have a type of procedure associated with it if the net is used as a controller.

A *status place* is equivalent to a place in an original Petri net. Its only procedure is the enable check for the associated transitions. A *simple place* has a simple procedure associated with it in addition to the transition-enable check. An *action place* is used to represent procedures that take a long time to be executed. Usually, these procedures are spawned off as subprocesses that are executed externally in parallel with the Petri net-based model, for example, on other control computers

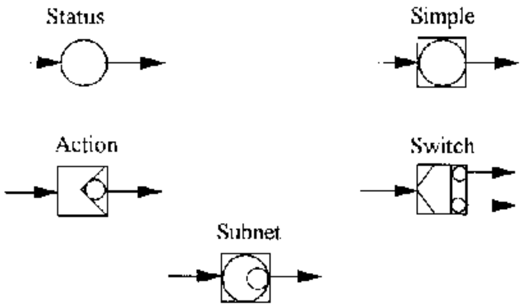


Figure 13 Place types in extended Petri nets.

or by other equipment. A *subnet place* denotes a subnet that represents a system abstraction, a subsystem area, etc. With subnet places, it is possible to model a complex system with Petri net in a hierarchical fashion. A *switch place* denotes a decision being taken. It may use the external information to resolve a conflict.

4.10.2 Multiple Tokens and Arcs

The concept of multiple classes of tokens and multiple arcs has been proposed in EPN to simultaneously indicate the flow of different objects (e.g., control, information, resources, etc.) within a system. A different set of tokens is defined and initiated to indicate each class of objects at a place. Associated with each class of tokens defined, there is an associated set of arcs. This facilitates tracing the flow of a particular type of token during the modeling of a system. The flow of different classes of tokens through the net is synchronized by utilizing the concept of *stream* through a transition t_i. A stream through t_i covers all places that are incident to $t_{i.}$ Since only tokens of a specific class flow through the transition t_i, this represents a monochromatic flow through the transition t_i. Arcs to the transition t_i, or from transition t_i, have to be labeled to indicate the classes of tokens that may flow along them. This leads to the restrictions that (1) an arc carries tokens of a specific class, and (2) the total number of arcs entering the transition equals the number of arcs leaving the same transition.

4.10.3 Colored Petri Nets

Similar to the idea used in EPNs, more mathematical and compact EPNs called colored Petri nets (CPNs) has been developed [18]. The differences between CPNs and original Petri nets are that:

1. A set of *colors* associated with the places P and transition T is defined.
2. $I(p, t)$ and $O(t, p)$ define functions that are expressed in a matrix.
3. The tokens in a place become a summation of place colors.

CPNs are an important extension to original Petri nets, providing for compact models with a higher level of abstraction and an improved graphical representation capability. However, the increasing modeling power of CPNs also increases the difficulties in analyzing and determining various properties. The analysis of CPNs is mainly done with matrix-based techniques (invariant analysis) rather than the coverability graph.

REFERENCES

1. CA Petri. Kommunication mit Automaten. PhD dissertation, University of Bonn, West Germany, 1962.
2. AA Desrochers, RY Al-Jaar. Applications of Petri Net in Manufacturing Systems. New York: IEEE Press, 1995.
3. CG Cassandras. Discrete Event Systems: Modeling and Performance Analysis. Richard D. Irwin, Inc., and Aksen Associate, Inc., Homewood, IL, 1993.
4. JL Peterson. Petri Net Theory and the Modeling of the System. Englewood Clifs, NJ: Prentice-Hall, 1981.
5. K Lautenbach. Linear algebraic techniques for place/transition nets. In: Advances in Petri nets 1986 (part I) Petri Nets: Central Models and Their Properties, vol 254. New York: Springer-Verlag, 1987, pp 142–167.
6. T Murata. Petri nets: Properties, analysis, and applications. Proc IEEE 71: 541–580, 1989.
7. PA Remy. On the generation of organizational architectures using Petri net. Tech Rep LIDS-TH-1630, Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, Cambridge, MA, Dec 1986.
8. T Murata. State equation, controllability and maximal matchings of Petri nets. IEEE Trans Autom Control AC-22: 412–415, 1977.
9. J Martine, M Silva. A simple and fast algorithm to obtain all invariants of a generalized Petri net. In: C Girault, W Reisig, eds. Application and Theory of Petri Nets, vol 52. New York: Springer-Verlag, 1982, pp 301–311.
10. J Sifakis. Use of Petri nets for performance evaluation. In: H Beilner, E Gelenbe, eds. Measuring, Modeling, and Evaluating Computer Systems. Amsterdam: North-Holland, 1977, pp 75–93.
11. C Ramamoorthy and G Ho. Performance evaluation of asynchronous concurrent systems using Petri nets. IEEE Trans Soft Engng SE-6: 440–449, 1980.

12. JB Dugan. Extended Stochastic Petri nets: applications and analysis. PhD thesis, Duke University, July 1984.
13. MK Molloy. Performance analysis using stochastic Petri nets. IEEE Trans Computers C-31: 913–917, 1982.
14. MA Marsan, G Conte, G Balbo. A class of generalized stochastic Petri nets for the performance evaluation of multiprocessor systems. ACM Trans Comput Syst 2: 93–122, 1984.
15. R Valette. Analysis of Petri nets by stepwise refinements. J Computer Syst Sci 18: 35–46, 1979.
16. M Zhou, F DiCesare. Parallel and sequential mutual exclusions for Petri net modeling of manufacturing systems with shared resources. IEEE Trans Robot Autom 7: 515–527, 1991.
17. M Zhou, F DiCesare, AA Desrochers. A hybrid methodology for synthesis of Petri net models for manufacturing systems. IEEE Trans Robot Autom 8: 350–361, 1992.
18. K Jenson. Colored Petri Nets: Basic Concepts, Analysis Methods and Practical Use, vol 1. New York: Springer-Verlag, 1993.

Chapter 4.5

Decision Analysis

Hiroyuki Tamura
Osaka University, Toyonaka, Osaka, Japan

5.1 INTRODUCTION

This chapter attempts to show the central idea and results of decision analysis and related decision-making models without mathematical details. Utility theory and value theory are described for modeling value perceptions of a decision maker under various situations, risky or riskless situations, and situation of single or multiple attributes. An analytic hierarchy process (AHP) is also included, taking into account the behavioral nature of multiple criteria decision making.

5.2 UTILITY THEORY

Multiattribute utility theory is a powerful tool for multiobjective decision analysis, since it provides an efficient method of identifying von Neumann–Morgernstern utility functions of a decision maker. The book by Keeney and Raiffa [1] describes in detail the standard approach. The significant advantage of the multiattribute utility theory is that it can handle both uncertainty and multiple conflicting objectives: the uncertainty is handled by assessing the decision maker's attitude towards risk, and the conflicting objectives are handled by making the utility function multidimensional (multiattribute).

In many situations, it is practically impossible to assess directly a multiattribute utility function, so it is necessary to develop conditions that reduce the dimensionality of the functions that are required to be assessed. These conditions restrict the form of a multiattribute utility function in a decomposition theorem.

In this section, after a brief description of an expected utility model of von Neumann and Morgenstern [2], additive, multiplicative, and convex decompositions are described for multiattribute utility functions [1,3].

5.2.1 Expected Utility Model

Let $A = \{a, b, \ldots\}$ be a set of alternative actions from which a decision maker must choose one action. Suppose the choice of $a \in A$ results in a consequence x_i with probability p_i and the choice of $b \in A$ results in a consequence x_i with probability q_i, and so forth. Let

$$X = \{x_1, x_2, \ldots\}$$

be a set of all possible consequences. In this case

$$p_i \geq 0, \qquad q_i \geq 0, \ldots \forall i$$

$$\sum_i p_i = \sum_i q_i = \cdots = 1$$

Let a real function u be a utility function on X. Then the expected utilities of actions $a, b, \ldots$ are written, respectively, as

$$E_a = \sum_i p_i u(x_i), \qquad E_b = \sum_i q_i u(x_i), \ldots \tag{1}$$

The assertion that the decision maker chooses an alternative action as if he maximizes his expected utility is called the expected utility hypothesis of von Neumann and Morgenstern [2]. In other words, the decision maker chooses an action according to the normative rule

$$a \succ b \Leftrightarrow E_a > E_b \qquad a \sim b \Leftrightarrow E_a = E_b \tag{2}$$

where $a \succ b$ denotes "a is preferred to b," and $a \sim b$ denotes "a is indifferent to b." This rule is called the expected utility rule. A utility function which satisfies Eqs. (1) and (2) is uniquely obtained within the class of positive linear transformations.

Figure 1 shows a decision tree and lotteries which explain the above-mentioned situation, where $\ell_a, \ell_b, \ldots$ denote lotteries which the decision maker comes across when he chooses the alternative action $a, b, \ldots$, respectively, and described as

$$\ell_a = (x_1, x_2, \ldots; p_1, p_2, \ldots)$$
$$\ell_b = (x_1, x_2, \ldots; q_1, q_2, \ldots)$$

Definition 1. *A certainty equivalent of lottery ℓ_a is an amouint $\hat{x}$ such that the decision maker is indifferent between ℓ_a and $\hat{x}$.*

From the expected untility hypothesis we obtain

$$u(\hat{x}) = u(E_a) = \sum_i p_i u(x_i) \tag{3}$$

In a set X of all possible consequences, let x^0 and x^* be the worst and the best consequences, respectively. Since the utility function is unique within the class of positive linear transformation, let us normalize the utility function as

$$u(x^0) = 0 \qquad u(x^*) = 1$$

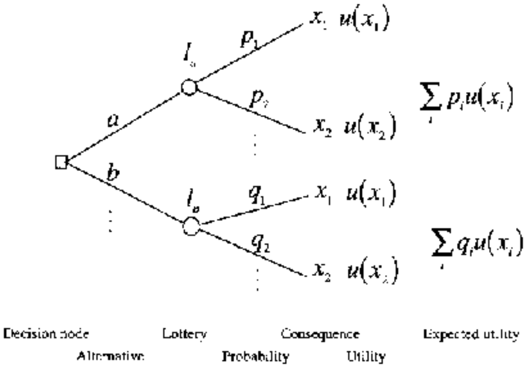


Figure 1 A decision tree and lotteries.

Let $\langle x^*, p, x^0 \rangle$ be a lottery yielding consequences x^* and x^0 with probabilities p and $(1-p)$, respectively. In particular, when $p = 0.5$ this lottery is called the 50–50 lottery and is denoted as $\langle x^*, x^0 \rangle$. Let x be a certainty equivalent of lottery $\langle x^*, p, x^0 \rangle$, that is,

$$x \sim \langle x^*, p, x^0 \rangle$$

Then

$$u(x) = pu(x^*) + (1-p)\,u(x^0) = p$$

It is easy to identify a single-attribute utility function of a decision maker by asking the decision maker about the certainty equivalents of some 50–50 lotteries and by means of a curve-fitting technique.

The attitude of a decision maker toward risk is described as follows.

Definition 2. *A decision maker is risk averse if he prefers the expected consequence $\bar{x}(= \sum_i p_i x_i)$ of any lotteries to the lottery itself.*

In this case

$$u(\bar{x}) > \sum_i p_i u(x_i) \tag{4}$$

If a decision maker is risk averse, his utility function is concave. The converse is also true. A decision maker is risk neutral (prone) if and only if his utility function is linear (convex).

5.2.2 Multiattribute Utility Function

The following results are the essential summary of Refs. 1 and 3. Let a specific consequence $x \in X$ be characterized by n attributes (performance indices) $X_1, X_2, \ldots, X_n$ (e.g., price, design, performance, etc., of cars, productivity, flexibility, reliability, etc., of manufacturing systems, and so on). In this case a specific consequence $x \in X$ is represented by

$$x = (x_1, x_2, \ldots, x_n) \quad x_1 \in X_1, x_2 \in X_2, \ldots, x_n \in X_n$$

A set of all possible consequences X can be written as a subset of an n-dimensional Euclidean space as $X = X_1 \times X_2 \times \cdots \times X_n$. This consequence space is called n-attribute space. An n-attribute utility function is defined on $X = X_1 \times X_2 \times \cdots \times X_n$ as $u : X_1 \times X_2 \times \cdots \times X_n \to R$.

Let I be a subset of $\{1, 2, \ldots, n\}$ with r $(1 \leq r < n)$ elements, and J be a complementary subset of I with $(n-r)$ elements. Suppose a set of n attributes $\{X_1, X_2, \ldots, X_n\}$ is divided into two subsets $\{X_i, i \in I\}$ and $\{X_i, i \in J\}$. Let X_I be an r-attribute space composed of $\{X_i, i \in I\}$, and X_J be an $(n-r)$-attribute space composed of $\{X_i, i \in J\}$. Then $X = X_I \times X_J$.

Definition 3. *Attribute X_I is utility independent of attribute X_J, denoted $X_I(\mathrm{UI})X_J$, if conditional preferences for lotteries on X_I given $x_J \in X_J$ do not depend on the conditional level $x_J \in X_J$.*

Let us assume that x_I^0 and x_I^* are the worst level and the best level of the attribute X_I, respectively.

Definition 4. *Given an arbitrary $x_J \in X_J$, a normalized conditional utility function $u_I(x_I \mid x_J)$ on X_I is defined as*

$$u_I(x_I \mid x_J) := \frac{u(x_I, x_J) - u(x_I^0, x_J)}{u(x_I^*, x_J) - u(x_I^0, x_J)} \tag{5}$$

where it is assumed that $u(x_I^*, x_J) > u(x_I^0, x_J)$.

From Definition 4 it is obvious that

$$u_I(x_I^* \mid x_J) = 1 \qquad u_I(x_I^0 \mid x_J) = 0 \qquad \forall x_J \in X_J$$

From Definitions 3 and 4 the following equation holds, if $X_I(\mathrm{UI})X_J$:

$$u_I(x_I \mid x_J) = u_I(x_I \mid x_J^0) \qquad \forall x_J \in X_J$$

In other words, utility independence implies that the normalized conditional utility functions do not depend on the different conditional levels.

Definition 5. *Attributes $X_1, X_2, \ldots, X_n$ are mutually utility independent, if $X_I(\mathrm{UI})X_J$ for any $I \subset \{1, 2, \ldots, n\}$ and its complementary subset J.*

Theorem 1. *Attributes $X_1, X_2, \ldots, X_n$ are mutually utility independent, if and only if*

$$u(x) = u(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} k_i u_i(x_i) \quad \text{if} \quad \sum_{i=1}^{n} k_i = 1 \tag{6}$$

or

$$ku(x) + 1 = \prod_{i=1}^{n} \{kk_i u_i(x_i) + 1\} \qquad \text{if} \quad \sum_{i=1}^{n} k_i \neq 1 \tag{7}$$

holds, where

$$u(x_1^0, x_2^0, \ldots, x_n^0) = 0 \qquad u(x_1^*, x_2^*, \ldots, x_n^*) = 1$$

$$u_i(x_i) := u_i(x_i \mid x_{i^c}^0) \quad i^c = \{1, \ldots, i-1, i+1, \ldots, n\}$$

$$k_i = u(x_i^*, x_{i^c}^0)$$

and k is a solution of

$$k + 1 = \prod_{i=1}^{n} (kk_i + 1)$$

Definition 6. *Attributes $X_1, X_2, \ldots, X_n$ are additive independent if preferences over lotteries on $X_1, X_2, \ldots, X_n$ depend only on their marginal probability distributions, not on their joint probability distribution.*

Theorem 2. *Attributes $X_1, X_2, \ldots, X_n$ are additive independent if and only if Eq. (6) holds.*

From Theorems 1 and 2 the additive independence is a special case of mutual utility independence.

For notational simplicity we deal only with the two-attribute case ($n = 2$) in the following discussion. The cases with more than two attributes are discussed in Tamura and Nakamura [3]. We deal with the case where

$$u_1(x_1 \mid x_2) \neq u_1(x_1) \qquad \text{for some } x_2 \in X_2$$
$$u_2(x_2 \mid x_1) \neq u_2(x_2) \qquad \text{for some } x_1 \in X_1$$

that is, utility independence does not hold between the attributes X_1 and X_2.

Definition 7. *Attribute X_1 is mth-order convex dependent on attribute X_2, denoted $X_1(\mathrm{CD}_m)X_2$, if there exist distinct $x_2^j \in X_2$ $(j = 0, 1, \ldots, m)$ and real functions $\lambda_j : X_2 \to R$ $(j = 0, 1, \ldots, m)$ on X_2 such that the normalized conditional utility function $u_1(x_1 \mid x_2)$ can be written as*

$$u_1(x_1 \mid x_2) = \sum_{j=0}^{m} \lambda_j(x_2) u_1(x_1 \mid x_2^j) \qquad \sum_{j=1}^{m} \lambda_j(x_2) = 1 \tag{8}$$

for all $x_1 \in X_1$, $x_2 \in X_2$, where m is the smallest nonnegative integer for which Eq. (8) holds.

This definition says that, if $X_1(\mathrm{CD}_m)X_2$, then any normalized conditional utility function on X_1 can be described as a convex combination of $(m+1)$ normalized conditional utility functions with different conditional levels where the coefficients $\lambda_j(x_2)$ are not necessarily nonnegative.

In Definition 7, if $m = 0$, then $u_1(x_1 \mid x_2) = u_1(x_1 \mid x_2^0)$ for all $x_2 \in X_2$. This implies

$$X_1(\mathrm{CD}_0)X_2 \Rightarrow X_1(\mathrm{UI})X_2$$

that is, zeroth-order convex dependence is nothing but the utility independence. This notion shows that the property of convex dependence is a natural extension of the property of utility independence.

For $m = 0, 1, \ldots,$ if $X_1(\mathrm{CD}_m)X_2$, then X_2 is at most $(m+1)$th-order convex dependent on X_1. If $X_1(\mathrm{UI})X_2$, then $X_2(\mathrm{UI})X_1$ or $X_2(\mathrm{CD}_1)X_1$. In general, if $X_1(\mathrm{CD}_m)X_2$, then X_2 satisfies one of the three properties: $X_2(\mathrm{CD}_{m-1})X_1$, $X_2(\mathrm{CD}_m)X_1$ or $X_2(\mathrm{CD}_{m+1})X_1$.

Theorem 3. *For $m = 1, 2, \ldots,$ $X_1(\mathrm{CD}_m)X_2$ if and only if*

$$\begin{aligned} u(x_1, x_2) &= k_1 u_1(x_1 \mid x_2^0) + k_2 u_2(x_2 \mid x_1^0) \\ &\quad + u_1(x_1 \mid x_2^0) f(x_1^*, x_2) + \sum_{i=1}^{m^*} \sum_{j=1}^{m} \qquad (9) \\ &\quad c_{ij} G(x_1, x_2^i) G(x_1^j, x_2) \end{aligned}$$

where

$$\begin{aligned} f(x_1, x_2) &= u_1(x_1 \mid x)_2)[(1-k_1)u_2(x_2 \mid x_1^*) \\ &\quad + k_1 - k_2 u_2(x_2 \mid x_1^0)] - k_1 u_1(x_1 \mid x_2^0) \\ G(x_1, x_2) &= k_1[(1-k_1)u_2(x_2 \mid x_1^*) \\ &\quad + k_1 - k_2 u_2(x_2 \mid x_1^0)] \\ &\quad \times [u_1(x_1 \mid x_2) - u_1(x_1 \mid x_2^0)] \\ u(x_1^0, x_2^*) &= 0, \qquad u(x_1^*, x_2^*) = 1 \\ k_1 &= u(x_1^*, x_2^0), \qquad k_2 = u(x_1^0, x_2^*) \end{aligned}$$

c_{ij} is a constant, and summation $i = 1$ to m^* means $i = 1, 2, \ldots, m-1, {}^*$.

Theorem 4. *For $m = 1, 2, \ldots$ $X_1(\mathrm{CD}_m)X_2$ and $X_2(\mathrm{CD}_m)X_1$, that is, X_1 and X_2 are mutually mth-order convex dependent, denoted $X_1(\mathrm{MCD}_m)X_2$, if and only if*

$$\begin{aligned} u(x_1, x_2) &= k_1 u_1(x_1 \mid x_2^0) + k_2 u_2(x_2 \mid x_1^0) \\ &\quad + \sum_{i=1}^{m^*} \sum_{j=1}^{m^*} d_{ij} f(x_1, x_2^i) f(x_1^j, x_2) \\ &\quad + \sum_{i=1}^{m^*} \sum_{j=1}^{m^*} d'_{ij} G(x_1, x_2^i) H(x_1^j, x_2) \qquad (10) \end{aligned}$$

where

$$\begin{aligned} H(x_1, x_2) &= k_2[(1-k_2)u_1(x_1 \mid x_2^*) \\ &\quad + k_2 - k_1 u_1(x_1 \mid x_2^0)] \\ &\quad \times [u_2(x_2 \mid x_1) - u_2(x_2 \mid x_1^0)] \end{aligned}$$

and d_{ij} and d'_{ij} are constants.

We have obtained two main convex decomposition theorems which can represent a wide range of utility functions. Moreover, when the utility on the arbitrary point has a particular value, that is, $d'_{ij} = 0$ for all i, j in Eq. (10), we can obtain one more decomposition of utility functions which does not depend on that point. This decomposition still satisfies $X_1(\mathrm{CD}_m)X_2$ and $X_2(\mathrm{CD}_m)X_1$, so we call this new property reduced mth-order convex dependence and denote it by $X_1(\mathrm{RCD}_m)X_2$.

We note that when $d'_{ij} = 0$ for all i, j and $m = 1$, Eq. (10) reduces to Fishburn's bilateral decomposition [4]:

$$\begin{aligned} u(x_1, x_2) &= k_1 u_1(x_1 \mid x_2^0) + k_2 u_2(x_2 \mid x_1^0) \\ &\quad + f(x_1, x_2^*) f(x_1^*, x_2) / f(x_1^*, x_2^*) \end{aligned}$$

When $m = 1$ and $d'_{ij} \neq 0$, that is, $X_1(\mathrm{MCD}_1)X$, Eq. (10) reduces to

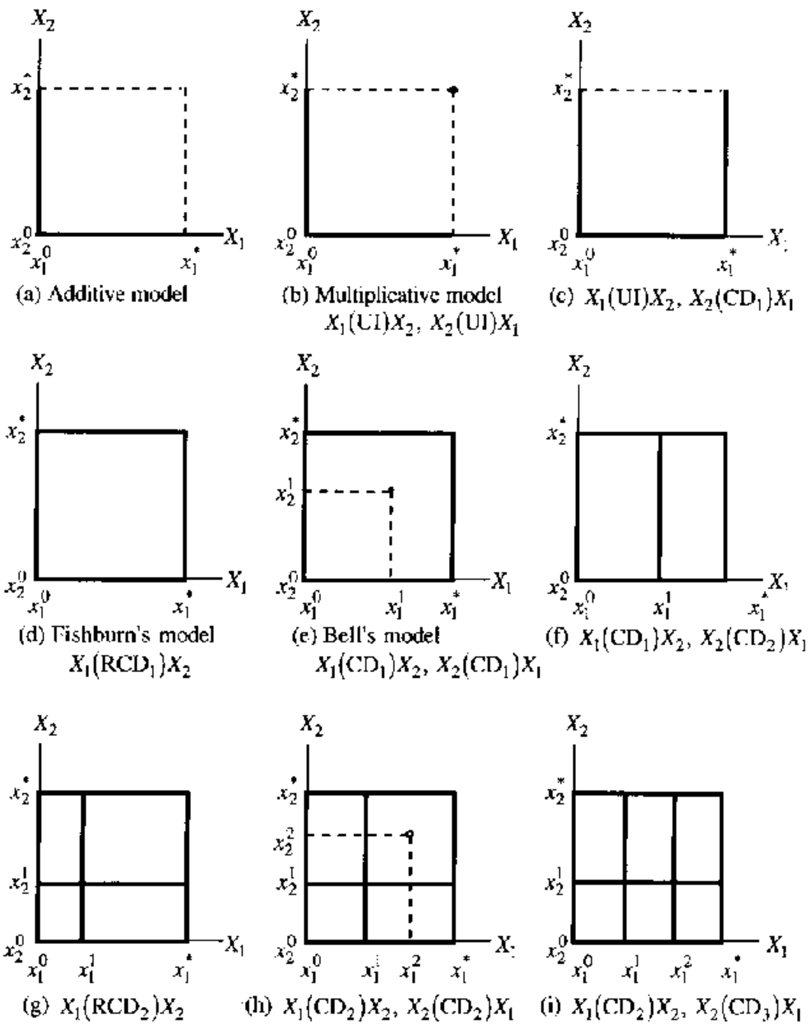
$$\begin{aligned} u(x_1, x_2) &= k_1 u_1(x_1 \mid x_2^0) + k_2 u_2(x_2 \mid x_1^0) \\ &\quad + f(x_1, x_2^*) f(x_1^*, x_2) / f(x_1^*, x_2^*) \\ &\quad + d' G(x_1, x_2^*) H(x_1^*, x_2) \end{aligned}$$

which is Bell's decomposition under the interpolation independence [5].

On two scalar attributes the difference between the conditional utility functions necessary to construct the previous decomposition models and the convex decomposition models is shown in Fig. 2. By assessing utilities on the lines and points shown bold, we can completely specify the utility function in the cases indicated in Fig. 2. As seen from Fig. 2. an advantage of the convex decomposition is that only single-attribute conditional utility functions need be assessed even for high-order convex dependent cases. Therefore, it is relatively easy to identify the utility functions.

5.3 MEASURABLE VALUE THEORY

Measurable value functions in this section are based on the concept of "difference in the strength-of-preference" [6] between alternatives. In this section we discuss such measurable value functions under certainty, under risk where the probability of each event occurring is known, and under uncertainty where the probability of each event occurring is unknown but the probability of a set of events occurring is known.



(a) Additive model

(b) Multiplicative model $X_1(\mathrm{UI})X_2$, $X_2(\mathrm{UI})X_1$

(c) $X_1(\mathrm{UI})X_2$, $X_2(\mathrm{CD}_1)X_1$

(d) Fishburn's model $X_1(\mathrm{RCD}_1)X_2$

(e) Bell's model $X_1(\mathrm{CD}_1)X_2$, $X_2(\mathrm{CD}_1)X_1$

(f) $X_1(\mathrm{CD}_1)X_2$, $X_2(\mathrm{CD}_2)X_1$

(g) $X_1(\mathrm{RCD}_2)X_2$

(h) $X_1(\mathrm{CD}_2)X_2$, $X_2(\mathrm{CD}_2)X_1$

(i) $X_1(\mathrm{CD}_2)X_2$, $X_2(\mathrm{CD}_3)X_1$

Figure 2 Consequences (bold lines) for which we need to assign utilities to specify the utility function in each case indicated.

5.3.1 Measurable Value Function Under Certainty

Measurable value functions provide an interval scale of measurement for preferences. However, practically, it is too difficult to directly identify a multiattribute measurable value function. Therefore, it is necessary to develop conditions that reduce the dimensionality of the functions that are required to be identified. These conditions restrict the form of a multiattribute measurable value function in a decomposition theorem. Dyer and Sarin [7] presented conditions for additive and multiplicative forms of the multiattribute measurable value function. These conditions are called difference independence and weak difference independence. They correspond to additive independence and utility independence, respectively, described in Sec. 5.2.

In this section, Dyer and Sarin's [7] difference independence and weak difference independence are briefly reviewed. Then, the condition of weak difference independence is extended to describe a new concept of finite-order independence of structural difference [8] for constructing multiattribute measurable value functions under certainty. This concept corresponds to convex dependence described in the previous section.

Let X be the set of all consequences in a decision problem. In n-attribute problem X is described as $X = X_1 \times X_2 \times \cdots \times X_n$ where X_i denotes the set of possible consequences for the ith attribute. Let $x^1, x^2, x^3, x^4 \in X$ where $x^k = (x_1^k, x_2^k, \ldots, x_n^k)$, $k = 1, 2, 3, 4$. Define X^*

as a nonempty subset of $X \times X$ and $\succsim^*$ as a weak order on X^*. Describe

$$x^1x^2 \succsim^* x^3x^4$$

to mean that the difference of the strength-of-preference for x^1 over x^2 is greater than or equal to the difference of the strength-of-preference for x^3 over x^4. If it is assumed that $(X, X^*, \succsim^*)$ denotes a positive difference structure [9], there exists a real-valued function v on X such that, for all $x^1, x^2, x^3, x^4 \in X$, if x^1 is preferred to x^2 and x^3 to x^4 then

$$x^1x^2 \succsim^* x^3x^4, \Leftrightarrow \; v(x^1) - v(x^2) \geq v(x^3) - v(x^4) \qquad (11)$$

Furthermore, since v is unique up to positive linear transformation, it is a cardinal function, and v provides an interval scale of measurement.

Define the binary relation $\succsim$ on X by

$$x^1x^3 \succsim^* x^2x^3 \Leftrightarrow x^1 \succsim x^2 \qquad (12)$$

then

$$x^1 \succsim x^2 \Leftrightarrow v(x^1) \geq v(x^2) \qquad (13)$$

Thus, v provides a measurable value function on X.

For $I \subset \{1, 2, \ldots, n\}$, partition X with n attributes into two sets X_I with r attributes and X_J with $(n - r)$ attributes, that is, $X = X_I \times X_J$. For $x_I \in X_I$, $x_J \in X_J$, write $x = (x_I, x_J)$.

Definition 8 [7]. *The attribute set X_I is difference independent of X_J, denoted $X_I(\mathrm{DI})X_J$, if for all $x_I^1, x_I^2 \in X$ such that $(x_I^1, x_J') \succsim (x_I^2, x_J')$ for some $x_J' \in X_J$*

$$(x_I^1, x_J')(x_I^2, x_J') \sim^* (x_I^1, x_J)(x_I^2, x_J) \qquad \forall x_J \in X_J \qquad (14)$$

This definition says that if $X_I(\mathrm{DI})X_J$ the difference in the strength of preference between (x_I^1, x_J) and (x_I^2, x_J) is not affected by $x_J \in X_J$. The property of this difference independence under certainty corresponds to the property of additive independence under uncertainty shown in Definition 6, and the decomposition theorem is obtained as a theorem as follows.

Theorem 5. *Suppose there exists a multiattribute measurable value function v on X. Then a multiattribute measurable value function $v(x)$ can be written as the same additive form shown in Eq. (6) if and only if $X_i(\mathrm{DI})X_{i^c}$, $i = 1, 2, \ldots, n$ where*

$$i^c = \{1, \ldots, i-1, i+1, \ldots, n\} \qquad X = X_i \times X_{i^c}$$

Dyer and Sarin [7] introduced a weaker condition than difference independence, which is called weak difference independence. This condition plays a similar role to the utility independence condition in multiattribute utility functions.

Definition 9 [7]. *X_I is weak difference independent of X_J, denoted $X_I(\mathrm{WDI})X_J$, if, for given $x_I^1, x_I^2, x_I^3, x_I^4 \in X_I$ and some $x_J' \in X_J$,*

$$(x_I^1, x_J')(x_I^2, x_J') \succsim^* (x_I^3, x_J')(x_I^4, x_J')$$

then

$$(x_I^1, x_J)(x_I^2, x_J) \succsim^* (x_I^3, x_J)(x_I^4, x_J) \qquad \forall x_J \in X_J \qquad (15)$$

This definition says that if $X_I(\mathrm{WDI})X_J$ the ordering of difference in the strength of preference depends only on the values of the attributes X_I and not on the fixed values of X_J. The property of the weak difference independence can be stated more clearly by using the normalized conditional value function, defined as follows.

Definition 10. *Given an arbitrary $x_J \in X_J$, define a normalized conditional value function $v_I(x_I \mid x_J)$ on X_I as*

$$v_I(x_I \mid x_J) := \frac{v(x_I, x_J) - v(x_I^0, x_J)}{v(x_I^*, x_J) - v(x_I^0, x_J)} \qquad (16)$$

where

$$v(x_I^*, x_J) > v(x_I^0, x_J)$$

and $x_I^* \in X$ and $x_I^0 \in X_I$ denote the best and the worst consequences, respectively.

Normalized conditional value function $v_I(x_I \mid x_J)$ denotes the ordering of preference on X_I, which is called preference structure here, under the given conditional level $x_J \in X_J$. From Definitions 9 and 10 we obtain a theorem as follows.

Theorem 6. *$X_I(\mathrm{WDI})X_J \Leftrightarrow v_I(x_I \mid x_J) = v_I(x_I \mid x_J^0)$, $\forall x_J \in X_J$.*

This theorem shows that the property of weak difference independence is equivalent to the independence of normalized conditional value functions on the conditional level. Hence, this theorem is often used for assuring the property of weak difference independence.

Definition 11. *The attributes $X_1, X_2, \ldots, X_n$ are said to be mutually weak difference independent, if for every $I \subset \{1, 2, \ldots, n\}$, $X_I(\mathrm{WDI})X_J$.*

The basic decomposition theorem of the measurable additive/multiplicative value functions is now stated.

Theorem 7. *If there exists a measurable value function v on X and if $X_1, X_2, \ldots, X_n$ are mutually weak difference independent, then a multiattribute measurable value function $v(x)$ can be written as the same additive form as Eq. (6), or multiplicative form, as shown in Eq. (7).*

Dyer and Sarin [7] stated this theorem under the condition of mutual preferential independence plus one weak difference independence instead of using the condition of mutual weak difference independence. For practical applications it is easier to assess mutual preferential independence than to assess mutual weak difference independence.

For notational simplicity we deal only with the two-attribute case ($n = 2$) in the following discussions. We deal with the cases where

$$v_1(x_1 \mid x_2) \neq v_1(x_1 \mid x_2^0) \qquad \text{for some } x_2 \in X_2$$

that is, weak difference independence does not hold between X_1 and X_2.

Definition 12. *X_1 is mth-order independent of structural difference with X_2, denoted $X_1(\mathrm{ISD}_m)X_2$, if for given $x_1^1, x_1^2, x_1^3, x_1^4 \in X_1$ and some $x_2 \in X_2$ such that*

$$(x_1^1, x_2)(x_1^2, x_2) \succsim^* (x_1^3, x_2)(x_1^4, x_2) \tag{17}$$

there exist $x_2^j \in X_2$ ($j = 0, 1, \ldots, m$) and $\theta_j : X_2 \to R$ ($j = 0, 1, \ldots, m$) such that

$$\sum_{j=0}^{m} \theta_j(x_2)(x_1^1, x_2^j) \sum_{j=0}^{m} \theta_j(x_2)(x_1^2, x_2^j) \succsim^* \sum_{j=0}^{m} \theta_j (x^2)(x_1^3, x_2^j) \sum_{j=0}^{m} \theta_j(x_2)(x_1^4, x_2^j) \tag{18}$$

This definition represents the ordering of difference in the strength of preference between the linear combinations of consequences on X_1 with $(m+1)$ different conditional levels. If $m = 0$ in Eq. (18), we obtain Eq. (15), and hence

$$X_1(\mathrm{ISD}_0)X_2 \Rightarrow X_1(\mathrm{WDI})X_2 \tag{19}$$

This notion shows that the property of independence of structural difference is a natural extension of the property of weak difference independence.

Definition 12 shows that there exists $v(x_1, x_2^j)$ ($j = 0, 1, \ldots, m$) such that

$$v(x_1, x_2) = \alpha(x_2) \sum_{j=0}^{m} \theta_j(x_2) v(x_1, x_2^j) + \beta(x_2) \tag{20}$$

$$\alpha(x_2) > 0$$

holds. If we define

$$\lambda_j(x_2) := \frac{\theta_j(x_2)\left\{v(x_1^*, x_2^j) - v(x_1^0, x_2^j)\right\}}{\sum_{i=0}^{m} \theta_i(x_2)\left\{v(x_1^*, x_2^i) - v(x_1^0, x_2^i)\right\}} \tag{21}$$

$$j = 0, 1, \ldots, m$$

we obtain

$$\sum_{j=0}^{m} \lambda_j(x_2) = 1$$

If we rewrite Eq. (16) using Eqs. (20) and (21), we obtain

$$v_1(x_1 \mid x_2) = \sum_{j=0}^{m} \lambda_j(x_2) v_1(x_1 \mid x_2^j) \tag{22}$$

If $m = 0$, we obtain Theorem 6 for $n = 2$. From this notion we also find that the property of independence of structural difference is a natural extension of the property of weak difference independence. Furthermore, Eq. (22) corresponds the definition of mth-order convex dependence in multiattribute utility theory shown in Eq. (8).

If we define

$$d_1(x_1 \mid x_2) := v_1(x_1 \mid x_2) - v_1(x_1 \mid x_2^0) \tag{23}$$

$d_1(x_1 \mid x_2)$ shows the difference of the preference structures for the conditional level of $x_2 \in X_2$ and $x_2^0 \in X_2$, and it is called the structural difference function. Then we obtain

$$X_1(\mathrm{ISD}_0)X_2 \Rightarrow d_1(x_1 \mid x_2) = 0 \qquad \forall x_2 \in X_2$$

$$X_1(\mathrm{ISD}_m)X_2 \Rightarrow d_1(x_1 \mid x)_2 = \sum_{j=0}^{m} \lambda_j(x_2)\, d_1(x_1 \mid x_2^j)$$

Since mth-order independence of structural difference in this measurable multiattribute value theory corresponds to mth-order convex dependence in multiattribute utility theory, the decomposition theorems described in Sec. 5.2 are valid if the expression "utility function" is replaced by the expression "measurable value function" in Theorems 3 and 4.

Multiattribute measurable value functions can be identified if we know how to obtain:

1. Single-attribute value functions
2. The order of structural difference independence

3. The scaling coefficients appearing in the decomposition forms.

For identifying single-attribute measurable value functions, we use the equal-exchange method based on the concept of equal difference points [7].

Definition 13. *For $[x^0, x^*] \subset X$, if there exists $x^1 \in X$ such that*

$$x^* x^1 \sim^* x^1 x^0 \tag{24}$$

for given $x^0 \in X$ and $x^* \in X$, then x^1 is the equal difference point for $[x^0, x^*] \subset X$.

From Eq. (24) we obtain

$$v(x^*) - v(x^1) = v(x^1) - v(x^0) \tag{25}$$

Since $v(x^0) = 0$, $v(x^*) = 1$, we obtain $v(x^1) = 0.5$. Let x^2 and x^3 be the equal difference points for $[x^0, x^1]$ and $[x^1, x^*]$, respectively. Then we obtain

$$v(x^2) = 0.25 \qquad v(x^3) = 0.75$$

It is easy to identify a single-attribute measurable value function of a decision maker from these five points and a curve-fitting technique.

How to find the order of structural difference independence and the scaling coefficients appearing in the decomposition forms is omitted here. Detailed discussion on this topic can be found in Tamura and Hikita [8].

5.3.2 Measurable Value Function Under Risk

The expected utility model described in Sec. 5.2 has been widely used as a normative model of decision analysis under risk. But, as seen in Refs. 10–12, various paradoxes for the expected utility model have been reported, and it is argued that the expected utility model is not an adequate descriptive model.

In this section a descriptive extension of the expected utility model to account for various paradoxes is discussed using the concept of strength of preference.

Let X be a set of all consequences, $x \in X$, and A a set of all risky alternatives; a risky alternative $\ell \in A$ is written as

$$\ell = (x_1, x_2, \ldots, x_n; p_1, p_2, \ldots, p_n) \tag{26}$$

which yields consequence $x_i \in X$ with probability p_i, $i = 1, 2, \ldots, n$, where $\sum p_i = 1$.

Let A^* be a nonempty subset of $A \times A$, and $\succsim$ and $\succsim^*$ be binary relations on A and A^*, respectively. Relation $\succsim$ could also be a binary relation on X. We interpret $\ell_1 \succsim \ell_2 (\ell_1, \ell_2 \in A)$ to mean that ℓ_1 is preferred or indifferent to ℓ_2, and $\ell_1 \ell_2 \succsim^* \ell_3 \ell_4 (\ell_1, \ell_2, \ell_3, \ell_4 \in A)$ to mean that the strength of preference for ℓ_1 over ℓ_2 is greater than or equal to the strength of preference for ℓ_3 over ℓ_4.

We postulate that $(A, A^*, \succsim^*)$ takes a positive difference structure which is based on the axioms described by Kranz et al. [9]. The axioms imply that there exists a real-valued function F on A such that for all $\ell_1, \ell_2, \ell_3, \ell_4 \in A$, if $\ell_1 \succsim \ell_2$ and $\ell_3 \succsim \ell_4$, then

$$\ell_1 \ell_2 \succsim^* \ell_3 \ell_4 \Leftrightarrow F(\ell_1) - F(\ell_2) \geq F(\ell_3) - F(\ell_4) \tag{27}$$

Since F is unique up to a positive linear transformation, it is a cardinal function. It is natural to hold for $\ell_1, \ell_2, \ell_3 \in A$ that

$$\ell_1 \ell_3 \succsim^* \ell_2 \ell_3 \Leftrightarrow \ell_1 \succsim \ell_2$$

Then from Eq. (27) we obtain

$$\ell_1 \succsim \ell_2 \Leftrightarrow F(\ell_1) \geq F(\ell_2) \tag{28}$$

Thus, F is a value function on A and, in view of Eq. (27), it is a measurable value function.

We assume that the decision maker will try to maximize the value (or utility) of a risky alternative $\ell \in A$, which is given by the general form as follows:

$$\max_{\ell \in A} F(\ell) = \max_{\ell \in A} \sum_i f(x_i, p_i) \tag{29}$$

where $f(x, p)$ denotes the value (strength of preference) for a consequence x which comes out with probability p. This function is called the measurable value function under risk. The main objectives here are to give an appropriate decomposition and interpretation of $f(x, p)$ and to explore its descriptive implications to account for the various paradoxes.

The model Eq. (29) is reduced to the expected utility form by setting

$$f(x, p) = pu(x) \tag{30}$$

when $u(x)$ is regarded as a von Neumann–Morgenstern utility function, described in Sec. 5.2. The prospect theory of Kahneman and Tversky [11] is obtained by setting

$$f(x, p) = \pi(p)\, v(x) \tag{31}$$

where $\pi(p)$ denotes a weighting function for probability and $v(x)$ a value function for consequence. In this model the value of each consequence is multiplied by a decision weight for probability (not by probability itself).

Extending this Kahneman–Tversky model we obtain a decomposition form [13]

$$f(x, p) = w(p \mid x)\, v(x) \tag{32}$$

where

$$w(p \mid x) := \frac{f(x, p)}{f(x, 1)} \tag{33a}$$

$$v(x) := v(x \mid 1) \tag{33b}$$

$$v(x \mid p) := \frac{f(x, p)}{f(x^*, p)} \tag{33c}$$

and x^* denotes the best consequence. In our model, Eq. (32), the expected utility model, Eq. (30), and Kahneman–Tversky model, Eq. (31) are included as special cases. Equation (33b) implies that $v(x)$ denotes a measurable value function under certainty described in Sec. 5.3.1. Therefore, our model, Eq. (32), also includes Dyer and Sarin's model [7] as a special case.

The model Eq. (32) could also be written as

$$f(x, p) = w(p)\, v(x \mid p) \tag{34}$$

where

$$w(p) := w(p \mid x^*) \tag{35}$$

We assume that

$$f(x, 0) = 0 \qquad \forall x \in X \tag{36a}$$

$$f(x^R, p) = 0 \qquad \forall p \in [0, 1] \tag{36b}$$

where $x^R \in X$ denotes the reference point (e.g., status quo). The better region on X compared with x^R is called the gain domain and the worse region the loss domain. We also assume that

$$f(x, p) \geq 0 \qquad \text{on the gain domain}$$
$$f(x, p) < 0 \qquad \text{on the loss domain}$$

It will be shown that the conditional weighting function $w(p \mid x)$ describes the strength of preference for probability under the given conditional level of consequence, and $v(x \mid p)$ describes the strength of preference for consequence under the given conditional level of probability.

For interpreting the descriptive model $f(x, p)$ we need to interpret F such that Eq. (27) holds. Dyer and Sarin [14] and Harzen [15] have discussed the strength of preference under risk where a certainty equivalent of a risky alternative is used to evaluate the strength of preference.

For all $x_1, x_2, x_3, x_4 \in X$, $\alpha \in [0, 1]$, and $y \in X$ such that $x_1 \succsim x_2 \succsim x_3 \succsim x_4$, we consider four alternatives:

$$\ell_1 = (x_1, y; \alpha, 1-\alpha) \qquad \ell_2 = (x_2, y; \alpha, 1-\alpha)$$
$$\ell_3 = (x_3, y; \alpha, 1-\alpha) \qquad \ell_4 = (x_4, y; \alpha, 1-\alpha) \tag{37}$$

In this case we obtain

$$\ell_1, \ell_2 \succsim^* \ell_3 \ell_4 \Leftrightarrow f(x_1, \alpha) - f(x_2, \alpha) \geq f(x_3, \alpha) - f(x_4, \alpha) \tag{38a}$$

$$\Leftrightarrow v(x_1 \mid \alpha) - v(x_2 \mid \alpha) \geq v(x_3 \mid \alpha) - v(x_4 \mid \alpha) \tag{38b}$$

Therefore, the value function $v(x \mid p)$ defined by Eq. (33c) represents the strength of preference for the four risky alternatives in Eq. (37).

For all $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \in [0, 1]$, $x \in X$ and $x^R \in X$, we consider four alternatives:

$$\ell_1' = (x, x^R; \alpha_1, 1-\alpha_1) \quad \ell_2' = (x, x^R; \alpha_2, 1-\alpha_2) \tag{39a}$$

$$\ell_3' = (x, X^R; \alpha_3, 1-\alpha_3) \quad \ell_4' = (x, x^R; \alpha_4, 1-\alpha_4) \tag{39b}$$

then we obtain

$$\ell_1' \ell_2' \succsim^* \ell_3' \ell_4' \Leftrightarrow f(x, \alpha_1) - f(x, \alpha)_2) \geq f(x, \alpha_3) - f(x, \alpha_4) \tag{40a}$$

$$\Leftrightarrow w(\alpha_1 \mid x) - w(\alpha_2 \mid x) \geq w(\alpha_3 \mid x) - w(\alpha_4 \mid x) \tag{40b}$$

Therefore, the weighting function defined by Eq. (33a) represents the strength of preference for the four risky alternatives in Eq. (39).

The above discussions assert that the descriptive model $f(x, p)$ represents the measurable value function under risk to evaluate the consequence $x \in X$ which comes out with probability p.

In the expected utility model it assumes invariance of preference between certainty and risk when other things are equal. The Kahneman–Tversky model of Eq. (31) could explain a so-called certainty effect to resolve the Allais paradox [10]. Our descriptive model $f(x, p)$ could also resolve the Allais paradox, as shown below.

As an example, consider the following two situations in gain domain:

1. $\ell_1 = (10M; 1)$,
   $\ell_2 = (50M, 10M, 0; 0.1, 0.89, 0.01)$.
2. $\ell_3 = (50M, 0; 0.1, 0.9)$,
   $\ell_4 = (10M, 0; 0.11, 0.89)$.

where $1M = \$1000$. Most people show the preference

$$\ell_1 = (10M; 1) \succ \ell_2 = (50M, 10M, 0; 0.1, 0.89, 0.01) \tag{41a}$$

$$\ell_3 = (50M, 0; 0.1, 0.9) \succ \ell_4 = (10M, 0; 0.11, 0.89) \tag{41b}$$

This preference violates the expected utility model as follows: Eq. (41a) implies

$$u(10M) > 0.1u(50M) + 0.89u(10M) + 0.01u(0) \tag{42a}$$

whereas Eq. (41b) implies

$$0.1u(50M) + 0.9u(0) > 0.11u(10M) + 0.89u(0) \tag{42b}$$

where u denotes a von Neumann–Morgenstern utility function. Equations (42a) and (42b) show the contradiction. This phenomenon is called the Allais paradox.

The descriptive model $f(x, p)$ could properly explain the preference of Eq. (41) as follows. Let

$$v(50M) = 1 \quad v(10M) = \theta \quad v(0) = 0 \quad 0 < \theta < 1$$

Then, using our descriptive model $f(x, p)$, the preference of Eq. (41) can be written as

$$\begin{aligned} v(10M) &> f(50M, 0.1) + f(10M, 0.89) \\ &= v(50M)\, w(0.1 \mid 50M) + v(10M)\, w(0.89 \mid 10M) \end{aligned}$$

and

$$\begin{aligned} f(50M, 0.1) > f(10M, 0.11) &\Rightarrow v(50M)\, w(0.1 \mid 50M) \\ &> v(10M)\, w(0.11 \mid 10M) \end{aligned}$$

Hence we obtain

$$\frac{w(0.1 \mid 50M)}{1 - w(0.89 \mid 10M)} < \theta < \frac{w(0.1 \mid 50M)}{w(0.11 \mid 10M)} \tag{43}$$

If we could find θ such that Eq. (43) holds, our descriptive model $f(x, p)$ could resolve the Allais paradox properly.

5.3.3 Measurable Value Function Under Uncertainty

In this section we deal with the case where probability of occurrence for each event is unknown. When we describe the degree of ignorance and uncertainty by the basic probability of Dempster–Shafer theory [16], the problem is how to represent the value of a set element which consists of multiple elements. We will try to construct a measurable value function under uncertainty based on this concept.

Conventional probability theory is governed by Bayes' rule, which is called Bayes' theory of probability. Such probability is called Bayes' probability. Let $p(A)$ be the Bayes' probability of occurring an event A. Given two events A and B, we denote by $A + B$ the event that occurs when A or B or both occur. We say that A and B are mutually exclusive if the occurrence of one at a given trial excludes the occurrence of the other. If A and B are mutually exclusive, then we obtain

$$p(A + B) = p(A) + p(B)$$

in Bayes' theory of probability. This implies that if $p(A) = 0.3$ then $p(\bar{A}) = 1 - p(A) = 0.7$ where $\bar{A}$ denotes the complement of A.

In Dempster–Shafer theory of probability [16] let $\mu(A_i)$ be basic probability which could be assigned by any subset A_i of Θ, where Θ dentoes a set containing every possible element. The basic probability $\mu(A_i)$ can be regarded as a semimobile probability mass. Let $\Lambda = 2^\Theta$ be a set containing every subset of Θ. Then, the basic probability $\mu(A_i)$ is defined on Λ and takes a value contained in [0, 1]. When $\mu(A_i) > 0$, A_i is called the focal element or the set element and the following conditions hold:

$$\mu(\emptyset) = 0$$

$$\sum_{A_i \in \Lambda} \mu(A_i) = 1$$

In general the Dempster–Shafer basic probability does not hold the additivity. As a special case, if the probability is assigned only for each element, the basic probability is reduced to the Bayes' probability.

Let the value function under uncertainty based on this Dempster–Shafer basic probability be

$$f^*(B, \mu) = w'(\mu)\, v^*(B \mid \mu) \tag{44}$$

where B denotes a set element, μ denotes the basic probability, w' denotes the weighting function for the basic probability, and v^* denotes the value function with respect to a set element. The set element B is a subset of $\Lambda = 2^\Theta$. Equation (44) is an extended version of the value function, Eq. (34), where an element is extended to a set element and the Bayes' probability is extended to the Dempster–Shafer basic probability.

For identifying v^*, we need to find the preference relations among set elements, which is not an easy task. If the number of elements contained in the set Θ is getting larger, it is not practical to find v^*. To cope with this difficulty we introduce an axiom of dominance as follows.

Axiom of Dominance 1. *In the set element B let the worst consequence be m_B and the best consequence be M_B. For any $B', B'' \in \Lambda \in 2^\Theta$*

$$m_{B'} \precsim m_{B''},\ M_{B'} \precsim M_{B''} \Rightarrow B' \precsim B'' \tag{45}$$

and

$$m_{B'} \sim m_{B''}, M_{B'} \sim M_{B''} \Rightarrow B' \sim B'' \tag{46}$$

Using Axiom of Dominance 1, we will restrict a set element B to

$$\Omega = \left\{(m, M) \in \Theta \times \Theta : m \lesssim M\right\} \tag{47}$$

where m and M denote the worst and the best consequence in the set element B, respectively. Then, Eq. (44) is reduced to

$$f^*(\Omega, \mu) = w'(\mu)v^*(\Omega \mid \mu) \tag{48}$$

Suppose we look at an index of pessimism $\alpha(m, M)$, such that the following two alternatives are indifferent [17].

Alternative 1. One can receive m for the worst case and M for the best case. There exists no other information.

Alternative 2. One receives m with probability $\alpha(m, M)$ and receives M with probability $1 - \alpha(m, M)$, where $0 < \alpha(m, M) < 1$.

If one is quite pessimistic, $\alpha(m, M)$ becomes nearly equal to 1, and if one is quite optimistic $\alpha(m, M)$ becomes nearly equal to zero. If we incorporate this pessimism index $\alpha(m, M)$ in Eq. (48), the value function is obtained as

$$\begin{aligned} v^*(\Omega \mid \mu) &= v^*((m, M) \mid \mu) \\ &= \alpha(m, M)v'(m \mid \mu) \\ &\quad + (1 - \alpha(m, M))v'(M \mid \mu) \end{aligned} \tag{49}$$

where v' denotes a value function for a single element.

Incorporating the Dempster–Shafer probability theory in the descriptive model $f^*(\Omega, \mu)$ of a value function under uncertainty, we could model the lack of belief which could not be modeled by the Bayes' probability theory. As the result our descriptive model $f^*(\Omega, \mu)$ could resolve the Ellsburg paradox [18,19] as follows.

Suppose an urn contains 30 balls of red, black, and white. We know that 10 of 30 balls are red, but for the other 20 balls we know only that each of these balls is either black or white. Suppose we pick a ball from this urn, and consider four events as follows:

a. We will get 100 dollars if we pick a red ball.
b. We will get 100 dollars if we pick a black ball.
c. We will get 100 dollars if we pick a red or white ball.
d. We will get 100 dollars if we pick a black or white ball.

Many people show the preference [9,20],

$$a \succ b \qquad d \succ c$$

The probability of picking up a red ball is 1/3. Let p_b and p_w be the probability of picking up a black ball and a white ball, respectively. Then

$$p_b + p_w = \tfrac{2}{3}$$

The expected utility theory says that

$$a \succ b \Rightarrow \tfrac{1}{3}u(1M) > p_b u(1M) \Rightarrow p_b < \tfrac{1}{3} \tag{50}$$

$$\begin{aligned} d \succ c &\Rightarrow \tfrac{2}{3}u(1M) > \tfrac{1}{3}u(1M) + p_w u(1M) \\ &\Rightarrow p_w < \tfrac{1}{3} \Rightarrow p_b > \tfrac{1}{3} \end{aligned} \tag{51}$$

where u denotes von Neumann–Morgenstern utility function and $1M = 100$ dollars. Equations (50) and (51) are obviously contradictory. This phenomenon is called the Ellsburg paradox [18,19]. Therefore, the expected utility theory cannot represent the preference when the probability of each event is not known but only basic probability for set of events is known. This phenomenon shows that one prefers the events with known probability rather than the events with unknown probability and is called the sure-thing principle [9].

How could we explain the preference of this Ellsburg paradox by using the descriptive model $f^*(\Omega, \mu)$ of a value function under uncertainty? Let $\{R\}$ be the event of picking a red ball and $\{B, W\}$ be the set element of picking a black or a white ball. Then the basic probability is written as

$$\mu(\{R\}) = \tfrac{1}{3}$$

$$\mu(\{B, W\}) = \tfrac{2}{3}$$

In this case a set Θ containing every possible event is written as

$$\Theta = \{0, 1M\}$$

Table 1 shows the basic probability of getting each event for each alternative. The value for each alternative is given by

Table 1 Basic Probability for Each Event

	Event		
Alt.	{0}	{1M}	{0, 1M}
a	2/3	1/3	0
b	1/3	0	2/3
c	0	1/3	2/3
d	1/3	2/3	0

$$V(a) = w'\left(\tfrac{2}{3}\right)v'\left(\{0\} \mid \tfrac{2}{3}\right) + w'\left(\tfrac{1}{3}\right)v'\left(\{1M\} \mid \tfrac{1}{3}\right) \tag{52a}$$

$$V(b) = w'\left(\tfrac{1}{3}\right)v'\left(\{0\} \mid \tfrac{1}{3}\right) + w'\left(\tfrac{2}{3}\right)v^*\left(\{0, 1M\} \mid \tfrac{2}{3}\right) \tag{52b}$$

$$V(c) = w'\left(\tfrac{1}{3}\right)v'\left(\{1M\} \mid \tfrac{1}{3}\right) + w'\left(\tfrac{2}{3}\right)v^*\left(\{0, 1M\} \mid \tfrac{2}{3}\right) \tag{52c}$$

$$V(d) = w'\left(\tfrac{1}{3}\right)v'\left(\{0\} \mid \tfrac{1}{3}\right) + w'\left(\tfrac{2}{3}\right)v'\left(\{1M\} \mid \tfrac{2}{3}\right) \tag{52d}$$

In the set Θ let x^0 and x^* be the worst consequence and the best consequence, then

$$x^0 = 0 \qquad x^* = 1M$$

Therefore we obtain

$$v'(\{0\} \mid \mu) = 0 \qquad v'(\{1M\} \mid \mu) = 1 \qquad \forall \mu$$

Let an index of pessimism be $\alpha = \alpha(0, 1M)$, then

$$\begin{aligned} a \succ b \Rightarrow V(a) &> V(b) \\ \Rightarrow w'\left(\tfrac{1}{3}\right) &> w'\left(\tfrac{2}{3}\right)v^*\left(\{0, 1M\} \mid \tfrac{2}{3}\right) \\ \Rightarrow w'\left(\tfrac{1}{3}\right) &> (1-\alpha)w'\left(\tfrac{2}{3}\right) \\ d \succ c \Rightarrow V(d) &> V(c) \\ \Rightarrow w'\left(\tfrac{2}{3}\right) &> w'\left(\tfrac{1}{3}\right) + w'\left(\tfrac{2}{3}\right)v^*\left(\{0, 1M\} \mid \tfrac{2}{3}\right) \\ \Rightarrow w'\left(\tfrac{1}{3}\right) &< \alpha w'\left(\tfrac{2}{3}\right) \end{aligned}$$

To hold these preference relations we need to have $\alpha = \alpha(0, 1M)$ such that

$$\frac{w'(2/3) - w'(1/3)}{w'(2/3)} < \alpha \qquad \frac{w'(1/3)}{w'(2/3)} < \alpha \tag{53}$$

If $\alpha = \alpha(0, 1M) > 0.5$, Eq. (53) holds. This situation shows that, in general, one is pessimistic about events with unknown probability. The Ellsburg paradox is resolved by the descriptive model $f^*(\Omega, \mu)$ of a value function under uncertainty.

5.4 BEHAVIORAL ANALYTIC HIERARCHY PROCESS

The analytic hierarchy process (AHP) [21] has been widely used as a powerful tool for deriving priorities or weights which reflect the relative importance of alternatives for multiple criteria decision making, because the method of ranking items by means of pairwise comparison is easy to understand and easy to use compared with other methods (e.g., Keeney and Raiffa [1]) of multiple criteria decision making. That is, AHP is appropriate as a normative approach which prescribes optimal behavior how decision should be made. However, there exist difficult phenomena to model and to explain by using conventional AHP. Rank reversal is one of these phenomena. That is, conventional AHP is inappropriate as a behavioral model which is concerned with understanding how people actually behave when making decisions.

In AHP, rank reversal has been regarded as an inconsistency in the methodology. When a new alternative is added to an existing set of alternatives, several attempts have been made to preserve the rank [22–24]. However, the rank reversal could occur in real world as seen in the well-known example of a person ordering a meal in a restaurant, shown by Luce and Raiffa [25].

In this section we show a behavioral extension [26] of a conventional AHP, such that the rank reversal phenomenon is legitimately observed and explanatory. In general, it is pointed out that the main causes of rank reversal are violation of transitivity and/or change in decision-making structure [27]. In AHP these causes correspond to inconsistency in pairwise comparison and change in hierarchical structure, respectively. Without these causes, AHP should not lead to rank reversal. But if we use inappropriate normalization procedure such that the entries sum to 1, the method will lead to rank reversal even when the rank should be preserved [24,28]. Some numerical examples which show the inconsistency in the conventional AHP and which show the legitimacy of the rank reversal in the behavioral AHP, are included.

5.4.1 Preference Characteristics and Status Characteristics

We show two characteristics in the behavioral AHP: preference characteristics and status characteristics. The preference characteristics represent the degree of satisfaction of each alternative with respect to each criterion. The status characteristics represent the evaluated value of a set of alternatives. The evaluation of each alternative for multiple criteria is performed by integrating these two characteristics.

In a conventional AHP it has been regarded that the cause of rank reversal lies in inappropriate normalization procedure such that entries sum to 1 [22]. Here we add a hypothetical alternative such that it gives the aspiration level of the decision maker for each criterion, and the (ratio) scale is determined by normalizing the eigenvectors so that the entry for this hypothetical alternative is equal to 1. Then, the weighting coefficient for the satisfied alternative will become more than or equal to 1, and the weighting coefficient for the dissatisfied alternative will become less than 1. That is, the weighting coefficient of each alternative under a concerning criterion represents the decision maker's degree of satisfaction. Unless the aspiration level of the decision maker changes, the weighting coefficient for each alternative does not change even if a new alternative is added or an existing alternative is removed from a set of alternatives.

The status characteristics represent the determined value of a set of alternatives under a criterion. If the average importance of all alternatives in the set is far from aspiration level 1 under a criterion, the weighting coefficient for this criterion is increased. Furthermore, the criterion which gives larger consistency index can be regarded that the decision maker's preference is fuzzy under this criterion. Thus, the importance of such criterion is decreased.

Let A be an $n \times n$ pairwise comparison matrix with respect to a criterion. Let $A = (a_{ij})$, then

$$1/\rho \leq a_{ij} \leq \rho \tag{54}$$

Usually, $\rho = 9$. Since $a_{ij} = w_i/w_j$ for priorities w_i and w_j

$$1/\rho \leq w_i/w_j \leq \rho \tag{55}$$

Equation (55) is satisfied when item j is at the aspiration level. In this case $w_j = 1$, then

$$1/\rho \leq w_i \leq \rho \tag{56}$$

Then, we obtain

$$-1 \leq \log_\rho w_i \leq 1 \tag{57}$$

Taking the geometrical mean of w_i's, we still obtain

$$1/\rho \leq \left(\prod_{i=1}^{n} w_i\right)^{1/n} \leq \rho \tag{58}$$

$$-1 \leq \log_\rho \left(\prod_{i=1}^{n} w_i\right)^{1/n} \leq 1 \tag{59}$$

Let

$$C := \left| \log_\rho \left(\prod_{i=1}^{n} w_i\right)^{1/n} \right| \tag{60}$$

then we obtain

$$0 \leq C \leq 1 \tag{61}$$

We call C the status characteristics which denote the average importance of n alterntives. If $C = 0$, the average importance of n alternatives is at the aspiration level. For larger C the importance of the concerning criterion is increased.

Let w_i^B be basic weights obtained from preference characteristics, CI be the consistency index, and $f(\text{CI})$ be a function of CI, which is called reliability function. We evaluate the revised weight w_i by integrating preference characteristics w_i^B and status characteristics C as

$$w_i = w_i^B \times C^{f(\text{CI})} \tag{62}$$

$$0 \leq C \leq 1$$

$$0 \leq f(\text{CI}) \leq 1$$

where

$$f(\text{CI}) = 0 \qquad \text{for CI} = 0$$

If $\sum_{i=1}^{n} w_i \neq 1$, then w_i is normalized to sum to 1. The same procedure is repeated when there exist many levels in the hierarchical structure.

If the priority of an alternative is equal to 1 under every criterion, the alternative is at the aspiration level. In this case the overall priority of this alternative is obtained as 1. Therefore, the overall priority of each alternative denotes the satisfaction level of each alternative. If this value is more than or equal to 1, the corresponding alternative is satisfactory, and conversely, if it is less than 1, the corresponding alternative is unsatisfactory. The behavioral AHP gives not only the ranking of each alternative, but it gives the level of satisfaction.

5.4.2 Algorithm of Behavioral AHP

Step 1. Multiple criteria and multiple alternatives are arranged in a hierarchical structure.

Step 2. Compare the criteria pairwise which are arranged in the one level–higher level of alternatives. Eigenvector corresponding to the maximum eigenvalue of the pairwise comparison matrix is normalized to sum to 1. The priority obtained is set to be preference characteristics which represent basic priority.

Step 3. For each criterion the decision maker is asked for the aspiration level. A hypothetical alternative which gives the aspiration level for all the criteria is added to a set of alternatives. Including this hypothetical alternative, a pairwise comparison matrix for each criterion is evaluated. The eigenvector corresponding to the maximum eigenvalue is normalized so that the entry for this hypothetical alternative is equal to 1.

Step 4. If CI $= 0$ for each comparison matrix, preference characteristics, that is, the basic priority is used as the weighting coefficient for each criterion. If CI $\neq 0$ for some criteria the priority for these criteria is revised by using Eq. (62), taking into account the status characteristics.

Step 5. If some priorities are revised taking into account the status characteristics, the priority for each criterion is normalized to sum to 1.

Step 6. The overall weight is evaluated. If there exists an upper level in the hierarchy, go to Step 7. Otherwise, stop.

Step 7. Evaluate the pairwise comparison matrix of criteria with respect to each criterion in the higher level. If some pairwise comparison matrices are not consistent, evaluate status characteristics and revise the priority. Go to Step 6.

5.4.3 Numerical Examples

5.4.3.1 Rank Reversal Should Not Be Observed

When pairwise comparisons are consistent and there exists no change in the hierarchical structure of decision making, rank reversal should not be observed. This example deals with such a case, where by using conventional AHP, rank reversal will be observed.

Suppose there exist four criteria (C_1, C_2, C_3, C_4) and four alternatives (A_1, A_2, A_3, A_4) where each criterion is equally weighted. Table 2 shows the result of the direct rating of each alternative under each criterion.

From this table we can describe a consistent pairwise comparison matrix, and thus overall weighting and ranking of four alternatives can be obtained by using AHP. Table 3 shows the overall weighting and ranking obtained by Saaty's AHP and our behavioral AHP where in the behavioral AHP it is assumed that alternative A_4 is at the aspiration level for all the four criteria.

Suppose we eliminate alternative A_4 from a set of alternatives. Table 4 shows overall weighting and ranking obtained by Saaty's AHP and our behavioral AHP.

In Saaty's conventional AHP, rank reversal is observed even though the rank reversal should not be observed.

Table 2 Direct Rating of Alternatives

	C_1	C_2	C_3	C_4
A_1	1	9	1	3
A_2	9	1	9	1
A_3	8	1	4	5
A_4	4	1	6	6

Table 3 Overall Weighting and Ranking Before Eliminating an Alternative

	Saaty's AHP		Behavioral AHP	
	Weight	Rank	Weight	Rank
A_1	0.261	1	2.479	1
A_2	0.252	2	1.229	2
A_3	0.245	3	1.125	3
A_4	0.241	4	1.000	4

Table 4 Overall Weighting and Ranking After Eliminating an Alternative

	Saaty's AHP		Behavioral AHP	
	Weight	Rank	Weight	Rank
A_1	0.320	3	2.479	1
A_2	0.336	2	1.229	2
A_3	0.344	1	1.125	3

5.4.3.2 Rank Reversal Could Be Observed (I)

When a pairwise comparison matrix is inconsistent under a criterion, the consistency index will become large. Under this criterion the decision maker's preference is fuzzy. In this case it is considered that the decision-making process will proceed, taking into account the preference characteristics and the status characteristics. This example deals with such a case.

Suppose three alternatives (a, b, c) are evaluated under two criteria (X, Y). Table 5 shows pairwise comparison of two criteria. As this result the basic weight for criteria (X, Y) is obtained as

$$w_X^B = 0.333 \qquad w_X^B = 0.667$$

Suppose the aspiration levels for the criteria (X, Y) are s_X, s_Y). Suppose we obtained pairwise comparison matrices as shown in Table 6 in which the aspiration level is included for each criterion.

The pairwise comparison matrix under criterion X is completely consistent, therefore the status characteristics do not affect the preference characteristics at all. But the pairwise comparison matrix under criterion Y is somewhat inconsistent, therefore we need to take into account the status characteristics. If we compute status characteristics C based on Eq. (60), we obtain 0.147. Suppose a reliability function is written simply as $f(\text{CI}) = 10 \times \text{CI}$, we obtain revised weights for criteria X and Y as

Table 5 Pairwise Comparison of Multiple Criteria

	X	Y	Weight
X	1	1/2	0.333
Y	2	1	0.667

$$w_x = 0.333$$

$$w_Y = 0.667 \times 0.147^{10\times 0.014} = 0.510$$

Since pairwise comparison matrix under criterion Y is somewhat inconsistent, the weight for criterion Y is decreased. Then, the weight for these criteria is normalized to sum to 1 as

$$w_X = 0.395 \qquad w_Y = 0.605$$

As the result the weight for the consistent criterion is increased and the weight for the inconsistent criterion is decreased.

Table 7 shows overall weighting and ranking for four alternatives, including a hypothetical alternative s, such that it gives the aspiration level for each criterion.

Table 6 Pairwise Comparison Matrix of Each Alternative

X	a	b	c	s_X	Weight
a	1	1/6	1/2	1/2	0.50
b	6	1	3	3	3.00
c	2	1/3	1	1	1.00
s_X	2	1/3	1	1	1.00

CI = 0.

Y	a	b	c	s_y	Weight
a	1	7	2	3	2.97
b	1/7	1	1/3	1/3	0.57
c	1/2	3	1	2	1.56
s_Y	1/3	3	1/2	1	1.00

CI = 0.014.

Table 7 Overall Weighting and Ranking

	X	Y	Weight	Rank
	0.395	0.605		
a	0.500	2.976	2.00	1
b	3.000	0.423	1.44	2
c	1.000	1.560	1.34	3
s	1.000	1.000	1.00	4

Table 8 **Revised** Result of Pairwise Comparison Matrix Under Criterion Y

Y	a	b	c	s_Y	Weight
a	1	8	2	3	3.11
b	1/8	1	1/4	1/3	0.38
c	1/2	4	1	2	1.68
s_Y	1/3	3	1/2	1	1.00

CI = 0.005.

Table 9 Overall Weighting and Ranking After Revising the Pairwise Comparison Matrix

	X	Y	Weight	Rank
	0.354	0.646		
a	0.500	3.111	2.19	1
b	3.000	0.377	1.27	3
c	1.000	1.679	1.45	2
s	1.000	1.000	1.00	4

Suppose the decision maker was told that the pairwise comparison matrix under criterion Y is somewhat inconsistent. If they revised the pairwise comparison matrix to decrease the consistency index as shown in Table 8, the weight for each criterion would be revised as

$$w_X = 0.333$$

$$w_Y = 0.667 \times 0.163^{10\times 0.005} = 0.609$$

and the normalized weight to sum to 1 is obtained as

$$w_X = 0.354 \qquad w_Y = 0.646$$

The overall weighting and ranking are obtained as shown in Table 9.

In this example we can find that rank reversal could be observed when there exist inconsistency and ambiguity of pairwise comparison.

5.4.3.3 Rank Reversal Could Be Observed (II)

In general, criteria included in the hierarchical structure of AHP are such that alternatives to be evaluated are discriminated and ranked under each criterion. A common criterion such that it takes the same value for all the alternatives cannot be included in the hierarchical structure. Existence of such criterion can be found in the example of a man who orders a meal in a restaurant, as shown in Luce and Raiffa [25].

Suppose a man ordered beef steak when he found beef steak and salmon steak on the menu. But when he found that escargot was also on the menu, he changed his mind and ordered salmon steak. This is a typical rank reversal phenomenon when an alternative is added to the existing set of alternatives. How could we explain his preference?

The reason why rank reversal is observed in his preference is as follows. By recognizing that the restaurant could serve escargot, he perceived that the quality of the restaurant is very high. As a result, what he wants in this restaurant has changed. By adding an alternative "escargot" to the menu a new attribute "quality of restaurant" is added to the criteria. If we were to add this new attribute to the criteria of the conventional AHP, could we model a proper decision-making process of rank reversal? The answer to this question is no, since the attribute "quality of restaurant" is common to all the alternatives "beef," "salmon," and "escargot." That is, it is not possible to do pairwise comparison of these three alternatives under the criterion "quality of restaurant."

By perceiving a different quality of restaurant, his preference changed in that he thought that all the alternatives would taste better, he could increase his budget, even an inexpensive alternative must be tasty, and so forth. The change of aspiration level could model these phenomena.

Suppose he evaluates two alternatives "beef steak" and "salmon steak" under two criteria "taste" and "price" at the beginning. He describes his aspiration level for each criterion. Then, he will start to do a pairwise comparison. Suppose Table 10 is obtained for pairwise comparison of two criteria and Table 11 is obtained for pairwise comparison of two alternatives under each criterion where it is assumed that the aspiration level for taste is the same as salmon steak and prices for beef steak, salmon steak, and aspiration level are 2000 yen, 1000 yen, and 800 yen, respectively.

Taking into account that each pairwise comparison matrix is completely consistent, we will obtain overall weighting and ranking for beef steak and salmon steak, as shown in Table 12.

Table 10 Pairwise Comparison of Criteria

	Taste	Price	Weight
Taste	1	1/2	0.333
Price	2	1	0.667

Table 11 Pairwise Comparison of Alternatives Under Each Criterion

	Beef	Salmon	AL	Weight
Taste				
Beef	1	2	2	2.0
Salmon	1/2	1	1	1.0
AL	1/2	1	1	1.0
Price				
Beef	1	0.5	0.4	0.4
Salmon	2	1	0.8	0.8
AL	2.5	1.25	1	1.0

Suppose he found a new alternative, escargot, for 4800 yen, and perceived that this restaurant is a high-quality restaurant. Under this circumstance, suppose his aspiration level is changed so that the aspiration level for taste is the same as beef steak and the aspiration level for price is raised up to 1200 yen. Table 13 shows the resulting weighting and ranking for beef steak, salmon steak, and escargot. Rank reversal is observed in comparing Tables 12 and 13. This rank reversal phenomenon can be interpreted as, by the change of aspiration level, the degree of predominance of beef steak in taste is decreased and the degree of predominance of salmon steak in price is increased. As a result, the overall weighting of salmon steak is increased and that of beef steak is decreased.

It is thus shown that a conventional AHP is not an adequate behavioral model of multiple criteria decision making, although it has been used as a powerful tool for a normative model. In this section a behavioral analytical hierarchy process has been described. This behavioral AHP could properly model the legitimacy of rank reversal when an alternative is added or removed from an existing set of alternatives. Key ideas are to include a hypothetical alternative which gives an aspiration level to each criterion when a pairwise comparison matrix is composed and to take into account the status characteristics to modify the preference characteristics.

Table 12 Overall Weighting and Ranking Before Adding an Alternative

	Taste	Price	Weight	Rank
	0.333	0.667		
Beef	2.000	0.400	0.933	1
Salmon	1.000	0.800	0.867	2

Table 13 Overall Weighting and Ranking After Adding an Alternative

	Taste	Price	Weight	Rank
	0.333	0.667		
Beef	1.000	0.600	0.733	3
Salmon	0.500	1.200	0.967	1
Escargot	2.000	0.250	0.833	2

5.5 CONCLUDING REMARKS

As a powerful tool of decision analysis, single-attribute and multiattribute utility theory based on the expected utility hypothesis was described. Measurable value function under certainty based on the concept of strength of preference was then described. This value function was then extended to measurable value functions under risk and also under uncertainty to resolve various paradoxes arising in the utility theory based on the expected utility hypothesis.

Since the analytic hierarchy process (AHP) has been widely used for decision analysis of multiple criteria decision making, we discussed the deficiency of conventional AHP and described a revised behavioral model of AHP which deals with understanding how people actually behave when making decisions.

REFERENCES

1. RL Keeney, H Raiffa. Decisions with Multiple Objectives. Cambridge, UK: Cambridge University Press, 1993. (first published by Wiley in 1976.)
2. J von Neumann, O Morgenstern. Theory of Games and Economic Behavior. 3rd ed. New York: Wiley, 1953.
3. H Tamura, Y Nakamura. Decompositions of multiatribute utility functions based on convex dependence. Operat Res 31:488, 1983.
4. PC Fishburn. von Neumann–Morgenstern utility functions. Operat. Res 22:35, 1974.
5. DE Bell. Multiattribute utility functions: Decompositions using interpolation. Manag Sci 25: 744, 1979.
6. PC Fishburn. Utility Theory for Decision Making. New York: Wiley, 1970.
7. JS Dyer, RK Sarin. Measurable multiattribute value functions. Operat Res 27:810, 1979.
8. H Tamura, S Hikita. On measurable multiattribute value functions based on finite-order independence of structural difference. In: G. Fandel et al., eds. Large-Scale Modelling and Interactive Decision Analysis. Lecture Notes in Economics and Mathematical Systems 273, Springer, Berlin, 1986, p 1.
9. DH Kranz, RD Luce, P Suppes, A Tversky. Foundations of Measurement. New York: Academic Press, 1971.
10. M Allais, O Hagen, eds. Expected Utility Hypothesis and the Allais Paradox. Dordrecht, Holland: D Reidel, 1979.
11. D Kahneman, A Tversky. Prospect theory: an analysis of decision under risk. Econometrica 47:263, 1979.
12. BP Stigum, F Wenstop, eds. Foundations of Utility and Risk Theory with Applications. Dordrecht, Holland: D Reidel, 1983.
13. H Tamura, Y Mori, Y Nakamura. On a measurable value function under risk: A descriptive model of preference resolving the expected utility paradoxes: In: Y Sawaragi, K Inoue, H Nakayama, eds. Toward Interactive and Intelligent Decision Support Systems, vol 2. Lecture Notes in Economics and Mathematical Systems 286, Springer: Berlin, 1987, p 210.
14. JS Dyer, RK Sarin. Relative risk aversion. Manag Sci 28:875, 1982.
15. GB Hazen. Independence assumptions regarding strength of preference in risky decision making. Department of Industrial Engineering and Management Science, Northwestern University, 1982.
16. G Shafer. A Mathematical Theory of Evidence. Princeton, NJ: Princeton University Press, 1976.
17. J Jaffray. Application of linear utility theory to belief functions. Proceedings of 2nd International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU '88), 1988 pp 1–8.
18. PJH Schoemaker. Experiments on Decision Under Risk: The Expected Utility Hypothesis. Boston: Kluwer-Nijhoff, 1982.
19. PJH Schoemaker. The expected utility model: its variants, purposes, evidence and limitations. J Econ Lit XX:529, 1982.
20. D Ellsburg. Risk, ambiguity and the Savage axioms. Q J Econ 75:643, 1961.
21. TL Saaty. The Analytic Hierarchy Process. New York: McGraw-Hill, 1980.
22. V Belton, T Gear. On a shortcoming of Saaty's method of analytic hierarchies. OMEGA Int J Manag Sci 11:228–230, 1983.
23. J Barzilai, WD Cook, B. Golany. Consistent weights for judgements matrices of relative importance of alternatives. Operat Res Lett 6:131, 1987.
24. JS Dyer. Remarks on the analytic hierarchy process. Manag Sci 36:249, 1990.

25. RD Luce, H Raiffa. Games and Decisions. New York: Wiley, 1957.

26. H Tamura, S Takahashi, I Hatono, M Umano. On a descriptive analytic hierarchy process (D-AHP) for modeling the legitimacy of rank reversal. Proceedings of the International Conference on Methods and Applications of Multicriteria Decision Making, Mons, Belgium, 1997, pp 82–85.

27. A Tversky, P Slovic, D Kahneman. The cause of preference reversal. Am Econ Rev 80:204, 1990.

28. AA Salo, RP Hamalainen. Preference assessment by imprecise ratio statements. Operat Res 40:1053, 1992.

Chapter 5.1

Sensors: Touch, Force, and Torque

Richard M. Crowder
University of Southampton, Southampton, England

1.1 INTRODUCTION

The objective of any robotic sensing system is to acquire knowledge and resolve uncertainty about the robot's environment, including its relationship with the workpiece. Prior to discussing the requirements and operation of specific sensors, the broad objectives of sensing need to be considered. The control of a manipulator or industrial robot is based on the correct interpretation of sensory information. This information can be obtained either internally to the robot (for example, joint positions and motor torques) or externally using a wide range of sensors. The sensory information can be obtained from both vision and nonvision sensors. A vision system allows the position and orientation of the workpiece to be acquired; however, its performance is dependent on lighting, perspective distortion, and the background. A touch, force, or torque sensor will provide information regarding the contact between the sensor and workpiece, and is normally localized in nature. It is recognized that these sensors will not only complement vision sensing, but offer a powerful sensing capability in their own right. Vision may guide the robot arm through many manufacturing operations, but it is the sense of touch that will allow the robot to perform delicate manipulations and assembly tasks.

1.2 TOUCH AND TACTILE SENSING

Touch and tactile sensors are devices which measure the parameters of a contact between the sensor and an object. This interaction obtained is confined to a small defined region. This contrasts with a force and torque sensor, which measures the total forces being applied to an object. In the consideration of tactile and touch sensing, the following definitions are commonly used:

Touch sensing. This is the detection and measurement of a contact force at a defined point. A touch sensor can also be restricted to binary information, namely, touch and no touch.

Tactile sensing. This is the detection and measurement of the spatial distribution of forces perpendicular to a predetermined sensory area, and the subsequent interpretation of the spatial information. A tactile sensing array can be considered to be a coordinated group of touch sensors.

Slip. This is the measurement and detection of the movement of an object relative to the sensor. This can be achieved either by a specially designed slip sensor or by the interpretation of the data from a touch sensor or a tactile array.

Tactile sensors can be used to sense a diverse range of stimuli, from detecting the presence or absence of a grasped object to a complete tactile image. A tactile sensor consists of an array of touch-sensitive sites; the sites may be capable of measuring more than one property. The contact forces measured by a sensor are able to convey a large amount of information about the state of a grip. Texture, slip, impact, and other contact conditions generate force and position signatures that can be used to identify the state of a manipulation. This information can be determined by examination of the frequency domain, and is fully discussed in the literature [1].

As there is no comprehensive theory available that defines the sensing requirements for a robotic system, much of the knowledge is drawn from investigation of human sensing, and the analysis of grasping and manipulation. Study of the human sense of touch suggests that creating a gripper incorporating tactile sensing requires a wide range of sensors to fully determine the state of a grip. The detailed specification of a touch sensor will be a function of the actual task it is required to perform. Currently, no general specification of a touch or tactile sensor exists. Reference 2, though dated, can be used as an excellent basis for defining the desirable characteristics of a touch or tactile sensor suitable for the majority of industrial applications:

- A touch sensor should ideally be a single-point contact, though the sensory area can be any size. In practice, an area of 1–2 mm^2 is considered a satisfactory compromise between the difficulty of fabricating a subminiature sensing element and the coarseness of a large sensing element.
- The sensitivity of the touch sensor is dependent on a number of variables determined by the sensor's basic physical characteristic. In addition the sensitivity may also be the application, in particular any physical barrier between the sensor and the object. A sensitivity within the range 0.4 to 10 N, together with an allowance for accidental mechanical overload, is considered satisfactory for most industrial applications.
- The minimum sensor bandwidth should be 100 Hz.
- The sensor's characteristics must be stable and repeatable with low hysteresis. A linear response is not absolutely necessary, as information-processing techniques can be used to compensate for any moderate nonlinearities.
- As the touch sensor will be used in an industrial application, it will need to be robust and protected from environmental damage.
- If a tactile array is being considered, the majority of applications can be undertaken by an array 10–20 sensors square, with a spatial resolution of 1–2 mm.

In a dexterous end effector, the forces and relative motions between the grasped object and the fingers need to be controlled. This can be achieved by using a set of sensors capable of determining in real time the magnitude, location, and orientation of the forces at the contact point. This problem has been approached by using miniature force and torque sensors inside the fingertips, to provide a robot with an equivalent to the kinesthetic sense found in humans. The integration of skinlike and kinesthetic-like sensing will result in robots being equipped with artificial haptic perceptions [3].

The study of human touch and the use of perceived information indicates that other variables, such as hardness and thermal properties, can also be measured, and this allows greater flexibility in an automated process. Human touch is of considerable complexity, with sensors that respond to a range of stimuli including temperature, pain, acceleration, velocity, and intensity. The human touch sensors in the skin may have many purposes, but are predominantly protective to prevent self-inflicted damage to the body. The human touch sense is obtained by a combination of four sensors: a transient load detector, a continuous force sensor, a position transducer to give proprioceptive data, and an overload sensor (i.e., pain) reacting both to force and other external environmental conditions. This combination of sensors is very sensitive, e.g., a fine surface texture can be detected, but there is poor spatial resolution; the difficulty in reading Braille is readily apparent. Humans are very good at learning about an unknown object from touch. The information from the sensors is brought together through the nervous system to give us the sense of feel. It should be noted that the sensory information is processed and interpreted both locally (peripheral nervous system) and centrally (spinal cord and the brain).

1.2.1 Touch Sensor Technology

Many physical principles have been exploited in the development of tactile sensors. As the technologies involved are very diverse, this chapter can only consider the generalities of the technology involved. In most cases, the developments in tactile sensing technologies are application driven. It should be recognized that the operation of a touch or tactile sensor is very dependent on the material of the object being gripped.

The sensors discussed in this chapter are capable of working with rigid objects. However, if nonrigid material is being handled, problems may arise. Work has shown that conventional sensors can be modified to operate with nonrigid materials [4].

1.2.1.1 Mechanically Based Sensors

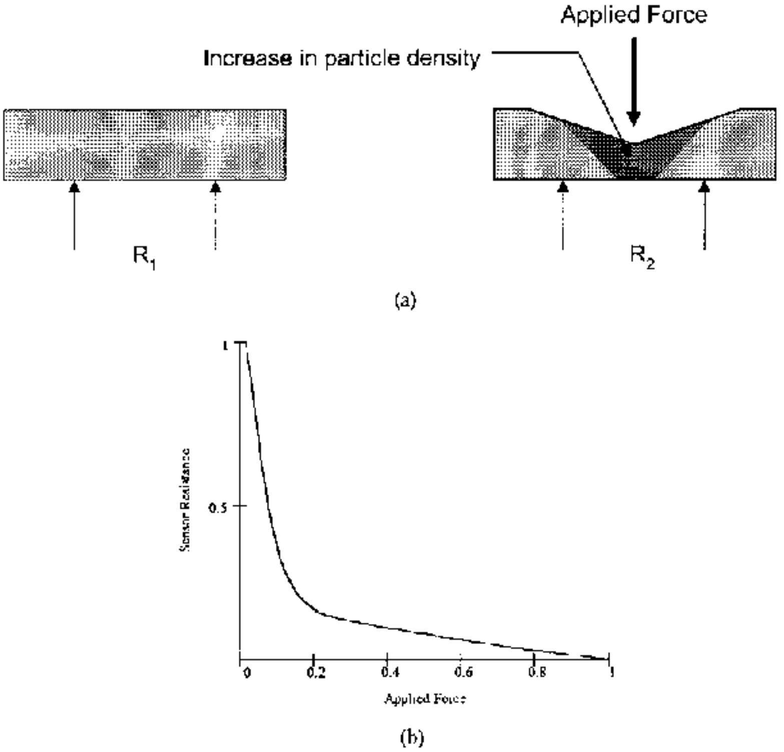
The simplest form of touch sensor is one where the applied force is applied to a conventional mechanical microswitch to form a binary touch sensor. The force required to operate the switch will be determined by its actuating characteristics and any external constraints. Other approaches are based on a mechanical movement activating a secondary device, such as a potentiometer or displacement transducer.

1.2.1.2 Resistive-Based Sensors

The use of compliant materials that have a defined force-resistance characteristics have received considerable attention in touch and tactile sensor research [5]. The basic principle of this type of sensor is the measurement of the resistance of a conductive elastomer or foam between two points. The majority of the sensors use an elastomer that consists of a carbon-doped rubber. The resistance of the elastomer changes with the application of force, resulting from the deformation of the elastomer altering the particle density (Fig. 1). If the resistance measurement is taken between opposing surfaces of the elastomer, the upper contacts have to be made using a flexible printed circuit to allow movement under the applied force. Measurement from one side can easily be achieved by using a dot-and-ring arrangement on the substrate (Fig. 2). Resistive sensors have also been developed using elastomer cords laid in a grid pattern, with the resistance measurements being taken at the points of intersection. Arrays with 256 elements have been constructed [6]. This type of sensor easily allows the construction of a tactile image of good resolution.

The conductive elastomer or foam-based sensor, while relatively simple, does suffer from a number of significant disadvantages:

An elastomer has a long nonlinear time constant. In addition the time constant of the elastomer, when force is applied, is different from the time con-



(a)

(b)

Figure 1 Resistive sensor based on a conductive foam or elastomer. (a) Principle of operation. (b) Normalized resistance against applied force.

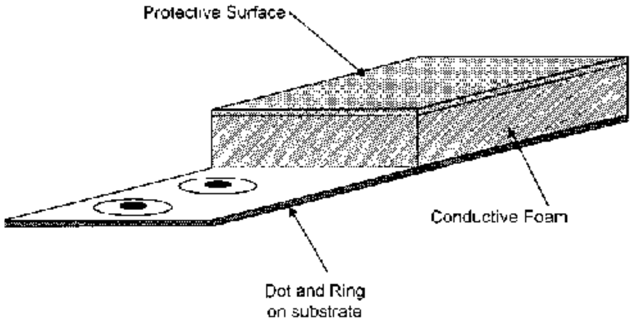


Figure 2 A resistive tactile sensor based on a dot-and-ring approach.

stant when the applied force is removed.

The force–resistance characteristics of elastomer-based sensors are highly nonlinear, requiring the use of signal-processing algorithms.

Due to the cyclic application of forces experienced by a tactile sensor, the resistive medium within the elastomer will migrate over a period of time. Additionally, the elastomer will become permanently fatigued, leading to permanent deformation of the sensor. This will give the sensor a poor long-term stability and will require replacement after an extended period of use.

Even with the electrical and mechanical disadvantages of conductive elastomers and foams, the majority of industrial analog touch or tactile sensors have been based on the principle of resistive sensing. This is due to the simplicity of their design and interface to the robotic system.

1.2.1.3 Force-Sensing Resistor

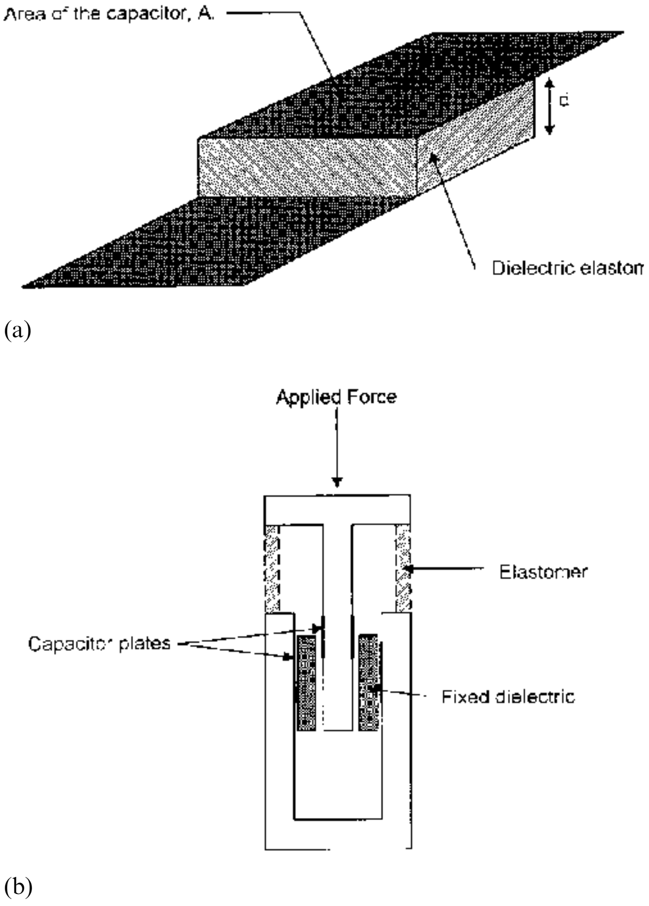
A force-sensing resistor is a piezoresistive conductive polymer, which changes resistance in a predictable manner following application of force to its surface. It is normally supplied as a polymer sheet which has had the sensing film applied by screen printing. The sensing film consists of both electrically conducting and nonconducting particles suspended in a matrix. The particle sizes are of the order of fractions of microns, and are formulated to reduce the temperature dependence, improve mechanical properties and increase surface durability. Applying a force to the surface of a sensing film causes particles to touch the conducting electrodes, changing the resistance of the film. As with all resistive-based sensors the force-sensitive resistor requires a relatively simple interface and can operate satisfactorily in moderately hostile environments.

1.2.1.4 Capacitive-Based Sensors

The capacitance between two parallel plates is given by

$$C = \frac{\varepsilon A}{d} \tag{1}$$

where A is the plate area, d the distance between the plates, and ε the permittivity of the dielectric medium. A capacitive touch sensor relies on the applied force either changing the distance between the plates or the effective surface area of the capacitor. In Fig. 3a, the two conductive plates of the sensor are separated by a dielectric medium, which is also used as the elastomer to give the sensor its force-to-capacitance characteristics.



(a)

(b)

Figure 3 (a) Parallel plate capacitive sensor. (b) Principal components of a coaxial force sensor.

To maximize the change in capacitance as force is applied, it is preferable to use a high-permittivity dielectric in a coaxial capacitor design. Figure 3b shows the cross-section of the capacitive touch transducer in which the movement of one set of the capacitor's plates is used to resolve the displacement and hence applied force. The use of a highly dielectric polymer such as polyvinylidene fluoride maximizes the change in capacitance. From an application viewpoint, the coaxial design is better as its capacitance will give a greater increase for an applied force than the parallel plate design. In both types of sensors, as the size is reduced to increase the spatial resolution, the sensor's absolute capacitance will decrease. With the limitations imposed by the sensitivity of the measurement techniques, and the increasing domination of stray capacitance, there is an effective limit on the resolution of a capacitive array.

To measure the change in capacitance, a number of techniques can be, the most popular is based on the use of a precision current source. The charging characteristic of the capacitive sensor is given by

$$I = \frac{C\,dv}{dt} = \frac{\varepsilon A}{d}\frac{dV}{dt} \tag{2}$$

hence, the voltage across the sensor over a period of time is defined as

$$dV = \frac{I\,dt\,d}{\varepsilon A} \tag{3}$$

As the current source, I, and sampling period, dt, are defined, the capacitance and hence the applied force can be determined [7]. A second approach is to use the sensor as part of a tuned or LC circuit, and measure the frequency response. Significant problems with capacitive sensors can be caused if they are in close proximity with the end effector's or robot's earthed metal structures, as this leads to stray capacitance. This can be minimized by good circuit layout and mechanical design of the touch sensor. It is possible to fabricate a parallel plate capacitor on a single silicon slice [8]. This can give a very compact sensing device; this approach is discussed in Sec. 1.2.1.10.

1.2.1.5 Magnetic-Based Sensor

There are two approaches to the design of touch or tactile sensors based on magnetic transduction. Firstly, the movement of a small magnet by an applied force will cause the flux density at the point of measurement to change. The flux measurement can be made by either a Hall effect or a magnetoresistive device. Second, the core of the transformer or inductor can be manufactured from a magnetoelastic material that will deform under pressure and cause the magnetic coupling between transformer windings, or a coil's inductance, to change. A magnetoresistive or magnetoelastic material is a material whose magnetic characteristics are modified when the material is subjected to changes in externally applied physical forces [9]. The magnetorestrictive or magnetoelastic sensor has a number of advantages that include high sensitivity and dynamic range, no measurable mechanical hysteresis, a linear response, and physical robustness.

If a very small permanent magnet is held above the detection device by a compliant medium, the change in flux caused by the magnet's movement due to an applied force can be detected and measured. The field intensity follows an inverse relationship, leading to a nonlinear response, which can be easily linearized by processing. A one-dimensional sensor with a row of 20 Hall-effect devices placed opposite a magnet has been constructed [10]. A tactile sensor using magnetoelastic material has been developed [11], where the material was bonded to a substrate, and then used as a core for an inductor. As the core is stressed, the material's susceptibility changes; this is measured as a change in the coil's inductance.

1.2.1.6 Optical Sensors

The rapid expansion of optical technology in recent years has led to the development of a wide range of tactile sensors. The operating principles of optical-based sensors are well known and fall into two classes:

Intrinsic, in which the optical phase, intensity, or polarization of transmitted light are modulated without interrupting the optical path

Extrinsic, where the physical stimulus interacts with the light external to the primary light path.

Intrinsic and extrinsic optical sensors can be used for touch, torque, and force sensing. For industrial applications, the most suitable will be that which requires the least optical processing. For example, the detection of phase shift, using interferometry, is not considered a practical option for robotic touch and force sensors. For robotic touch and force-sensing applications, the extrinsic sensor based on intensity measurement is the most widely used due to its simplicity of construction and the subsequent information processing. The potential benefits of using optical sensors can be summarized as follows:

Immunity to external electromagnetic interference, which is widespread in robotic applications.

Intrinsically safe.

The use of optical fiber allows the sensor to be located some distance from the optical source and receiver.

Low weight and volume.

Touch and tactile optical sensors have been developed using a range of optical technologies:

Modulating the intensity of light by moving an obstruction into the light path. The force sensitivity is determined by a spring or elastomer. To prevent crosstalk from external sources, the sensor can be constructed around a deformable tube, resulting in a highly compact sensor (Fig. 4a).

A design approach for a reflective touch sensor is shown in Fig. 4b, where the distance between the reflector and the plane of source and the detector is the variable. The intensity of the received light is a function of distance, and hence the applied force. The U-shaped spring was manufactured from spring steel, leading to a compact overall design. This sensor has been successfully used in an anthropomorphic end effector [12]. A reflective sensor can be constructed with source–receiver fiber pairs embedded in a solid elastomer structure. As shown in Fig. 4c, above the fiber is a layer of clear elastomer topped with a reflective silicone rubber layer. The amount of light reflected to the receiver is determined by an applied force that changes the thickness of the clear elastomer. For satisfactory operation the clear elastomer must have a lower compliance than the reflective layer. By the use of a number of transmitter–receiver pairs arranged in a grid, the tactile image of the contact can be determined [13].

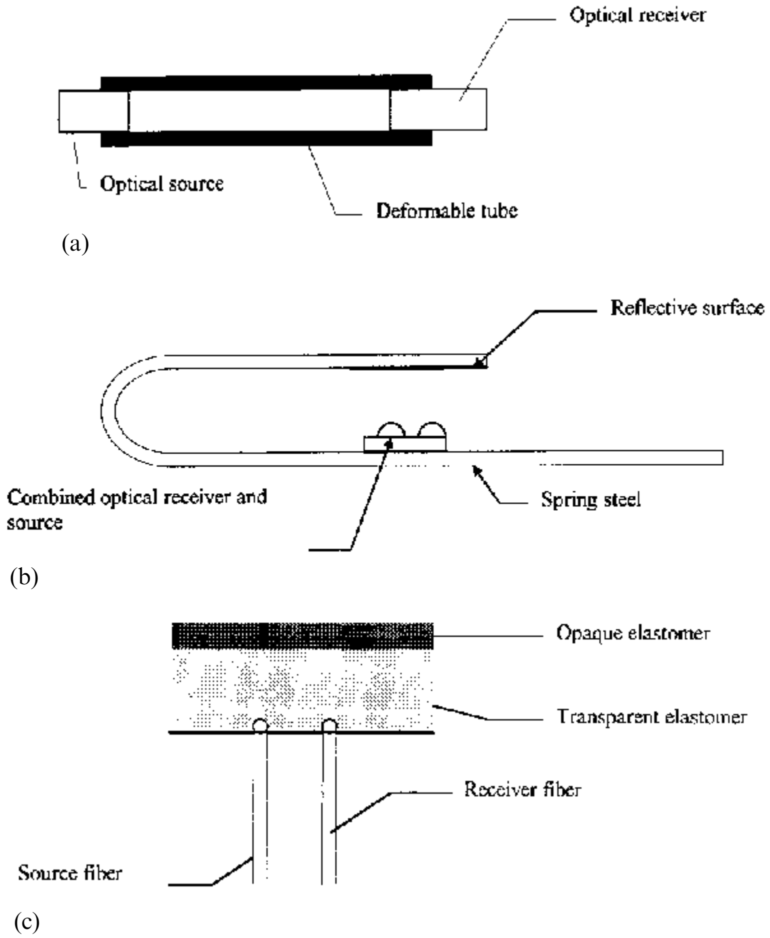


Figure 4 (a) Optical touch sensor based on obstructing the light path by a deformable tube. (b) Optical reflective touch sensor. (c) Optical reflective sensor based on two types of elastomer.

Photoelasticity is the phenomenon where stress or strain causes birefringence in optically transparent materials. Light is passed through the photoelastic medium. As the medium is stressed, it effectively rotates the plane of polarization, and hence the intensity of the light at the detector changes as a function of the applied force [14]. A suitable sensor is discussed in Section 1.2.2.2.

A change in optical density occurs at a boundary, and determines if total internal reflection may occur. As shown in Fig. 5, an elastomer membrane is separated by air from a rigid translucent medium that is side illuminated. If the elastomer is not in contact with the surface, total internal reflection will occur and nothing will be visible to the detector. However, as the membrane touches the top surface of the lower medium, the boundary conditions will change, thus preventing total internal reflection, and the light will be scattered. Hence an image will be seen by the detector. The generated image is highly suitable for analysis by a vision system [15].

1.2.1.7 Optical-Fiber-Based Sensors

In the previous section, optical fibers were used solely for the transmission of light to and from the sensor; however, tactile sensors can be constructed from the fiber itself. A number of tactile sensors have been developed using this approach. In the majority of cases either the sensor structure was too big to be attached to the fingers of a robotic hand or the operation was too complex for use in the industrial environment. A suitable design can be based on internal-state microbending of optical fibers. Microbending is the process of light attenuation in the core of fiber where a mechanical bend or perturbation (of the order of few microns) is applied to the outer surface of the fiber. The degree of attenuation depends on the fiber parameters as well as radius of curvature and spatial wavelength of the bend. Research has demonstrated the feasibility of effecting microbending on an optical fiber by the application of a force to a second orthogonal optical fiber [16]. One sensor design comprises four layers of fibers, each layer overlapping orthogonally to form a rectangular grid pattern. The two active layers are sandwiched between two corrugation layers, where the fibers in adjcent layers are slightly staggered from each other for better microbending effect. When the force is applied to a fiber intersection, microbending appears in the stressed fibers, attenuating the transmitted light. The change in the light intensity provides the tactile information.

1.2.1.8 Piezoelectric Sensors

Although quartz and some ceramics have piezoelectric properties, polymeric materials that exhibit piezoelectric properties are suitable for use as touch or tactile sensors; polymers such as polyvinylidene fluoride

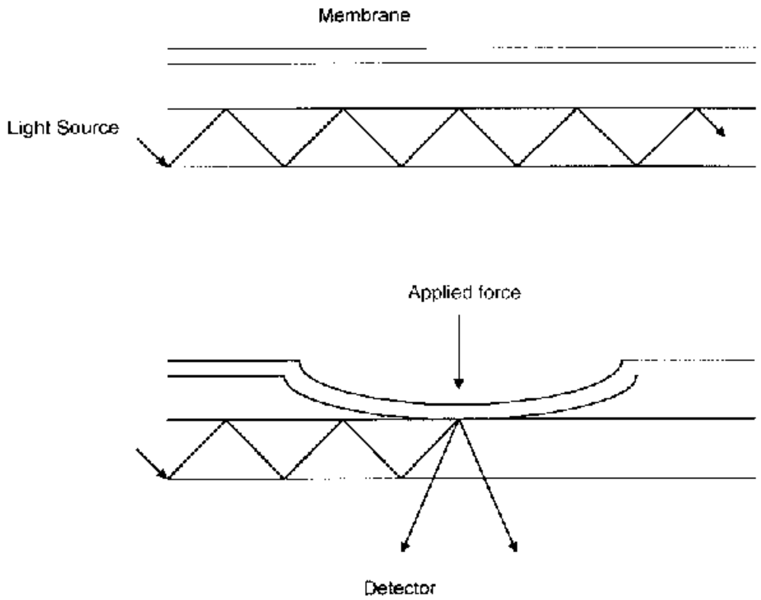


Figure 5 Optical boundary sensor.

(PVDF) are normally used [17]. Polyvinylidene fluoride is not piezoelectric in its raw state, but can be made piezoelectric by heating the PVDF within an electric field. Polyvinylidene fluoride is supplied as sheets between 5 μm and 2 mm thick, and has good mechanical properties. A thin layer of metalization is applied to both sides of the sheet to collect the charge and permit electrical connections to be made. In addition it can be molded, hence PVDF has number of attractions when considering tactile sensor material as an artificial skin.

As a sensing element the PVDF film acts as a capacitor on which charge is produced in proportion to the applied stress. The charge developed can be expressed in terms of the applied stress, $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \sigma_3]^T$, the piezoelectric constant, $\mathbf{d} = [d_1, d_2, d_3]^T$, and the surface area, giving

$$\mathbf{q} = A\delta \cdot \boldsymbol{\sigma} \tag{4}$$

The piezoelectric touch transducer is most often used in conjunction with a charge amplifier; this results in an output voltage that is proportional to the applied stress. Using a high-impedance field-effect transistor (FET) input amplifier (Fig. 6), the amplifier's output voltage is given by

$$v = \frac{dq}{dt} R_f = AR_f \mathbf{d} \cdot \frac{d\boldsymbol{\sigma}}{dt} \tag{5}$$

which can be calibrated to give a force measurement.

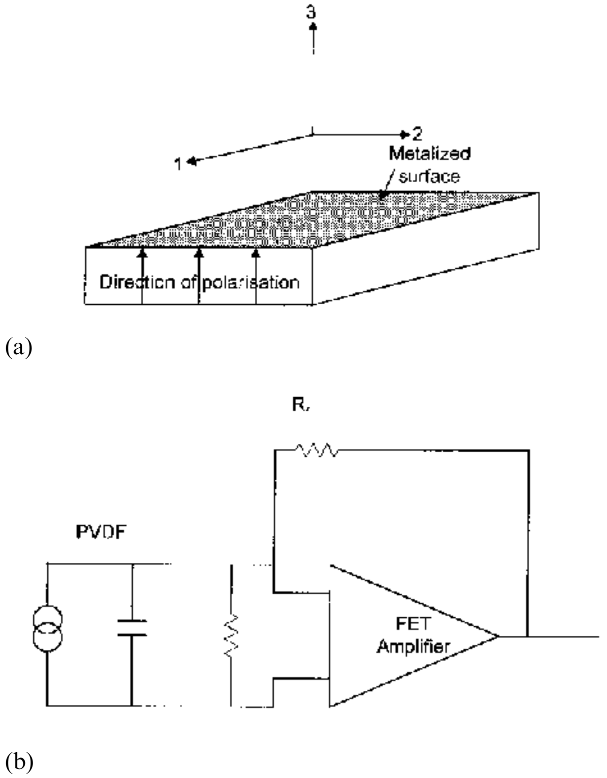


Figure 6 PVDF touch sensor. (a) Definition used in the polarization of PVDF film. (b) Equivalent circuit of a sensor.

The piezoelectric sensors are essentially dynamic, and are not capable of detecting static forces. In practice their use is restricted to specialist applications such as slip and texture detection. The use of PVDF in piezoelectric sensors causes difficulty in scanning an array of sensing elements, as PVDF exhibits pyroelectric effects. Therefore some applications require a reference sensor of unstressed PVDF to allow the separation of the piezoelectric effect from the pyroelectric signal.

1.2.1.9 Strain Gages in Tactile Sensors

A strain gage, when attached to a surface, will detect the change in length of the material as it is subjected to external forces. The strain gage is manufactured from either resistive elements (foil, wire, or resistive ink) or from semiconducting material. A typical resistive gage consists of a resistive grid bonded to an epoxy backing film. If the strain gage is prestressed prior to the application of the backing medium, it is possible to measure both tensile and compressive stresses. The semiconducting strain gage is fabricated from a suitably doped piece of silicone; in this case the mechanism used for the resistance change is the piezoresistive effect [18].

When applied to robotic touch applications, the strain gage is normally used in two configurations: as a load cell, where the stress is measured directly at the point of contact, or with the strain gage positioned within the structure of the end effector.

1.2.1.10 Silicon-Based Sensors

Technologies for micromachining sensors are currently being developed worldwide. The developments can be directly linked to the advanced processing capabilities of the integrated circuit industry, which has developed fabrication techniques that allow the interfacing of the nonelectronic environment to be integrated through microelectromechanical systems [19]. Though not as dimensionally rigorous as the more mature silicon planar technology, micromachining is inherently more complex as it involves the manufacture of a three-dimensional object. Therefore the fabrication relies on additive layer techniques to produce the mechanical structure.

The excellent characteristics of silicon, which have made micromachined sensors possible, are well known [20], and include a tensile strength comparable to steel, elasticity to breaking point, very little mechanical hysteresis in devices made from a single crystal, and a low thermal coefficient of expansion.

To date it is apparent that microengineering has been applied most successfully to sensors. Some sensor applications take advantage of the device-to-device or batch-to-batch repeatability of wafer-scale processing to remove expensive calibration procedures. Current applications are restricted largely to pressure and acceleration sensors, though these in principle can be used as force sensors. As the structure is very delicate, there are still problems in developing a suitable tactile sensor for industrial applications [21].

1.2.1.11 Smart Sensors

The most signfiicant problem with the sensor systems discussed so far is that of signal processing. Researchers are therefore looking to develop a complete sensing system rather than individual sensors, together with individual interfaces and interconnections. This allows the signal processing to be brought as close as possible to the sensor itself or integrated with the sensor. Such sensors are generally termed *smart sensors*. It is the advances in silicon fabrication techniques which have enabled the recent developments in smart sensors. There is no single definition of what a smart sensor should be capable of doing, mainly because interest in smart sensors is relatively new. However, there is a strong feeling that the minimum requirements are that the sensing system should be capable of self-diagnostics, calibration, and testing. As silicon can be machined to form moving parts such as diaphragms and beams, a tactile sensor can, in principle, be fabricated on single piece of silicon. Very little commercial success has been obtained so far, largely due to the problems encountered in transferring the technology involved from the research laboratory to industry.

In all tactile sensors there is a major problem of information processing, and interconnection. As an $n \times n$ array has $2n$ connections and individual wires, any reduction in interconnection requirements is welcomed for ease of construction and increased reliability. A number of researchers have been addressing the problem of integrating a tactile sensor with integral signal processing. In this design the sensor's conductive elastomer sheet was placed over a substrate. The significant feature of this design is that the substrate incorporates VLSI circuitry so that each sensing element not only measures its data but processes it as well. Each site performs the measurements and processing operations in parallel. The main difficulty with this approach was the poor discrimination, and susceptibility to physical damage. However, the VLSI approach was demonstrated to be viable, and alleviated the problems of wiring up each site and processing the data serially.

1.2.1.12 Multistimuli Touch Sensors

It has been assumed that all the touch sensors discussed in this section respond only to a force stimulus. However, in practice most respond to other external stimuli, in particular, temperature. If PVDF has to be used as a force sensor in an environment with a widely varying ambient temperature, there may be a requirement for a piece of unstressed PVDF to act as a temperature reference. It is possible for a sensor to respond both to force and temperature changes. This has a particular use for object recognition between materials that have different thermal conductivity, e.g., between a metal and a polymer [22]. If the complexity of the interpretation of data from PVDF is unsuitable for an application, touch sensors incorporating a resistive elastomer for force, and thermistors for temperature measurement can be constructed. By the use of two or more force-sensitive layers on the sensor, which have different characteristics (e.g., resistive elastomer and PVDF), it is possible to simulate the epidermal and dermal layers of human skin.

1.2.2 Slip Sensors

Slip may be regarded as the relative movement of one object's surface over another when in contact. The relative movement ranges from simple translational motion to a combination of translational and rotational motions. When handling an object, the detection of slip becomes necessary so as to prevent the object being dropped due to the application of a low grip force. In an assembly operation, it is possible to test the occurrence of slip to indicate some predetermined contact forces between the object and the assembled part. For the majority of applications some qualitative information on object slip may be sufficient, and can be detected using a number of different approaches.

1.2.2.1 Interpretation of Tactile-Array Information

The output of a tactile-sensing array is the spatial distribution of the forces over the measurement area. If the object is stationary, the tactile image will also remain stationary. However, if the pattern moves with time, the object can be considered to be moving; this can be detected by processing the sensor's data.

1.2.2.2 Slip Sensing Based on Touch-Sensing Information

Most point-contact touch sensors are incapable of discrimination between relative movement and force. However, as the surfaces of the tactile sensor and the object are not microscopically smooth, the movement of an object across the sensor will cause a high-frequency, low-amplitude vibration to be set up, which can be detected and interpreted as movement across the sensor. This has been achieved by touch sensors based on the photoelastic effect [23] and piezoelectric [24] sensors. In a photoelastic material the plane of polarization is a function of the material stress. Figure 7a shows a sensor, developed at the University of Southampton, to detect slip. The sensor uses the property of photoelastic material, where the plane of the material's polarization is rotated as the material is stressed. In the sensor, light is first passed through a polarizing film (polarizer), the material, then a second polarizing film (analyzer). As the stress applied to the material changes, the amount of received light varies.

Typical results are shown in Fig. 7b; the changes in stress are caused by vibrations due to the photoelastic material slipping–sticking as the object moves relative to the sensor. The sensitivity of the sensor can be increased by artificially roughening the surface area of the sensor. In addition to slip detection, the information from the sensor can be used to determine information about the surface roughness of the gripped object by measurement of the vibration characteristics.

1.2.2.3 Sensors to Specifically Detect Slip

It is possible to develop sensors that will respond only to relative movement. They are normally based on the principle of transduction discussed for touch sensors, but the sensors' stimulus comes from the relative movement of an area of the gripper.

Several methods to detect slip have been reported. One sensor requries a sapphire needle protruding from a sensor surface to touch the slipping object; this generates vibrations which in turn stimulate a piezoelectric crystal. The disadvantage of this approach is that it picks up external vibrations from the gripper and robot mechanics, and the needle frequently wears out. The improved version of this sensor uses a steel ball at the end of the probe, with the piezoelectric crystal replaced by a permanent magnet and a coil enclosed in a damping medium. To avoid the problem of interference signals from external vibrations, a range of interrupt-type slip sensors have been designed. In one design, a rubber roller has a permanent magnet passing over a magnetic head which generates a voltage when slip occurs. In a similar design the roller has a number of slits which interrupt an optical path; this allows an indication of slip to be obtained. Though these sensors give a very good indication of the speed and direction of slip there are disadvantages with poor slip resolution and the possibility of jamming of the roller.

1.2.3 Summary

This section has discussed the technology available for touch, tactile, and slip sensors. In the interpretation of a sensor's information, consideration should be taken of its design and use. One aspect that is often overlooked is the mechanical filtering of the sensory information caused by the sensor's protective covering material. Work has shown [25] that a cover of as little as 0.2 mm thick will degrade a sensor that is required to have a spatial resolution of less than 1 mm. As shown in Fig. 8, a point contact is diffused so that a number of sensors are stimulated. The degree of filtering is a function of the covering material, its thickness, and its physical properties, and requires the use of finite-element tecniques to be analyzed fully.

For any application the characteristics of a number of sensors may need to be compared. Table 1 presents a summary of the major advantages and disadvantages, allowing this comparison to be made between the transduction techniques.

1.3 FORCE AND TORQUE MEASUREMENT

As noted earlier, touch sensors operate at the point of contact. If, however, it is required to measure the global forces and torques being exerted on an object by a robotic system, a multiaxis force measurement system is needed. If an object fixed in space is considered and

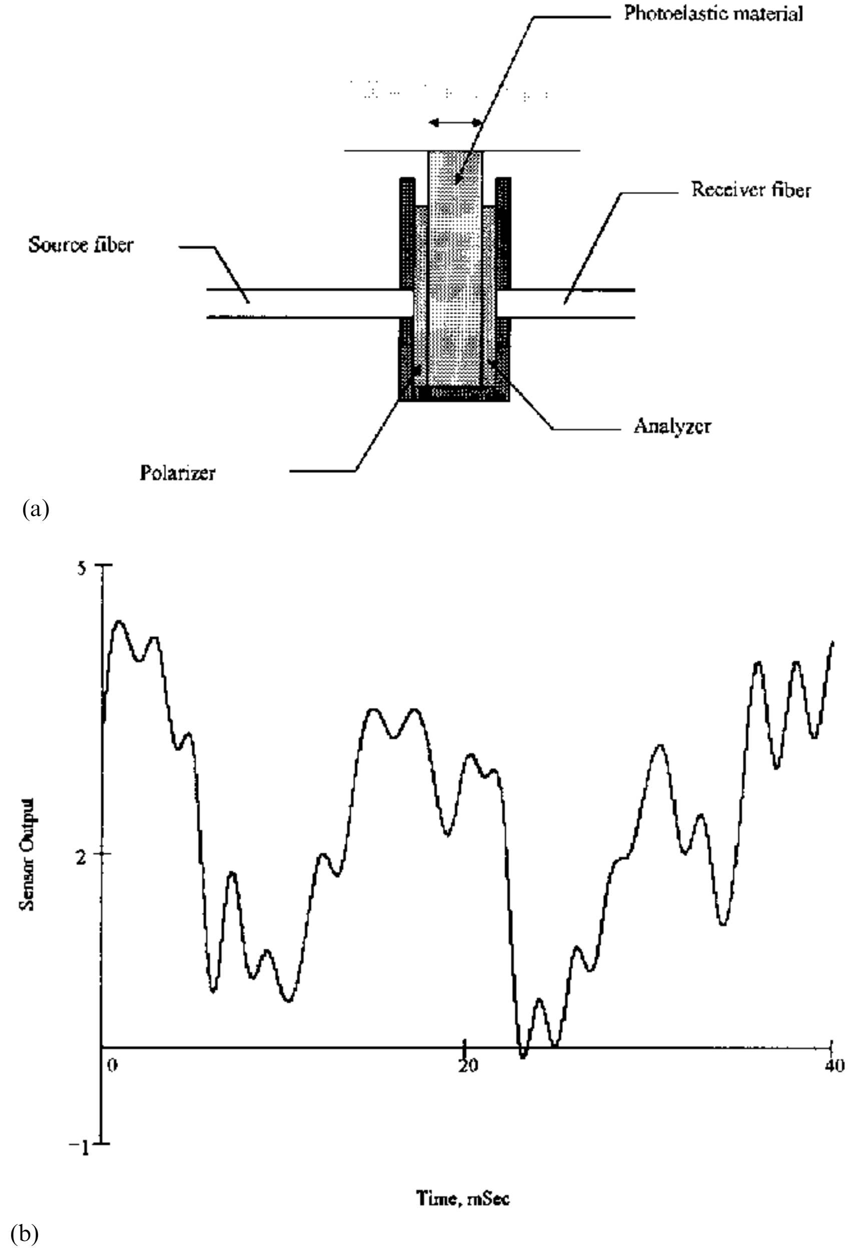


Figure 7 (a) Optical slip sensor. (b) Typical results from a photoelastic slip sensor.

subjected to a combination of the six possible orthogonal forces and torque, the object will deform. This can be measured by suitable sensors. In general a multiaxis force sensor incorporates a number of transducers measuring the load applied to a mechanical structure. These sensors greatly enhance the capabilities of a robotic system, by incorporating compliant damping control, active stiffness control and hybrid force control.

For a generalized multiaxis force sensor, it is possible to write

$$[\mathbf{S}] = [\mathbf{C}][\mathbf{F}] \tag{6}$$

where [**S**] is a single-column matrix of the sensor outputs, [**F**] the applied forces, and [**C**] the $n \times 6$ coupling matrix, n being the number of sensors. Therefore,

$$[\mathbf{F}] = [\mathbf{D}][\mathbf{S}] \tag{7}$$

Table 1 Summary of Sensor Characteristics

Sensor	Advantage	Disadvantage
Mechanical	Simple	Poor spatial resolution Arrays complex to manufacture
Resistive (including FSR)	Wide dynamic range Durable Good overload tolerance	Hysteresis Limited spatial resolution
Capacitive	Wide dynamic range Robust	Susceptible to EMI Temperaturer sensitive Limitation in spatial resolution
Magnetic	Wide dynamic range Robust	Poor spatial resolution Susceptible to EMI
Optical (intrinsic and extrinsic)	Intrinsically safe Very high resolution possible	Electronics can be complex
Piezoelectric	Wide dynamic range Good mechanical properties	Pyroelectric effect Dynamic response only Scanning of charge amplifier

where $[\mathbf{D}] = [\mathbf{C}]^{-1}$ and is known as the decoupling matrix. Thus each of the six forces can be calculated from

$$F_i = \sum_{j=1}^{n} D_{ij} S_j \tag{8}$$

To measure the forces and torques that the robot is subjecting an object to, it is necessary to design a suitable structure to mount within the robot's wrist [26,27]. Figure 9 shows a typical four-beam structure, with eight strain gages [28]. Even though the use of strain gages is the most common method of measurement, there is no reason why other sensing techniques cannot be used. In addition, the structure could be constructed from magnetoelastic material, or use can be made of photoelastic or capacitive sensors. In order to compute the forces, the coefficients of the [**D**] matrix have to be determined. For a four-beam structure, the relationship is

$$[\mathbf{D}] = \begin{bmatrix} 0 & 0 & D_{13} & 0 & 0 & 0 & D_{17} & 0 \\ D_{21} & 0 & 0 & 0 & D_{25} & 0 & 0 & 0 \\ 0 & D_{32} & 0 & D_{34} & 0 & D_{36} & 0 & D_{38} \\ 0 & 0 & 0 & D_{44} & 0 & 0 & 0 & D_{48} \\ 0 & D_{52} & 0 & 0 & 0 & D_{56} & 0 & 0 \\ D_{61} & 0 & D_{63} & 0 & D_{65} & 0 & D_{67} & 0 \end{bmatrix}$$

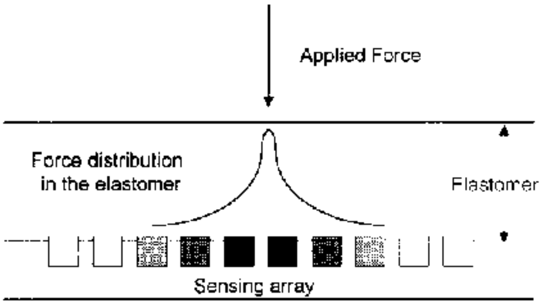


Figure 8 Diffusion of a point force due to the mechanical properties of the elastomer.

This decoupling matrix will only be valid if the cross-coupling between the eight gages are neglible. In practice, cross-coupling will occur and the use of the above [**D**] matrix will result in errors of approximately 5%. To achieve negligible errors, the [**D**] matrix will need to contain 48 nonzero elements. The value of the elements can only be determined by calibration, using published methods [29].

There are a number of possible locations within a robot system where information regarding the applied force or torque can be measured, but in general the closer the sensor is located to the point of interest, the more accurate the measurement. Possible sensor locations are:

At the interface between the end effector and the robot, as part of the robot's wrist. The location requires little re-engineering but the sensor must

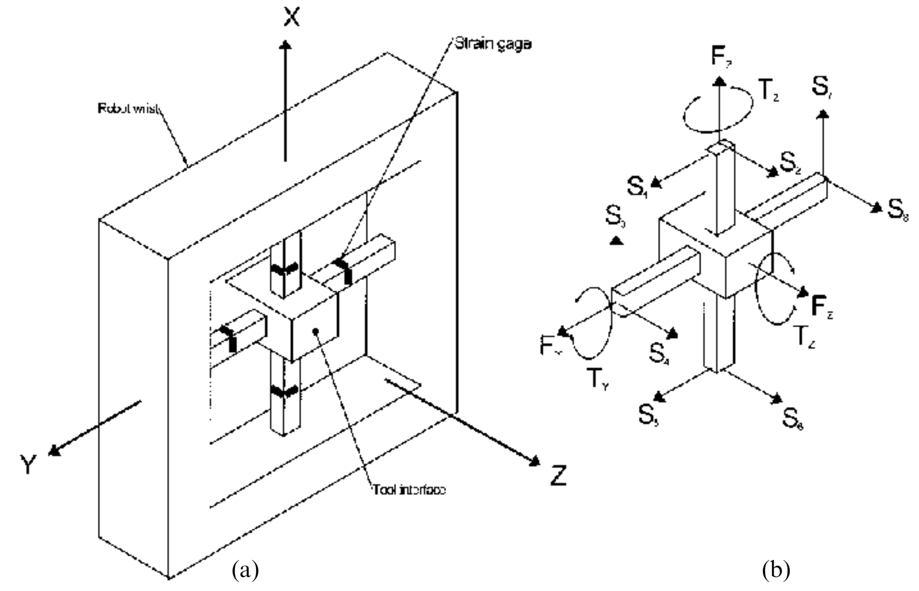


Figure 9 Four-beam wrist sensor. (a) The generalized construction, showing the location of the strain gages and the robot and tool interfaces. (b) The relationship between the strain gages and the six forces.

have high stiffness, to ensure that the disturbing forces are damped, allowing a high sampling rate. The high stiffness will minimize the deflections of the wrist under the applied forces that lead to positional errors. The sensor needs to be small so as not to restrict the movements of the robot within the workspace.

Within the individual joints of the robots by using joint torque sensing; however, inertia, gravity loading and joint friction present in a manipulator will complicate the determination of the forces.

The wrist sensor structure discussed above is effectively rigid; however, a force and torque sensor can be designed to have the fast error absorption of a passive compliance structure and the measurement capabilities of a multiaxis sensor. The structure, including its sensing package, is normally termed the instrumented remote center compliance (IRCC) [30]. In the IRCC, some or all of the remote center compliance deflections are measured. From the deflections, the applied forces and torques can be determined and, if required, used in the robot control algorithms. In the conventional IRCC the end effector interface plate is mounted on three shear blocks, with the base connected to the robot. The applied forces cause the platform to move relative to the base and this movement gives the IRCC its remote center characteristics. The movements can be detected by the use of displacement transducers or by the use of two-dimensional optical position sensors. In the position sensor, a light source fixed to the platform illuminates the sensing element, allowing the position of the light source can be measured in two axes. This, when combined with the outputs from the other position sensors, is used to calculate the applied force, both magnitude and direction. It should be noted that the calculation of the applied forces using an IRCC is nontrivial due to the complex movement of the platform under load.

1.3.1 External to the Robot

The touch, torque, and force sensors that have been discussed are also suitable for mounting external to the robot. In most assembly tasks, the robot is used to make two parts and the applied force can be measured by sensors attached to the workpiece. This information can be used either to modify the robot's position, the applied forces, or to adjust the position of workpiece. To achieve the latter, the workpiece has to be mounted on a table capable of multiaxis movement. The construction of the table is similar to an IRCC and the same form of algorithms can be used.

It is possible to use any of the touch and tactile sensors external to the robot, the only limitation being the accuracy and resolution of the task being performed. While the Maltese cross sensor is usually placed within the robot's wrist, it can be placed at almost any point in the robot's structure. After the wrist, the most common place is integral to the robot's base, where it is termed a pedestal sensor. However, it is of only limited use at this point due to the complexity of the transformations.

1.4 CONCLUDING COMMENTS

The last decade has seen considerable effort applied to research and development activities related to the design of touch, force, and torque sensors, primarily for robotics applications. This brief survey has not considered the processing of the measured data, sensory data fusion, and sensory-motor integration. Research on these topics is rapidly expanding. Most of the work related to the processing methodologies and algorithms have been focused on the analysis of static tactile images, following the lines developed in the field of machine vision, some of which have been reported in the literature [31]. This approach is limited in scope and does not consider a major asset of tactile sensing which lies in the processes of active touch and manipulation. Planning active touch procedures and analyzing the pertinent sensor data was recognized early on as being important, but progress in this area has been quite slow.

As a final observation, it should be noted that although the rapid growth in interest in this field of sensing has initiated significant progress in haptic technology, very little movement to real applications has occurred. At present the market for these devices is still very marginal, despite some early optimistic forecasts. Future widespread use of tactile and haptic systems is still foreseen, but the time scale for these events to occur should be realistically correlated with the great theoretical and technical difficulties associated with this field, and with the economic factors that ultimately drive the pace of its development.

REFERENCES

1. B Eberman, JK Salisbury. Application of dynamic change detection to dynamic contact sensing. Int J Robot Res 13(5): 369–394, 1994.
2. LD Harmon. Automated tactile sensing. Int J Robot Res 1(2): 3–32, 1982.
3. AM Okamura, ML Turner, MR Cutkosky. Haptic exploitation of objects with rolling and sliding. IEEE Conference on Robotics and Automation, 1997, New York, p 2485–2490.
4. R Stone, P Brett. A sensing technique for the measurement of tactile forces in the gripping of dough like material. Proc Inst Mech Engrs 210(B3): 309–316, 1996.
5. HR Nicholls. Advanced Tactile Sensing for Robotics. Singapore: World Scientific Publishing, 1992.
6. BE Robertson, AJ Walkden. Tactile sensor system for robotics. In: A Pugh, ed. Robot Sensors, vol 2: Tactile and Non-Vision. Bedford, UK: IFS (Publications), 1986, p 89–97.
7. BV Jayawant, MA Onori, J Watson. Robotic tactile sensing: a new array sensor: robot sensors. In: A Pugh, ed. Robot Sensors, vol 2: Tactile and Non-Vision. Bedford, UK: IFS (Publications), 1986, p 120–125.
8. Y Lee, K Wise. A batch fabricated silicone capacitive pressure transducer with low temperature sensitivity. IEEE Trans Electron Dev ED-29(1): 1982, p 42–48.
9. J Vranish, R Demoyer. Outstanding potential shown by magnetoelastic force feedback sensors for robots. Sensor Rev 2(4): 1982.
10. G Kinoshita, T Hajika, K Hattori. Multifunctional tactile sensors with multi-elements for fingers. Proceedings of the International Conference on Advanced Robotics, 1983, pp 195–202.
11. EE Mitchell, J Vranish. Magnetoelastic force feedback sensors for robots and machine tools—an update. Proceedings, 5th International Conference on Robot Vision and Sensory Controls, 1985, p 200–205.
12. RM Crowder. An anthropomorphic robotic end effector. Robot Autonom Syst 1991, p 253–268.
13. J Schneiter, T Sheridan. An optical tactile sensor for manipulators. Robot Computer Integr Manuf 1(1): 65–71, 1984.
14. W Splillman, D McMahan. Multimode Fibre Optic Sensor Based on Photoelastic Effect. Sudby, MA: Sperry Research Centre, 1985.
15. P Dario, D De Rossi. Tactile sensors and the gripping challenge. IEEE Spectrum 1985, p 46–52.
16. D Jenstrom, C Chen. A fibre optic microbend tactile array sensor. Sensors Actuators 20(3): 239–248, 1989.
17. P Dario, R Bardelli, D de Rossi, L Wang. Touch sensitive polymer skin uses piezoelectric properties to recognise orientation of objects. Sensor Rev 2(4): 194–198, 1982.
18. JM Borky, K Wise. Integrates signal conditioning for silicone pressure sensing. IEEE Trans Electron Dev ED26(12): 1906–1910, 1979.
19. J Bryzek, K Petersen, W McCalley. Micromachines on the march. IEEE Spectrum 31(5): 20–31, 1994.
20. L Vinto. Can micromachining deliver? Solid State Technol 38: 57–59, 1995.
21. J Smith, C Baron, J Fleming, S Montague, J Rodriguez, B Smith, J Sniegowski. Micromachined sensors and actuator research at a microelectronics development laboratory. Proceedings of the American Conference on Smart Structures and Materials, San Diego, 1995, pp 152–157.

22. D Siegel, I Garabieta, J Hollerbach. An integrated tactile and thermal sensor. IEEE Conference on Robotics and Automation, San Francisco, California, 1996, pp 1286–1291.
23. F Kvansik, B Jones, MS Beck. Photoelastic slip sensors with optical fibre links for use in robotic grippers. Institute of Physics Conference on Sensors, Southampton, UK, 1985, pp 58–59.
24. R Howe, M Cutkosky. Dynamic tactile sensing, perception of fine surface features with stress rate sensing. IEEE Trans Robot Autom 9(2): 145–150, 1993.
25. AM Shimojo. Mechanical filtering effect of elastic covers for tactile sensing. IEEE Trans Robot Autom 13(1): 128–132, 1997.
26. H Van Brussel, H Berlien, H Thielemands. Force sensing for advanced robotic control. Robotics 2(2): 139–148, 1986.
27. PM Will, D Grossman. An experimental system for computer controlled assembly. IEEE Trans Computers C-24(9) 1975, pp 879–87.
28. B Roth, B Shimans. On force sensing information and its use in controlling manipulators, Information Control problems in Manufacturing, Tokyo 1980, p 119–26.
29. K Bejczy. Smart sensors for smart hands. Progr Astronaut Aeronaut 1980.
30. T DeFazio, D Seltzer, DE Whitney. The IRCC instrumented remote center compliance. Rob Sens 2; 33–44, 1986.
31. HR Nicholls, MH Lee. A survey of robot tactile sensing technology. Int J Robot Res 8(3): 3–30, 1989.

Chapter 5.2

Machine Vision Fundamentals

Prasanthi Guda, Jin Cao, Jeannine Gailey, and Ernest L. Hall
University of Cincinnati, Cincinnati, Ohio

2.1 INTRODUCTION

The new machine vision industry that is emerging is already generating millions of dollars per year in thousands of successful applications. Machine vision is becoming established as a useful tool for industrial automation, where the goal of 100% inspection of manufactured parts during production is becoming a reality. The purpose of this chapter is to present an overview of the fundamentals of machine vision. A review of human vision is presented first to provide an understanding of what can and cannot be easily done with a machine vision system.

2.2 HUMAN VISUAL SYSTEM

Human beings receive at least 75% of their total sensory input through visual stimuli; our vision processes are executed in billions of parallel neural networks at high speeds, mainly in the visual cortex of the brain. Even with the fastest supercomputers, it is still not possible for machines to duplicate all of the functions of human vision. However, an understanding of the fundamentals of human image formation and perception provides a starting point for developing machine vision applications.

The human visual system comprises three main organs: the eyes, the optic nerve bundle, and the visual cortex of the brain. This complex system processes a large amount of electrochemical data and performs its tasks in a highly efficient manner that cannot yet be duplicated by machine vision. However, the human visual system can also be confused by illusions. The key element to vision is light, which is radiant energy in the narrow range of the electromagnetic spectrum, from about 350 nm (violet) to 780 nm (red). This energy, upon stimulation of the retina of the human eye, produces a visual sensation we call visible light. Photometry and colorimetry are sciences that describe and quantify perceived brightness and color, and objective measurements of radiant energy.

The visual cortex of the human brain, as shown in Fig. 1, is the central location for visual processing. We believe that the inner layer, or white matter, of the brain consists mainly of connections, while the outer layer, or gray matter, contains most of the interconnections that provide neural processing. The eyes function as an optical system whose basic components consist of the lens, the iris, and the retina. The lens of the eye focuses the incoming light to form an inverted image on the retina at the rear wall of the eye. The amount of light entering the eye is controlled by a muscle group called the iris. The retina, as shown in Fig. 2, consists of about 125 million light-sensitive receptors that, because of the many-to-one connections, have some processing capabilities. These receptors consist of color-sensitive "cones" and brightness-sensitive "rods." The central part of the retina is called the fovea. It contains a dense cluster of between six and seven million cones that are sensitive to color and are connected directly to the brain via individual nerves.

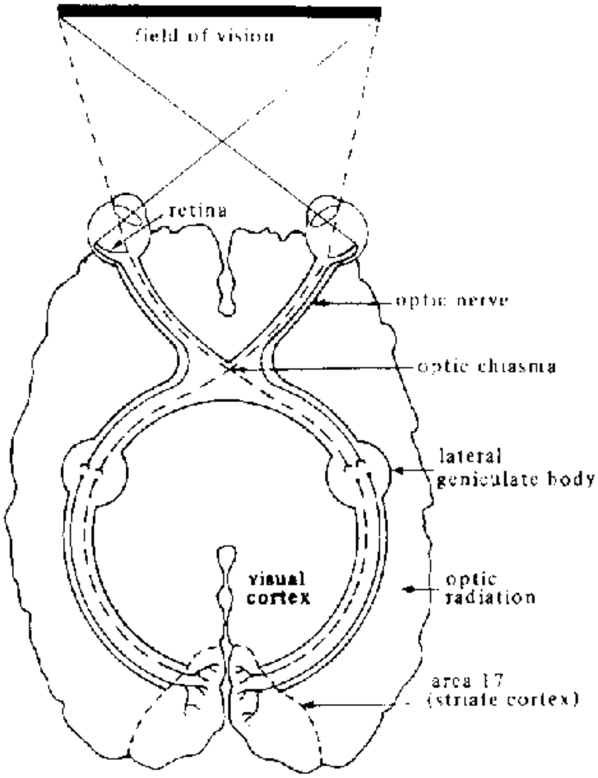


Figure 1 The visual cortex is the center for high-level processing in the human brain.

When an image is projected onto the retina, it is converted into electrical impulses by the cones and then transmitted by the optic nerves into the brain. The optic nerve has between one and two million neurons. Around the periphery and distributed across the surface of the retina are the rods. Unlike cones, rods share nerve endings and are sensitive to light and dark but are not involved in color vision. The rods in the human eye can adapt to a range of light intensities over several orders of magnitude. This permits humans to see not only outside in the bright sunlight but also in a darkened room.

Most of the neural processing of image impulses carried via the optic nerve bundle takes place in the visual cortex. Various theories have been suggested to attempt to describe visual cortical processing, including: edge enhancement, computing correlation, Fourier transforms, and other higher-level operations [1]. The basic architecture of our organic neural networks is now being used as a template for developing neural network computer algorithms. These artificial neural algorithms can perform tasks, such as recognizing objects, making decisions, function approximations, and even sytem identification and discovery. Nonetheless, many capabilities of human beings, such as visual understanding and description, still present challenging problems.

An example of early visual system neural processing is shown in Fig. 3. The neural network in this experiment, known as the *backward-inhibition model*, consists of a linear recurrent network followed by a nonlinear

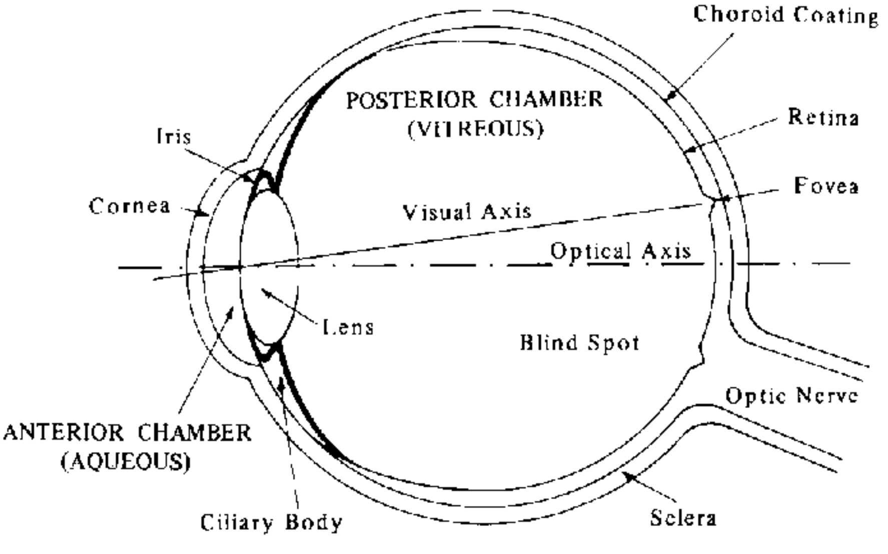
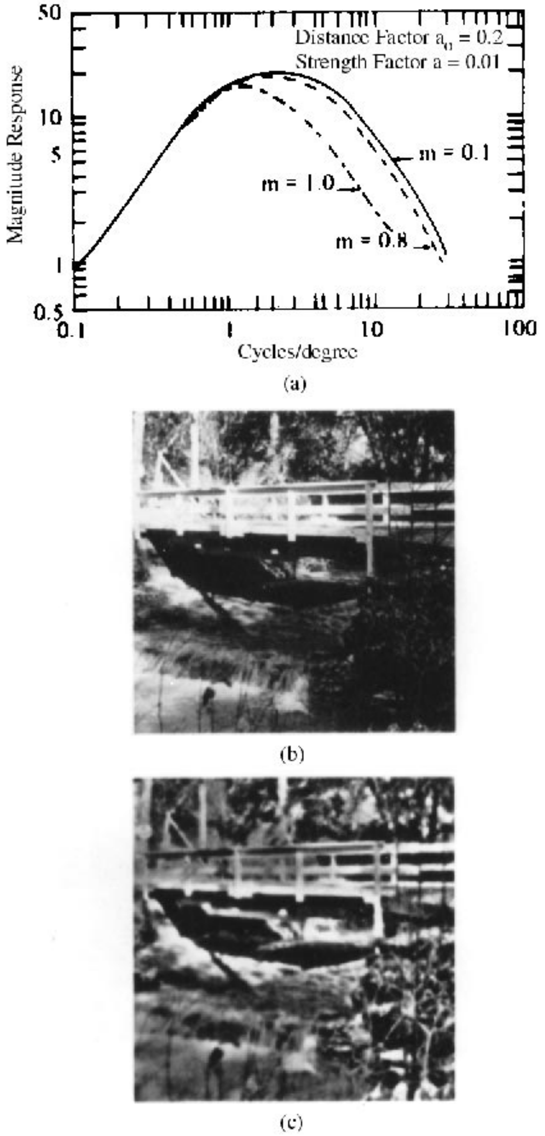


Figure 2 Anatomy of the eye showing location of retina.



(a)

(b)

(c)

Figure 3 Example of neural processing in the early stages of vision.

element followed by another network. The inhibition equation may be written as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 0 & w_{12} \\ w_{21} & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \tag{1}$$

where the output responses are y_i, the input is x_i and the coefficients, w_{ij}, regulate the amount of inhibition. This model was used to demonstrate the nonlinear nature of the frequency sensitivity of the human visual system [2]. The human visual system was found to respond in a nonlinear manner to the contrast of an input signal as shown in Eq. (1). By the combination of linear and nonlinear processing, a model was developed which showed similar characteristics, as shown in Fig. 3a. An original image and the result of processing by the nonlinear model are shown in Fig. 3b and c. Note that the image is blurred slightly, but that a considerable edge enhancement is produced. As may be observed in Fig. 3, the reduction of irrelevant detail and enhancement of important edges in the dark regions of the image was achieved. This effect could be analogous to night vision. Perhaps early humans, living in caves or hunting and foraging at night, had required this survival ability to discern the outlines of predators in the dark shadows.

The neural network model may also be written to emphasize the fact that the response must be greater than a threshold, b, to produce an output:

$$y = f(g[w^T x - b]) \tag{2}$$

In this case, g represents the nonlinear, zero–one step function whose value is 1 when the threshold, b, is exceeded, and 0 otherwise. Additionally, the function f may be another function that determines frequency sensitivity or recognition selectivity. The overall neural network function is a composite function of a basic linear decision element combined with nonlinear function mappings that are characteristic of modern multilayer neural networks.

2.3 MACHINE VISION HARDWARE COMPONENTS

A machine vision system consists of hardware and software components. The basic hardware components are of a light source, a solid state camera and lens, and a vision processor. The usual desired output is data that is used to make an inspection decision or to permit a comparison with other data. The key considerations for image formation are lighting and optics.

One of the first considerations in a machine vision application is the type of illumination to be used. Natural, or ambient, lighting is always available but rarely sufficient. Point, line, or area lighting sources may be used as an improvement over ambient light. Spectral considerations should be taken into account in order to provide a sufficiently high contrast between the objects and background. Additionally, polarizing filters may be required to reduce glare or undesirable spectral reflections. If a moving object is involved, a rapid shutter or *strobe* illumination can be used to capture an image without motion blur. To obtain an excellent outline of an object's boundary, back lighting can provide an orthogonal projection used to silhouette an object. Line illumination, produced with a cylindrical lens, has proven useful in many vision systems. Laser illumination must be used with proper safety precautions, since high-intensity point illumination of the retina can cause permanent damage.

Another key consideration in imaging is selecting the appropriate camera and optics. High-quality lenses must be selected for proper field of view and depth of field; automatic focus and zoom controls are available. Cameras should be selected based on scanning format, geometrical precision, stability, bandwidth, spectral response, signal-to-noise ratio, automatic gain control, gain and offset stability, and response time. A shutter speed or frame rate greater than one-thirtieth or one-sixtieth of a second should be used. In fact, the image capture or digitization unit should have the capability of capturing an image in one frame time. In addition, for camera positioning, the position, pan and tilt angles can be servo controlled. Robot-mounted cameras are used in some applications. Fortunately, with recent advances in solid-state technology, solid-state cameras are now available at a relatively lower cost.

Since the advent of the Internet and the World Wide Web (WWW), a great variety of images are now available to anyone. This has also led to an increase in the variety of formats for image data interchange [3]. Some of the most common image formats now are bitmaps (BMP), data-compressed JPEGs (JPG), and the GIF87a (GIF) file format. The Graphics Interchange Format (GIF), shown in Table 1, was developed by CompuServe, and is used to store multiple bitmap images in a single file for exchange between platforms. The image data is stored in a bitmap format in which numbers represent the values of the picture elements or pixels. The bit depth determines the number of colors a pixel can represent. For example, a 1-bit pixel can be one of two colors, whereas an 8-bit pixel can be one of 256 colors. The maximum image size with the GIF format is 64,000 by 64,000 pixels. The image data stored in a GIF file is always compressed using the Lempel–Ziv–Welch (LZW) technique. The GIF data can also be interlaced up to 4:1 to permit images to display progressively instead of top down.

Table 1 GIF Image Format Characteristics

Header and color table information	Header
	Logical screen descriptor
	Global color table
Image 1	Local image descriptor
	Local color table
	Image data
Image 2	Local image descriptor
	Local color table
	Image data
Image *N*	Local image descriptor
	Local color table
	Image data
	Trailer

There are literally hundreds of various image file formats. Table 2 lists some common formats as well as the extension, creator, and conversion filter(s) for each format.

2.4 MACHINE VISION ALGORITHMS AND TECHNIQUES

2.4.1 Image Functions and Characteristics

As mentioned by Wagner [4], in manufacturing, human operators have traditionally performed the task of visual inspection. Machine vision for automatic inspection provides relief to workers from the monotony of routine visual inspection, alleviates the problems due to lack of attentiveness and diligence, and in some cases improves overall safety. Machine vision can even expand the range of human vision in the following ways:

Improving resolution from optical to microscopic or electron microscopic

Extending the useful spectrum from the visible to the x-ray and infrared or the entire electromagnetic range, and improving sensitivity to the level of individual photons

Enhancing color detection from just red, green, and blue spectral bands to detecting individual frequencies

Improving time response from about 30 frames per second to motion stopping strobe-lighted frame rates or very slow time lapse rates

Modifying the point of view from the limited perspective of a person's head to locations like Mars, the top of a fixture, under a conveyor or inside a running engine.

Another strength of machine vision systems is the ability to operate consistently, repetitively, and at a high rate of speed. In addition, machine vision system components with proper packaging, especially solid-state

Table 2 Other Image Formats

Image type	Extensions	Creator	Filters
BMP (Microsoft Windows bitmap image file)	bmp, dib, vga, bga, rle, rl4, rl8	Microsoft	Imconv, imagemagick, xv, Photoshop, pbmplus
CGM (Computer Graphics Metafile)	cgm	American National Standards Institute, Inc	
DDIF (DEC DDIF file)	ddif	Digital Equipment Co. (DEC)	pbmplus
GIF (Graphics Interchange Format file)	gif, giff	CompuServe Information Service	Imconv, imagemagick, xv, PhotoShop, pbmplus, Utah Rasterfile Toolkit
ICO (Microsoft Windows icon image file)	ico	Microsoft	imconv
IMG (Img-whatnot file)	img		pbmplus
JPEG (Joint Photographic Experts Group compressed file)	jpeg, jpg, jfif	The Independent JPEG Group	Imagemagick, xv, JPEGsoftware, PhotoShop, DeBabelizer
MPNT (Apple Macintosh MacPaint file)	mpnt, macp, pntg, mac, paint	Apple Computer, Inc.	Imconv, sgitools, pbmplus, PhotoShop, Utah Rasterfile Toolkit
PCD (Photo-CD file)	pcd	Kodak	Imagemagick, pbmplus, PhotoShop, DeBabelizer
PCX (ZSoft IBM PC Paintbrush file)	pcx, pcc	ZSoft Corp. (PC Paintbrush)	Imagemagick, xv, PhotoShop, DeBabelizer
PDF (Adobe Acrobat PDF file)	pdf	ADEX Corp.	
Photoshop (Adobe PhotoShop file)	photoshop	Adobe	PhotoShop, DeBabelizer
Pix (Alias image file)	pix, alias	Alias Research, Inc.	Imconv, sgitools, DeBabelizer
PS (Adobe PostScript file)	ps, ps2, postscript, psid	Adobe	Imconv, sgitools, xv, pbmplus, PhotoShop, Utah Rasterfile Toolkit
TIFF (Tagged-Image File Format)	tiff, tif	Aldus, MicroSoft, and NeXT	Imconv, sgitools, xv, pbmplus, PhotoShop, Utah Rasterfile Toolkit
XWD (X Window System window dump image file)	xwd, x11	X Consortium/MIT	Imconv, sgitools, pbmplus, Utah Rasterfile Toolkit

cameras, can be used in hostile environments, such as outer space or in a high-radiation hot cell. They can even measure locations in three dimensions or make absolute black-and-white or color measurements, while humans can only estimate relative values.

The limitations of machine vision are most apparent when attempting to do a taks that is either not fully defined or that requries visual learning and adaptation. Clearly it is not possible to duplicate all the capabilities of the human with a machine vision system. Each process and component of the machine vision system must be carefully selected, designed, interfaced, and tested. Therefore, tasks requiring flexibility, adaptability, and years of training in visual inspection are still best left to humans.

Machine vision refers to the science, hardware, and software designed to measure, record, process, and display spatial information. In the simplest two-dimensional case, a digital black-and-white image function, I, as shown in Fig. 4, is defined as

$$I = \{f(x, y) : x = 0, 1, \ldots, N-1; \\ y = 0, 1, \ldots, N-1\} \tag{3}$$

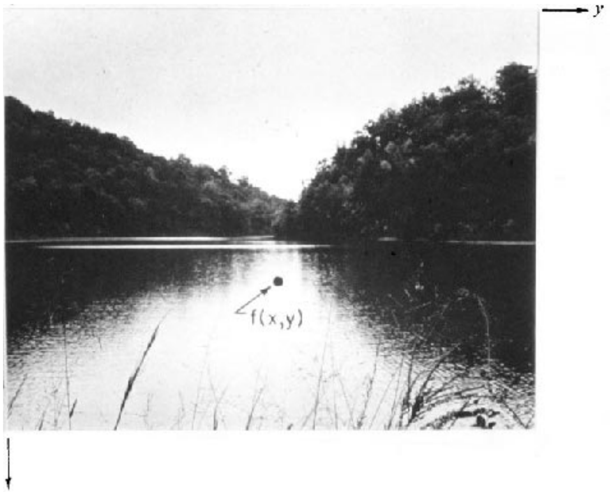


Figure 4 Black-and-white image function.

Each element of the image $f(x, y)$ may be called a picture element or *pixel*. The value of the function $f(x, y)$ is its *gray-level* value, and the points where it is defined are called its domain, window or *mask*.

The computer image is always *quantized* in both spatial and gray-scale coordinates. The effects of spatial resolution are shown in Fig. 5.

The gray-level function values are *quantized* to a discrete range so that they can be stored in computer memory. A common set of gray values might range from 0 to 255 so that the value may be stored in an 8-bit byte. Usually 0 corresponds to dark and 255 to white. The effects of gray-level quantization are shown in Fig. 6, which shows the same image displayed at $1, 2, 4$ and 8 bits per pixel. The 1-bit, or binary, image shows only two shades of gray, black for 0 and white for 1. The binary image is used to display the silhouette of an object if the projection is nearly orthographic. Another characteristic of such an image is the occurrence of false contouring. In this case, contours may be produced by coarse quantization that are not actually present in the original image. As the number of gray shades is increased, we reach a point where the differences are indistinguishable. A conservative estimate for the number of gray shades distinguishable by the normal human viewer is about 64. However, changing the viewing illumination can significantly increase this range.

In order to illustrate a simple example of a digital image, consider a case where a digitizing device, like an optical scanner, is required to capture an image of 8 × 8 pixels that represents the letter "V." If the lower intensity value is 0 and higher intensity value is 9, then the digitized image we expect should look like Fig. 7. This example illustrates how numbers can be assigned to represent specific characters.

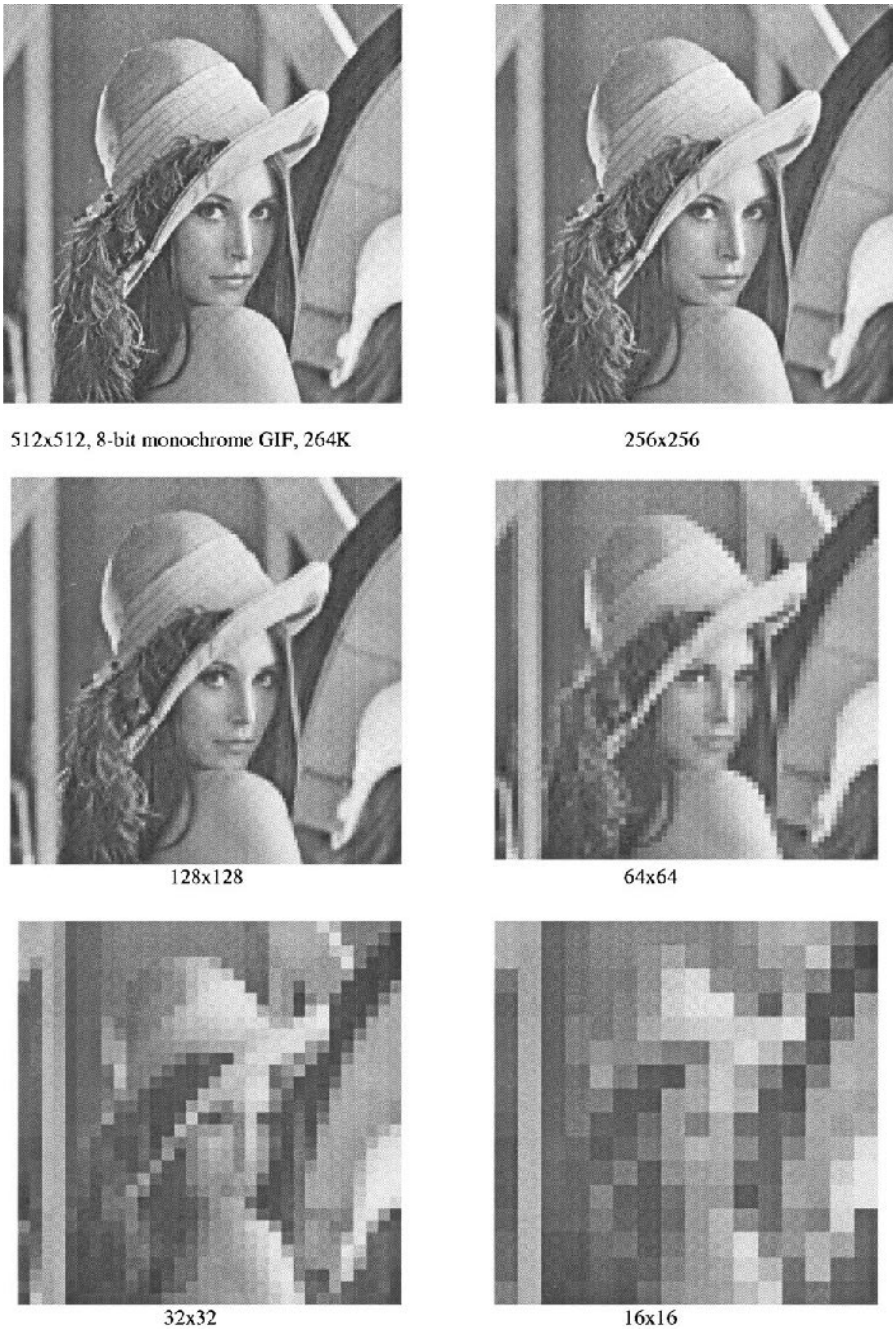
On the other hand, in a color image, the image function is a vector function with color components such as red, green, and blue defined at each point. In this case, we can assign a particular color to every gray-level value of the image. Assume that the primary colors, red, green, and blue, are each scaled between 0 and 1 and that for a given gray level a proportion of each of the primary color components can be appropriately assigned. In this case, the three primary colors comprise the axes of a unit cube where the diagonal of the cube represents the range of gray-level intensities, the origin of the cube corresponds to black with values $(0, 0, 0)$, and the opposite end of the diagonal $(1, 1, 1)$ represents white. The color cube is shown in Fig. 8.

In general, a three-dimensional color image function at a fixed time and with a given spectral illumination may be written as a three-dimensional vector in which each component is a function of space, time, and spectrum:

$$\mathit{Colorimage} = \left\{ \begin{array}{c} r(x, y, z, t, \lambda) \\ g(x, y, z, t, \lambda) \\ b(x, y, z, t, \lambda) \end{array} \right\} \tag{4}$$

Images may be formed and processed through a continuous spatial range using optical techniques. However, machine vision refers to digital or computer processing of the spatial information. Therefore, the range of the spatial coordinates will be discrete values. the necessary transformation from a continuous range to the discrete range is called *sampling*, and it is usually performed with a discrete array camera or the discrete array of photoreceptors of the human eye.

The viewing geometry, shown in Fig. 9, is also an important factor. Surface appearance can vary from diffuse to specular with quite different images. Lambert's law for diffuse surface reflection surfaces demonstrates that the angle between the light source location and the surface normal is most important in determining the amount of reflected light, as shown in Fig. 10. For specular reflection, shown in Fig. 11, both the angle between the light source and surface normal and the angle between the viewer and surface normal are important in determining the appearance of the reflected image.



512x512, 8-bit monochrome GIF, 264K

256x256

128x128

64x64

32x32

16x16

Figure 5 Image at various spatial resolutions.

2.4.2 Frequency Space Analysis

Since an image can be described as a spatial distribution of density in a space of one, two, or three dimensions, it may be transformed and represented in a different way in another space. One of the most important transformations is the Fourier transform. Historically, computer vision may be considered a new branch of signal processing, an area where Fourier analysis has had one of its most important applications. Fourier analysis gives us a useful representation of a signal because signal properties and

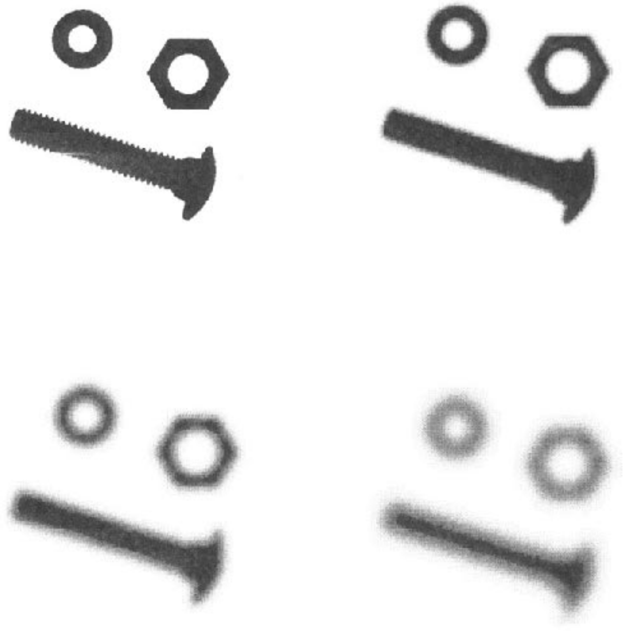


Figure 6 Images at various gray-scale quantization ranges.

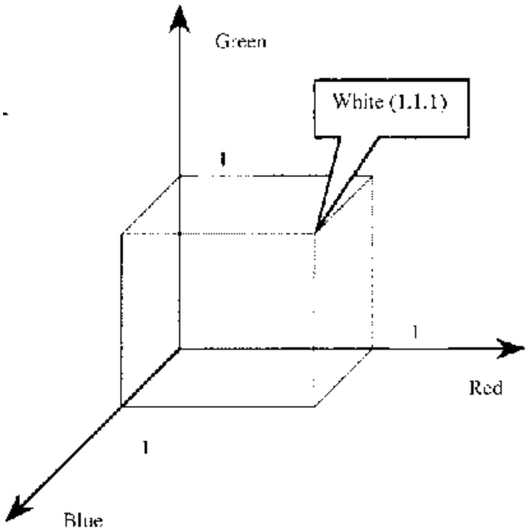


Figure 8 Color cube shows the three-dimensional nature of color.

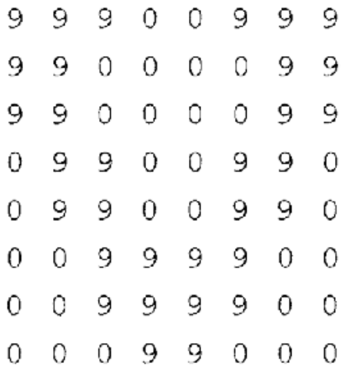
basic operations, like linear filtering and modulations, are easily described in the Fourier domain. A common example of Fourier transforms can be seen in the appearance of stars. A star lools like a small point of twinkling light. However, the small point of light we observe is actually the far-field Fraunhoffer diffraction pattern or Fourier transform of the image of the star. The twinkling is due to the motion of our eyes. The moon image looks quite different, since we are close enough to view the near-field or Fresnel diffraction pattern.

While the most common transform is the Fourier transform, there are also several closely related transforms. The *Hadamard*, *Walsh*, and discrete cosine transforms are used in the area of image compression. The *Hough* transform is used to find straight lines in a binary image. The *Hotelling* transform is commonly used to find the orientation of the maximum dimension of an object [5].

2.4.2.1 Fourier Transform

The one-dimensional Fourier transform may be written as

$$F(u) = \int_{-\infty}^{\infty} f(x)e^{-iux}\,dx \tag{5}$$



9	9	9	0	0	9	9	9
9	9	0	0	0	0	9	9
9	9	0	0	0	0	9	9
0	9	9	0	0	9	9	0
0	9	9	0	0	9	9	0
0	0	9	9	9	9	0	0
0	0	9	9	9	9	0	0
0	0	0	9	9	0	0	0

Figure 7 Digitized image.

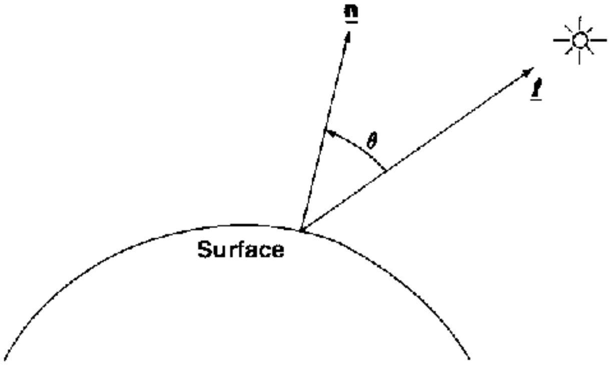


Figure 9 Image surface and viewing geometry effects.

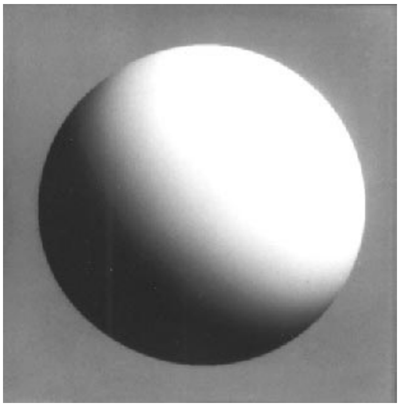


Figure 10 Diffuse surface reflection.

In the two-dimensional case, the Fourier transform and its corresponding inverse representation are:

$$F(u,v)=\int\int_{-\infty}^{\infty} f(x,y)e^{-i2\pi(ux+vy)}\,dx\,dy$$
$$f(x,y)=\int\int_{-\infty}^{\infty} F(u,v)e^{i2\pi(ux+vy)}\,du\,dv \qquad (6)$$

The discrete two-dimensional Fourier transform and corresponding inverse relationship may be written as

$$F(u,v)=\frac{1}{N^2}\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} f(x,y)e^{-i2\pi(ux+vy)/N}$$
$$f(x,y)=\sum_{u=0}^{N-1}\sum_{v=0}^{N-1} F(u,v)e^{i2\pi(ux+vy)/N} \qquad (7)$$

for $x=0,1,\ldots,N-1$; $y=0,1,\ldots,N-1$ and $u=0,1,\ldots,N-1$; $v=0,1,\ldots,N-1$.

2.4.2.2 Convolution Algorithm

The convolution theorem, that the input and output of a linear, position invariant system are related by a convolution, is an important principle. The basic idea of convolution is that if we have two images, for example, pictures A and B, then the convolution of A and B means repeating the whole of A at every point in B, or vice versa. An example of the convolution theorem is shown in Fig. 12. The convolution theorem enables us to do many important things. During the Apollo 13 space flight, the astronauts took a photograph of their damaged spacecraft, but it was out of focus. Image processing methods allowed such an out-of-focus picture to be put back into focus and clarified.

2.4.3 Image Enhancement

Image enhancement techniques are designed to improve the quality of an image as perceived by a human [1]. Some typical image enhancement techniques include *gray-scale conversion*, *histogram*, *color composition*, etc. The aim of image enhancement is to improve the interpretability or perception of information in images for human viewers, or to provide "better" input for other automated image processing techniques.

2.4.3.1 Histograms

The simplest types of image operations are point operations, which are performed identically on each point in an image. One of the most useful point operations is based on the histogram of an image.

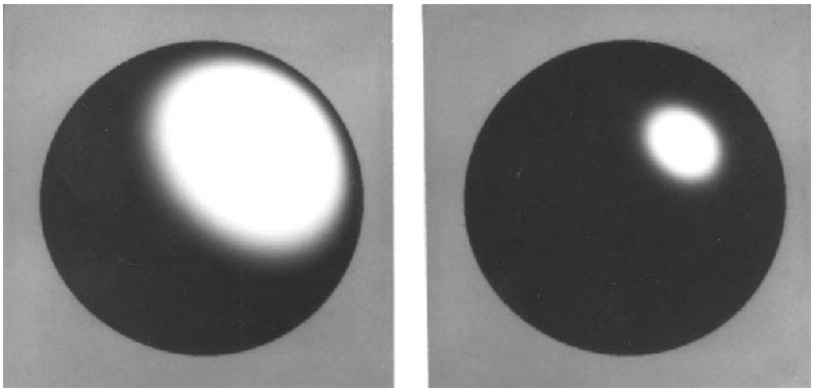
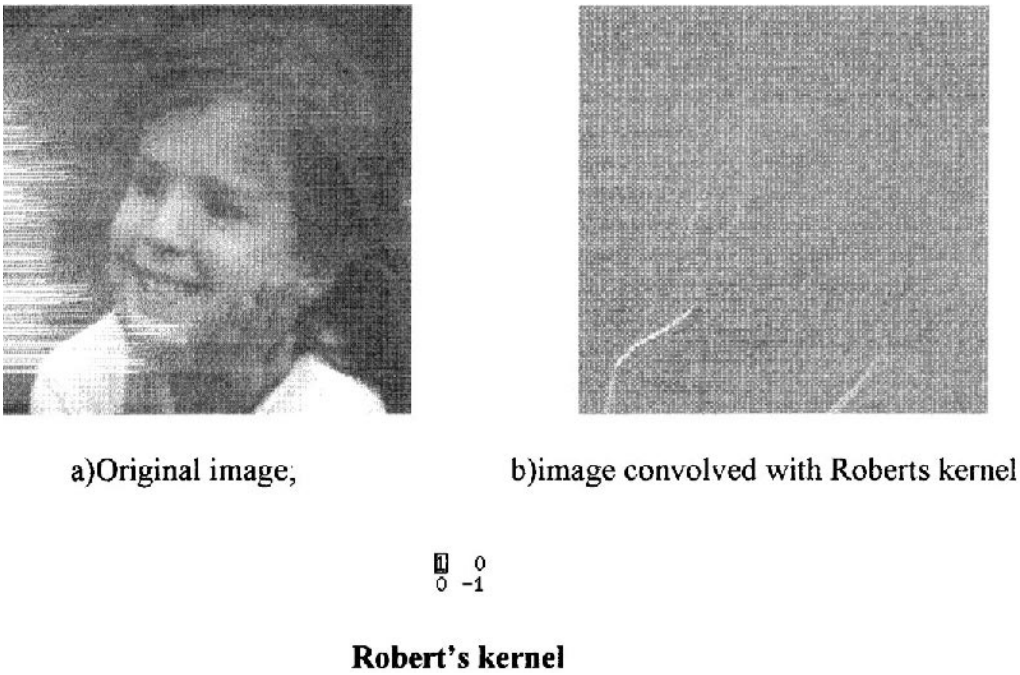


Figure 11 Specular reflection.



a)Original image;

b)image convolved with Roberts kernel

$$\begin{matrix} \boxed{1} & 0 \\ 0 & -1 \end{matrix}$$

Robert's kernel

Figure 12 An example of the convolution theorem.

Histogram Processing. A histogram of the frequency that a pixel with a particular gray level occurs within an image provides us with a useful statsitical representation of the image. Consider the image shown in Fig. 13 as an example. It represents a square on a light background. The object is represented by gray levels greater than 4. Figure 14 shows its histogram, which consists of two peaks.

In the case of complex images like satellite or medical images that may consists of up to 256 gray levels and 3000 × 3000 pixels, the resulting histograms will have many peaks. The distribution of those peaks and their magnitude can reveal significant information about the information content of the image.

Histogram Equalization. Although it is not generally the case in practice, ideally the image histogram should be distributed across the range of gray-scale values as a uniform distribution. The distribution, as shown by the example in Fig. 15, can be dominated by a few values spanning only a limited range. Statistical theory shows that using a transformation function equal to the cumulative distribution of the gray-level intensities in

2	2	4	2	4	2	5	2
4	2	2	4	2	5	2	4
2	4	7	6	6	6	2	4
2	2	6	6	6	7	5	2
4	4	7	7	7	6	2	2
2	5	6	6	6	6	5	4
4	4	5	4	4	2	2	2
4	2	5	5	4	2	4	5

Figure 13 A square on a light background.

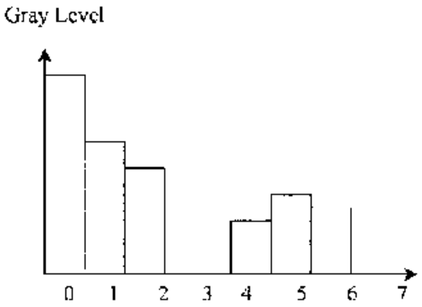


Figure 14 Histogram with bimodal distribution containing two peaks.

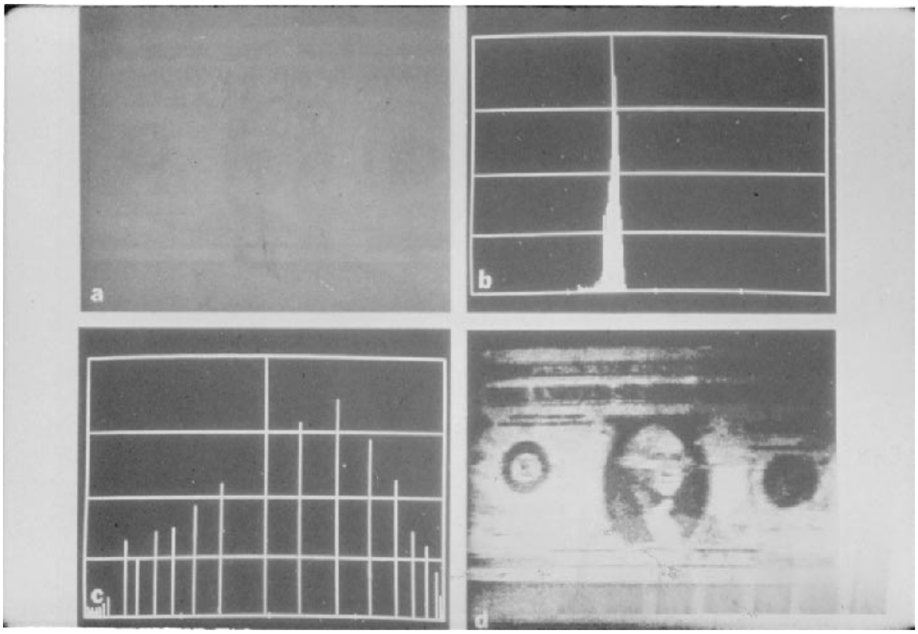


Figure 15 An example of histogram equalization. (a) Original image, (b) histogram, (c) equalized histogram, (d) enhanced image.

the image enables us to generate another image with a gray-level distribution having a uniform density.

This transformation can be implemented by a three-step process:

1. Compute the histogram of the image.
2. Compute the cumulative distribution of the gray levels.
3. Replace the original gray-level intensities using the mapping determined in 2.

After these processes, the original image, shown in Fig. 13, can be transformed, and scaled and viewed as shown in Fig. 16. The new gray-level value set S_k, which represents the cumulative sum, is

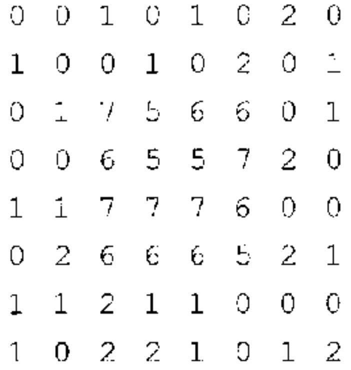
$$S_k = (1/7, 2/7, 5/7, 5/7, 5/7, 6/7, 6/7, 7/7) \quad \text{for } k = 0, 1, \ldots, 7 \tag{8}$$

Histogram Specification. Even after the equalization process, certain levels may still dominate the image so that the eye cannot interpret the contribution of the other levels. One way to solve this problem is to specify a histogram distribution that enhances selected gray levels relative to others and then reconstitutes the original image in terms of the new distribution. For example, we may decide to reduce the levels between 0 and 2, the background levels, and increase the levels between 5 and 7 correspondingly. After the similar step in histogram equalization, we can get the new gray levels set S'_k:

$$S'_k = (1/7, 5/7, 6/7, 6/7, 6/7, 6/7, 7/7, 7/7) \quad \text{for } k = 0, 1, \ldots, 7 \tag{9}$$

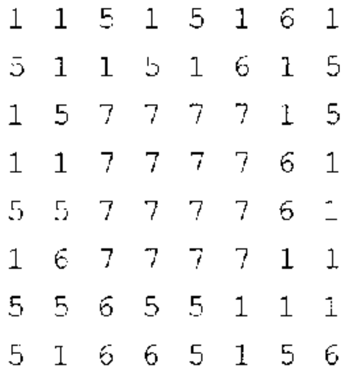
By placing these values into the image, we can get the new histogram-specified image shown in Fig. 17.

Image Thresholding. This is the process of separating an image into different regions. This may be based upon its gray-level distribution. Figure 18 shows how an image looks after thresholding. The percentage



0	0	1	0	1	0	2	0
1	0	0	1	0	2	0	1
0	1	7	5	6	6	0	1
0	0	6	5	5	7	2	0
1	1	7	7	7	6	0	0
0	2	6	6	6	5	2	1
1	1	2	1	1	0	0	0
1	0	2	2	1	0	1	2

Figure 16 Original image before histogram equalization.



1	1	5	1	5	1	6	1
5	1	1	5	1	6	1	5
1	5	7	7	7	7	1	5
1	1	7	7	7	7	6	1
5	5	7	7	7	7	6	1
1	6	7	7	7	7	1	1
5	5	6	5	5	1	1	1
5	1	6	6	5	1	5	6

Figure 17 New image after histogram equalization.

threshold is the percentage level between the maximum and minimum intensity of the initial image.

2.4.4 Image Analysis and Segmentation

An important area of electronic image processing is the segmentation of an image into various regions in order to separate objects from the background. These regions may roughly correspond to objects, parts of objects, or groups of objects in the scene represented by that image. It can also be viewed as the process of identifying edges that correspond to boundaries between objects and regions that correspond to surfaces of objects in the image. Segmentation of an image typically precedes semantic analysis of the image. Their purposes are [6]:

Data reduction: often the number of important features, i.e., regions and edges, is much smaller than the number of pixels.

Feature extraction: the features extracted by segmentation are usually "building blocks" from which object recognition begins. These features are subsequently analyzed based on their characteristics.

A region in an image can be seen as a significant change in the gray level distribution in a specified direction. As a simple example, consider the single line of gray levels below:

0 0 0 0 0 1 0 0 0 0 0

The background is represented by gray level with a zero value. Since the sixth pixel from the left has a different level that may also characterize a single point. This sixth point represents a discontinuity in that all the other levels. The process of recognizing such discontinuities may be extended to the detection of lines within an image when they occur in groups.

2.4.4.1 Edge Detection

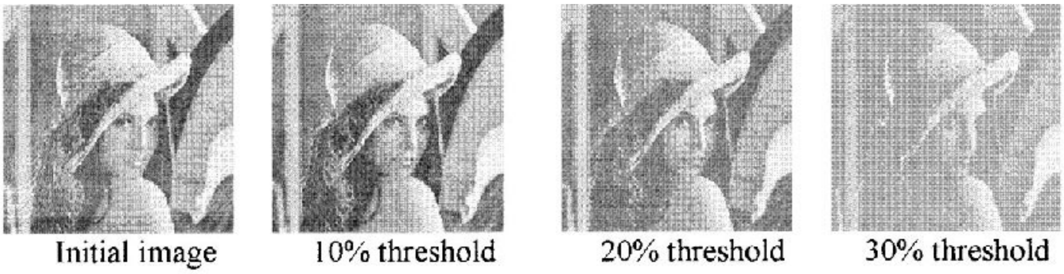
In recent years, a considerable number of edge- and line-detecting algorithms have been proposed, each being demonstrated to have particular merits for particular types of images [7]. One popular technique is called the parallel processing, template-matching method, which involves a particular set of windows being swept over the input image in an attempt to isolate specific edge features. Another widely used technique is called sequential scanning, which involves an ordered heuristic search to locate a particular feature.

Consider the example of a convolution mask or matrix, given below:

$$\begin{matrix} a1 & a2 & a3 \\ a4 & a5 & a6 \\ a7 & a8 & a9 \end{matrix} \qquad (10)$$

It consists of a 3 × 3 set of values. This matrix may be convolved with the image. That is, the matrix is first located at the top left corner of the image. If we denote the gray levels in the picture corresponding to the matrix values $a1$ to $a9$ by $v1$ to $v9$, then the product is formed:

$$T = a1 * v1 + a2 * v2 + \cdots + a9 * v9 \qquad (11)$$



Initial image 10% threshold 20% threshold 30% threshold

Figure 18 Image thresholding.

Next, we shift the window one pixel to the right and repeat the calculation. After calculating all the pixels in the line, we then reposition the matrix one pixel down and repeat this procedure. At the end of the entire process, we have a set of T values, which enable us to determine the existence of the edge. Depending on the values used in the mask template, various effects such as smoothing or edge detection will result.

Since edges correspond to areas in the image where the image varies greatly in brightness, one idea would be to differentiate the image, looking for places where the magnitude of the derivative is large. The only drawback to this approach is that *differentiation enhances noise*. Thus, it needs to be combined with *smoothing*.

Smoothing Using Gaussians. One form of smoothing the image is to convolve the image intensity with a gaussian function. Let us suppose that the image is of infinite extent and that the image intensity is $I(x, y)$. The Gaussian is a function of the form

$$G_\sigma(x, y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \tag{12}$$

The result of convolving the image with this function is equivalent to lowpass filtering the image. The higher the sigma, the greater the lowpass filter's effect. The filtered image is

$$I_\sigma(x, y) = I(x, y) * G_\sigma(x, y) \tag{13}$$

One effect of smoothing with a Gaussian function is a reduction in the amount of noise, because of the low pass characteristic of the Gaussian function. Figure 20 shows the image with noise added to the original, Fig. 19.

Figure 21 shows the image filtered by a lowpass Gaussian function with $\sigma = 3$.

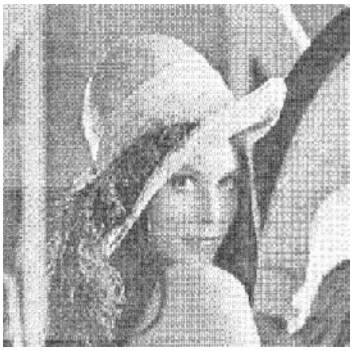


Figure 19 A digital image from a camera.

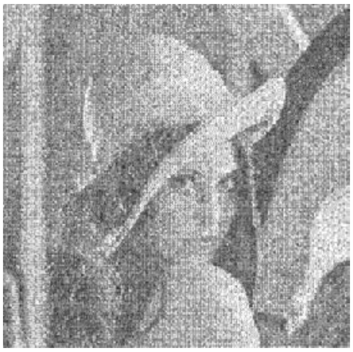


Figure 20 The original image corrupted with noise.

Vertical Edges. To detect vertical edges we first convolve with a Gaussian function and then differentiate

$$I_\sigma(x, y) = I(x, y) * G_\sigma(x, y) \tag{14}$$

the resultant image in the x-direction. This is the same as convolving the image with the derivative of the gaussian function in the x-direction that is

$$-\frac{x}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \tag{15}$$

Then, one marks the peaks in the resultant images that are above a prescribed threshold as edges (the threshold is chosen so that the effects of noise are minimized). The effect of doing this on the image of Fig. 21 is shown in Fig. 22.

Horizontal Edges. To detect horizontal edges we first convolve with a Gaussian and then differentiate the resultant image in the y-direction. But this is the same as convolving the image with the derivative of the gaussian function in the y-direction, that is

$$-\frac{y}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \tag{16}$$

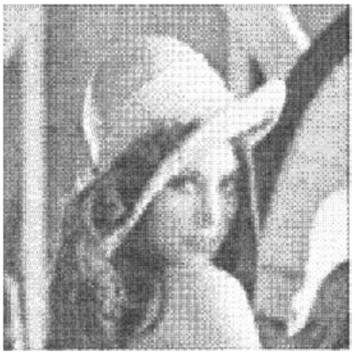


Figure 21 The noisy image filtered by a Gaussian of variance 3.

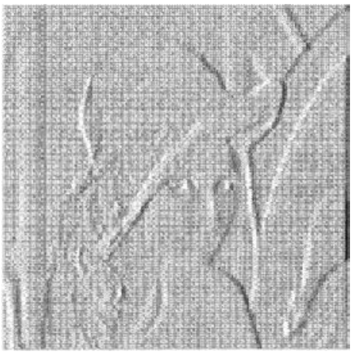


Figure 22 The vertical edges of the original image.

Then, the peaks in the resultant image that are above a prescribed threshold are marked as edges. The effect of this operation is shown in Fig. 23.

Canny Edges Detector. To detect edges at an arbitrary orientation one convolves the image with the convolution kernels of vertical edges and horizontal edges. Call the resultant images $R_1(x, y)$ and $R_2(x, y)$. Then forms the square root of the sum of the squares:

$$R = R_1^2 + R_2^2 \tag{17}$$

This edge detector is known as the *Canny edge detector*, as shown in Fig. 24, which was proposed by Canny [8]. Now set the thresholds in this image to mark the peaks as shown in Fig. 25. The result of this operation is shown in Fig. 26.

2.4.4.2 Three Dimensional—Stereo

The two-dimensional digital images can be thought of as having gray levels that are a function of two spatial variables. The most straightforward generalization to three dimensions would have us deal with images having gray levels that are a function of three spatial variables. The more common examples are the three-dimensional images of transparent microscope specimens or larger objects viewed with x-ray illumination. In these images, the gray level represents some local property, such as optical density per millimeter of path length.

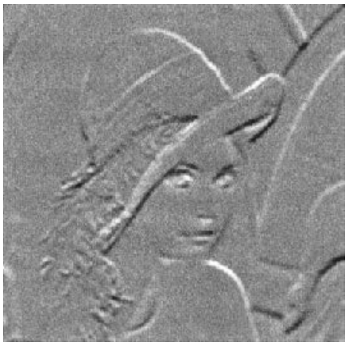


Figure 23 The horizontal edges of the original image.

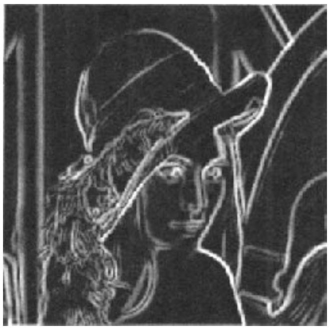


Figure 24 The magnitude of the gradient.

Most humans experience the world as three-dimensional. In fact, most of the two-dimensional images we see have been derived from this three-dimensional world by camera systems that employ a perspective projection to reduce the dimensionality from three to two [9].

Spatially Three-Dimensional Image. Consider a three-dimensional object that is not perfectly transparent, but allows light to pass through it. We can think of a local property that is distributed throughout the object in three dimensions. This property is the local optical density.

CAT Scanners. Computerized axial tomography (CAT) is an x-ray technique that produces three-dimensional images of a solid object.

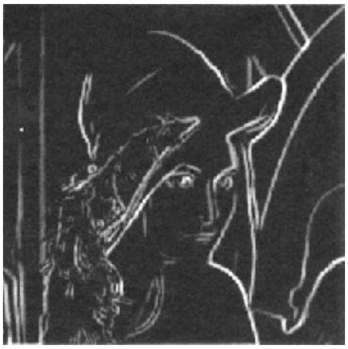


Figure 25 Threshold of the peaks of the magnitude of the gradient.

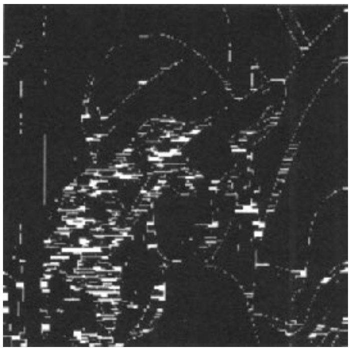


Figure 26 Edges of the original image.

Stereometry. This is the technique of deriving a range image from a stereo pair of brightness images. It has long been used as a manual technique for creating elevation maps of the earth's surface.

Stereoscopic Display. If it is possible to compute a range image from a stereo pair, then it should be possible to generate a stereo pair given a single brightness image and a range image. In fact, this technique makes it possible to generate stereoscopic displays that give the viewer a sensation of depth.

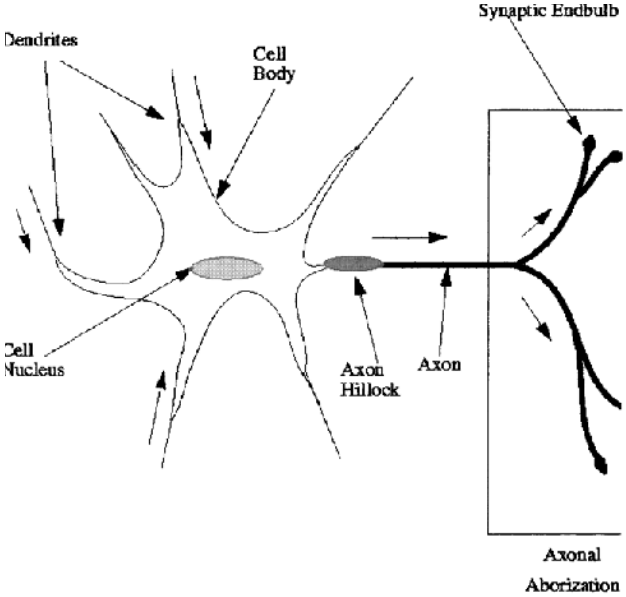
Shaded Surface Display. By modeling the imaging system, one can compute the digital image that would result if the object existed and if it were digitized by conventional means. Shaded surface display grew out of the domain of computer graphics and has developed rapidly in the past few years.

2.4.5 Image Recognition and Decisions

2.4.5.1 Neural Networks

Artificial neural networks (ANNs) can be used in image processing applications. Initially inspired by biological nervous systems, the development of artificial neural networks has more recently been motivated by their applicability to certain types of problem and their potential for parallel processing implementations.

Biological Neurons. There are about a hundred billion neurons in the brain, and they come in many different varieties, with a highly complicated internal structure. Since we are more interested in large networks of such units, we will avoid a great level of detail, focusing instead on their salient computational features. A schematic diagram of a single biological neuron is shown in Fig. 27.



Figurer 27 A schematic diagram of a single biological neuron.

The cells at the neuron connections, or synapses, receive information in the form of electrical pulses from the other neurons. The synapses connect to the cell inputs, or dendrites, and form an electrical signal output of the neuron is carried by the axon. An electrical pulse is sent down the axon, or the neuron "fires," when the total input stimuli from all of the dendrites exceeds a certain threshold. Interestingly, this local processing of interconnected neurons results in self-organized emergent behavior.

Artificial Neuron Model. The most commonly used neuron model, depicted in Fig. 28, is based on the

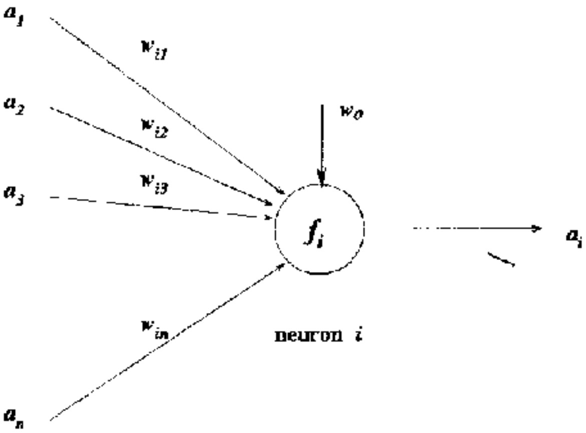


Figure 28 ANN model proposed by McCulloch and Pitts in 1943.

model proposed by McCulloch and Pitts in 1943 [11]. In this model, each neuron's input, a_1–a_n, is weighted by the values w_{i1}–w_{in}. A bias, or offset, in the node is characterized by an additional constant input w_0. The output, a_i, is obtained in terms of the equation

$$a_i = f\left(\sum_{j=1}^{N} a_j w_{ij} + w_0\right) \tag{18}$$

Feedforward and Feedback Networks. Figure 29 shows a feedforward network in which the neurons are organized into an input layer, hidden layer or layers, and an output layer. The values for the input layer are set by the environment, while the output layer values, analogous to a control signal, are returned to the environment. The hidden layers have no external connections, they only have connections with other layers in the network. In a feedforward network, a weight w_{ij} is only nonzero if neuron i is in one layer and neuron j is in the previous layer. This ensures that information flows forward through the network, from the input layer to the hidden layer(s) to the output layer. More complicated forms for neural networks exist and can be found in standard textbooks. Training a neural network involves determining the weights w_{ij} such that an input layer presented with information results in the output layer having a correct response. This training is the fundamental concern when attempting to construct a useful network.

Feedback networks are more general than feedforward networks and may exhibit different kinds of behavior. A feedforward network will normally settle into a state that is dependent on its input state, but a feedback network may proceed through a sequence of states, even though there is no change in the external inputs to the network.

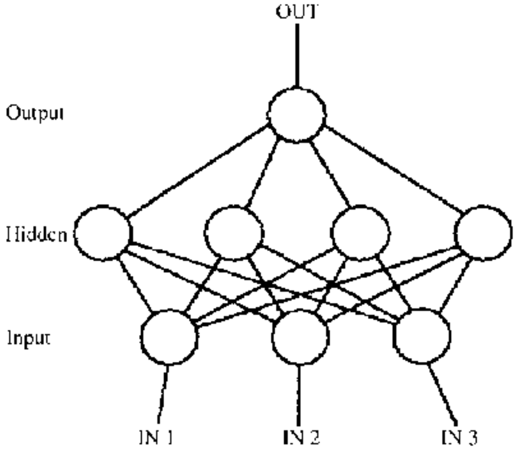


Figure 29 A feedforward neural network.

2.4.5.2 Supervised Learning and Unsupervised Learning

Image recognition and decision making is a process of discovering, identifying, and understanding patterns that are relevant to the performance of an image-based task. One of the principal goals of image recognition by computer is to endow a machine with the capability to approximate, in some sense, a similar capability in human beings. For example, in a system that automatically reads images of typed documents, the patterns of interest are alphanumeric characters, and the goal is to achieve character recognition accuracy that is as close as possible to the superb capability exhibited by human beings for performing such tasks.

Image recognition systems can be designed and implemented for limited operational environments. Research in biological and computational systems is continually discovering new and promising theories to explain human visual cognition. However, we do not yet know how to endow these theories and applications with a level of performance that even comes close to emulating human capabilities in performing general image decision functionality. For example, some machines are capable of reading printed, properly formatted documents at speeds that are orders of magnitude faster than the speed that the most skilled human reader could achieve. However, systems of this type are highly specialized and thus have little extendibility. That means that current theoretical and implementation limitations in the field of image analysis and decision making imply solutions that are highly problem dependent.

Different formulations of learning from an environment provide different amounts and forms of information about the individual and the goal of learning. We will discuss two different classes of such formulations of learning.

Supervised Learning. For supervised learning, a "training set" of inputs and outputs is provided. The weights must then be determined to provide the correct output for each input. During the training process, the weights are adjusted to minimize the difference between the desired and actual outputs for each input pattern.

If the association is completely predefined, it is easy to define an error metric, for example mean-squared error, of the associated response. This is turn gives us the possibility of comparing the performance with the predefined responses (the "supervision"), changing the learning system in the direction in which the error diminishes.

Unsupervised Learning. The network is able to discover statistical regularities in its input space and can automatically develop different modes of behavior to represent different classes of inputs. In practical applications, some "labeling" is required after training, since it is not known at the outset which mode of behavior will be associated with a given input class. Since the system is given no information about the goal of learning, all that is learned is a consequence of the learning rule selected, together with the individual training data. This type of learning is frequently referred to as self-organization.

A particular class of unsupervised learning rule which has been extremely influential is Hebbian learning [12]. The Hebb rule acts to strengthen often-used pathways in a network, and was used by Hebb to account for some of the phenomena of classical conditioning.

Primarily some type of regularity of data can be learned by this learning system. The associations found by unsupervised learning define representations optimized for their information content. Since one of the problems of intelligent information processing deals with selecting and compressing information, the role of unsupervised learning principles is crucial for the efficiency of such intelligent systems.

2.4.6 Image Processing Applications

Artificial neural networks can be used in image processing applications. Many of the techniques used are variants of other commonly used methods of pattern recognition. However, other approaches of image processing may require modeling of the objects to be found within an image, while neural network models often work by a training process. Such models also need attention devices, or invariant properties, as it is usually infeasible to train a network to recognize instances of a particular object class in all orientations, sizes, and locations within an image.

One method commonly used is to train a network using a relatively small window for the recognition of objects to be classified, then to pass the window over the image data in order to locate the sought object, which can then be classified once located. In some engineering applications this process can be performed by image preprocessing operations, since it is possible to capture the image of objects in a restricted range of orientations with predetermined locations and appropriate magnification.

Before the recognition stage, the system has to be determined such as which image transform is to be used. These transformations include Fourier transforms, or using polar coordinates or other specialized coding schemes, such as the chain code. One interesting neural network model is the neocognition model of Fukushima and Miyake [13], which is capable of recognizing characters in arbitrary locations, sizes and orientations, by the use of a multilayered network.

For machine vision, the particular operations include setting the quantization levels for the image, normalizing the image size, rotating the image into a standard orientation, filtering out background detail, contrast enhancement, and edge direction. Standard techniques are available for these and it may be more effective to use these before presenting the transformed data to a neural network.

2.4.6.1 Steps in Setting Up an Application

The main steps are shown below.

Physical setup: light source, camera placement, focus, field of view
Software setup: window placement, threshold, image map
Feature extraction: region shape features, gray-scale values, edge detection
Decision processing: decision function, training, testing.

2.4.7 Future Development of Machine Vision

Although image processing has been successfully applied to many industrial applications, there are still many definitive differences and gaps between machine vision and human vision. Past successful applications have not always been attained easily. Many difficult problems have been solved one by one, sometimes by simplifying the background and redesigning the objects. Machine vision requirements are sure to increase in the future, as the ultimate goal of machine vision research is obviously to approach the capability of the human eye. Although it seems extremely difficult to attain, it remains a challenge to achieve highly functional vision systems.

The narrow dynamic range of detectable brightness causes a number of difficulties in image processing. A novel sensor with a wide detection range will drastically change the impact of image processing. As microelectronics technology progreses, three-dimensional compound sensor, large scale integrated circuits (LSI) are also anticipated, to which at least preprocessing capability should be provided.

As to image processors themselves, the local parallel pipelined processor may be further improved to proved higher processing speeds. At the same time, the multiprocessor image processor may be applied in industry when the key-processing element becomes more widely available. The image processor will become smaller and faster, and will have new functions, in response to the advancement of semiconductor technology, such as progress in system-on-chip configurations and wafer-scale integration. It may also be possible to realize one-chip intelligent processors for high-level processing, and to combine these with one-chip rather low-level image processors to achieve intelligent processing, such as knowledge-based or model-based processing. Based on these new developments, image processing and the resulting machine vision improvements are expected to generate new values not merely for industry but for all aspects of human life.

2.5 MACHINE VISION APPLICATIONS

Machine vision applications are numerous as shown in the following list.

- Inspection:
  - Hole location and verification
  - Dimensional measurements
  - Part thickness
  - Component measurements
  - Defect location
  - Surface contour accuracy
- Part identification and sorting:
  - Sorting
  - Shape recognition
  - Inventory monitoring
  - Conveyor picking—nonoverlapping parts
  - Conveyor picking—overlapping parts
  - Bin picking
- Industrial robot control:
  - Tracking
  - Seam welding guidance
  - Part positioning and location determination
  - Collision avoidance
  - Machining monitoring
- Mobile robot applications:
  - Navigation
  - Guidance
  - Tracking
  - Hazard determination
  - Obstacle avoidance.

2.5.1 Overview

High-speed production lines, such as stamping lines, use machine vision to meet online, real time inspection needs. Quality inspection involves deciding whether parts are acceptable or defective, then directing motion control equipment to reject or accept them. Machine guidance applications improve the accuracy and speed of robots and automated material handling equipment. Advanced systems enable a robot to locate a part or an assembly regardless of rotation or size. In gaging applications, a vision system works quickly to measure a variety of critical dimensions. The reliability and accuracy achieved with these methods surpasses anything possible with manual methods.

In the machine tool industry, applications for machine vision include sensing tool offset and breakage, verifying part placement and fixturing, and monitoring surface finish. A high-speed processor that once cost $80,000 now uses digital signal processing chip technology and costs less than $10,000. The rapid growth of machine vision usage in electronics, assembly systems, and continuous process monitoring created an experience base and tools not available even a few years ago.

2.5.2 Inspection

The ability of an automated vision system to recognize well-defined patterns and determine if these patterns match those stored in the system's CPU memory makes it ideal for the inspection of parts, assemblies, containers, and labels. Two types of inspection can be performed by vision systems: quantitative and qualitative. Quantitative inspection is the verification that measurable quantities fall within desired ranges of tolerance, such as dimensional measurements and the number of holes. Qualitative inspection is the verification that certain components or properties are present and in a certain position, such as defects, missing parts, extraneous components, or misaligned parts.

Many inspection tasks involve comparing the given object with a reference standard and verifying that there are no discrepancies. One method of inspection is called template matching. An image of the object is compared with a reference image, pixel by pixel. A discrepancy will generate a region of high differences. On the other hand, if the observed image and the reference are slightly out of registration, differences will be found along the borders between light and dark regions in the image. This is because a slight misalignment can lead to dark pixels being compared with light pixels.

A more flexible approach involves measuring a set of the image's properties and comparing the measured values with the corresponding expected values. An example of this approach is the use of width measurements to detect flaws in printed circuits. Here the expected width values were relatively high; narrow ones indicated possible defects.

2.5.2.1 Edge-Based Systems

Machine vision systems, which operate on edge descriptions of objects, have been developed for a number of defense applications. Commercial edge-based systems with pattern recognition capabilities have reached markets now. The goal of edge detection is to find the boundaries of objects by marking points of rapid change in intensity. Sometimes, the systems operate on edge descriptions of images as "gray-level" image systems. These systems are not sensitive to the individual intensities of patterns, only to changes in pixel intensity.

2.5.2.2 Component or Attribute Measurements

An attribute measurement system calculates specific qualities associated with known object images. Attributes can be geometrical patterns, area, length of perimeter, or length of straight lines. Such systems analyze a given scene for known images with predefined attributes. Attributes are constructed from previously scanned objects and can be rotated to match an object at any given orientation. This technique can be applied with minimal preparation. However, orienting and matching are used most efficiently in aplications permitting standardized orientations, since they consume significant processing time. Attribute measurement is effective in the segregating or sorting of parts, counting parts, flaw detection, and recognition decisions.

2.5.2.3 Hole Location

Machine vision is ideally suited for determining if a well-defined object is in the correct location relative to some other well-defined object. Machined objects typically consist of a variety of holes that are drilled, punched, or cut at specified locations on the part. Holes may be in the shape of circular openings, slits, squares, or shapes that are more complex. Machine vision systems can verify that the correct holes are in the correct locations, and they can perform this operation at high speeds. A window is formed around the hole to be inspected. If the hole is not too close to another hole or to the edge of the workpiece, only the image of the hole will appear in the window and the measurement process will simply consist of counting pixels. Hole inspection is a straightforward application for machine vision. It requires a two-dimensional binary image and the ability to locate edges, create image segments, and analyze basic features. For groups of closely located holes, it may also require the ability to analyze the general organization of the image and the position of the holes relative to each other.

2.5.2.4 Dimensional Measurements

A wide range of industries and potential applications require that specific dimensional accuracy for the finished products be maintained within the tolerance limits. Machine vision systems are ideal for performing 100% accurate inspections of items which are moving at high speeds or which have features which are difficult to measure by humans. Dimensions are typically inspected using image windowing to reduce the data processing requirements. A simple linear length measurement might be performed by positioning a long width window along the edge. The length of the edge could then be determined by counting the number of pixels in the window and translating into inches or millimeters. The output of this dimensional measurement process is a "pass–fail" signal received by a human operator or by a robot. In the case of a continuous process, a signal that the critical dimension being monitored is outside the tolerance limits may cause the operation to stop, or it may cause the forming machine to automatically alter the process.

2.5.2.5 Defect Location

In spite of the component being present and in the correct position, it may still be unacceptable because of some defect in its construction. The two types of possible defects are functional and cosmetic.

A functional defect is a physical error, such as a broken part, which can prevent the finished product from performing as intended. A costmetic defect is a flaw in the appearance of an object, which will not interfere with the product's performance, but may decrease the product's value when perceived by the user. Gray-scale systems are ideal for detecting subtle differences in contrast between various regions on the surface of the parts, which may indicate the presence of defects. Some examples of defect inspection include the inspection of:

Label position on bottles
Deformations on metal cans
Deterioration of dies
Glass tubing for bubbles
Cap seals for bottles
Keyboard character deformations.

2.5.2.6 Surface Contour Accuracy

The determination of whether a three-dimensional curved surface has the correct shape or not is an important area of surface inspection. Complex manufactured parts such as engine block castings or aircraft frames have very irregular three-dimensional shapes. However, these complex shapes must meet a large number of dimensional tolerance specifications. Manual inspection of these shapes may require several hours for each item. A vision system may be used for mapping the surface of these three-dimensional objects.

2.5.3 Part Identification and Sorting

The recognition of an object from its image is the most fundamental use of a machine vision system. Inspection deals with the examination of objects without necessarily requiring that the objects be identified. In part recognition however, it is necessary to make a positive identification of an object and then make the decision from that knowledge. This is used for categorization of the objects into one of several groups. The process of part identification generally requires strong geometrical feature interpretation capabilities. The applications considered often require an interface capability with some sort of part-handling equipment. An industrial robot provides this capability.

There are manufacturing situations that require that a group of varying parts be categorized into common groups and sorted. In general, parts can be sorted based on several characteristics, such as shape, size, labeling, surface markings, color, and other criteria, depending on the nature of the application and the capabilities of the vision system.

2.5.3.1 Character Recognition

Usually in manufacturing situations, an item can be identified solely based on the identification of an alphanumeric character or a set of characters. Serial numbers on labels identify separate batches in which products are manufactured. Alphanumeric characters may be printed, etched, embossed, or inscribed on consumer and industrial products. Recent developments have provided certain vision systems with the capability of reading these characters.

2.5.3.2 Inventory Monitoring

Categories of inventories, which can be monitored for control purposes, need to be created. The sorting process of parts or finished products is then based on these categories. Vision system part identification capabilities make them compatible with inventory control systems for keeping track of raw material, work in process, and finished goods inventories. Vision system interfacing capability allows them to command industrial robots to place sorted parts in inventory storage areas. Inventory level data can then be transmitted to a host computer for use in making inventory-level decisions.

2.5.3.3 Conveyor Picking—Overlap

One problem encountered during conveyor picking is overlapping parts. This problem is complicated by the fact that certain image features, such as area, lose meaning when the images are joined together. In cases of a machined part with an irregular shape, analysis of the overlap may require more sophisticated discrimination capabilities, such as the ability to evaluate surface characteristics or to read surface markings.

2.5.3.4 No Overlap

In manufacturing environments with high-volume mass production, workpieces are typically positioned and oriented in a highly precise manner. Flexible automation, such as robotics, is designed for use in the relatively unstructured environments of most factories. However, flexible automation is limited without the addition of the feedback capability that allows it to locate parts. Machine vision systems have begun to provide the capability. The presentation of parts in a random manner, as on a conveyor belt, is common in flexible automation in batch production. A batch of the same type of parts will be presented to the robot in a random distribution along the conveyor belt. The robot must first determine the location of the part and then the orientation so that the gripper can be properly aligned to grip the part.

2.5.3.5 Bin Picking

The most common form of part representation is a bin of parts that have no order. While a conveyor belt insures a rough form of organization in a two-dimensional plane, a bin is a three-dimensional assortment of parts oriented randomly through space. This is one of the most difficult tasks for a robot to perform. Machine vision is the most likely tool that will enable robots to perform this important task. Machine vision can be used to locate a part, identify orientation, and direct a robot to grasp the part.

2.5.4 Industrial Robot Control

2.5.4.1 Tracking

In some applications like machining, welding, assembly, or other process-oriented applications, there is a need for the parts to be continuously monitored and positioned relative to other parts with a high degree of precision. A vision system can be a powerful tool for controlling production operations. The ability to measure the geometrical shape and the orientation of the object coupled with the ability to measure distance is important. A high degree of image resolution is also needed.

2.5.4.2 Seam Welding Guidance

Vision systems used for this application need more features than the systems used to perform continuous welding operations. They must have the capability to maintain the weld torch, electrode, and arc in the proper positions relative to the weld joint. They must also be capable of detecting weld joints details, such as widths, angles, depths, mismatches, root openings, tack welds, and locations of previous weld passes. The capacity to perform under conditions of smoke, heat, dirt, and operator mistreatment is also necessary.

2.5.4.3 Part Positioning and Location Determination

Machine vision systems have the ability to direct a part to a precise position so that a particular machining operation may be performed on it. As in guidance and control applications, the physical positioning is performed by a flexible automation device, such as a robot. The vision system insures that the object is correctly aligned. This facilitates the elimination of expensive fixturing. The main concern here would be how to achieve a high level of image resolution so that the position can be measured accurately. In cases in which one part would have to touch another part, a touch sensor might also be needed.

2.5.4.4 Collision Avoidance

Occasionally, there is a case in industry, where robots are being used with flexible manufacturing equipment, when the manipulator arm can come in contact with another piece of equipment, a worker, or other obstacles, and cause an accident. Vision systems may be effectively used to prevent this. This application would need the capability of sensing and measuring relative motion as well as spatial relationships among objects. A real-time processing capability would be required in order to make rapid decisions and prevent contact before any damage would be done.

2.5.4.5 Machining Monitoring

The popular machining operations like drilling, cutting, deburring, gluing, and others, which can be programmed offline, have employed robots successfully. Machine vision can greatly expand these capabilities in applications requiring visual feedback. The advantage of using a vision system with a robot is that the vision system can guide the robot to a more accurate position by compensating for errors in the robot's positioning accuracy. Human errors, such as incorrect positioning and undetected defects, can be overcme by using a vision system.

2.5.5 Mobile Robot Applications

This is an active research topic in the following areas.

Navigation
Guidance
Tracking
Hazard determination
Obstacle avoidance.

2.6 CONCLUSIONS AND RECOMMENDATIONS

Machine vision, even in its short history, has been applied to practically every type of imagery with various degrees of success. Machine vision is a multidisciplinary field. It covers diverse aspects of optics, mechanics, electronics, mathematics, photography, and computer technology. This chapter attempts to collect the fundamental concepts of machine vision for a relatively easy introduction to this field.

The declining cost of both processing devices and required computer equipment make it likely to have a continued growth for the field. Several new technological trends promise to stimulate further growth of computer vision systems. Among these are:

Parallel processing, made practical by low-cost microprocessors
Inexpensive charge-coupled devices (CCDs) for digitizing
New memory technologies for large, low-cost image storage arrays
Inexpensive, high-resolution color display systems.

Machine vision systems can be applied to many manufacturing operations where human vision is traditional. These systems are best for applications in which their speed and accuracy over long time periods enable them to outperform humans. Some manufacturing operations depend on human vision as part of the manufacturing process. Machine vision can accomplish tasks that humans cannot perform due to hazardous conditions and carry out these tasks at a higher confidence level than humans. Beyond inspecting products, the human eye is also valued for its ability to make measurement judgments or to perform calibration. This will be one of the most fruitful areas for using machine vision to replace labor. The benefits involved include:

Better-quality products
Labor replacement
Warranty reduction
Rework reduction
Higher machine productivity.

REFERENCES

1. EL Hall. Computer Image Processing and Recognition. Academic Press, Orlando, FL, 1979.
2. CF Hall, EL Hall. A nonlinear model for the spatial characteristics of the human visual system. IEEE Trans Syst Man Cybern SMC-7(3): 161–170, 1978.
3. JD Murray, W Van Ryper. Encyclopedia of Graphic File Formats. Sebastopol, CA: O'Reilly and Associates, 1994.
4. G Wagner. Now that they're cheap, we have to make them smart. Proceedings of the SME Applied Machine Vision' 96 Conference, Cincinnati, OH, June 3–6, 1996, pp 463–485.

5 RC Gonzalez and RE Woods, Digital Image Processing, Addison-Wesley, Reading, MA, 1992, pp. 81–157.

6. S Shager, T Kanade. Recursive region segmentation by analysis of histograms. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1982, pp 1166–1171.
7. MD Levine. Vision in Man and Machine. McGraw-Hill, New York, 1985, pp 151–170.
8. RM Haralick, LG Shapiro. Computer and Robot Vision. Addison-Wesley, Reading, MA, 1992, pp 509–553.
9. EL Hall. Fundamental principles of robot vision. In: Handbook of Pattern Recognition and Image Processing. Computer Vision, Academic Press, Orlando, FL, 1994, pp 542–575.
10. R Schalkoff, Pattern Recognition, John Wiley, NY, 1992, pp 204–263.
11. WS McCulloch and WH Pitts, "A Logical Calculus of the Ideas Imminent in Nervous Behavior," Bulletin of Mathematical Biophysics, Vol. 5, 1943, pp. 115–133.
12. D Hebb. Organization of Behavior, John Wiley & Sons, NY, 1949.
13. K Fukushima and S Miyake, "Neocognition: A New Algorithm for Pattern Recognition Tolerant of Deformations and Shifts in Position," Pattern Recognition, Vol. 15, No. 6, 1982, pp. 455–469.
14. M Sonka, V Klavec, and R Boyle, Image Processing, Analysis and Machine Vision, PWS, Pacific Grove, CA, 1999, pp. 722–754.

Chapter 5.3

Three-Dimensional Vision

Joseph H. Nurre
Ohio University, Athens, Ohio

3.1 INTRODUCTION

Three-dimensional vision concerns itself with a system that captures three-dimensional displacement information from the surface of an object. Let us start by reviewing dimensions and displacements. A displacement between two points is a one-dimensional measurement. One point serves as the origin and the second point is located by a displacement value. Displacements are described by a multiplicity of standard length units. For example, a displacement can be 3 in. Standard length units can also be used to create a co-ordinate axis. For example, if the first point is the origin, the second point may fall on the co-ordinate 3 which represents 3 in.

Determining the displacements among three points requries a minimum of two co-ordinate axes, assuming the points do not fall on a straight line. With one point as the origin, measurements are taken in perpendicular (orthogonal) directions, once again using a standard displacement unit.

Three-dimensional vision determines displacements along three co-ordinate axes. Three dimensions are required when the relationship among four points is desired that do not fall on the same plane. Three-dimensional sensing systems are usually used to acquire more than just four points. Hundreds or thousands of points are obtained from which critical spatial relationships can be derived. Of course, simple one-dimensional measurements can still be made point to point fronm the captured data.

The three-dimensional vision systems discussed in this chapter can also be referred to as triangulation systems. These systems typically consist of two cameras, or a camera and projector. The systems use geometrical relationships to calculate the location of a large number of points, simultaneously. Three-dimensional vision systems are computationally intensive. Advances in computer processing and storage technologies have made these systems economical.

3.1.1 Competing Technologies

Before proceeding, let us review other three-dimensional capture technologies that are available. Acquisition of three-dimensional data can be broadly categorized into contact and noncontact methods. Contact methods require the sensing system to make physical contact with the object. Noncontact methods probe the surface unobtrusively.

Scales and calipers are traditional contact measurement devices that require a human operator. When the operator is a computer, the measuring device would be a co-ordinate measuring machine (CMM). A CMM is a rectangular robot that uses a probe to acquire three-dimensional positional data. The probe senses contact with a surface using a force transducer. The CMM records the three-dimensional position of the sensor as it touches the surface point.

Several noncontact methods exist for capturing three-dimensional data. Each has its advantages and disadvantages. One method, known as time of flight, bounces a laser, sound wave, or radio wave off the surface of interest. By measuring the time it takes for the signal to return, one can calculate a position. Acoustical time-of-flight systems are better known as sonar, and can span enormous distances underwater. Laser time-of-flight systems, on the other hand, are used in industrial settings but also have inherently large work volumes. Long standoffs from the system to the measured surface are required.

Another noncontact technique for acquiring three-dimensional data is image depth of focus. A camera can be fitted with a lens that has a very narrow, but adjustable depth of field. A computer controls the depth of field and identifies locations in an image that are in focus. A group of points are acquired at a specific distance, then the lens is refocused to acquire data at a new depth.

Other three-dimensional techniques are tailored to specific applications. Interferometry techniques can be used to determine surface smoothness. It is frequently used in ultrahigh precision applications that require accuracies up to the wavelength of light. Specialized medical imaging systems such as magnetic resonance imaging (MRI) or ultrasound also acquire three-dimensional data by penetrating the subject of interest. The word "vision" usually refers to an outer shell measurement, putting these medical systems outside the scope of this chapter.

The competing technology to three-dimensional triangulation vision, as described in this chapter, are CMM machines, time-of-flight devices, and depth of field. Table 1 shows a brief comparison among different systems representing each of these technologies. The working volume of a CMM can be scaled up without loss of accuracy. Triangulation systems and depth-of-field systems lose accuracy with large work volumes. Hence, both systems are sometimes moved as a unit to increase work volume. Figure 1 shows a triangulation system, known as a laser scanner. Laser scanners can have accuracies of a thousandth of an inch but the small work volume requires a mechanical actuator. Triangulation systems acquire an exceptionally large number of points simultaneously. A CMM must repeatedly make physical contact with the object to acquire points and therefore is much slower.

3.1.2 Note on Two-Dimensional Vision Systems

Vision systems that operate with a single camera are two-dimensional vision systems. Three-dimensional information may sometimes be inferred from such a vision system. As an example, a camera acquires two-dimensional information about a circuit board. An operator may wish to inspect the solder joints on the circuit board, a three-dimensional problem. For such a task, lighting can be positioned such that shadows of solder joints will be seen by the vision system. This method of inspecting does not require the direct measurement of three-dimensional co-ordinate locations on the surface of the board. Instead the three-dimensional information is inferred by a clever setup. Discussion of two-dimensional image processing for inspection of three dimensions by inference can be found in Chap. 5.2. This chapter will concern itself with vision systems that capture three-dimensional position locations.

Table 1 Comparison of Three-Dimensional Technologies

System	Work volume (in.)	Depth resolution (in.)	Speed (points/sec)
Triangulation (DCS Corp.) 3D Areal Mapper^{TM}	12 × 12 × 12	0.027	~ 100,000
CMM (Brown & Sharp Mfg. Co.) MicroVal PFX^{TM}	14 × 16 × 12	0.0002	< 1
Laser time of flight (Perceptron, Inc.) LASAR^{TM}	9.8 ft × 9.8 ft × 6.5 ft	< 0.08	~ 200
Depth of focus (View Engineering, Inc.) Precis Series^{TM}	30 × 30 × 6	0.002	~ 700 after Z-travel of 4 in/sec

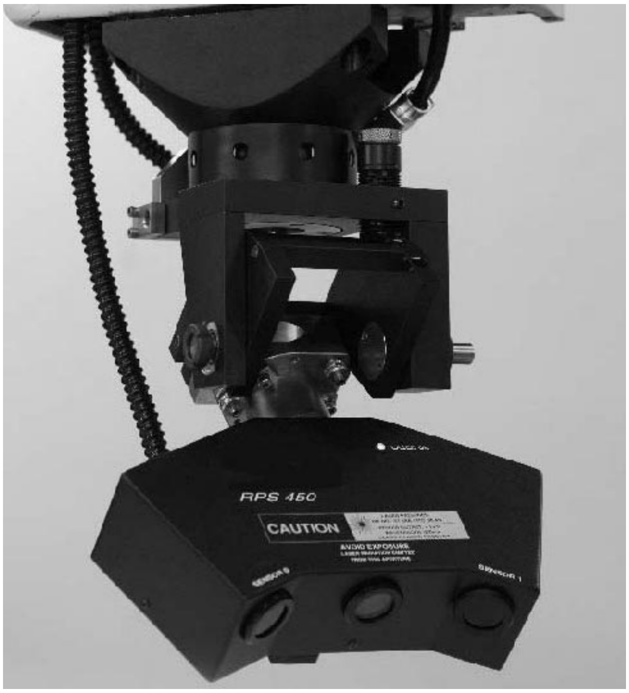


Figure 1 A laser scanner may be attached to a mechanical actuator to increase its work volume. (Photograph courtesy of Laser Design Inc., Minneapolis, MN.)

3.1.3 Chapter Outline

To reiterate, the purpose of this chapter is to discuss the acquisition and processing of three-dimensional surface data using three-dimensional vision or triangulation systems. Section 2 will discuss the design and availability of acquisition systems. Section 3 discussed data processing issues unique to three-dimensional data. Section 4 is a conclusion.

3.2 DATA CAPTURE

3.2.1 Triangulation Geometry

Three-dimensional vision systems use triangulation to obtain three-dimensional measurement data. Triangulation is simply the intersection on a surface of two lines of sight. Figure 2 shows a standard triangulation configuration. A triangle is created by the lines of sight from the two imaging devices and a baseline drawn between them. The location and viewing direction of each device must be well known. The equations for the lines of sight can then be determined, with the intersection giving a three-dimensional point in space.

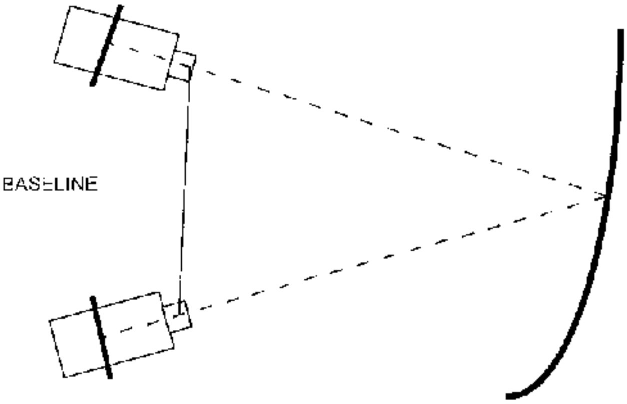


Figure 2 Stereo triangulation vision system. Points on the surface are found by the intersection of two lines of sight.

At least one line of sight is always passive (viewing) while the second line of sight may be passive or active (projecting). The lines of sight are usually calculated from the pinhole camera model described below.

3.2.2 Pinhole Camera Model

The pinhole camera model is widely accepted as a reasonable approximation of a true camera system. In a pinhole camera, light rays from a scene pass through a lens center or focal point onto the image plane as shown in Fig. 3. The distance between the focal point and image plane is the focal length of the device.

A pinhole camera is modeled mathematically by a perspective projection matrix. Consider the geometrical arrangement of the pinhole camera shown in Fig. 4. The camera focal point is assumed to be located at the origin. The image plane is parallel to the x–y-plane and is displaced from the origin by f, the focal length. In a physical camera, the image plane lies on the negative z-axis and hence the image is flipped. An image plane on the positive axis is used mathematically, however, in order to analyze the image in an upright orientation.

A line can be drawn through the focal point, a pixel point, and the imaged point in space. With the focal point at the origin, the b variable in the traditional line equation, $y = mx + b$, will equal zero. In three dimensions, the line of sight can be described as the intersection of the planes:

$$x = m_x z \tag{1}$$

$$y = m_y z \tag{2}$$

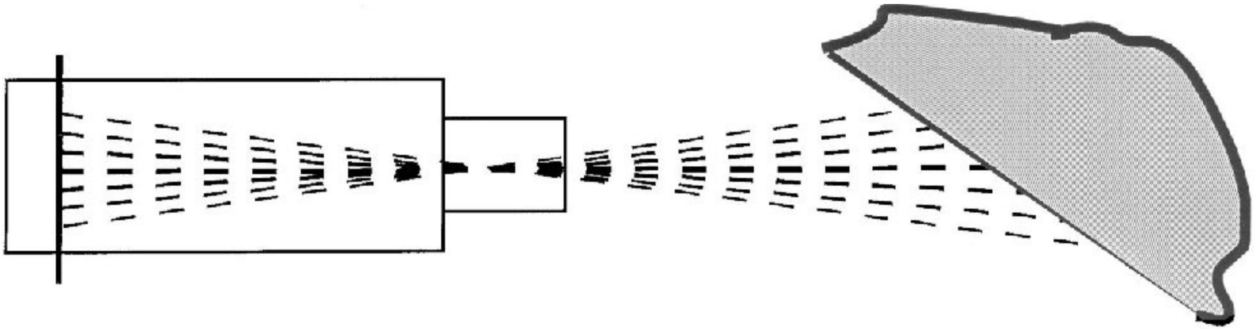


Figure 3 The pinhole camera is a widely used approximate for a camera or projector.

where the slope of the line is the pixel position divided by the focal length:

$$m_x = x_{\text{pixel}}/f \tag{3}$$

$$m_y = y_{\text{pixel}}/f \tag{4}$$

Hence

$$x_{\text{pixel}}(z/f) = x \tag{5}$$

$$y_{\text{pixel}}(z/f) = y \tag{6}$$

Define

$$w = z/f \tag{7}$$

Equations (5), (6), and (7) can be written in the matrix form

$$\begin{bmatrix} wx_{\text{pixel}} \\ wy_{\text{pixel}} \\ w \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1/f & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{8}$$

Equation (8) can be used to find a pixel location, $(x_{\text{pixel}}, y_{\text{pixel}})$, for any point (x, y, z) in space. Three-dimensional information is reduced to two-dimensional information by dividing wx_{pixel} by w. Equation (8) cannot be inverted. It is not possible to use a pixel location alone to determine a unique (x, y, z) point.

In order to represent the camera in different locations, it is helpful to define a z_{pixel} co-ordinate that will always have a constant value. The equation below is a perspective projection matrix that contains such a constant:

$$\begin{bmatrix} wx_{\text{pixel}} \\ wy_{\text{pixel}} \\ wz_{\text{pixel}} \\ w \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1/f & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{9}$$

In Eq. (9), z_{pixel} always evaluates to f, the focal length and the location of the image plane.

The camera model can now be displaced using standard homogeneous transformation matrices [1]. For

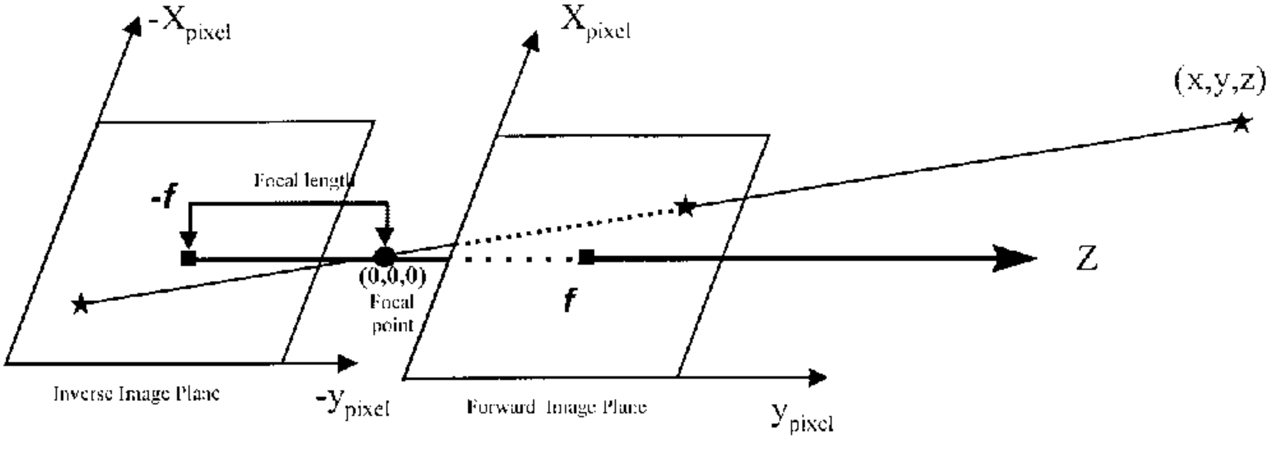


Figure 4 The pinhole camera model leads to the perspective projection matrix.

example, to simulate moving the focal point to a new location, d, on the z-axis, one would use the equation

$$\begin{bmatrix} wx_{\text{pixel}} \\ wy_{\text{pixel}} \\ wz_{\text{pixel}} \\ w \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1/f & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -d \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{10}$$

This equation subtracts a value of d in the z-direction from every point being viewed in the co-ordinate space. That would be equivalent to moving the camera forward along the z-direction by a value of d.

The co-ordinate space orientation, and hence the camera's viewing angle, can be changed using standard rotation matrices [1]. A pinhole camera, five units away from the origin, viewing the world space at a 45° angle with respect to x–z-axis would have the matrix

$$\begin{bmatrix} wx_{\text{pixel}} \\ wy_{\text{pixel}} \\ wz_{\text{pixel}} \\ w \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1/f & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -5 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos 45^\circ & 0 & \sin 45^\circ & 0 \\ 0 & 1 & 0 & 0 \\ -\sin 45^\circ & 0 & \cos 45^\circ & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{11a}$$

or, in simplified form:

$$\begin{bmatrix} wx_{\text{pixel}} \\ wy_{\text{pixel}} \\ wz_{\text{pixel}} \\ w \end{bmatrix} = \begin{bmatrix} \cos 45^\circ & 0 & \sin 45^\circ & 0 \\ 0 & 1 & 0 & 0 \\ -\sin 45^\circ & 0 & \cos 45^\circ & -5 \\ 0 & 0 & -(\cos 45^\circ)/f & 5/f \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{11b}$$

Once again, the world co-ordinates are changed to reflect the view of the camera, with respect to the pinhole model.

Accuracy in modeling a physical camera is important for obtaining accurate measurements. When setting up a stereo vision system, it may be possible to precisely locate a physical camera and describe that location with displacement and rotation transformation matrices. This will require precision fixtures and lasers to guide set up. Furthermore, special camera lenses should be used, as standard off-the-shelf lenses often deviate from the pinhole model. Rather than try to duplicate transformation matrices in the setup, a different approach can be taken.

Let us consider the general perspective projection matrix for a camera located at some arbitrary location and rotation:

$$\begin{bmatrix} wx_{\text{pixel}} \\ wy_{\text{pixel}} \\ wz_{\text{pixel}} \\ w \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{12}$$

Specialized fixtures are not required to assure a specific relationship to some physically defined origin. (Cameras, however, must always be mounted to hardware that prevents dislocation and minimizes vibration.) The location of the camera can be determined by the camera view itself. A calibration object, with known calibration points in space, is viewed by the camera and is used to determine the a_{ij} constants. Equation (12) has 16 unknowns. Sixteen calibration points can be located at 16 different pixel locations generating a sufficient number of equations to solve for the unknowns [2]. More sophisticated methods of finding the a_{ij} constants exist, and take into account lens deviations from the pinhole model [3,4].

3.2.3 System Types

3.2.3.1 Passive Stereo Imaging

Passive stereo refers to two cameras viewing the same scene from different perspectives. Points corresponding to the same location in space must be matched in the images, resulting in two lines of sight. Triangulation will then determine the (x, y, z) point location.

Assume the perspective projection transformation matrix of one of the cameras can be described by Eq. (12) where $(x_{\text{pixel}}, y_{\text{pixel}})$ is replaced by (x', y'). The two equations below can be derived by substituting for the term w and ignoring the constant z_{pixel}.

$$(a_{11} - a_{41}x')x + (a_{12} - a_{42}x')y + (a_{13} - a_{43}x')z = (a_{44}x' - a_{14}) \tag{13}$$

$$(a_{21} - a_{41}y')x + (a_{22} - a_{42}y')y + (a_{23} - a_{43}y')z = (a_{44}y' - a_{24}) \tag{14}$$

Similarly, a second pinhole camera defines another pair of equations:

$$(b_{11} - b_{41}x'')x + (b_{12} - b_{42}x'')y + (b_{13} - b_{43}x'')z = (b_{44}x'' - b_{14}) \tag{15}$$

$$(b_{21} - b_{41}y'')x + (b_{22} - b_{42}y'')y + (b_{23} - b_{43}y'')z = (b_{44}y'' - b_{24}) \tag{16}$$

where the a_{ij} constants of Eq. (12) have been replaced with b_{ij}. Equations (13)–(16) can be arranged in matrix form to yield

$$\begin{bmatrix} a_{11} - a_{41}x' & a_{12} - a_{42}x' & a_{13} - a_{43}x' \\ a_{21} - a_{41}y' & a_{22} - a_{42}y' & a_{23} - a_{43}y' \\ b_{11} - b_{41}x'' & b_{12} - b_{42}x'' & b_{13} - b_{43}x'' \\ b_{21} - b_{41}y'' & b_{22} - b_{42}y'' & b_{23} - b_{43}y'' \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} a_{44}x' - a_{14} \\ a_{44}y' - a_{24} \\ b_{44}x'' - b_{14} \\ b_{44}y'' - b_{24} \end{bmatrix} \tag{17}$$

The constants a_{ij} and b_{ij} will be set based on the position of the cameras in world space. The cameras view the same point in space at locations (x', y') and (x'', y'') on their respective image planes. Hence, Eqs. (13)–(16) are four linearly independent equations with only three unknowns, (x, y, z). A solution for the point of triangulation, (x, y, z), can be achieved by using least-squares regression. However, more accurate results may be obtained by using other methods [4,5].

Passive stereo vision is interesting because of its similarity to human vision, but it is rarely used by industry. Elements of passive stereo can be found in photogrammetry. Photogrammetry is the use of passive images, taken from aircraft, to determine geographical topology [6]. In the industrial setting, determining points that correspond in the two images is difficult and imprecise, especially on smooth manufactured surfaces. The uncertainty of the lines of sight from the cameras result in poor measurements. Industrial systems usually replace one camera with a projection system, as described in the section below.

3.2.3.2 Active Stereo Imaging (Moire Systems)

In active stereo imaging, one camera is replaced with a projector. Cameras and projectors can both be simulated with a pinhole camera model. For a projector, the focal point of the pinhole camera model is replaced with a point light source. A transmissive image plane is then placed in front of this light source.

A projector helps solves the correspondence problem of the passive system. The projector projects a shadow from a known pixel location on its image plane. The shadow falls on a surface that may have been smooth and featureless. The imaging camera locates the shadow in the field of view using algorithms especially designed for the task. The system actively modifies the scene of inspection to simplify and make more precise the correspondence task. Often the projector projects a simple pattern of parallel stripes known as a Ronchi pattern, as shown in Fig. 5.

Let us assume that the a_{ij} constants in Eq. (17) correspond to the camera. The b_{ij} constants would describe the location of the projector. Equation (17) was overdetermined. The fourth equation, Eq. (16), which was generated by the y'' pixel position, is not needed to determine the three unknowns. A location in space can be found by

$$\begin{bmatrix} a_{11} - a_{41}x' & a_{12} - a_{42}x' & a_{13} - a_{43}x' \\ a_{21} - a_{41}y' & a_{22} - a_{42}y' & a_{23} - a_{43}y' \\ b_{11} - b_{41}x'' & b_{12} - b_{42}x'' & b_{13} - b_{43}x'' \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} a_{44}x' - a_{14} \\ a_{44}y' - a_{24} \\ b_{44}x'' - b_{14} \end{bmatrix} \tag{18}$$

All pixels in the y''-direction can be used to project a single shadow, since the specific y'' pixel location is not needed. Hence, a pattern of striped shadows is logical.

Active stereo systems use a single camera to locate projected striped shadows in the field of view. The stripes can be found using two-dimensional edge detection techniques described in Chap. 5.2. The image processing technique must assign an x'' location to the shadow. This can be accomplished by encoding the stripes [7,8]. Assume a simplified Ronchi grid as

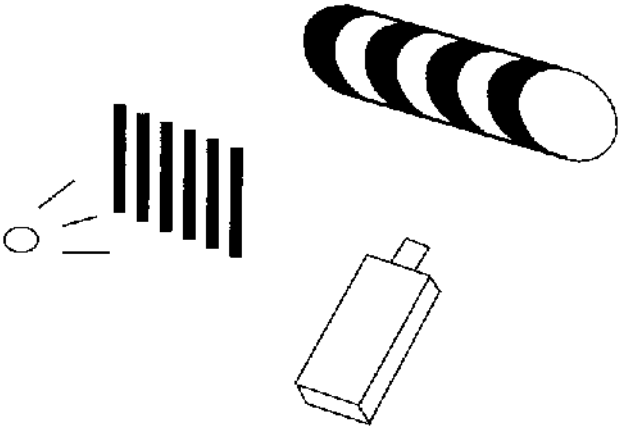


Figure 5 Example of an active stereo vision system.

shown in Fig. 6. Each of the seven stripe positions is uniquely identified by a binary number. The camera images the field of view three times. Stripes are turned on–off with each new image, based on the 3-bit numerical representation. The camera tracks the appearance of shadows in the images and determines the x'' position based on the code.

Prior to the advent of active stereo imaging, moire fringe patterns were used to determine three-dimensional surface displacements. When a sinusoidal grid is projected on a surface, a viewer using the same grid will see fringes that appear as a relief map of the surface [9–11]. Figure 7 shows a conceptual example using a sphere. The stripe pattern is projected and an observer views the shadows as a contour map of the sphere. In order to translate the scene into measurements, a baseline fringe distance must be established. The moire fringe patterns present an intuitive display for viewing displacements.

The traditional moire technique assumes the lines of sight for light source and camera are parallel. As discussed in Sec. 3.2.4, shadow interference occurs at discrete distances from the grid. This is the reason for the relief mapping. Numerous variations on the moire system have been made including: specialized projection patterns, dynamically altering projection patterns, and varying the relationship of the camera and projector. The interested reader should refer to the many optical journals available.

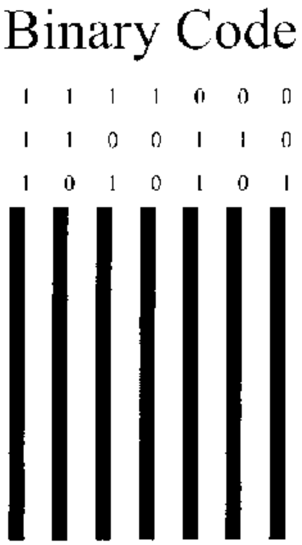


Figure 6 Ronchi grid stripes are turned on (value 1) and off (value 0) to distinguish the x'' position of the projector plane.

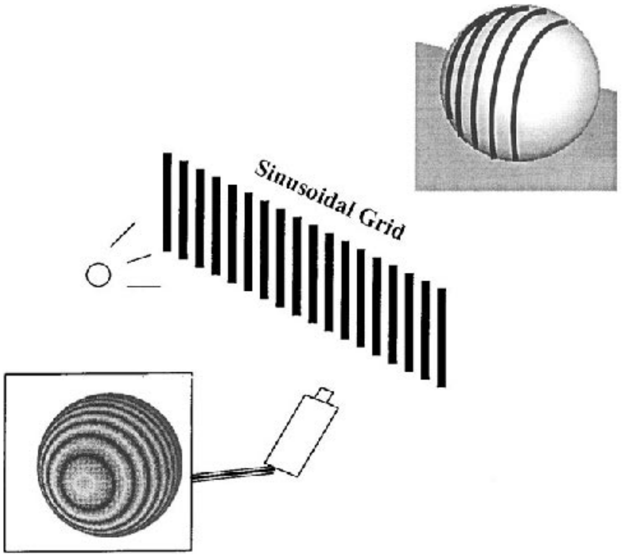


Figure 7 Classic system for moire topology measurements.

The moire technique is a triangulation technique. It is not necessary for the camera to view the scene through a grid. A digital camera consists of evenly spaced pixel rows, that can be modeled as a grid. Active stereo imaging could be described as a moire system using a Ronchi grid projection and a digital camera.

3.2.3.3 Laser Scanner

The simplest and most popular industrial triangulation system is the laser scanner. Previously, active stereo vision systems were described as projecting several straight-line shadows simultaneously. A laser scanner projects a single line of light onto a surface, for imaging by a camera. Laser scanners acquire a single slice of the surface that intersects the laser's projected plane of light. The scanner, or object, is then translated and additional slices are captured in order to obtain three-dimensional information.

For the laser scanner shown in Fig. 8, the laser plane is assumed to be parallel to the x–y-axis. Each pixel on the image plane is represented on the laser plane. The camera views the laser light reflected from the surface, at various pixel locations. Since the z-coordinate is constant, Eq. (18) reduces to

$$\begin{bmatrix} a_{11} - a_{41}x' & a_{12} - a_{42}x' \\ a_{21} - a_{41}y' & a_{22} - a_{41}y' \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a_{44}x' - a_{14} \\ a_{44}y' - a_{24} \end{bmatrix} \tag{19}$$

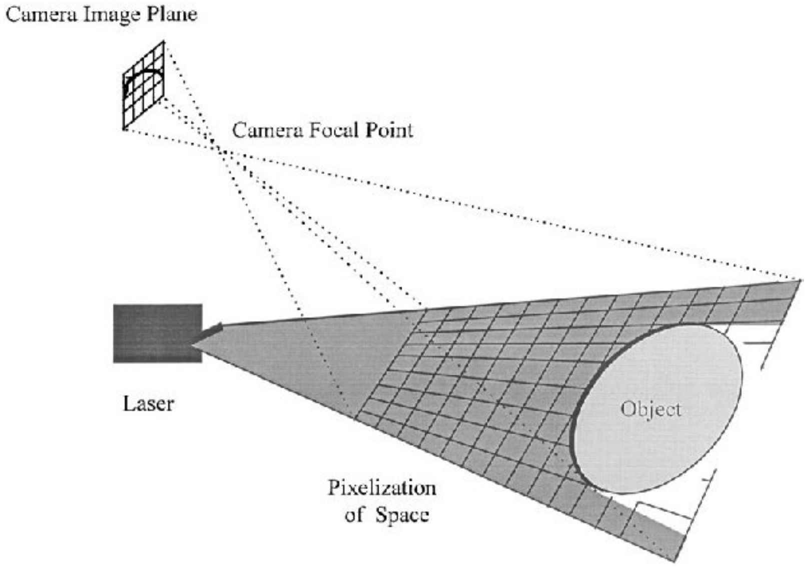


Figure 8 A laser scanner consists of a laser and camera. It is the most popular industrial triangulation system.

Determining the a_{ij} constants in Eq. (19) is unnecessary. The system maps (x', y') pixel locations to (x, y) co-ordinate points. This mapping can be determined with careful measurements and the results can be placed in a lookup table. Deviations from a pinhole camera model, due to lens nonlinearities, can be ignored.

Laser scanners are simple and inexpensive. A standard camera lens can be used. The laser projects a high intensity light with excellent linear characteristics and nearly unlimited depth of field. Projectors, on the other hand, often require high intensity light sources and uniquely designed projection plane optics to achieve a desired accuracy.

3.2.4 Triangulation-Resolution

Resolution of three-dimensional vision systems is determined by geometry. Consider two cameras systems, or camera and projector systems. Such systems are usually placed on the same horizontal plane, at an angle. Shown in Fig. 9 are two pinhole models with the same focal length at an angle γ to the z-axis. Consider the bottom most pixel on each imaging plane. The difference between pixel locations of the two cameras $(x_1 - x_2)$ would be zero. These two pixels view a point indicated by p in Fig. 9. If the view correspondence between pixels of the cameras is always zero, then the surface of an ellipse (shown in the figure) is being viewed [12]. As the difference between pixel locations changes, the distance from the camera changes to a new ellipse. Notice that this is not a linear relationship.

To simplify the resolution analysis, one of two configuration approximations are usually made. First, the cameras can be placed parallel to the z-axis, not at an angle γ. The resulting resolution map is shown in Fig. 10. Now when pixels correspond at a set difference $(x_1 - x_2 = \text{constant})$, the viewing surface is a straight line, not an ellipse. The relation between pixel differences and distance from the cameras is given by the equation:

$$z = \frac{fb}{x_1 - x_2} \tag{20}$$

where f is the focal length of both cameras and b is their displacement. Equation (20) is often presented in literature for triangulation systems (including moire), and is a useful approximation.

A second method of simplifying the stereo configuration is to assume the focal length of both cameras is infinity. A focal length of infinity is a mathematical convention for describing a camera with parallel lines of sight. As shown in Fig. 11, the lines of sight no longer converge to a single focal point. The resolution map is once again simplified to straight lines. The relation between pixel difference and distance becomes linear and is described by the equation

$$z = (x_1 - x_2)\cos\gamma \tag{21}$$

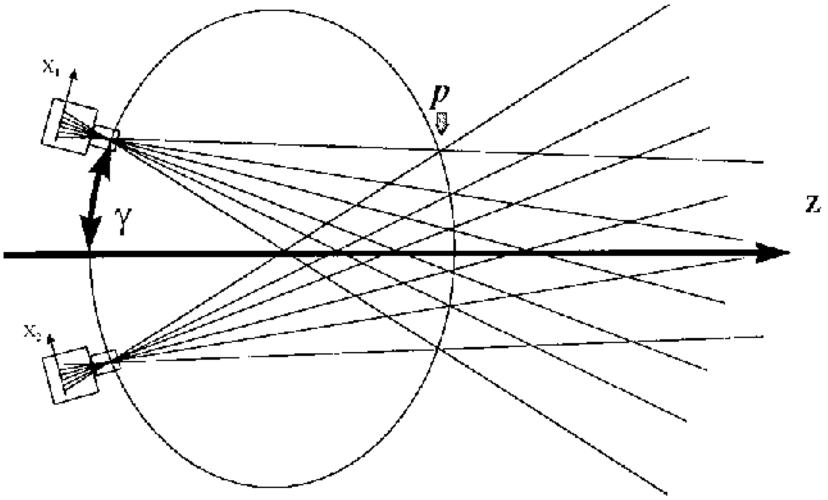


Figure 9 A stereo vision system modeled with two pinhole cameras. Matching the lines of sight (such as point p) from each camera results in an elliptical mapping.

For a laser scanner, the camera views a single line of sight from a laser. The simplest setup available is for the laser and camera to be parallel, as shown in Fig. 12. As the surface reflecting the laser is displaced in the z-direction, the image of the laser point shifts on the image plane. Assuming the focal length of the camera is f and the displacement between camera and laser is b, then by trigonometry it can be shown that

$$z = \frac{fb}{x} \tag{22}$$

The relationship between x and z is not linear. Resolution along the z-axis decreases as the object moves away from the system.

The configuration shown in Fig. 12 uses less than half of the camera's image plane. The camera will likely be at some angle γ to the laser line of sight, as shown in Fig. 13. This results in the equation

$$z = \frac{xb}{f\sin^2\gamma + x\sin\gamma\cos\gamma} \tag{23}$$

In most of the equations of this section, pixel difference and distance are inversely related. Equation (21) is the only exception, but it models a system that is very difficult to achieve precisely. The inverse relation means resolution decreases as the surface is moved away from the system. Hence, the accuracy of a three-dimensional vision system is quoted at a nominal distance.

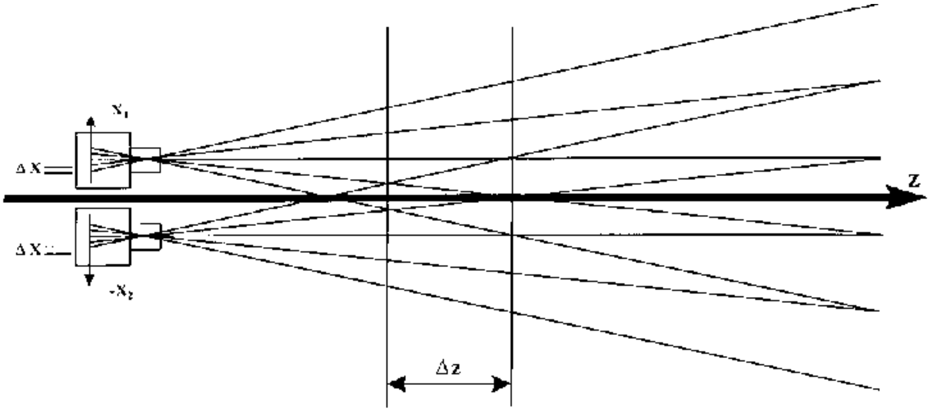


Figure 10 Parallel pinhole cameras have a simpler measurement mapping.

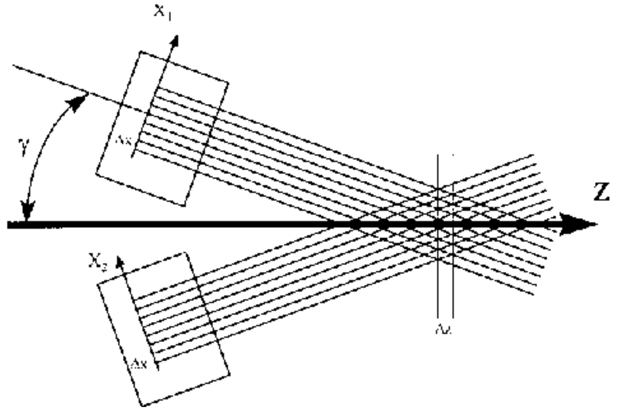


Figure 11 A stereo vision system with parallel lines of sight.

All the equations are scalable. Large systems can be built with large resolutions and large field of views. Large systems are usually limited by the amount of luminance that can be generated by the projection system. Smaller systems can also be built. Smaller systems are limited by interference due to the wave nature of light. Small laser scanners can achieve accuracies of a thousands of an inch for surface inspection.

In addition to the geometrical equations presented above, two other factors that influence accuracy must be kept in mind. First, uncertainties in locating a triangulation point in the camera must be accounted for. Steep surface slopes can be especially problematic as projected shadows may be difficult to find. Second, the position of all focal points and image planes must be maintained. The system components should be locked down and protected. Vibrations to the unit should be minimized.

3.2.5 Triangulation—Missing Data

All triangulation systems can suffer from occlusion. In Fig. 14, the upper camera's line of sight is blocked by the surface from viewing the laser projection. Viewing angle γ, which sets the baseline distance, is an important parameter for determining occlusion characteristics.

A small angle results in a narrow probing triangle. Such a probe is difficult to block and is useful for sharp surface concavities. From the equations of the previous section, decreasing the baseline distance, b, decreases the resolution on the z-axis. As the viewing angle is increased, the path of the light is more likely to be occluded by concave surfaces but accuracy improves. A balance between measurement accuracy and probing width must always be considered in triangulation systems. Adding a second camera, as shown in Fig. 14, is often another method to address the occlusion issue.

3.3 DATA PROCESSING

3.3.1 Goals for Three-Dimensional Data

A three-dimensional vision system can acquire a large number of data points from a surface. The data points represent a physically realizable object. In today's computer systems, manipulation of surface data is usually accomplished using a mechanical CAD (computer-aided design) package. Commercially available CAD packages are fast and efficient for modeling physical objects. Acquiring a CAD model from a physical object directly is referred to as reverse engineering.

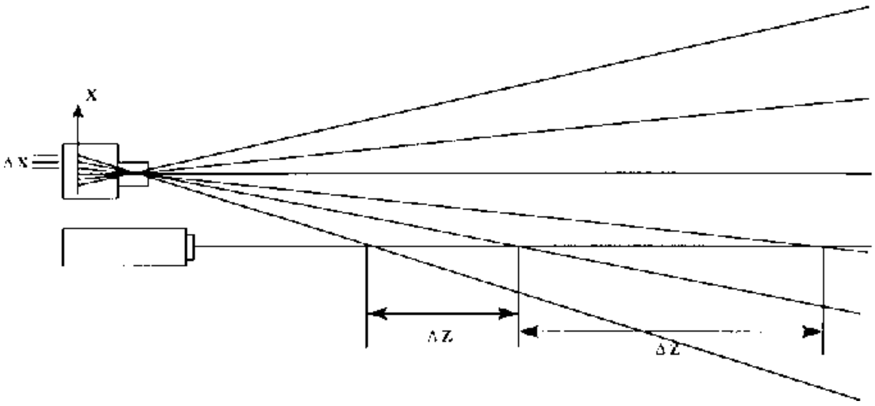


Figure 12 Resolution of a laser scanner is nonlinear in the z (depth) direction.

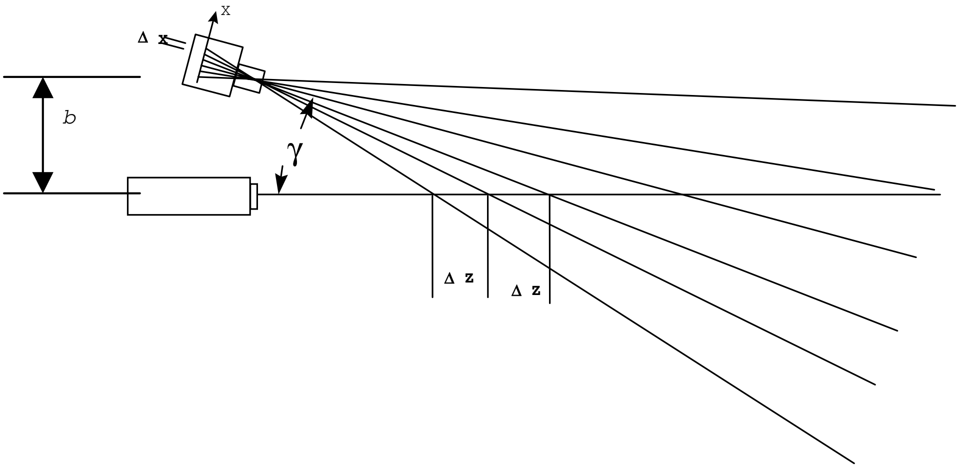


Figure 13 The field of view of the camera is better used by placing it at an angle.

CAD systems operate by representing an object as a collection of cubic B-spline surfaces. B-spline surfaces are portable to nearly all commercially available CAD software via a standard protocol known as the IGES standard. A collection of data points, such as those provided by a three-dimensional vision system, must be converted into a B-spline surface to allow for CAD processing.

Interpolating is one method of finding a surface which passes through a given set of three-dimensional data points. Interpolation assumes the data points are exact, uncorrupted by errors or noise. As a practical matter, the interpolation problem can be easily solved, but the resulting surfaces are not always fair [13]. Furthermore, interpolation requires the number of B-spline control points be equal to the number of original data points. Hence, if 20 data points fall on a plane, interpolation requries 20 control points, whereas a B-spline surface could represent a plane with as few as four control points. Interpolation results in redundant control points that slow a CAD system down.

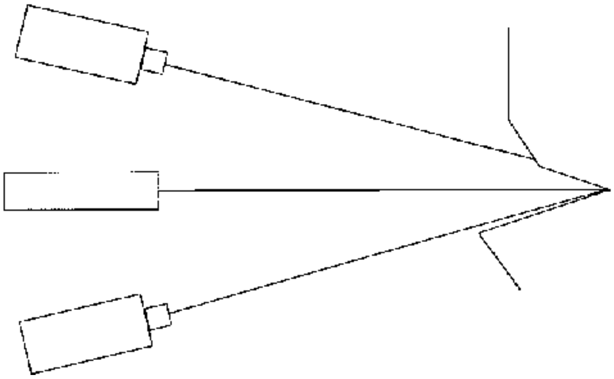


Figure 14 The upper camera is occluded from viewing the laser, while the lower camera would still have an unobstructed view.

Approximation is the process of finding a curve or surface which passes near the data, usually in a least-squares sense. Approximation requires additional analysis to fit the data but fewer control points are needed to represent the surface. In the sections below, an approximation technique for cubic B-splines will be discussed. Other approximation techniques to a wide variety of surface models can be found in Bolle and Vemuri [14].

3.3.2 Review of Cubic Splines

The surface to be fitted to the three-dimensional point data is the two-parameter cubic spline. To simplify the discussion, a one-parameter cubic spline will first be considered. The one-parameter cubic spline is a third-degree polynomial that describes the path of a particle through time and space. The trace of this particle is the desired curve.

The particle path and hence a two-dimensional curve can be represented by two polynomials:

$$x(t) = K_{0x} + K_{1x}t + K_{2x}t^2 + K_{3x}t^3 \tag{24}$$

$$y(t) = K_{0y} + K_{1y}t + K_{2y}t^2 + K_{3y}t^3 \tag{25}$$

where $0 \leq t \leq 1.0$. Equations (24) and (25) can be written as

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} K_{0x} \\ K_{0y} \end{bmatrix} + \begin{bmatrix} K_{1x} \\ K_{1y} \end{bmatrix} t + \begin{bmatrix} K_{2x} \\ K_{2y} \end{bmatrix} t^2 + \begin{bmatrix} K_{3x} \\ K_{3y} \end{bmatrix} t^3 \tag{26}$$

Equation (26) represents four points in two-dimensional space that are operated on to give a two-dimensional curve. The parameter t and its coefficient points can be rewritten as

$$\begin{aligned} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} B_{0x} \\ B_{0y} \end{bmatrix} (1-t)^3 + \begin{bmatrix} B_{1x} \\ B_{1y} \end{bmatrix} 3t(1-t)^2 \\ + \begin{bmatrix} B_{2x} \\ B_{2y} \end{bmatrix} 3t^2(1-t) + \begin{bmatrix} B_{3x} \\ B_{3y} \end{bmatrix} t^3 \end{aligned} \tag{27}$$

where Eqs. (26) and (27) can be made equivalent for the appropriate coefficient values.

Equation (27) is defined as the Bezier curve and is controlled by points B_0, B_1, B_2, and B_3. The curve always interpolates points B_0 and B_3, and is tangential to vectors $\overline{B_0B_1}$ and $\overline{B_2B_3}$, as shown in Fig. 15.

A different set of functions can be used that define a curve called the uniform B-spline:

$$\begin{aligned} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \frac{1}{6}\Bigg\{ \begin{bmatrix} C_{0x} \\ C_{0y} \end{bmatrix} (1-t)^3 + \begin{bmatrix} C_{1x} \\ C_{1y} \end{bmatrix} (3t^3 - 6t^2 + 4) \\ + \begin{bmatrix} C_{2x} \\ C_{2y} \end{bmatrix} (-3t^3 + 3t^2 + 3t + 1) \\ + \begin{bmatrix} C_{3x} \\ C_{3y} \end{bmatrix} t^3 \Bigg\} \end{aligned} \tag{28}$$

To understand the B-spline, one can find Bezier control points from the control points in the equation above [15]. First, draw a line between C_0–C_1, C_1–C_2 and C_2–C_3. These lines are then divided into three equal parts. Two additional lines are drawn from points on these divided lines, as shown in Fig. 16. Bezier control points B_0 and B_3 fall on the respective bisectors of these two new lines. Control points B_1 and B_2 are located on the hash marks of the line between C_1–C_2.

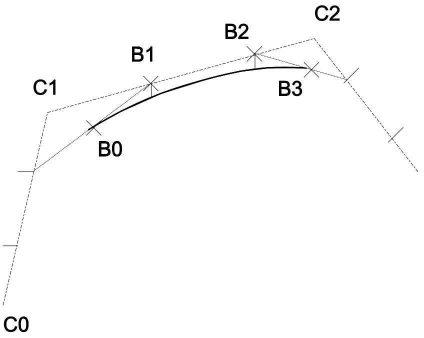


Figure 16 A Bezier curve can be obtained from the control points of a uniform B-spline as shown.

Control points of multiple B-splines always overlap. In Fig. 17, two cubic B-splines are controlled by the points C_0, C_1, C_2, C_3, and C_4. The first spline uses C_0, C_1, C_2, and C_3, while the second spline is controlled by C_1, C_2, C_3, and C_4. If one derives the Bezier control points for each of these splines, it is clear that the B_3 control point of the first spline is equal to B_0 of the second spline. This is known as C^0 continuity. In other words, the two splines always touch. Furthermore, the $\overline{B_2B_3}$ vector of the first spline is equal to the $\overline{B_0B_1}$ vector of the second spline. This is known as C^1 continuity. The tangent of the two splines is identical where they meet. It can be shown that in fact B-splines have C^2 continuity where they meet. The polynomials of the splines must be differentiated at least three times for a discontinuity between the functions to appear.

A two-parameter cubic spline can be used to describe a surface. Analysis of the surface spline proceeds in an analogous fashion to the curve, with the

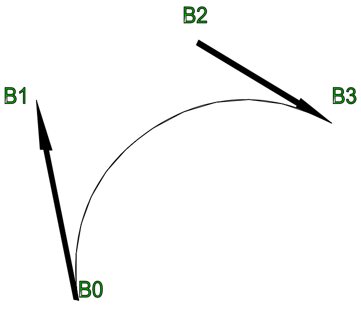


Figure 15 The influence of control points on a Bezier curve.

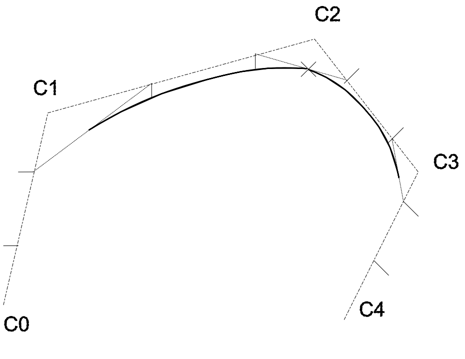


Figure 17 Two uniform B-splines exhibit C^2 continuity.

resulting surface once again demonstrating C^2 continuity.

3.3.3 Fitting Data to Splines

Splines can be fitted to data in a least-squares sense [13,16]. In linear least-squares regression, the straight-line function

$$d_i = f(t_i) = k_1 t_i + k_0 \tag{29}$$

is to be fitted to the data, where t_i is a known ordinate and d_i is the measured data. The values of k that best generate a line which fit the data in a least-squares sense, will minimize the functional

$$\tau = \sum_{i=0}^{N-1} (d_i - (k_1 t_i + k_0))^2 \tag{30}$$

From calculus, we know the minimum occurs when

$$\frac{\partial \tau}{\partial k_j} = 0 \qquad j = 0, 1 \tag{31}$$

Expressing Eq. (30) term by term gives

$$\tau = \{d_0 - (k_1 t_0 + k_0)\}^2 + \{d_1 - (k_1 t_1 + k_0)\}^2 + \{d_2 - (k_1 t_2 + k_0)\}^2 \dots$$

which gives the matrix form

$$\tau = \sum \left(\begin{bmatrix} d_0 \\ d_1 \\ \vdots \\ d_{N-1} \end{bmatrix} - \begin{bmatrix} t_0 & 1 \\ t_1 & 1 \\ \vdots & \vdots \\ t_{N-1} & 1 \end{bmatrix} \begin{bmatrix} k_1 \\ k_0 \end{bmatrix} \right)^2 \tag{32a}$$

or simply

$$\tau = \sum_{0}^{N-1} (\mathbf{d} - \mathbf{Hk})^2 \tag{32b}$$

where $N - 1$ is the number of data points. When Eq. (32) is differentiated with respect to the constants, k_j, and set to zero, we get

$$0 = 2\mathbf{H}^T(\mathbf{d} - \mathbf{Hk}) \tag{33a}$$

$$\mathbf{H}^T\mathbf{Hk} = \mathbf{H}^T\mathbf{d} \tag{33b}$$

This gives

$$\begin{bmatrix} \sum t_i^2 & \sum t_i \\ \sum t_i & N \end{bmatrix} \begin{bmatrix} k_1 \\ k_0 \end{bmatrix} = \begin{bmatrix} \sum d_i t_i \\ \sum d_i \end{bmatrix} \tag{34a}$$

$$\mathbf{Ak} = \mathbf{b} \tag{34b}$$

If Eq. (34) is nonsingular, the unknown $\mathbf{k}$ can be solved for using gaussian elimination or other appropriate methods [17]. An equation, similar to Eq. (34), can be derived for spline functions. The vector of constants $\mathbf{k}$ becomes a vector of spline control points, $\mathbf{z}$. Data points and appropriate functions of t from Eq. (28), can be used to construct an equation similar to Eq. (29). Proceeding in a similar manner as given above, creates a matrix equation:

$$\mathbf{A_B z} = \mathbf{b_B} \tag{35}$$

Equation (35) is nonsingular and can be solved for the B-spline control points. However, the resulting spline tends to fluctuate due to the noise in the measured data. To obtain a smooth curve, an additional "regularization" term is added to Eq. (35) to further constrain the spline.

Regularization is a theory developed early this century which has been used to solve ill-posed problems. An ill-posed problem may have multiple solutions. Regularization restricts the class of admissible solutions, creating a well-posed problem, by introducing a "stabilizing function" [18]. An integral part of the regularization process is a parameter which controls the tradeoff between the "closeness" of the solution to the data and its degree of "smoothness" as measured by the stabilizing function. The mathematical expression of regularization in one dimension is given as follows:

$$E = \int_\Omega (d - f)^2\, dx + \lambda \int_\Omega S(f)^2\, dx \tag{36}$$

In this functional, the first term is the continuous least squares measure of the closeness of the solution function f to the data. The second term is considered to be the stabilizer and assures the smoothness of f.

For surface data, a particularly useful stabilizer is to minimize the first and second partial derivative of the function [19,20]. The first partial derivative gives the surface elastic properties which become taut with increased emphasis. The second partial derivative causes the surface to act like a thin sheet of metal. Equation (36) for a one-parameter spline with first- and second-derivative stabilizers is given as

$$\tau = \sum_{i=0}^{N-1} (d_i - f(t_i))^2 + \lambda \int (\rho f'(u) + (1 - \rho) f''(u))^2\, du \tag{37}$$

The integral in Eq. (37) can be solved exactly due to the underlying simplicity of a polynomial. Minimizing just the integral part of Eq. (37) with respect to the control points gives a matrix:

$$\mathbf{A_2 z} \tag{38}$$

Finding the minimum of Eq. (37) with Eq. (35) is a matter of finding the solution below using standard numerical techniques:

$$(\mathbf{A_B} + \mathbf{A_2})\mathbf{z} = \mathbf{b_B} \tag{39}$$

3.3.4 Example

The noncontact feature of three-dimensional vision systems make them appropriate for safely inspecting delicate surfaces. One novel use of these systems is the scanning of humans. In this example, data captured from a human head is fitted with a *B*-spline surface and can be exported into a CAD system.

Surface data recorded from a human head is captured by a Cyberware Digitizer shown in Fig. 18. The laser scanner rotates about the subject's head. Scan data is captured and represented in cylindrical co-ordinates with a sampling resolution of 1.563 mm in the latitudinal direction and 0.7031° of arc in the longitudinal direction. This results in a 256 × 512 array of radius values. An example data set is shown in Fig. 19.

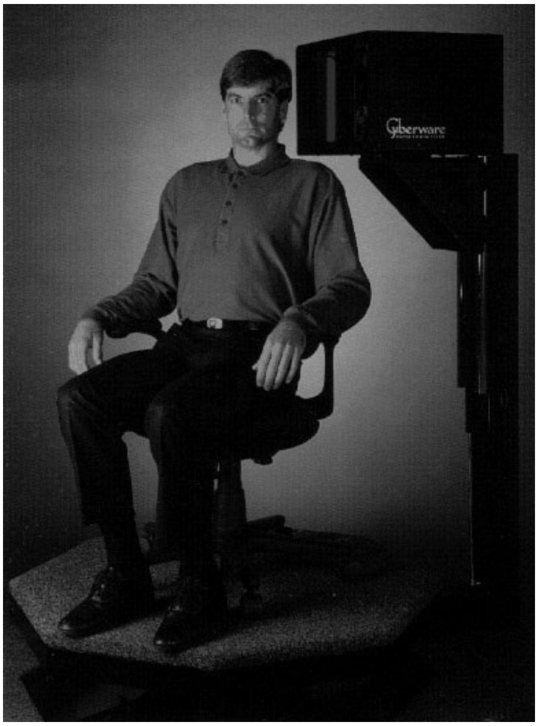


Figure 18 A laser scanner is used in a human head scanner configuration. (Photograph courtesy of Cyberware Inc., Monterey, CA.)

The head data points are fitted to a surface by minimizing Eq. (37). The first set of parameters to choose is the number of control points. Figure 20 shows the data set fitted with surfaces using 128 × 32 control points and 281 × 74 control points. Control points are evenly distributed throughout the *B*-spline surface. Ideally, control points should be concentrated in areas of rapid surface change such as the nose, but techniques for achieving this distribution automatically are still under research.

Secondly, regularization parameters λ and ρ must also be chosen. Reducing surface smoothness by decreasing the size of regularization parameter, λ, causes the fitted splines to oscillate in an effort to fit

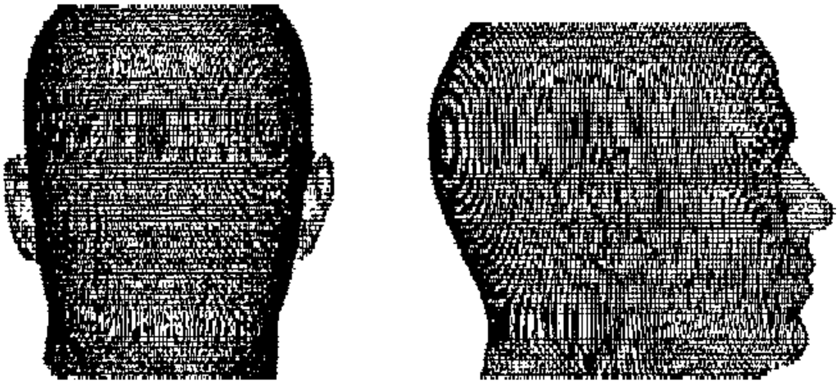


Figure 19 Various views of a head scan data set. (Data set courtesy of the Computerized Anthropometric Research and Design Laboratory at Wright Patterson Air Force Base, Dayton, OH.)

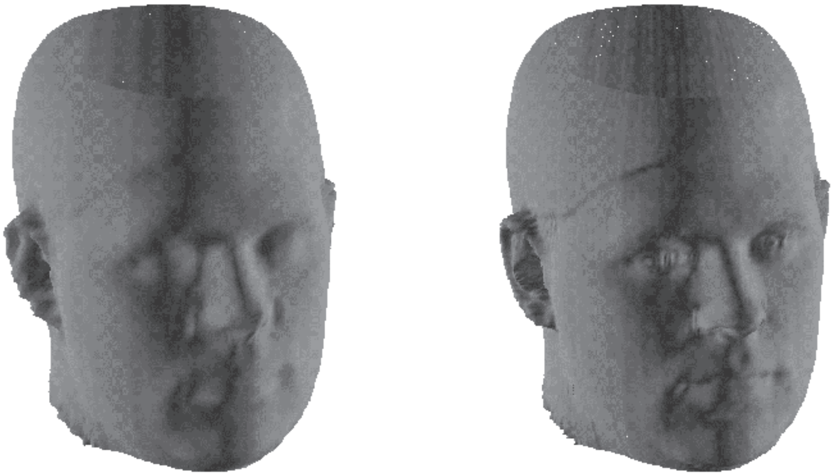


Figure 20 Two different surface fits applied to the same data set.

the surface to respective data points. Severe ripples result near corners around the nose. A value should be chosen that is as small as possible without causing oscillations. In Fig. 20, a value of $\lambda = 0.1$ yielded good results. The parameter ρ also affects fit. For head scans, experience indicates that less error can be achieved by biasing the fit toward first-degree information. However, if ρ is too small, surface oscillations will once again appear. For this example, a value of $\rho = 0.25$ yielded good results.

3.4 CONCLUSION

Advances in computer processing and storage technologies have made three-dimensional vision systems a reality. This chapter reviewed the various system designs available. For industrial settings, laser scanners are most popular, providing accuracy to one-thousandth of an inch. Active stereo imaging is also popular and is often described as a moire system.

Three-dimensional vision systems can acquire a large point data set of the surface under inspection. To use this data effectively, it should be imported into a CAD system. Capturing surface data of an object for CAD manipulation is referred to as reverse engineering. The data points can be fitted with *B*-spline surfaces and imported into a CAD system. Fitting surfaces is not a fully automated task. It requires understanding of and experimentation with several parameters that affect the final surface.

Three-dimensional vision systems continue to be an active area of research. Further improvements in computer hardware, as well as advances in software, will continue to enhance its application in the market of industrial inspection.

REFERENCES

1. JD Foley, A vanDam, SK Feiner, JF Hughes. Computer Graphics, Principles and Practice. 2nd ed. Reading, MA: Addison-Wesley, 1990.
2. EL Hall, MBK Tio, CA McPherson, FA Sadjadi. Measuring curved surfaces for robot vision. Computer 15(12): 42–54, 1982.
3. RY Tsai. A versatile camera calibration technique for high-accuracy 3D machine metrology using off-the-shelf TV cameras and lenses. IEEE J Robot Autom 3(4): 323–344, 1987.
4. O Faugeras. Three-Dimensional Computer Vision. Cambridge, MA: The MIT Press, 1996.
5. K Be. Determination of the most probable point from nonconcurrent lines. SPIE Technical Symposium Southeast, Orlando, FL, vol 635, 1986, p 552.
6. PR Wolf. Elements of Photogrammetry. New York: McGraw-Hill, 1983.
7. HS Yang, KL Boyer, AC Kak. Range data extraction and interpretation by structured light. Proceedings 1st IEEE Conference on Artificial Intelligence Applications, December, 1984, pp 199–205.
8. KL Boyer, AC Kak. Color-encoded structured light for rapid active ranging. IEEE-PAMI 9(1): 14–28, 1987.

9. JC Perrin, A Thomas. Electronic processing of moire fringes: application to moire topography and comparison with photogrammetry. Appl Optics 18(4): 563–574, 1979.
10. DM Meadows, WO Johnson, JB Allen. Generation of surface contours by Moire patterns. Appl Optics 9(4): 942–947, 1970.
11. CA Sciammarella. Moire method—a review. Exp Mech 22(11): 418–433, 1982.
12. JH Nurre, EL Hall. Positioning quadric surfaces in an active stereo imaging system. IEEE-PAMI 13(5): 491–495, 1991.
13. DF Rogers, JA Adams. Mathematical Elements for Computer Graphics, 2nd ed. New York: McGraw-Hill, 1990.
14. RM Bolle, BC Vemuri. On Three-Dimensional Surface Reconstruction Methods. IEEE Trans PAMI 13(1), 1–13, 1991.
15. G Farin. Curves and Surfaces for Computer aided Geometric Design: A Practical Guide. Boston, MA: Academic Press, 1993.
16. P Lancaster, K Salkauskas. Curve and Surface Fitting: An Introduction. Boston, MA: Academic Press, 1990.
17. WH Press, SA Teukolsky, WT Vetterling, BP Flannery. Numerical Recipes in C. New York: Cambridge University Press, 1992.
18. T Poggio, V Torre, C Koch. Computational vision and regularization theory. Nature 317: 314–319, 1985.
19. D Terzopoulos. Regularization of inverse visual problems involving discontinuities. IEEE Trans Patt Anal Mach Intell 8(6): 413–424, 1986.
20. SS Sinha, BG Schunck. A two-stage algorithm for discontinuity-preserving surface reconstruction. IEEE Trans Patt Anal Mach Intell 14(1): 36–55, 1992.

Chapter 5.4

Industrial Machine Vision

Steve Dickerson
Georgia Institute of Technology, Atlanta, Georgia

4.1 INTRODUCTION

The *Photonics Dictionary* defines machine vision (MV), as "interpretation of an image of an object or scene through the use of optical noncontact sensing mechanisms for the purpose of obtaining information and/or controlling machines or processes." Fundamentally, a machine vision system is a computer with an input device that gets an image or picture into memory in the form of a set of numbers. Those numbers are processed to obtain the information necessary for controlling machines or processes.

This chapter is intended to be a practical guide to the application of machine vision in industry. Chapter 5.2 provides the background to machine vision in general, which includes a good deal of image processing and the relationship of machine vision and human sight. Machine vision in the industrial context is often less a problem of image processing than of image acquisition and is much different than human visual function.

The Automated Imaging Association (AIA) puts the 1998 market for the North American machine vision industry at more than $1 billion with growth rates exceeding 10%.

4.1.1 MV in the Production of Goods and Services—a Review

Machine vision is not a replacement for human vision in the production of goods and services. Like nearly all engineering endeavors designed to increase productivity, the technology *does not* emulate human or nature's methods, although it performs functions similar to those of humans or animals. Normally, engineers and scientists have found ways to accomplish tasks far better than any natural system, but for very specific tasks and in ways quite different than nature. As examples, no person or animal can compete with the man-made transport system. Is there anything in nature comparable in performance to a car, a truck, a train, or a jet aircraft? Do any natural systems use wheels or rotating machinery for power? Are the materials in animals as strong and tough as the materials in machinery? Can any person compete with the computing power of a simple microprocessor that costs less than $4? Communications at a gigabit per second on glass fibers a few microns in diameter without error is routine. Any takers in the natural world?

But clearly with all this capability, engineering has not eliminated the need for humans in the production of goods and services. Rather, engineering systems have been built that replace the mundane, the repetitive, the backbreaking, the tedious tasks; and usually with systems of far higher performance than any human or animal could achieve. The human is still the creative agent, the final maker of judgments, the master designer, and the "machine" that keeps all these engineered systems maintained and running.

So it is with industrial machine vision. It is now possible to build machine vision systems that in *very specific tasks* is much cheaper, much faster, more accurate, and much more reliable than any person. *However*, the vision system will not usually be able to directly replace a person in a task. Rather, a structure to support the machine vision system must be in place, just as such a structure is in place to support the human in his productive processes. Let us make this clear by two examples:

Example 1. *Nearly every product has a universal product code on the packaging. Take the standard Coke can as an example. Why is the UPC there? Can a person read the code? Why is roughly 95% of the can's exterior cylinder decorated with the fancy red, white, and black design? Why is there that rather unnatural handle on the top (the flip-top opener)?*

Answers. *The UPC is a structure to* support a machine. *People do not have the ability to reliably read the UPC, but it is relatively easy to build a machine to read the UPC; thus the particular design of the UPC. The exterior design and the flip-top support the particular characteristics of people, and are structures to support them. Coca-Cola wants to be sure you immediately recognize the can and they want to be sure you can open it easily. If this can was processed only by machine, we could reduce the packaging costs because the machine could read the UPC and open the can mechanically without the extra structure of the flip-top.*

Example 2. *Driving a car at night is made much easier by the inclusion of lights on cars and rather massive amounts of retroreflective markings on the roads and on signs. The State of Georgia uses about four million pounds of glass beads a year to paint retroreflective stripes on roads. Could we get by without these structures to support the human's driving? Yes, but driving at night would be slow and unsafe. If we ever get to machine-based driving, would you expect that some structure would need to be provided to support the machine's ability to control the car? Would that structure be different than that to support the human driver?*

Thus we get to the bottom line. Machine vision is a technology that can be used to replace or supplement human vision in many tasks and, more often, can be used to otherwise contribute to improved productivity. That is, it can do tasks we would not expect human vision to perform. A good example would be direct vision measurement of dimensions. *But*, the entire task must be structured to support the vision system, *and* if done right, the vision system will be much more reliable and productive than any human could be.

As a final note to illustrate the potential importance of MV it is suggested that you consider the activities in any factory and ask why people are even involved in the production process. When you go through a factory, you will find that the vast majority of the employees are there because they possess hand–eye coordination and can seemingly control motion using feedback from eyes (as well as touch and sound), with little effort. *It is technically challenging* to build a machine that can cost-effectively assemble a typical product, that can load and unload a machine and move the parts to the next location, or that can inspect an arbitrary part at high speed. Consider the problem of making a "Big Mac" from start to finish. However, it is clear that we are moving in the direction where machine vision can provide the *eye* function, if we take a systems prospective. Often the design of both the product and the process need to take advantage of the strengths of machines, and not the strength of people for this to be economical.

Of course, this raises the specter of unemployment with increased "automation." Actually, it raises the specter of ever higher living standards. Consider that once, 50% of the work force was in farming. Consider that today's level of phone service would require 50% of the work force if we used the manual methods of the 1930s. Consider that the United States has already reduced the workers actually in factories to less than 10% of the work force, yet they produce nearly as much in value as all products that are consumed.

4.1.2 The Structure of Industrial Machine Vision

Industrial machine vision is driven by the need to create a useful output, automatically, by acquiring and processing an image. A typical MV process has the following elements:

1. Product presentation
2. Illumination
3. Image formation through optics
4. Image digitization
5. Image processing
6. Output of signals.

Reliability of the result, and real-time control of the entire process are very important. A few of the most successful examples of machine vision, illustrate the point.

Figure 1 shows a bar code reader. In manual bar code scanning:

1. *Product presentation*. Manual positioning of the product or reader. The device may be constantly looking for a code instead of using a trigger.
2. *Illumination*. A scanning laser beam. This gives a structured light with narrow wavelength band.
3. *Optics*. Standard optics with a filter to remove all but the laser light wavelength.
4. *Image digitization*. A single photo cell or pixel for most bar code readers.
5. *Image processing*. Looks for spacing of rise and fall of the signal. When a valid pattern is found.
6. *Output of signals*. Serial to cash register, computer. Perhaps wireless, radio, or infrared (IR).

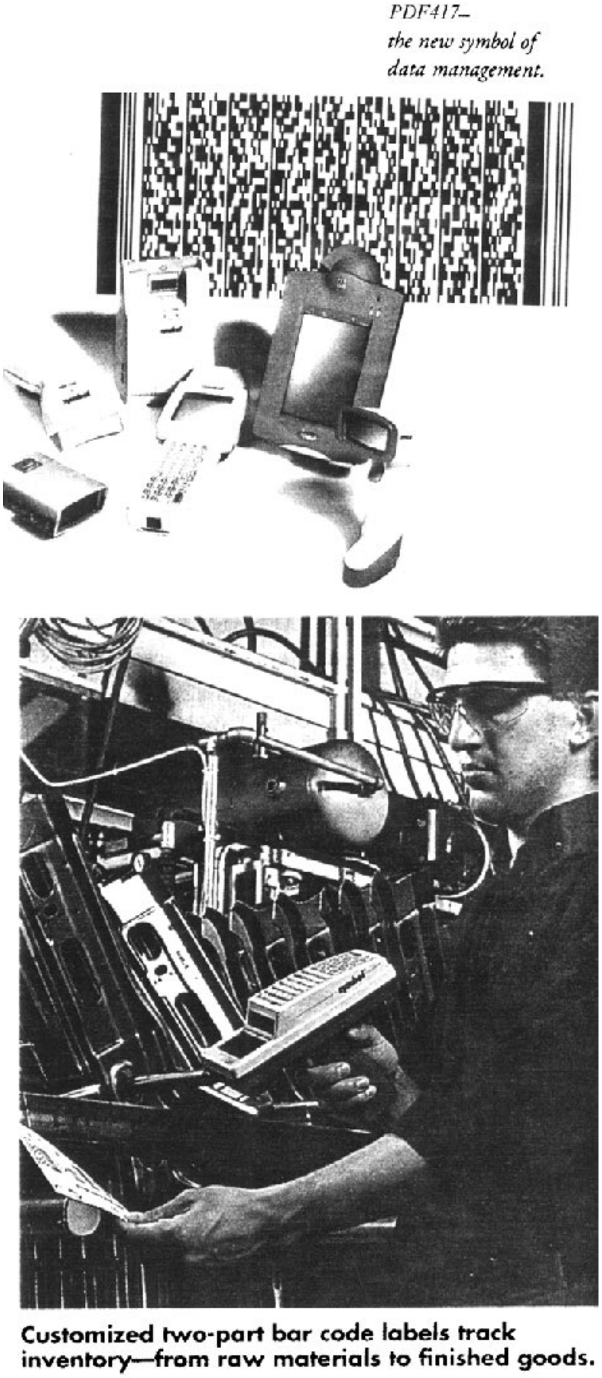

PDF417—
the new symbol of
data management.
Customized two-part bar code labels track inventory—from raw materials to finished goods.

Figure 1 A bar code reader. (Top) Symbol Technologies Readers and 2D bar code. (Bottom) Person using code reader.

Figure 2 shows an MV-driven circuit board assembly machine. In electronics assembly:

1. *Product presentation*. Circuit boards are indexed on a conveyor into an assembly position in the machine. Signals are sent to the MV system saying when to take and analyze images.
2. *Illumination*. Arrays of red light-emitting diodes (LEDs).
3. *Optics*. Standard but may include filter to mask non-LED wavelengths if background light is a problem (usually not, because LEDs overwhelm background).
4. *Image digitization*. Standard array charge-coupled device (CCD) of order of 512 × 512 pixels.
5. *Image processing*. Looking for position information of fiducials on circuit board, electronic parts, or naturally occurring edges on same. Will usually be able to reject bad parts based on incorrect outline geometry. Only the pixels on edges in predefined regions will be processed.
6. *Output of signals*. Sends the machine controller the locations. This is used to tailor the motion of the machine to the particular board position and the position of the part as picked up from the feeder.

Figure 3 shows a task of checking threads in a machined part. In checking the product for proper construction on a line:

1. *Product presentation*. Parts are brought to rest at a location just after the machine operation and close to a nominal position. An external signal tells the MV system to make a verification.
2. *Illumination*. LED illumination more or less from a single direction causes a portion of each thread and a reference surface on the part to be of particularly high brightness, taking advantage of the specular reflections of machined metal.

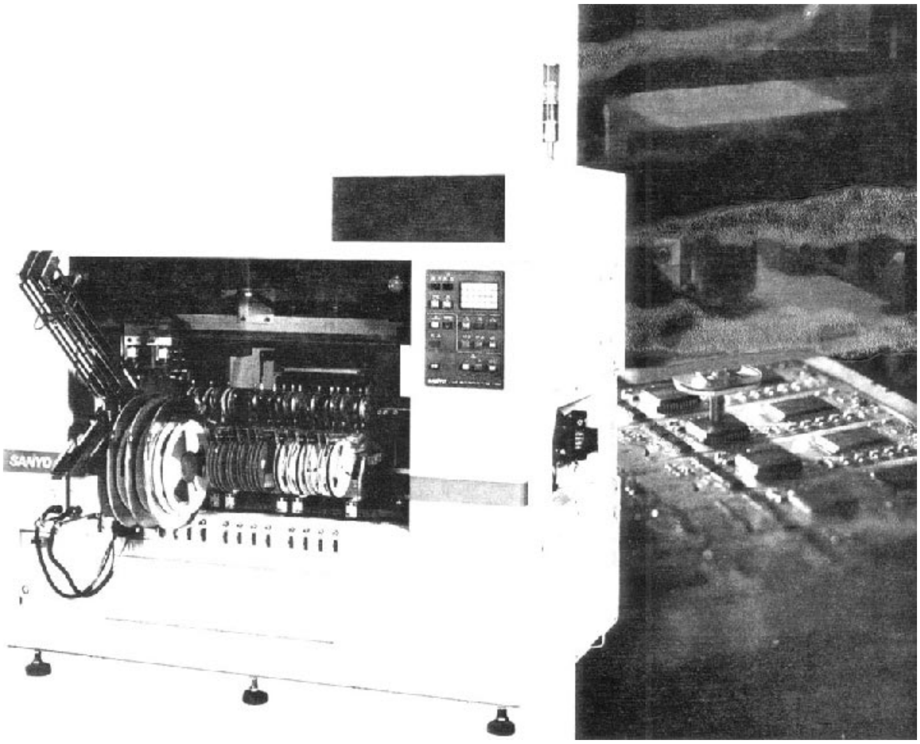


Figure 2 An MV-driven circuit board assembly machine. (From Sanyo General Catalog.)

3. *Optics*. Standard.
4. *Image digitization*. Standard array CCD of order of 256 × 256 pixels.
5. *Image processing*. Simple line scans are used to first find edges to precisely locate the part. Then a line scan is done to go through the region where the threads should be highlighted. The rising and falling intensity along that line is analyzed for the correct number and spacing of bright edges.
6. *Output of signals*. A good/bad signal could include feedback information on number and spacing of threads.

Note that, in each of these examples, the need to process an entire image is avoidable because in these applications, the scene observed is very structured; the system knows what to look for. The system then takes advantage of prior knowledge of what is expected in the scene. It is typical of industrial applications that a high degree of prior knowledge of the scene is available, otherwise the manufacturing process itself is likely to be out of control. If it does get out of control the MV system will sound the alert.

4.1.3 Generic Applications

The following five categories contain the bulk of current industrial applications. Several specific rather common examples are given.

Verifying presence and shape
- Label on a product? Must be correct.
- Product correct in a container? Correct before closure and distribution.
- Parts assembled correctly in a product? For instance, are all the rollers in a bearing?
- Product closure correct? For instance, a bottle top?
- Natural product OK and not contaminated? For instance, poultry inspection.
- Product looks good? For instance, chocolate bar color.

Measurement
- Parts on a conveyor or feeder are located so they can be handled.
- The position of part is determined so subsequent operations are OK. For instance, the position of a circuit board.

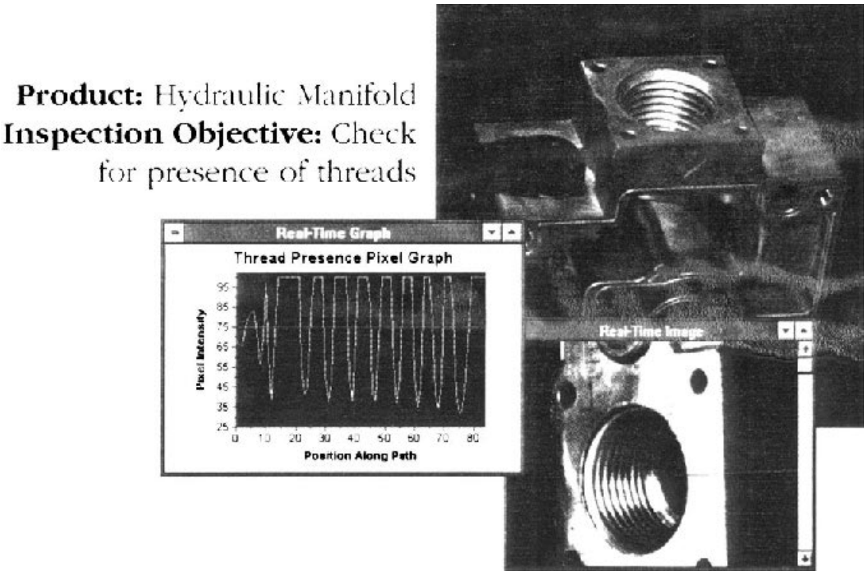

Product: Hydraulic Manifold
Inspection Objective: Check for presence of threads
Real-Time Graph
Thread Presence Pixel Graph
Pixel Intensity
95
85
75
65
55
45
35
25
0
10
20
30
40
50
60
70
80
Position Along Path
Real-Time Image

Figure 3 A task of checking theads in a machined part. (From DVT literature.)

Size and shape of a manufactured product measured. For instance, the shape of a contact lens.

Measure the dimensions of a natural product. For instance, the shape of a person for custom tailoring of clothes.

A critical dimension measured in a continuous product. For instance, diameter of glass fibers.

Code reading

Read a fingerprint or eye iris pattern for security purposes.

Read an intended code. For instance, bar codes, 2D codes, package sorting.

Text reading. For instance, license plates.

Web checking

Defects in a fabric. For instance, from airbag fabric to material for sheets.

Photographic film inspection.

Tracking

Landmarks for vehicle guidance.

Track end of a robot arm.

Missile guidance.

These are typical industrial applications today. Very similar applications can be expected to grow in the service sector which is cost sensitive. There are also a number of military applications where cost is less of an issue. Agricultural and food processing applications are getting a good deal of attention.

4.1.4 Some Rules of Thumb

The following is a modified list of several of the "proverbs" by Batchelor and Whelan.

1. Do not ask a machine vision to answer the question "What is this?" except in the case of a very limited well-defined set of alternatives, the Arabic letters, a predefined 2D code, or among a well-defined set of parts.
2. Consider machine vision for closed-loop feedback control of a manufacturing process rather than wait until all the value has been added to accept or reject. That is, put MV within or immediately after important processes to help maintain control of the process.
3. Do not reject out of hand a combined human-plus-machine vision solution. For example, some tasks, such as inspection of chickens, are not well suited to a completely automated system. Rather the automated system can be used to (1) reject far too many birds for subsequent human reinspection, (2) alert a human inspector to possible substandard characteristics, or (3) reject birds on the basis of characteristics not easily observed by humans, e.g., color characteristics outside of the range of human vision.
4. It is often better to work on improving the image rather than the algorithm that processes the image. Batchelor and Whelan say "if it mat-

ters that we use the Sobel edge detector rather than the Roberts operator, then there is something wrong, probably the lighting."

5. Often a machine vision system should have a minimum of manual field-adjustable components because adjustment will be too tempting or accidental. In particular, the focus and F-stop characteristic of cameras should usually be difficult to adjust once installed. A corollary of this is clearly that it should not need manual adjustment.
6. A benefit of machine vision is often improved quality. Improved quality in *components* of larger systems, e.g., automobiles, can have a tremendous effect on system reliability. A $1 roller bearing that fails inside of an engine can cause $2000 damage.
7. The actual hardware and included software costs can easily be a small fraction of the total MV system cost. A very large fraction of the cost can be in making the system work properly in the application. This argues for relatively simple applications and easy-to-use software, even if it means more MV systems will be needed.
8. Related to item 7, there is a tendency to add more and more requirements to an MV system that add less and less benefit, with the result of not being able to meet the specification for the MV system. This is a version of the usual 90:10 rule. In an inspection for example, 10% of the possible defects can account for 90% of the problems.
9. Be careful of things that could cause large changes in optical characteristics that cannot be handled automatically by the system. For example, the reflectivity of many surfaces is easily changed by a light coat of oil or moisture or by oxides from aging.
10. We quote, "a sales-person who says his company's vision system can operate in uncontrolled lighting is lying." As a rule, the lighting supplied for the purposes of the MV system must dominate over light that comes from the background so that changes in the background are not a factor.
11. If the human cannot see something, it is unlikely that the MV system can. This rule needs to be modified to take into account the fact that MV systems can see outside of the visible spectrum, in particular the near IR, and that with enough care and expense, an MV system can detect contrast variations that would be very difficult for a human. Most important an MV system can actually visually measure a dimension.
12. Inspection is not really a good human production task. An inspector is likely to have a high error rate due to boredom, dissatisfaction, distress, fatigue, hunger, alcohol, etc. Furthermore, in some cases the work environment is not appropriate for a human.
13. Lighting is difficult to get constant in time and space. Some variation in lighting should usually be tolerated by the rest of the MV system. The MV system might be part of a feedback system controlling lighting, including warning of degradation in lighting.
14. The minimum important feature dimension should be greater than two pixel dimensions. This is essentially a restatement of the Nyquist sampling theorem. Higher resolution, if needed, might well be achieved with multiple cameras rather than asking for a single camera of higher resolution.
15. Consider the effects of dirt and other workplace hostilities on a vision system. It may be necessary to take measures to keep the optics clean and the electronics cool. By the same token, consider the effect of the vision system, particularly strobing light, on workers. Strobes can be used to great advantage in MV but the workforce may need to be protected from this repetitive light.

4.2 IMAGE FORMATION

A machine vision system is a computer with an input device that gets an image or picture into memory. An image in memory is a two-dimensional array of numbers representing the amount of light at rows and columns, as shown in Fig. 4. Each number is a pixel or pixel value. The values of the numbers are proportional, or at least monotonically related, to the brightness of the corresponding points in the scene.

The minimum components to acquire the image are illumination, optics, array detector, and an analog to digital (A/D) converter as shown in Fig. 5. For historical reasons, the now typical machine vision system hardware items consist of illumination, a camera (includes the optics, detector, and electronics to convert to a TV signal), a frame grabber (that converts the TV signal to numbers, and a host computer as shown

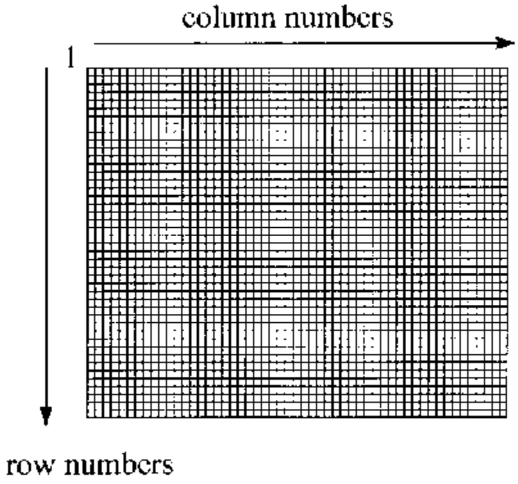


Figure 4 Image array.

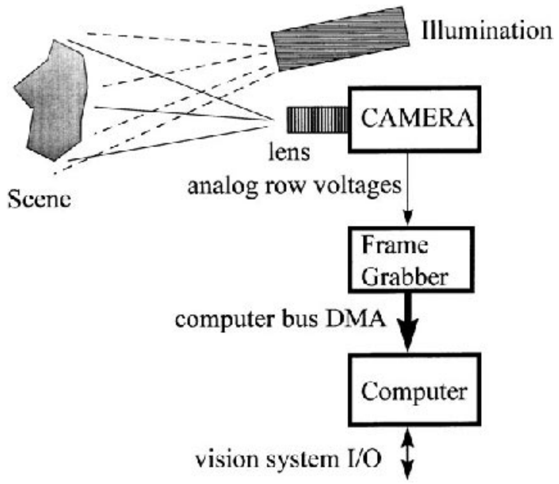


Figure 6 Block diagram of conventional MV system.

in Fig. 6. This is the result of using a TV camera as the original input device.

Television signals consist of an analog signal sequenced by rows. Each row has a voltage variation that represents the pattern of light variation across a row. In standard TV signals, the order of rows is "interlaced." That is, first the odd-numbered rows are presented and then the even-numbered rows. Many cameras for MV are "progressive scan" in which all rows are in sequence. The frame grabber uses a precisely timed sampling to convert the analog values in each row to numbers. This means that the number of pixels per row does not correspond directly to the photodetectors on the detector. Most cameras of the analog type also have a fixed imaging rate, typically 30 complete images per second. More modern cameras for MV do allow the rate to be varied and the timing of image acquisition to be controlled. Further information on TV style imaging processes and image processing is available in Jahne.

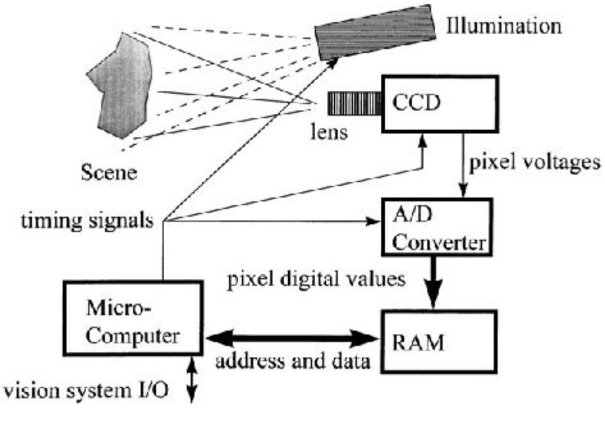


Figure 5 Block diagram of an integrated MV system.

Increasingly, MV systems do not use cameras that develop a television signal and the reader should assume that standard camera-based systems will decline drastically in use relative to the all-digital systems. In the more modern systems either a digital camera is used which outputs directly digital information for storage in RAM, or a direct coupling of the CCD, A/D converter, and a microcomputer bus is used. In the most highly coupled cases, the timing of illumination, light shuttering, and image downloading is controlled to a few microseconds. Furthermore, the images have no jitter because of the direct correspondence of the physical light-sensing element on the detector to a location in memory. Such integrated MV systems can also dynamically control the portion of the image that is downloaded from the detector, thus speeding image transfer to the computer. The highly integrated systems are less costly to build than the more traditional system. That, together with the higher performance in the geometrical and timing sense should result in their eventual dominance in terms of number of systems.

There is also a likely use of CMOS detectors in place of CCDs in the future for some applications. CMOS devices have the benefit of using the same manufacturing processes used to make many chips including microprocessors, RAM, and A/D converters. Thus there is the potential to integrate on a single chip most, if not all, of the components of an MV system.

The current shortcoming of CMOS devices is a lower uniformity in pixel values for a uniform illumination of the detector. The pattern of pixel values that results from uniform illumination is called "fixed pattern noise" and exists to a slight extent in CCD detectors.

4.2.1 Illumination

An understanding of illumination requires some minimal understanding of the characteristics of light. For our purposes it is convenient to consider light as being a collection of rays that have intensity (watts), wavelength (the visible is between 400 and 700 nm), and a direction in space. Light also has a property of polarization which is ignored here but can be important in MV. Illumination is usually designed to give a high-contrast image, often with a short exposure time. The contrast may be between of actual physical edges in the field of vision (FOV) or may be the result of shadows, glare, or light caused edges. In any case, some object(s) will be intended to be in the FOV and the reflection properties of the object(s) will be important. The exception is backlighting, where the object(s) themselves are not viewed but only their outline or in some cases transparency.

Reflection is characterized by the fraction of the energy of an incident light ray that is reradiated, R, the reflectivity, and the direction of such radiation. There are three idealized types of reflection: diffuse, specular, and retro, as shown in Fig. 7. A diffuse reflection distributes the energy in all directions equally in the sense that the surface is the same brightness, or same pixel values, regardless of the observation angle. A specular reflection is that of a mirror, and hence there is only one direction for the reflected light. A retroreflective surface ideally returns all the light in the direction from which it came. In practice, most surfaces are a combination of these. A strong retroreflection term, with few exceptions, is only found in cases where the surface is made intentionally retroreflective, e.g., road markings for night driving.

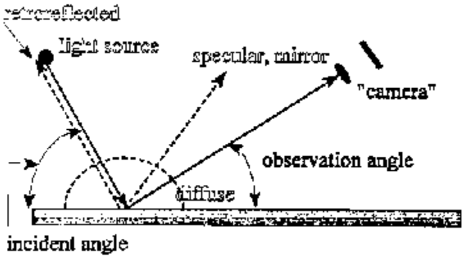


Figure 7 Types of reflection.

Real reflections are never ideal. A substantially diffuse surface will usually be slightly darker when viewed from a shallow angle and both specular and retro surfaces will tend to scatter the light about the ideal direction. Figure 7 shows the typical result of lighting from a single direction in the plane of the input light rays. The reflectivity R, is usually a weak function of the incident angle, but can be a strong function. R is often a strong function of wavelength in the visible region. Without this, everything would appear to the eye as shades of gray.

Now consider the practical implications of the properties of light in enhancing an image for MV purposes. Contrast between physical elements in the FOV can be enhanced by using:

1. *Wavelength*. The relative brightness of two adjacent surfaces and hence the contrast at the edge between the surfaces can be enhanced by a choice of light wavelength. Although here we assume that choice is made at the illumination source, it can be achieved or enhanced by a filter in the optics that blocks wavelengths that are not desired.
2. *Light direction*. This is perhaps the most useful of "tricks" used in MV illumination. A number of different lighting schemes are shown in Fig. 8, including:
   a. *Diffuse or "cloudy day" illumination*. This illumination, if done perfectly (a difficult task), has light intensity incoming at the same intensity from all directions. It almost eliminates the directional effects of reflection, specular or retro, and is extremely useful when viewing scenes with highly specular surfaces.
   b. *Dark-field illumination*. Here light rays are all close to the plane of the scene. It is used to highlight changes in elevation of the scene by creating bright areas where the surface is more nearly normal to the light and dark areas where the light is grazing the surface. The illumination can be all the way around the scene or not.
   c. *Directional illumination*. Similar to dark field but from one side only, at a higher elevation. Used to enhance shadows which can be used to estimate object height relative to a background.

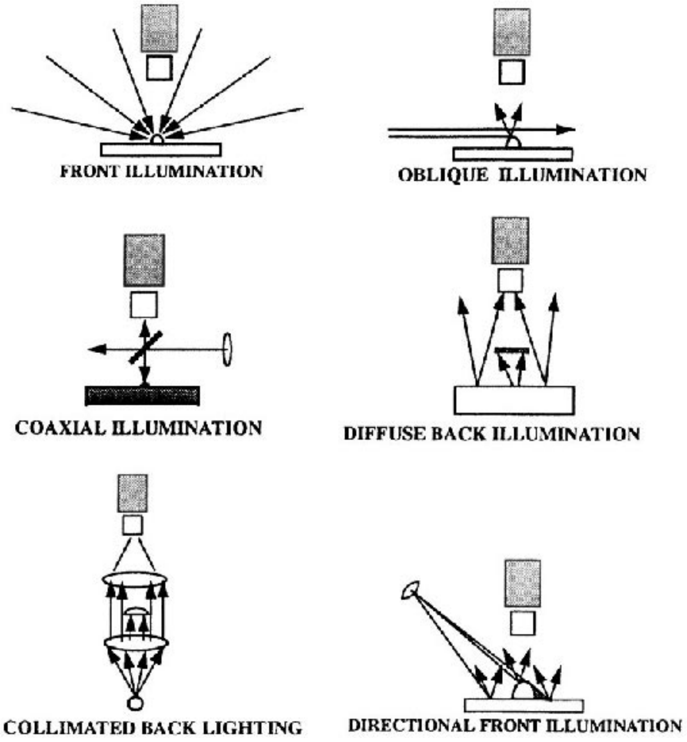


Figure 8 Lighting alternatives.

d. *Coaxial illumination or bright-field illumination*. True coaxial illumination requires a beam splitter in the optics system to direct light along the optical axis. But a ring of light around the optical axis is often used. This illumination is used to eliminate shadows.

e. *Backlighting*. A very effective way to eliminate the effect of reflection characteristics of the object, but only gives the outline of the object. It is also used in some cases where the object is translucent (cloth) or transparent (glass lenses). In the latter case, the backlighting may be very directional and the image depends on changes in the index of refraction to create a pattern of light at the video detector.

In cases b–e, the direction of light is important, but further refinement is possible regarding how much spread in light direction is used. For example, in c, a very collated light would sharpen all edges. In case d, a distributed light source is often desired rather than a very narrow beam or ring of light. This moves it closer to case a.

Useful contrasts can be created by "structuring light." As a last topic in illumination, consider the effect of a projected pattern of lines or points on a scene. If the pattern is projected off the optical axis, this illumination will create a pattern of lines and points in the image which depends on the elevation of the scene where the light strikes. This approach is often used in very sophisticated ways to make 3D profile of the surface. The illumination is often based on laser light.

One of the most significant differences between human vision and MV is the ability to use strobe lighting to advantage in MV. Strobing is useful primarily because it allows (1) greater light intensity and (2) a form of fast shuttering. The greater light intensity is achieved by having the illumination on only during light integration at the CCD. This reduces heating, allows greater energy without destruction of the source, and reduces overall size of the power source cooling, etc. reduction of heating for the same amount of light. Often the light will need to be on only a small fraction of the time. The xenon strobe in particular provides for shuttering down to fractions of a microsecond in extreme cases. This is not convenient mechanically or with the CCD itself.

4.2.1.1 Sources of Illumination

Light-emitting diodes have favorable characteristics for MV and are probably the leading industrial MV source. Their wavelengths tend to match that of CCDs and they have very long life if not overheated. Because they can be turned on and off quickly (microsecond rise and fall times), they are used at very high illumination during CCD light integration and are off otherwise. The current primary disadvantage of LEDs might be the current limitation on high outputs at shorter wavelengths, *if* shorter wavelengths are required for a good image. At shorter wavelengths, because of the CCD characteristics, more power is required and LEDs are less able to produce that power. Light-emitting diodes are normally packaged with an integral plastic lens that causes the light to be concentrated around a central axis. The cone of light is typically from 10° to 30°. This focusing can also be used to advantage, *but* care must be taken to arrange the LEDs for direct illumination in a way that gives uniform illumination. A single LED tends to have a pattern of light, so many may be needed, or some diffusing may be done, to make the illumination more uniform.

Illumination based on fluorescence is often used. This gives many wavelength combinations and is itself a distributed source, often useful in reducing glare, the effect of specular reflections. Most systems are based on fluorescent tubes, but there is the potential for use of flat panels either of the display technology or electroluminescent technology. When using fluorescent tubes, high-frequency drivers (10–30 kHz) are required in order to avoid the 60 Hz flicker of standard home tubes.

Incandescent light, both halogen and otherwise, are common. These lights are inexpensive and can be driven directly at 60 Hz because of the delay in heating and cooling of the filaments. Because they cannot effectively be strobed, these bulbs tend to be a heat source.

Xenon strobes are measured in joules per exposure with 1 J plus or minus a large range being feasible. Since the light output is often predominantly within 10 msec, the effective input wattage can be of the order of 100,000 W. This is one way to get very short effective exposure times. Exposure times of the order of 0.5 msec are possible.

The illumination sources listed above are feasible for industrial MV in terms of life. Care must be taken to maximize the life. Phosphor-based devices and xenon strobes have a finite life and slowly degrade. Light-emitting diodes seem to have very long service life if not driven beyond their specification. Filaments in incandescent bulbs last much longer if run at lower voltage, that is, cooler. Because CCDs are sensitive to longer wavelengths than people are, cooler operation does not bring the loss of illumination that might be expected.

Figure 9 shows wavelength distributions of interest.

4.2.1.2 Illumination Optics

Fiber optics provide a flexibility for spacial arrangements in the geometrical arrangement of light, including such things as back lights, ring lights, and lines of light. An additional advantage of fiber optics as the delivery source is that the size and heating of the illumination in the region of interest can be reduced. The actual source of the light is removed. However, such arrangements usually increase the total amount of equipment and energy used because of the losses between the actual light source and the output of the fibers.

The optical path of the light can include beam splitters to get input light on the same axis as the optical axis, filters to narrow the light wavelength spectrum, and reflectors. Reflectors with diffuse, highly reflective surfaces are often used to create a distributed uniform light source (much as is the commercial photographer's shop). Most "cloudy day illuminators" are based on such diffuse reflections to provide scene illumination.

4.2.2 Optics

Optics gather, filter, and focus light on the video array. The ability to collect light is usually expressed by the F-stop number, the ratio of the focal length to the opening diameter or "aperture," as illustrated in Fig. 10. Focal length is a function of the actual focusing of the lens. The focal length in use is equal to or greater than the focal length of the lens itself, which assumes the viewed scene is at infinity. The standard formula that allows calculation of the energy level on the detector is

$$E = \frac{\pi L (\cos\alpha)^4}{4(\text{F-stop})^2}$$

where E is the energy flux in watts/square meter, and L is the brightness of the surface being imaged in watts/steradian/square meter. For a perfectly diffuse reflection, L is given by

$$L = RE_s/\pi$$

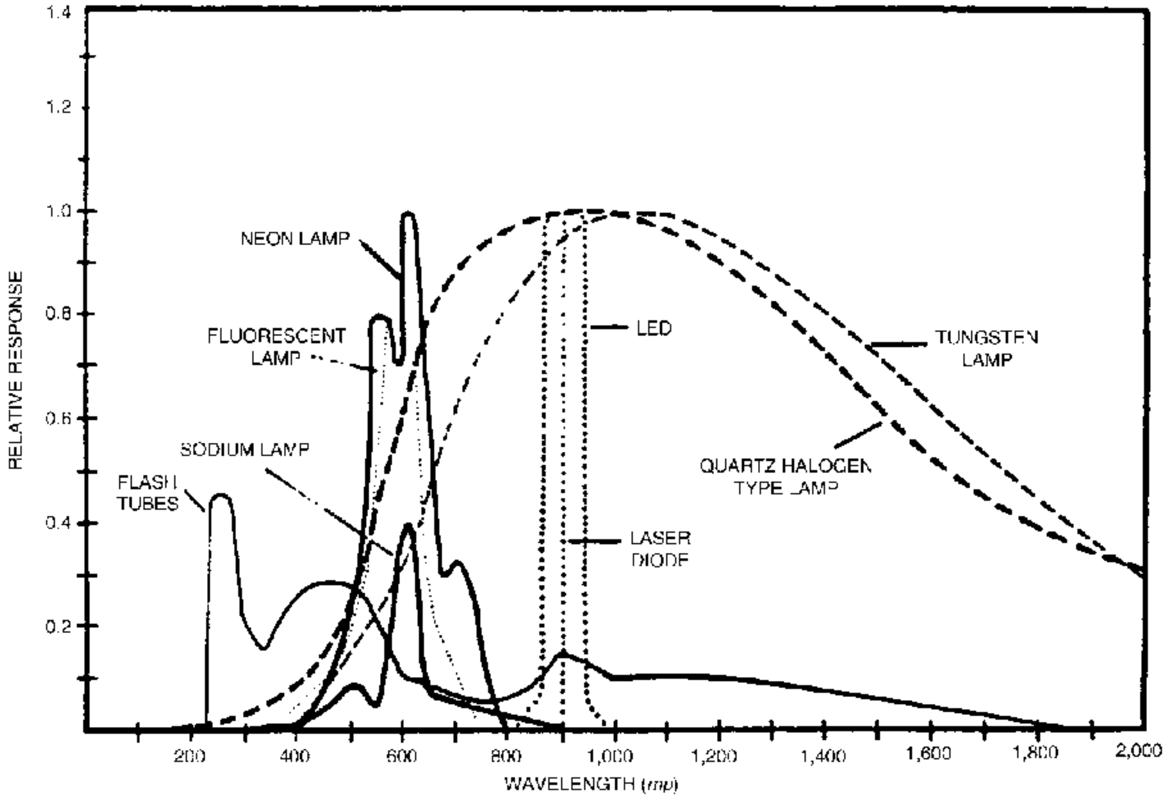


Figure 9 Wavelengths of interest. (From Photonics Handbook.)

where R is reflectivity a unitless ratio less than 1, and E_s is the incident energy level on the surface in watts/square meter. For retro and specular surfaces the effective L can be of the order of 1000 times greater than for diffuse surfaces if both the lighting and observation angles are favorable to high return. One normally wants as much energy as possible to reach the detector. This argues for strong illumination, discussed earlier, and a small F-stop. But a small F-stop leads to a small depth of field, which may or may not be acceptable. The approximate formulas for focus and depth of field are

$$\frac{1}{f} = \frac{1}{d_o} + \frac{1}{d_i}$$

$$\text{depth of field} = \frac{\epsilon(\text{F-stop})}{(\text{magnification})^2}$$

where ϵ is the allowed radius of the light spread of a point in an image. This can be taken as the pixel dimension for most MV applications. The depth of field is plus or minus for nominal object distance, d_o. Surprisingly, MV algorithms can often compensate for and even take advantage of what would appear to be a poorly focused image to the eye. Because of the above depth-of-field equation, high magnification images are harder to focus.

Naturally, the materials used for the optics must be transparent to the wavelengths used. This brings up the possible use of filters to prevent or substantially reduce the transmission to the detector of selected wavelengths. For this reason, filters are used together with illumination for effective wavelength control. They can be used at the illumination end and/or the optics end.

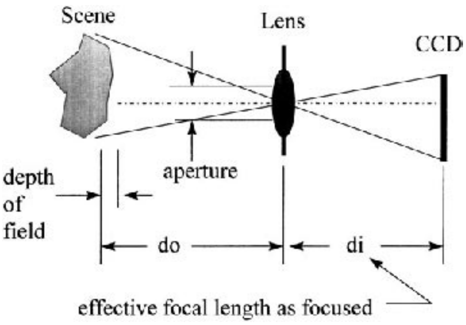


Figure 10 Simple lens optics and imaging.

Color vision based on a single CCD is achieved by filters on the CCD itself, one per pixel. Alternatively, there are commercially three CCD cameras where three different filters are used. This allows full image coverage in three colors, rather than spatially discrete pixels with different filters side by side. Theoretically, there could be two, three, or more different wavelengths used in color imaging, including outside of the visual spectrum. However, for cost reasons, most color systems are designed for human color separation.

There are some special characteristics of images related to optics that one might need to consider. First, there is an intrinsic darkening of an image around the edges, which results from the equation given earlier for light at the detector. Second, there are geometrical distortions caused by color and the overall geometry of the lens–detector combination that are corrected to the extent possible in photographic lenses but may not be totally corrected for in MV applications. These are discussed in Sec. 4.6 on calibration. Keep in mind that, in general, the index of refraction is a function of wavelength, so that lenses may focus light slightly differently at different wavelengths.

4.2.2.1 Telecentric Optics

Telecentric lens systems have the characteristic that the size of the image does not vary as the scene is moved toward or away from the lens, within limits. Thus they have an advantage in making measurements or matching reference shapes in those cases where the image is likely to vary in distance. They also have an advantage in that if the object of interest is in different places in the FOV there is no prospective distortion or occluding of the image. For instance, one can look right down a hole, even if the hole is not on the optical axis. These lens systems still have depth-of-field limitations and light-gathering limitations based on F-stop values. Because the optics diameter must encompass the entire FOV, telecentric optics are most practical from a cost standpoint for smaller FOV, including particularly microscopes, where the image on the CCD is larger than the actual scene.

4.2.3 Video Array

Charge-coupled device arrays in common practice are sensitive to wavelengths between 400 and 1000 nm with a peak sensitivity at approximately 750 nm. "Charge coupled" refers to the technology of the solid-state electronics used. The peak sensitivity of the eyes is about 550 nm and the total range is 350 nm to 750 nm. There is a movement to CMOS devices for low-cost consumer video systems which will probably have some effect on MV in the near future. CMOS devices can be built using the same processes used to make most consumer electronics today and allow the convenient inclusion of A/D converters, microprocessors, etc. on the same chip as the video array.

Some of the important features of a video array are:

1. The sensitivity to light
2. The ability to electronically shutter light without an actual mechanical or electronic shutter
3. The uniformity of light sensitivity between pixels
4. The rate at which pixels can be shifted out
5. The noise or randomness in pixel values.

Before making some generalizations in these regards it it useful to understand how a CCD works, or appears to work, for MV applications. Referring to Fig. 11, the process of creating an image is as follows.

1. Light strikes an array of photosensitive sites. These are typically of the order of 10 μm square. However, the actual photosensitive area may not really be square or cover 100% of the detector area.
2. The incoming photons are converted to electronc which, up to a limit, accumulate at the site. If the electrons spill over into adjacent sites, there is an effect called blooming. It is usually possible to electronically prevent blooming but the result is saturated portions

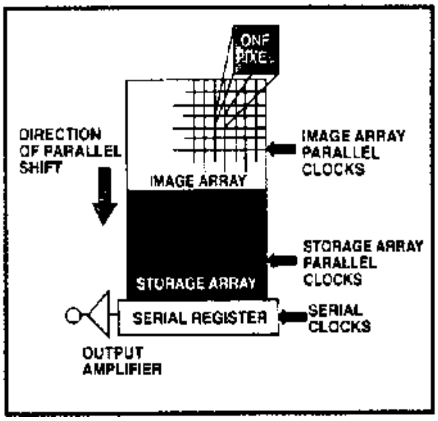

ONE PIXEL
DIRECTION OF PARALLEL SHIFT
IMAGE ARRAY
IMAGE ARRAY PARALLEL CLOCKS
STORAGE ARRAY PARALLEL CLOCKS
STORAGE ARRAY
SERIAL REGISTER
SERIAL CLOCKS
OUTPUT AMPLIFIER

Figure 11 CCD functional arrangement. (From *Sensors*, January 1998, p. 20.)

of the image where the pixel values are no longer proportional to intensity.

3. External signals are used to move the accumulated charges row by row into a single horizontal shift register.
4. The charges are shifted out of the shift register in response to external signals into an amplifier which converts charge to voltage for each pixel. These voltages appear external to the chip as pulses.
5. Because of the timing and number of external signals, the electronics is able to associate each pulse with a physical location and hence cause something like light intensity to be stored as a number.
6. At pixel sizes of 10 μm square and less, it is possible to have detectors that are very small in physical size. This leads to complete cameras, optics, detectors, and some electronics, in the 1 cm diameter range. These small video heads are sometimes useful in industrial processes.

There are complications in actual practice. In particular:

1. Electronic shuttering is achieved by first rapidly shifting out all rows without shifting out the pixel values, so that the light exposure can be started with near-zero charge. After exposure, the rows are rapidly shifted into a separate area of the CCD that is covered, so as to be dark. Then the process 3–5 above is applied to that separate area.
2. Charge does accumulate because of thermal activity. Thus CCDs work better if cooled, but for most MV applications this is not necessary. But it is also necessary to shift out images stored in the dark area in a timely manner to prevent additional charge accumulation. Cooling to 280 K will reduce the "dark current" or rate of charge accumulation to about 4% of that at 330 K. These thermally induced charges are a source of noise in the pixel values and thus a cooled detector can be made to operate effectively at lower light levels, at shorter exposure times, or with greater precision in light level measurement.
3. It is possible to download fractional images from predetermined segments of the image, thus increasing the effective frame rate at the expense of a smaller FOV. In the more sophisticated systems this smaller FOV can be dynamically adjusted, image by image.
4. The effect of row shifts can be thought of as moving a piece of film during exposure. Thus it is possible to compensate for motion in the scene in one direction. This is sometimes used in web inspection or where objects are on a conveyor belt where the scene is moving. The advantage is that effective exposure time is increased. Generally, the CCD in this case has relatively few rows compared to columns. The image is built up by adding rows of pixels in memory. This is the "line scan" approach to imaging, not discussed here.
5. There can be multiple output lines, so multiple rows or partial rows can be output at once. This allows very rapid successive images, sometimes thousands per second, for specialized applications.

Now we state some generalizations:

1. The sensitivity is typically good enough to allow us to build MV systems with exposure times to 1 msec without flashtube strobe lighting. With flashtube strobe lighting the exposure time is determined by the light source. However, even 1 msec exposures may cause considerable expense or care in the illumination. But using LEDs, for example, the light can be on only during exposure. An alternative approach to achieve high speed is very low noise levels from cooling and excellent electronics to allow relatively few accumulated electrons to be used.
2. Electronic shuttering typically is available with image shift times of the order of 1/10,000 sec, although much faster times are available. That is, the charges are cleared in this time and shifted out in this time. True exposure time needs to be several times this shift time.
3. Pixel uniformity is usually such that all pixels have an output less than 5% from nominal. Better or worse grades are achieved by screening the CCDs after production into classes with higher uniformity costing more. Alternatively, a vision system can be calibrated to digitally compensate for nonuniformity provided that there are no pixels that are "dead," interpreted here as 20% below nominal output.
4. Pixels can typically be shifted out of individual registers at 5–20 MHz. However, the A/D conversion electronics must keep up without adding noise. Many traditional cameras that use CCDs or CMOS devices are aimed at 30 Hz

television image rates. If such a camera is used in MV applications the timing and pixel rates are fixed.

5. The randomness in pixel values including the A/D conversion is typically about one level in 256 measured as three standard deviations.

All of the above generalizations are to be thought of as very approximate *and* (1) subject to rapid technological change and (2) may be of little consequence in a particular MV application. For example, if inspecting objects at the rate of one per second to match the production rate, a high-speed imaging and download may be unnecessary.

As a final note, A/D converters used tend to be 8-bit high-speed converters, although 10 and 12 bits are also used where fine gray-scale differentiation is required. There is a speed and cost penalty for more bits. Although A/D detectors may have excellent specifications in terms of linearity, one will find that some numbers are much more likely than nearby numbers. For example, 154 may occur much more frequently than 153 or 155. This is because the voltage input range needed to get 154 is wider than the others. The effect of this can be thought of as more noise in the pixel values.

4.3 IMAGE PROCESSING FOR MV

4.3.1 Objective

The objective of MV image processing is to convert the array of pixel values, the data in this process, to an output value or values. This process is often one of taking massive amounts of data and reducing it to a binary result, good or bad. To show how massive the data is consider an image that is 256 rows by 256 columns, where each value can be between 0 and 255. The resulting number of possible images is $256^{65,536}$, a number that exceeds the number of atoms in the solar system many times over.

Often the output is more complex than simply "good or bad," but by comparison to the input, still very simple. For example the UPC bar code reading function would give 10 numbers, each between 0 and 9, as the output. Other numerical results are things like dimensions (often expressed in fractional pixels for later conversion), and presence or absence of a number of items.

4.3.2 Fundamental Ideas—Segmentation, Feature Vector, Decision Making

The process of going from an image array to the output typically has the following elements:

1. Locate an object(s) in the image, called segmentation.
2. Create a set of numbers that describes the object(s), called a feature vector.
3. Calculate the implications of those features to create the output.

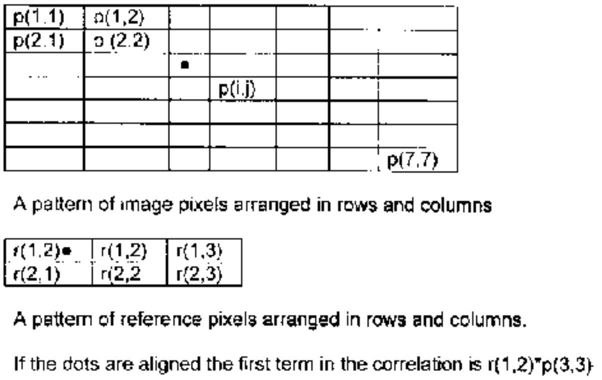
At the risk of what would appear to be terrible simplification there are only a few fundamental ideas that allow one to do these three steps. The first of these is correlation applied to patterns of pixels in an image. This is the main tool for segmentation. Once segmented, the properties are calculated from the lines and areas of pixels that define the objects using a number of measures that describe the object. Lastly, these numbers are used to make a decision or create on output based on either engineering judgment (an "expert" system built into the MV system) and/or on the basis of matching experimental data (of which a neural net would be an extreme example).

In choosing an algorithm to implement, speed of computation may be important. It is often tempting to use algorithms that have extremely high computational requirements partly because the programming is simpler, there is so much data, that is, hundreds of thousands of pixels values, and the generic task of pattern matching tends to blow up as the size of the pattern gets larger. Fortunately, the trend in microprocessors of all types is toward greater computational ability for the dollar. Thus computational cost is not the constraint that it has been in MV applications.

4.3.3 Correlation

Machine vision is all about "shape." That is, the information in an image is related to the patterns of pixel values that occur. Given a pattern of pixel values, such as shown in Fig. 12 the question is often, "does that pattern match a reference pattern?" The answer is based on the value of the correlation between the two patterns where $p(i,j)$ represents the pixel values and $r(i,j)$ represents the reference pattern. The simplest expression of correlation is

$$\text{Correlation} = \sum\sum p(i,j)\, r(i,j)$$



A pattern of image pixels arranged in rows and columns

A pattern of reference pixels arranged in rows and columns.

If the dots are aligned the first term in the correlation is r(1,2)*p(3,3)

Figure 12 Pattern of pixel values and reference pattern.

All double sums are taken over the rows and columns of the reference pattern, where the row and column numbers of pixel patterns are assumed shifted to match the current position of the reference pattern. The dots in the figures are points of alignment.

Suppose that $p(i,j) = a + b\ (r,j)$, where b is positive. Then one would say that both patterns had the same "shape," although we have changed average level and scaled. Under this circumstance the expression above has an absolute maximum relative to any rearrangement of the terms. For example,

$$\sum\sum p(j,i)r(i,j) \leq \sum\sum p(i,j)r(i,j)$$

Thus we can tell when one pattern matches another in a local sense when we have a relative maximum in the correlation. If the reference pattern were shifted a little on the pixel pattern we would expect the correlation to decrease. Correlation can be made more easy to interpret by the following form:

$$\text{Normalized correlation} = \frac{\sum\sum(p(i,j) - \boldsymbol{p})(r(i,j) - \boldsymbol{r})}{\sqrt{\sum\sum(p(i,j) - \boldsymbol{p})^2 \sum\sum(r(i,j) - \boldsymbol{r})^2}}$$

where $\boldsymbol{p}$ and $\boldsymbol{r}$ are the average values of the sets $p(i,j)$ and $r(i,j)$. This expression is exactly 1 when $p(i,j) = a + b\ r(i,j), b > 0$. It is zero when "there is no correlation" and -1 when the pixel pattern is exactly a "negative image," that is, the image values are least where the reference values are greatest or $p(i,j) = a - b\ r(i,j), b > 0$. The expression is called the normalized correlation and is the basis of many image processing algorithms because it seems to tell us when we match a shape. Note that the value of normalized correlation would not change if we left out either, but not both, of the terms $\boldsymbol{p}$ and $\boldsymbol{r}$ in the numerator.

Some comments are in order about what actually happens when this is used.

1. For practical purposes, the value of 1 is never achieved, so we must do a local optimization and then compare to a threshold or sometimes to all other local optimizations.
2. If the pattern is large, the computation is particularly intensive.
3. If the pattern is large, statement 1 above becomes more problematical because it is less likely in the real image that the correlation will be a good discriminator of shape. This happens because, over a larger region, intensity values vary from the norm, the shape may grow or shrink relative to the reference, and most important, a small feature within the shape may be out of tolerance but it contributes only a small amount to the correlation and is lost in the noise.
4. Although the normalized correlation detects shape, the normalization process itself eliminates the contrast of the shape with its background. This can result in very faint shapes having accidentally high correlations. This is especially likely to happen for small reference patterns.

None the less, the fundamental concept of correlation is extremely valuable and universally used in a local sense. Localized correlation can be used to detect edges quite reliably and it is usually the edges in an image that are the determinant of shape, including both edges that should be present and those that should not. Other local features can also be detected, e.g., corners, small holes, and "roughness" which help determine the features of the segmented image.

4.3.4 Edges—Location and Subpixelization

Edges are detected by simple correlations often using reference images, called masks (Fig. 13). These patterns are very small versions of a piece of an edge. They all have the property that the mean value is zero because that eliminates the need to use the mean value in the normalized correlation equation. They also use 0, 1, and 2 as that makes the multiplications not required (0), trivial (1), and a bit shift (2). When these simple masks are used, the denominator of the equation is not used, since we not only want to detect a strong correlation but a large value of contrast. The masks are actually giving a measure of gradient or derivative in a particular direction. Note also that there is no need to compute ***p***, since ***r*** is zero.

Figure 14 is a sample of a small portion of an image that contains an edge running more or less from lower left to upper right. The table in Fig. 14 shows the results of these masks applied in two places in the image. The dots in both the masks and image are aligned when making the calculations. As an example of the calculation.

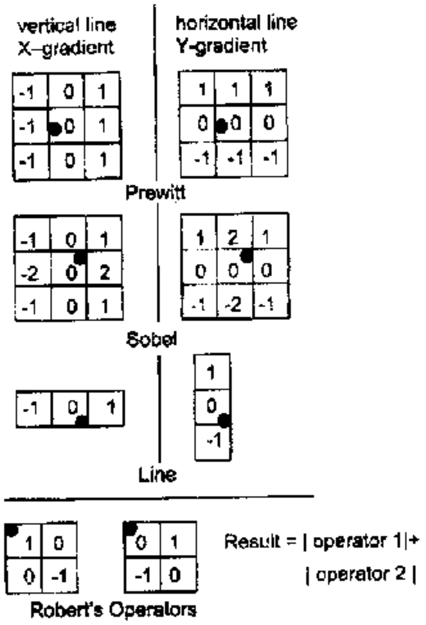
Now let us consider the possibility of correlation not just at the discrete pixel level but at a subpixel resolution. This goes by the name of subpixelization. This technique is valuable when the MV system is to make a precise measurement. This technique can be used at individual edge crossing as well for patterns in general. The basic steps are as follows.

1. Find a local maximum correlation.
2. Examine the value of correlation at nearby pixel values.
3. Fit the correlations to an assumed *continuous* function that is presumed to describe the local maximum value of correlation.
4. Calculate the maximum of the function which will be a subpixel value.
5. Correct if necessary or possible for known biases in the estimate.

Here is a very simple example, which is, even though simple, very useful. Assume the "[1 × 3] gradient operator" of Fig. 13. If that operator is moved to the points labeled 1, 2, and 3 in Fig. 14, the correlation values are. The middle value is the greatest. Assume the three points fit a parabola, the simplest possible polynomial which has a maximum. It is useful to make the row values considered temporarily to be labeled −1, 0, and 1. A formula for the maximum is of a parabola with three data points at −1, 0, and 1 is

$$x = \frac{g_1 - g_{-1}}{2(g_1 - 2g_0 + g_{-1})}$$

This formula for a sharp edge, substantially all the change in pixel values within five pixels, and a large contrast, greater than 100 on a scale of 0–255, would give a repeatability of about 0.02 pixels; however, it would have a systematic error with a period of one pixel that would cause the standard deviation about the true value to be about 0.03 pixel. Some of this systematic error could be removed using an experimentally learned correction or by a more sophisticated expression for the assumed shape of the correlation values.



-1	0	1
-1	0	1
-1	0	1

1	1	1
0	0	0
-1	-1	-1

-1	0	1
-2	0	2
-1	0	1

1	2	1
0	0	0
-1	-2	-1

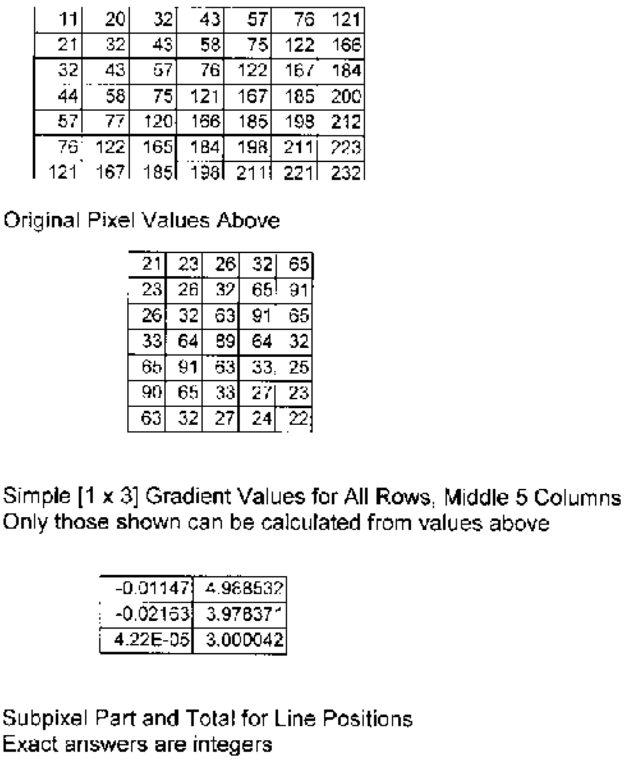
-1	0	1

1
0
-1

1	0
0	-1

0	1
-1	0

Figure 13 Various gradient masks: Prewitt, Sobel, Line, and Roberts.



11	20	32	43	57	76	121
21	32	43	58	75	122	166
32	43	57	76	122	167	184
44	58	75	121	167	185	200
57	77	120	166	185	198	212
76	122	165	184	198	211	223
121	167	185	198	211	221	232

Original Pixel Values Above

21	23	26	32	65
23	26	32	65	91
26	32	63	91	65
33	64	89	64	32
65	91	63	33	25
90	65	33	27	23
63	32	27	24	22

Simple [1 x 3] Gradient Values for All Rows, Middle 5 Columns
Only those shown can be calculated from values above

-0.01147	4.968532
-0.02163	3.978371
4.22E-05	3.000042

Subpixel Part and Total for Line Positions
Exact answers are integers

Figure 14 Image with an edge and results of gradient mask.

Some observations on the use of such techniques are:

1. If a large number of edge points are used to determine the location of line, center, and diameter of a circle, etc., then the error can be reduced significantly by an averaging effect. In particular, if each measurement of edge position is independent in its noise then a square-root law applies. For instance, for four averaged measurements the expected error is reduced by 2.
2. The measured edge position may not be the same as the edge one had in mind. For example, if a physical edge has a radius, the illumination and optics may result in the maximum rate of change of the light level, the optical edge, being different than an edge one has in mind, e.g., one of the two places where the arc stops. This can often be compensated in the illumination, or in the processing, or in a calibration, but must be remembered.
3. Position measurements are enhanced by contrast, the total change in light level between two areas or across and edge, and sharpness, the rate of change of the light level. It is theoretically possible to have too sharp an edge for good subpixelization, that is, the light level change is entirely from one pixel to the next.

A great deal of information can be derived from finding and locating edges. Generally a closed edge, often found by tracking around an object, can be used to quickly calculate various features of an object, such as area, moments, and perimeter length, and geometrical features, such as corners and orientation. Individual line segments have properties such as orientation, curvature, and length. Most image processing can be accomplished by judging which features, that is, numerical representation of shape, texture, contrast, etc., are important to arrive at results, and then using those features to make various conclusions. These conclusions are both numerical, the size of a part, and binary, good or no good.

4.3.5 Feature Vector Processing

In industrial MV systems, processing of features tends to have an engineering basis. That is, there is is a set of equations applied that often has a geometrical interpretation related to the image. Part of the reason for this is that applications tend to have requirements that they be correct a very high percentage of the time, e.g., reject a bad part with 99.9% reliability and reject a good part only 0.1% of the time (and in most of those cases only because the part was marginal). Likewise measurements need to have the tight tolerances associated with more traditional means.

However, most MV processing has a learned or experimental component. This is the result of the uncertainties in the system, particularly those associated with lighting, reflectivity, and optical arrangement. Thus almost all MV algorithms are dependent on observing sample scenes and using either expert systems or manual intervention to set thresholds and nominal values used in the processing.

One can expect that the use of sophisticated learning will be an increasing part of MV applications. These learning techniques have the objective of modifying the decision and measurement processes to converge on more reliable results, including adjustment for drift in MV system properties. Usually these learning algorithms take a model with unknown parameters and learn the parameters to use. Popular models are:

	Unknown parameters
Expert systems	Mostly gains and thresholds
Neural nets	Neuron gains and offsets
Nearest neighbor	Distance function parameters
Fuzzy logic	Thresholds

In every case, the designer or user of the system must make a judgment, hopefully backed by experience, as to what model will work well. For example, the neuron functions, number of layers, and number of neurons in each layer must be picked for a neural network. The same judgment must be made concerning what feature vector will be processed.

Refer to Chap. 5.2 for more advanced image processing techniques.

4.4 INTERFACING AN MV SYSTEM

An industrial MV system is intended to be used as a rather complex sensor, just like a photo-eye or linear encoder. As such it must communicate with the controller of a machine, a cell, or a production line. Traditionally, this communication has been through the PLC discrete I/O lines or as a discrete signal between elements of the production process.

For example, a classic implementation would have a photocell send a signal, e.g., one that pulls low a 20 V line, to the MV system to indicate that a part to be measured/inspected is in the FOV. The MV system would respond in a few milliseconds by acquiring an image. This may require the image system to energize illumination during image acquisition, done through a separate signal. The image would be analyzed and, if the part is bad, the MV system would output a 120 V AC signal, which causes the part to be pushed off the line into a rejection bin.

But if that was all the MV system did, it would not have been utilized to its potential. What was the actual defect? Even if not defective, are certain tolerances, e.g., dimensions, surface finish, quality of printing, deteriorating? This additional information is useful in providing feedback for control of the process both in the short run and as input to a statistical process control analysis. In addition, what if the parts to be inspected changed often, it would be useful to tell the MV system that the intended part has changed (although the MV system might be able to surmise from the image itself). All of this leads to the trend toward networked sensors. MV systems in particular require more data transfer, because they tend to be able to generate more information. In MV a picture is not worth a thousand words, but maybe 50.

The trend is clearly toward serial communications for networks in production facilities and in machines. A current difficulty is related to the number of different systems in use and rapid changes in the technology. Here are a few candidates:

	Bit/sec	No. of senders	No. of receivers	Distance (m)
Device Net	125K–500K	64	64	500–100
PROFIBUS	9.6K–12M	32	32	1200–100
CAN	(uses Device net)			
FBUS	2.5M	32	32	750
RS-232	115K	1	1	15
RS-422	10M	1	10	1200
RS-485	10M	10	10	1200
USB	12M	1	127	4
Ethernet				
Firewire				

These standards are not independent in implementation. That is, some build on others. The RS designations generally refer to the electrical specifications, while the others provide the structure of the messaging.

In production facilities a critical feature of communication system is low delays (latency) in communications. For many messages, the low delay is more important than data rates measured in bits per second. That is, unless real-time images are to be sent, the data rates associated with all of these standards would normally be adequate and the choice would depend on maximum delays in sending messages. An exception are those cases where for some reason, real-time or large numbers of images need to be transmitted. Then the higher data rates of Ethernet or Firewire become important.

4.5 SPEED AND COST CONSIDERATIONS

Here is a theoretical example of an MV industrial application that is configured to represent a highly integrated MV system with a fair amount of structure in the application. Structure is a knowledge of what is likely to be present and how to find the pixels of interest in making the inspection or measurement.

Light integration	2 ms	fair strobed illumination
Pixel download	6.6	256 × 256 pixels at 10 MHz, may not be full frame
Pixel processing	7.9	4% of all pixels, average of 30 ops/pixel, at 10 MIPS
Communications	1.7	20 bytes at 115 kbaud, standard max. rate for a PC RS-232
Total delay	18.2	from ready to image until result available
Max. frame rate	127 Hz	based on 7.9 msec pixel processing
Max. frame rate	55 Hz	based on 18.2 msec total delay

Note that with a typical CCD, it is possible to overlap the light integration, and the pixel downloading from the dark region. The processor though involved in controlling the CCD and communications, is able to spend negligible time on these tasks because other hardware is simply set up to do the tasks and generate interrupts when the processor must take an action. This example points out the distinct possibility that the maximum rate of image processing *could* be governed by any of the four processes and the possible advantage of overlapping when high frame rate is important.

The cost of MV systems for industrial applications can be expected to drop drastically in the future, to the point that it is a minor consideration relative to the implementation costs. Implementation costs include "programming" for the task, arranging for mounting, part presentation, illumination, and electronic interfacing with the production process.

Regarding the programming, almost all MV systems are programmed from a graphical interface which relieves the user from any need to program. However, there is always a tradeoff between generality of the tasks that can be done and the expertise needed by the user. Graphical interfaces and online documentation are being continuously improved. For the totally integrated MV systems that act as stand-alone sensors, a PC is usually used for setup and is then removed from direct connection. This is a trend in "smart sensors" generally.

Regarding the interfacing with the production process, there is a trend toward "plug and play" interfaces. Here the characteristics of the network protocol and the host computer software allow the identification and communications to be automatically set up by simply plugging in the device to the network. The USB and Firewire standards are examples.

4.6 VISION SYSTEM CALIBRATION

An MV system uses 2D images of 3D scenes, just as does the human eye. This brings up the need to relate the dimensions in the 3D "object space" to the 2D image in the "image space." This is a rather straightforward problem in geometrical transformations that will not be presented here. Two excellent papers are by Tsai [4] and Shih [5]. They both give a process which allows the analysis of an image in the object space to determine the relationship between the MV system pixel measurements and the object space. Usually techniques of this sophistication are not required. However, there are considerations that are of interest when *precision measurements* are made.

1. The vision array, e.g., CCD, has an image plane that is unknown with respect to the mounting surface of the camera. The same is true of the effective plane of the optics. The true optical axis is thus unknown. Vision systems could be furnished with a calibrated relationship but usually are not.
2. The effective focal length, which determines the image magnification, changes when the camera is focused. Thus part of the calibration process is only valid for a fixed focus. In many MV applications this is OK, since the application is dedicated and the focal length will not change.
3. An optics system may introduce a significant radial distortion. That is, the simple model of light passing through the center of the optics straight to the detector is often slightly flawed. The actual model is that the light ray bends slightly, and that bend angle is larger, positive or negative, as the incoming ray deviates further from the optical axis. This effect is generally larger with wide-angle optics. In real situations the total radial distortion is less than one pixel.

Chapter 6.1

The Future of Manufacturing

M. Eugene Merchant
Institute of Advanced Manufacturing Sciences, Cincinnati, Ohio

1.1 INTRODUCTION

Since the past is a springboard to the future, a brief review of major trends in manufacturing over the years, from the Industrial Revolution to the present, provides our springboard.

1.2 THE BEGINNINGS

The Industrial Revolution spawned organized manufacturing activity, in the form of *small* manufacturing companies. In such small, closely knit companies, every member of the organization could, face-to-face, communicate and co-operate quite freely and easily with every other member of that entity in carrying out the various functions involved in its overall operation. This situation was ideal for engendering manufacturing excellence. That is because the basic key to enabling a manufacturing organization to perform its product realization function most effectively is empowerment of every entity (people and equipment) in it to be *able* and *willing* to *communicate* and *co-operate* fully with every other entity in the organization.

However, as factories grew in size, operating a company in such a manner became more and more difficult, leading to the establishment of functional departments within a company. But the unfortunate result of this was that communication and co-operation between personnel in different departments was not only poor but difficult. Thus as companies grew in size, personnel in each department gradually became more and more isolated from those in the others. This situation finally led to a ‘‘bits-and-pieces’’ approach to the creation of products, throughout the manufacturing industry.

1.3 A WATERSHED EVENT

Then, in the 1950s, there occurred a technological event having major potential to change that situation, namely, the invention of the digital computer. This was indeed a watershed event for manufacturing, though not recognized as such at the time. However, by the 1960s, as digital computer technology gradually began to be applied to manufacturing in various ways (as, for example, in the form of numerical control of machine tools) the potential of the digital computer for aiding and perhaps revolutionizing manufacturing slowly began to be understood. It gradually began to be recognized as an extremely powerful tool—a *systems* tool—capable of integrating manufacturing's former ‘‘bits-and-pieces.’’ This recognition spawned a new understanding of the nature of manufacturing, namely that manufacturing *is* fundamentally a system. Thus, with the aid of the digital computer, it should be possible to operate it as such.

Out of this recognition grew a wholly new concept, namely that of the computer integrated manufacturing

(CIM) system, having the capability not only to flexibly *automate* and online *optimize* manufacturing, but also to *integrate* it and thus operate it as a system. By the end of the 1960s this concept had led to initial understanding of the basic components of the CIM system and their interrelationship, as illustrated, for example, in Fig. 1.

1.4 NEW INSIGHT EVOLVES

Emergence of such understanding as the above of the potential of digital computer technology to significantly improve manufacturing's productivity and capabilities resulted in generation of major activity aimed at developing and implementing the application of manufacturing-related computer technology and reducing it to practice in industry, thus reaping its inherent potential benefits. What followed during the 1970s and early 1980s was a long, frustrating struggle to accomplish just that. It is important to note, however, that the focus and thrust of this struggle was almost totally on the *technology* of the system (and not on its human-resource factors). As the struggle progressed, and the technology finally began to be implemented more and more widely in the manufacturing industry, observation of the most successful cases of its reduction to practice began to make clear and substantiate the very substantial benefits which digital computer technology has the potential to bring to manufacturing. The most significant of these were found to be greatly:

Increased product quality
Decreased lead times
Increased worker satisfaction
Increased customer satisfaction
Decreased costs
Increased productivity
Increased flexibility (agility)
Increased product producibility.

However, a puzzling and disturbing situation also emerged, namely, these potential benefits were able to be realized fully by only a few pioneering companies, worldwide! The reason why this should be so was not immediately evident. But by the late 1980s the answer to this puzzle, found by benchmarking the pioneering companies, had finally evolved. It had gradually become clear that while excellent engineering of the *technology* of a system of manufacturing is a *necessary* condition for enabling the system to fully realize the potential benefits of that technology, it is not a *sufficient* condition. The technology will only perform at its full potential if the utilization of the system's *human resources* also is so engineered as to enable all personnel to communicate and co-operate fully with one another. Further, the engineering of those resources must also be done *simultaneously* with the engineering of the application of the technology. Failure to meet any of these necessary conditions defeats the technology!

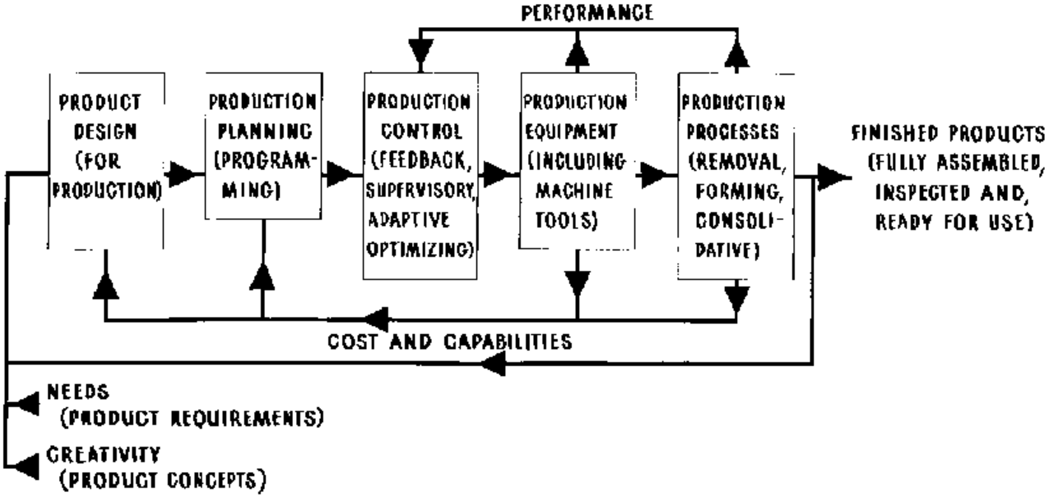

PERFORMANCE
PRODUCT DESIGN (FOR PRODUCTION)
PRODUCTION PLANNING (PROGRAMMING)
PRODUCTION CONTROL (FEEDBACK, SUPERVISORY, ADAPTIVE OPTIMIZING)
PRODUCTION EQUIPMENT (INCLUDING MACHINE TOOLS)
PRODUCTION PROCESSES (REMOVAL, FORMING, CONSOLIDATIVE)
FINISHED PRODUCTS (FULLY ASSEMBLED, INSPECTED AND, READY FOR USE)
COST AND CAPABILITIES
NEEDS (PRODUCT REQUIREMENTS)
CREATIVITY (PRODUCT CONCEPTS)

Figure 1 Initial concept of the computer-integrated manufacturing system, 1969.

1.5 A NEW APPROACH TO THE ENGINEERING OF MANUFACTURING EMERGES

It is evident that this finding requires a new approach to be taken to the overall process of the engineering of modern systems of manufacturing (i.e., manufacturing enterprises). This approach to such engineering requires that proper utilization of the human resources of the system must be engineered, along with the engineering of its technology. Further, the two must be engineered simultaneously. Many of the features of this new approach began to be recognized early on, as made evident in the "wheel-type" diagram of the computer integrated manufacturing enterprise developed by the Computer Automated Systems Association of the Society of Manufacturing Engineers in 1993, reproduced as Fig. 2. However, it is only in the years since 1996 that this new approach has emerged in full.

This emerging new approach to the engineering of manufacturing has brought with it a significant present and future challenge, namely that of developing methodology for accomplishment of effective engineering of the utilization of human resources in manufacturing enterprises. Efforts to develop such methodology are

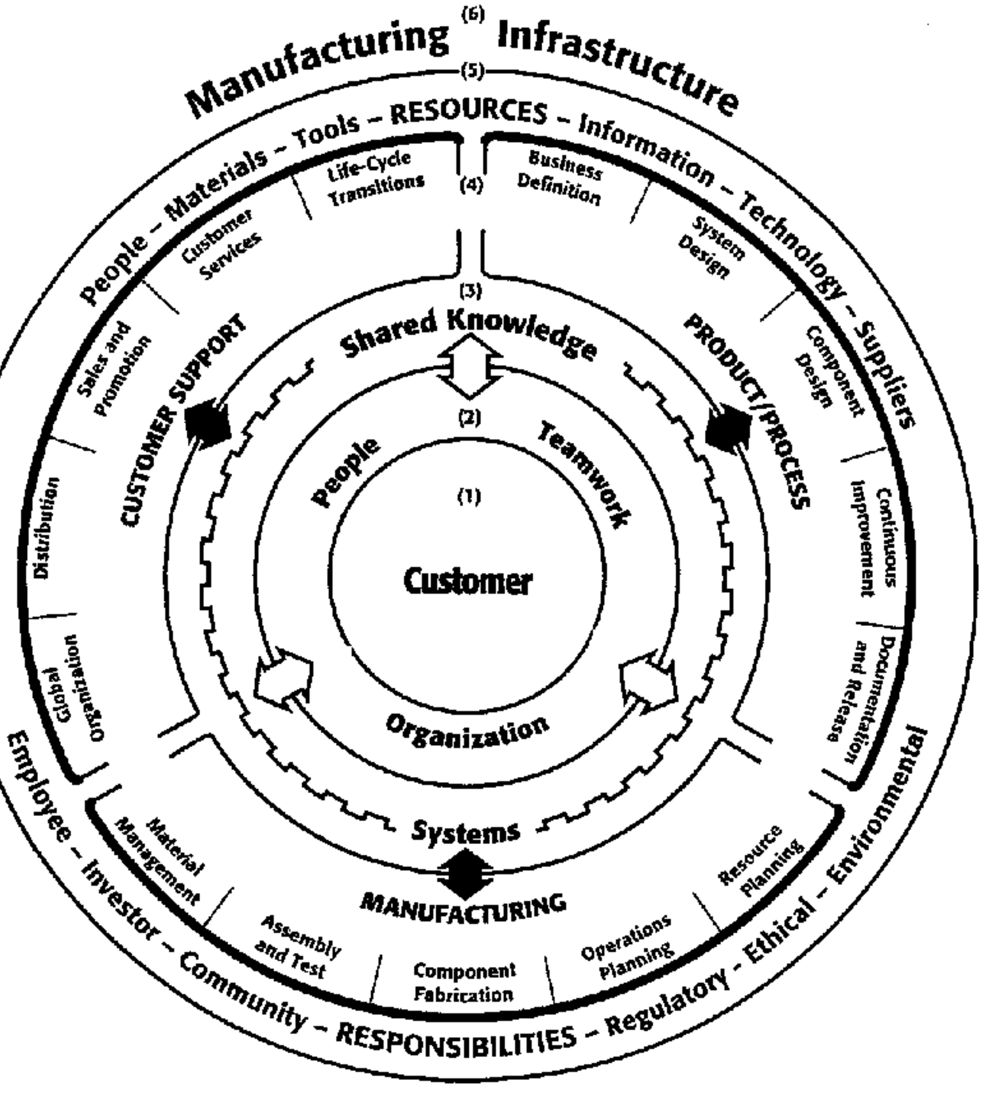


Figure 2 CASA/SME manufacturing enterprise wheel.

of course already underway. Some of the more effective methodologies which have already emerged and been put into practice include:

Empower individuals with the full authority and knowledge necessary to the carrying out of their responsibilities.

Use empowered *multidisciplinary teams* (both managerial and operational) to carry out the functions required to realize products.

Empower a company's collective human resources to fully *communicate* and *co-operate with* each other.

Further, an important principle underlying the *joint* engineering of the technology *and* the utilization of it in modern systems of manufacturing has recently become apparent. This can be stated as follows:

So develop and apply the technology that it will support the *user*, rather than, that the user will have to support the *technology*.

However, these methodologies have barely scratched the surface. Continuation and expansion of research on this subject is an ongoing and long-term need.

1.6 WHERE WE ARE TODAY

As a result of the evolution over the years of the technology and social structure of manufacturing, as briefly described in the foregoing, we are now at a state where:

1. Manufacturing enterprises, both large and small, are rapidly learning how to achieve a high degree of integration of their equipment, people and overall operation, both locally and globally, through utilization of advanced digital computer technology.
2. Such organizations are also beginning to discover how to so engineer both the technology and the utilization of their human resources in such integrated systems that both that technology and the organization's people are able to perform at their full potential.

Further, the development of digital computer technology is now advancing very rapidly in at least three main areas of major importance to the operation of future manufacturing enterprises, namely:

1. *Holonic systems*. These are systems of autonomous entities which, despite the fact that they are autonomous, are enabled to both *communicate* and *co-operate* with all the other entities in the system. The application of this technology to manufacturing systems is currently essentially experimental, but shows considerable potential for enhancing the performance of such systems.
2. *Virtual reality*. This technology is already being applied on a small scale in manufacturing systems, but still in only a rudimentary way. Even so, it shows great promise.
3. *Intelligent systems*. At this stage, the degree of intelligence which has been developed and demonstrated in manufacturing systems still represents only a very small fraction of its true potential. However, it is important to bear in mind that a large-scale international co-operative program, known as the intelligent manufacturing systems (IMS) program, is currently underway among the major industrialized countries of the world, aimed at significantly advancing that potential.

This overall picture of where we are today contains at least tenuous clues as to what manufacturing may be like in the future. Nevertheless, it has led us to the conclusion that future *manufacturing enterprises* and *manufacturing technologies* may well have such characteristics as are set forth in what follows below.

1.7 THE FUTURE MANUFACTURING ENTERPRISE

The manufacturing enterprise of the future will be a virtual enterprise comprising an integrated global *holonic* system of autonomous units, both large and small, located in various places throughout the world. The fact that the system is holonic is its key feature. That fact means that every entity of the system (people, machines, software elements, etc., including its external suppliers, customers, and other stakeholders) within or associated with each of its units will be enabled and empowered to both fully *communicate* and fully *co-operate* with one another, for the purpose of attaining a common goal (or goals).

The autonomous units making up such an enterprise will resemble conventional companies in a general way, but, in addition to a core unit, they will consist mainly of semispecialized units having special skills necessary to the attainment of one (or more) of the enterprise's current goals. Thus they will be the principal elements of the supply chain required for the attainment of those goals. However, the composition of the overall enterprise will be dynamic, changing as new goals are chosen. Furthermore, to be accepted as a "member" of a given enterprise, a unit will have to negotiate "employment" in it, based not only on its special skills but also on the entire spectrum of its capabilities. "Employment" will terminate if it fails to perform as required, or as new goals are chosen for the attainment of which it has failed to prepare itself.

The operation of the product realization process in such global manufacturing enterprises will be based on *concurrently* engineering both the *technology* required to carry out that product realization process and the *utilization of the human resources* required to carry out that same process, to enable both the technology and the human resources to perform at their full *joint* (synergistic) potential.

1.8 FUTURE MANUFACTURING TECHNOLOGIES

It is obvious from the foregoing that a wealth of new or improved technologies will be needed to accomplish full realization of the future manufacturing enterprise as described above. In particular, two main types of technology will need considerable development. These are:

1. Technologies to enable the enterprise to be holonic.
2. Technologies to enable the enterprise to be effectively *managed.*

Concerning the *first*, these are technologies needed to enable and empower every entity (persons, machines, software systems, etc.) to both fully communicate and fully co-operate online and in real time, with every other entity of the enterprise (including its external suppliers, customers and other stakeholders) and to do so wherever they are, worldwide. The ultimate need is to enable such communication and co-operation to be of a character that is equal to that possible if they are in the same room and face-to-face with each other. First of all, this will require that the technology have the capability to flawlessly transfer and share between persons not only information, but also knowledge, understanding and intent. Here (taking a "bluesky" type of approach for a moment), development of technology that can provide capability for mind-to-mind communication would be the ultimate goal. Secondly, to fully enable such communication and co-operation between all entities (persons, machines, software, etc.) will require capability to fully replicate the environment of a distant site at the site which must join in the physical action required for co-operation. Here, development of the emerging technologies associated with virtual reality is a must.

Concerning the *second* of the two types of needed technology, referred to above, the major problems to be dealt with in enabling the future enterprise to be effectively managed arise from two sources. The first of these is the uncertainty engendered by the sheer complexity of the system. The second is the fact that (like all sizable systems of manufacturing) the future enterprise is, inherently, a nondeterministic system. This comes about because systems of manufacturing have to interface with the world's economic, political, and social systems (as well as with individual human beings), all of which are nondeterministic. This results in a high degree of uncertainty in the performance of such systems, which, when no other measures prove able to handle it, is dealt with by exercise of human intuition and inference. The technology which shows greatest promise for dealing with this inherent uncertainty is that of artificial-intelligence-type technology. This will, in particular, need to be developed to provide capability for performance of powerful intuition and inference which far exceeds that of humans.

1.9 CONCLUSION

It seems evident from a review of the evolution of manufacturing from its beginnings to the present, that, under the impact of today's rapidly advancing computer technology, major changes for the better still lie ahead for manufacturing. It can be expected that the global manufacturing enterprises which are evolving today will unfold into holonic systems in which all entities (people, machines, software elements, etc.) will be enabled to communicate and co-operate with each other globally as fully as though they were in the same room together. Further, the composition of the enterprises themselves will consist of semispecialized units which compete and negotiate for "membership" in a given enterprise. The operation of the product realization process in such global manufacturing enterprises will be based on integration of the engineering of the technology required to carry out that process with the engineering of the utilization of the human resources required to carry out that same process, to enable both the technology and the human resources to perform at their full joint (synergized) potential.

Chapter 6.2

Manufacturing Systems

Jon Marvel
Grand Valley State University, Grand Rapids, Michigan

Ken Bloemer
Ethicon Endo-Surgery Inc., Cincinnati, Ohio

2.1 INTRODUCTION

This chapter provides an overview of manufacturing systems. This material is particularly relevant to organizations considering automation because it is always advisable to first streamline and optimize operations prior to automation. Many automation attempts have had less than transformational results because they focused on automating existing processes without re-engineering them first. This was particularly evident with the massive introduction of robots in the automobile industry in the 1970s and early 1980s. Automation, in the form of robots, was introduced into existing production lines, essentially replacing labor with mechanization. This resulted in only marginal returns on a massive capital investment. Therefore, the authors present manufacturing techniques and philosophies intended to encourage organizations to first simplify and eliminate non-value-added elements prior to considering automation.

This chapter begins with a categorization of the various types of manufacturing strategies from make-to-stock through engineer-to-order. This is followed by a discussion covering the spectrum of manufacturing systems including job shops, project shops, cellular manufacturing systems, and flow lines. The primary content of this chapter deals with current manufacturing techniques. Here readers are introduced to the concepts of push versus pull systems and contemporary manufacturing philosophies including just in time, theory of constraints, and synchronous and flow manufacturing. Lastly, the authors present several world-class manufacturing metrics which may be useful for benchmarking purposes.

It is important to note that the term manufacturing system, although sometimes used interchangeably with production system, consists of three interdependent systems. As seen in Fig. 1, the manufacturing system incorporates enterprise support, production, and production support. Production has the prime responsibility to satisfy customer demand in the form of high-quality low-cost products provided in timely manner. The enterprise and production support system provides the organization with the infrastructure to enable production to attain this goal. Many of the manufacturing strategies addressed in this chapter include all three interdependent systems.

2.1.1 Product Positioning Strategies

The manufacturing organization, operating within its manufacturing system, must determine which product positioning strategy is most appropriate to satisfy the market. The product positioning strategy is associated

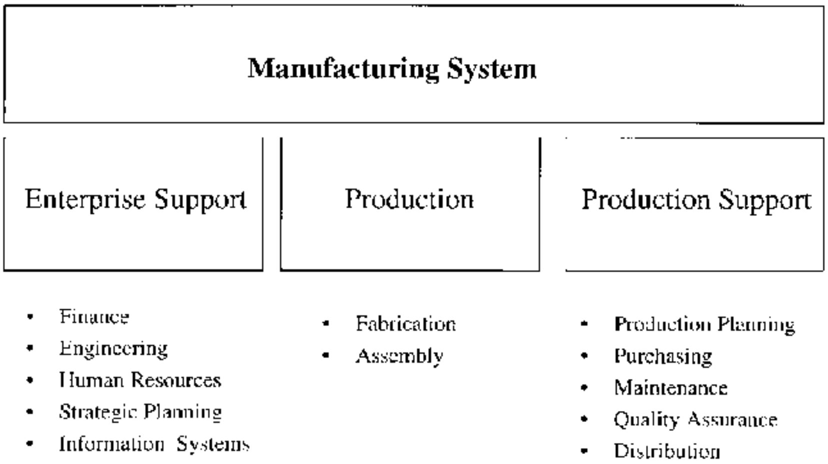


Figure 1 Manufacturing system components.

with the levels and types of inventories that the organization holds. Manufacturing lead time, level of product customization, delivery policy, and market demand are the typical factors which influence the choice of strategies. Organizations will usually follow one or any combination of the following strategies:

1. *Make-to-stock*: a manufacturing system where products are completed and placed into finished goods inventory or placed in a distribution center prior to receiving a customer order. This strategy highlights the immediate availability of standard items. The organization must maintain an adequate stock of finished goods in order to prevent stockouts, since the customers will not accept delays in product availability.
2. *Assemble-to-order*: a manufacturing system where products undergo final assembly after receiving a customer order. Components or subassemblies used for final assembly are purchased, stocked, or planned for production prior to receiving the customer order. This system is able to produce a large variety of final products from standard components and subassemblies with short lead times. This type of system is also known as finished-to-order or packaged-to-order.
3. *Make-to-order*: a manufacturing system where the product is manufactured after a customer has placed an order. In this environment, production must be able to satisfy the demands of individual customers. Longer lead times are usually tolerated, since the product is customized to the customer's specific needs.
4. *Engineer-to-order*: a manufacturing system where the customer order requires engineering design or other degrees of product specialization. A significant amount of the manufacturing lead time is spent in the planning or design stages. The organization receives customer orders based on technical ability to design and produce highly customized products. This type of system is also known as design-to-order.

2.1.2 Product Processing Strategies

2.1.2.1 Job Shop

Job shops (Table 1) are one of the most common types of product processing systems used in the United States today. Machines, typically general purpose, with similar functional or processing capabilities are grouped together as a department. Parts are routed through the different departments via a process plan. This environment satisfies a market for nonstandard or unique products. Products are manufactured in small volumes with high product variety. These types of functional layouts are also referred to as process layouts. Products manufactured in a job shop could include space vehicles, reactor vessels, turbines, or aircraft. An example of a job shop layout, also known as a process layout, is shown in Fig. 2.

As product volumes increase, job shops are transformed into *production job shops*. these types of environments typically require machines with higher production rates in order to regulate medium-size production runs. Machine shops and plastic molding plants are typically classified as production job shops.

Table 1 Job Shop Characteristics

People	Personnel require higher skill levels in order to operate a variety of equipment Personnel are responsible for a diversity of tasks Specialized supervision may be necessary
Machinery	Production and material-handling equipment are multipurpose Machine utilizations are maximized General-purpose equipment requires lower equipment investment Increased flexibility of machinery allows uncomplicated routing manipulation to facilitate even machine loading and accommodate breakdowns
Methods	Product diversity creates jumbled and spaghetti-like flow Lack of coordination between jobs prevents balanced product flow Low demand per product Detailed planning and production control is required to handle variety of products and volumes
Materials	Parts spending a long time in the process creating with high work-in-process inventory Low throughput rates Products run in batches Increased material-handling requirements

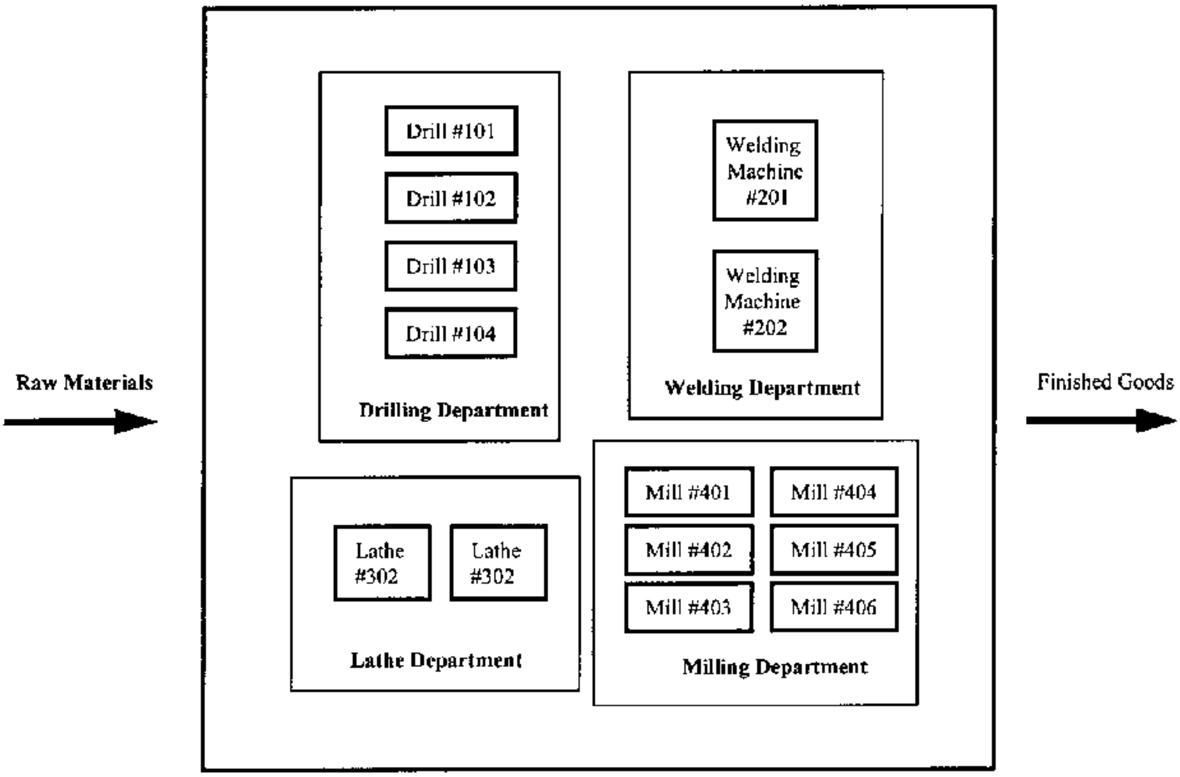


Figure 2 Job shop.

Table 2 Project Shop Characteristics

People	Personnel are highly trained and skilled Opportunities for job enrichment are available General supervision is required Pride and quality in job are heightened due to workers' ability to complete entire job
Machinery	Resources are required to be available at proper time in order to maintain production capacity Equipment duplication exists
Methods	General instructions provide work plans rather than detailed process plans Continuity of operations and responsibility exist Production process is flexible to accommodate changes in product design Tight control and coordination in work task scheduling is required
Materials	Material movement is reduced Number of end items is small but lot sizes of components or subassemblies ranges from small to large Increased space and work-in-process requirements exist

2.1.2.2 Project Shop

In a project shop (Table 2), the products position remains stationary during the manufacturing process due to the size, weight, and/or location of the product. Materials, people, and machinery are brought to the product or product site. This type of environment is also called a fixed-position or fixed-site layout. Products manufactured in a project shop could include aircraft, ships, locomotives, or bridge and building construction projects. An example of a project shop layout is shown in Fig. 3.

2.1.2.3 Cellular Manufacturing System

A cellular manufacturing system (Table 3) forms production cells by grouping together equipment that can process a complete family of parts. The production

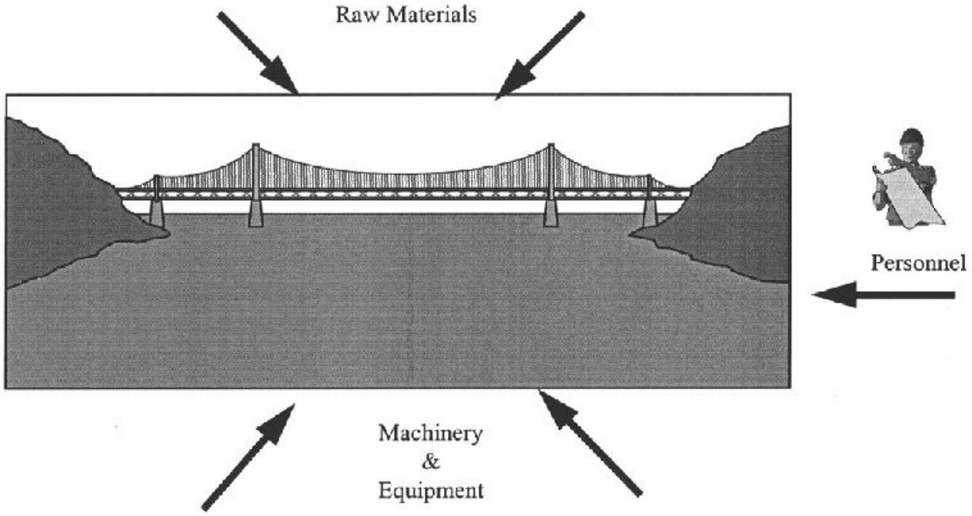


Figure 3 Project shop.

Table 3 Cellular Manufacturing Characteristics

People	Job enlargement and cross-training opportunities exist Labor skills must extend to all operations in cell Provides team atmosphere General supervision is required Personnel are better utilized Provides better communication between design and manufacturing engineering
Machinery	Increased machine utilization results from product groupings Standardization based on part families helps decrease machine setup times by 65–80% Required floor space is reduced 20–45% to produce same number of products as a job shop General-purpose rather than dedicated equipment is common
Methods	Smoother flow, reduced transportation time, less expediting, decreased paperwork, and simpler shop floor controls result Families of parts, determined through group technology, have same set or sequence of manufacturing operations Production control has responsibility to balance flow Cells are less flexible than job shop layouts
Materials	Material buffers and work-in-process are required if the flow is not balanced Reduction of 70–90% in production lead times and WIP inventories compared to job shops Parts move through fewer material-handling operations, 75–90% reduction compared to job shops Quality-related problems decrease 50–80%

environment contains one or more cells which are scheduled independently. The flow among the equipment in the cells can vary depending on the composition of parts within the part family. The family parts are typically identified using group technology techniques. An example of a project shop layout is shown in Fig. 4.

2.1.2.4 Flow Line

The last major style of configuring a manufacturing system is a flow line (Table 4). In a flow line, machines and other types of equipment are organized according to the process sequence and the production is rate based. These types of layout are also known as product or repetitive manufacturing layouts. Dedicated repetitive and mixed-model repetitive are the most common types of flow lines for discrete products. Dedicated repetitive flow lines produce only one product on the line or variations if no delay is incurred for changeover time. Mixed model repetitive refers to manufacturing two or more products on the same line. Changeover between products is minimized and mixed model heuristics determine the sequence of product variation that flow through the line. When the flow line produces liquids, gases, or powders, such as an oil refinery, the manufacturing process is referred to as a continuous system rather than a flow line. An example of a flow line layout is shown in Fig. 5.

A special type of flow line is the transfer line. Transfer lines utilize a sequence of machines dedicated to one particular part or small variations of that part. Usually the workstations are connected by a conveyor, setups take hours if not days, and the capacity is fully utilized. Examples of transfer lines include automotive assembly, beverage bottling or canning, and heat treating facilities. Automated transfer line which include NC or CNC machines, and a material handling system that enables parts to follow multiple routings, are generally referred to as flexible machining systems (FMS).

2.2 PUSH VERSUS PULL TECHNIQUES

A basic functional requirement in a production system is the ability to provide a constant supply of materials to the manufacturing process. The production system must not only ensure that there is a constant supply of materials but that these materials must be the correct materials supplied at the appropriate time in the correct quantity for the lowest overall cost. Generally, material release systems can be categorized as either "push" or "pull" systems. Push systems will normally schedule material release based on predetermined schedules, while pull systems utilize downstream demand to authorize the release of materials.

Traditional manufacturing environments, which normally utilize material requirements planning

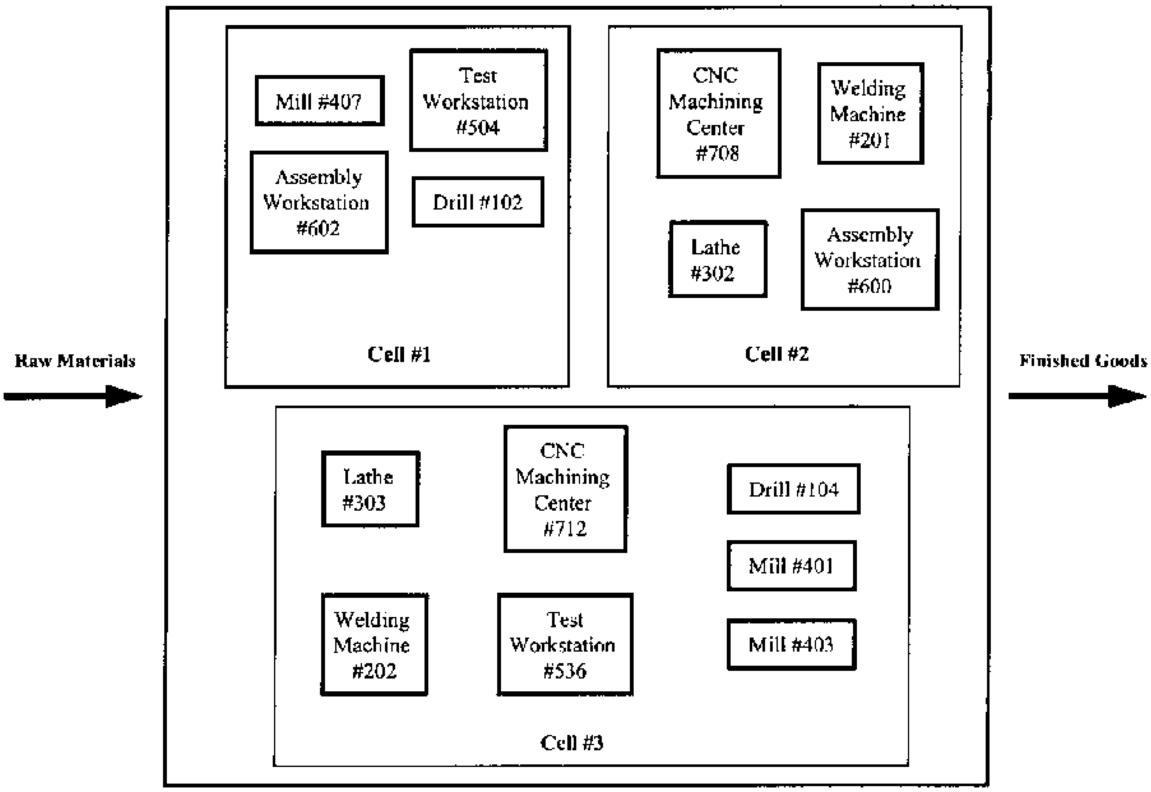


Figure 4 Cellular manufacturing system.

(MRP) or manufacturing resource planning (MRPII) systems, will schedule material releases to the production floor based on a predetermined capacity, batch size, and standard processing times. While there is a sufficient supply of acceptable quality materials and production time is within the standard allotted time, materials will flow smoothly through the system. However, if one operation in the process becomes unavailable due to downtime or other reasons, inventory will start to build up at this workcenter. This buildup

Table 4 Flow Line Characteristics

People	Less skill is needed for production line personnel General supervision is required
Machinery	Dedicated equipment is used to manufacture specific product One machine of each type is required unless redundancy is needed to balance flow Large capital investment Higher production rates Machine stoppage shuts down production Bottleneck station paces the line
Methods	High product volume provides low unit costs Standardized products are delivered at predictable output rates Ratio of value-added to non-value-added time in process is increased Simplified production control Product design changes can cause layout to become obsolete
Materials	Small amount of work-in-process in system Flow lines provide for direct, logical material flow Material-handling requirements are reduced

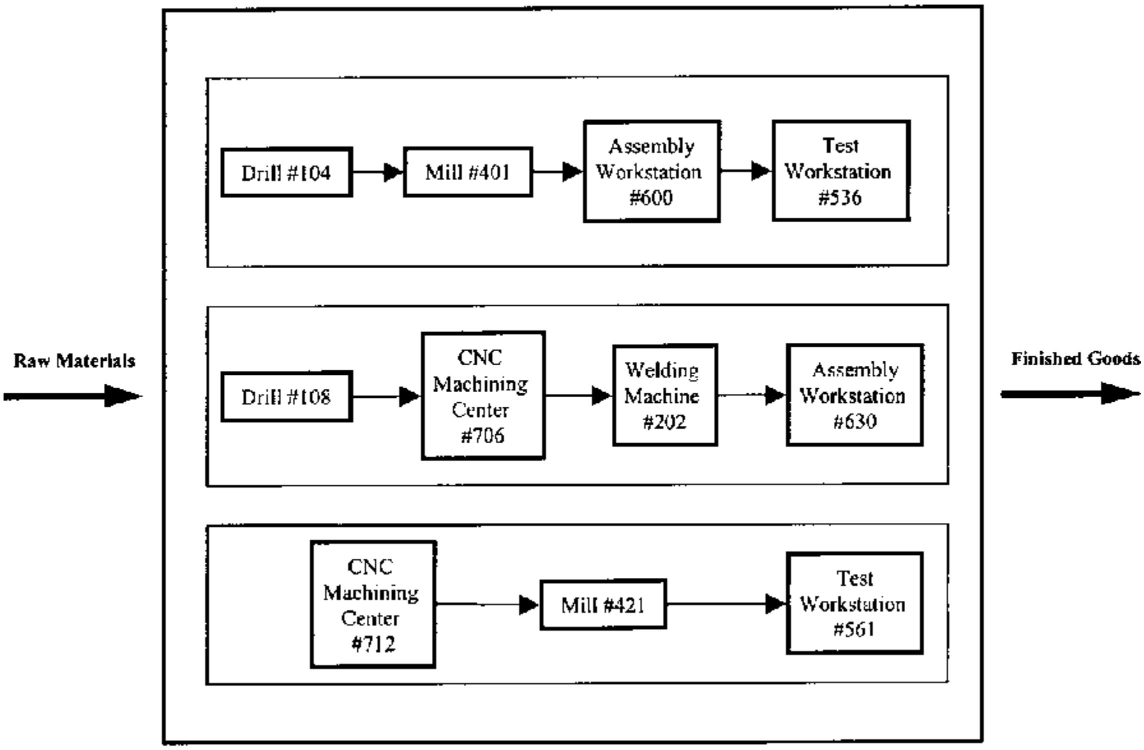


Figure 5 Flow Line.

occurs because the schedule dictates that these workcenters continue to produce as long as materials are available.

Pull systems differ from push systems, since they authorize the start of jobs rather than scheduling the start of these jobs. Pull systems are also known as "demand pull' systems because the authorization for work is triggered by the demand of the downstream customer. The downstream customer can be another workcenter, a distribution center, an original equipment manufacturer (OEM), or the final customer. After authorization for production, the workcenter performs its operations in order to satisfy the downstream demand. The usage by this workcenter of materials or components, which are produced by upstream processes, will in turn create new demand for these upstream processes. In this way, the downstream customers are pulling components and materials through the production system.

Characteristics of the most common push system, MRP, and pull system, kanban, are included in the following sections. Recent literature reflects the use of an alternative pull system, called CONWIP, which is briefly described. At the end of these discussions, comparisons between these systems provide the reader with an understanding of the capabilities and advantages of these systems.

2.2.1 Push

2.2.1.1 MRP Systems

The master production schedule (MPS) is a common mechanism utilized by firms to establish the production plan for the short-term horizon. This short-term horizon depends on the nature of the production process and typically varies from 6 months to 1 year. The MPS, based on market forecasts and firm customer orders, identifies the quantities and due dates for the production of end products. In order to satisfy the production requirements of the MPS, the components, assemblies, and raw materials used to manufacture the end products must be available in the correct quantities at the proper time. If any of the components are unavailable, production cannot meet the delivery schedule.

Material requirements planning (MRP) is the system that calculates the required quantities and dates for all materials (components, assemblies, raw materials) that need to be available to production in order to satisfy the MPS. The MRP system analyzes each level of the production process and, using lead time offsets, calculates the requirement dates for the materials. In addition to the MPS, the MRP system requires two other inputs; the inventory status of the materials and the bill of material of the end products (see Fig. 6). The inventory status contains details such as purchasing lead times and quantities on hand, information that is required to calculate the time-phased material requirements. The other input, the bill of materials, lists quantities of all required components, assemblies, etc. to produce a single end product.

The MRP system typically integrates this information in tableau form and is referred to as the MRP record. An example of a typical MRP record is shown in Table 5. This table represents a product which has a 3-week lead time and is replenished with a lot size quantity of 50 units. The MRP record is a time-phased block of information that is updated on a periodic basis. The record shows a specific number of future periods from the current period. As the current time period expires, the data for this period is removed from the record and the future periods all shift one time period. For example, when Week 22 has passed, it is removed from the MRP record and the new MRP record shows activities for Weeks 23 through 29.

Table 5 Basic MRP Record

		Week						
		22	23	24	25	26	27	28
Gross requirements		10	10	20	30	30	10	10
Scheduled receipts				30				
Projected on hand	25	15	5	15	35	5	45	35
Planned order receipts					50		50	
Planned order releases		50		50				

The top row represents the planning period and can range from days to months. The typical planning period, as shown below, is in weekly increments. The second row, titled "gross requirements," is the expected demand for this item during this specific period. The third row, titled "scheduled receipts," is an existing open order that has been released to manufacturing or a supplier prior to the first period

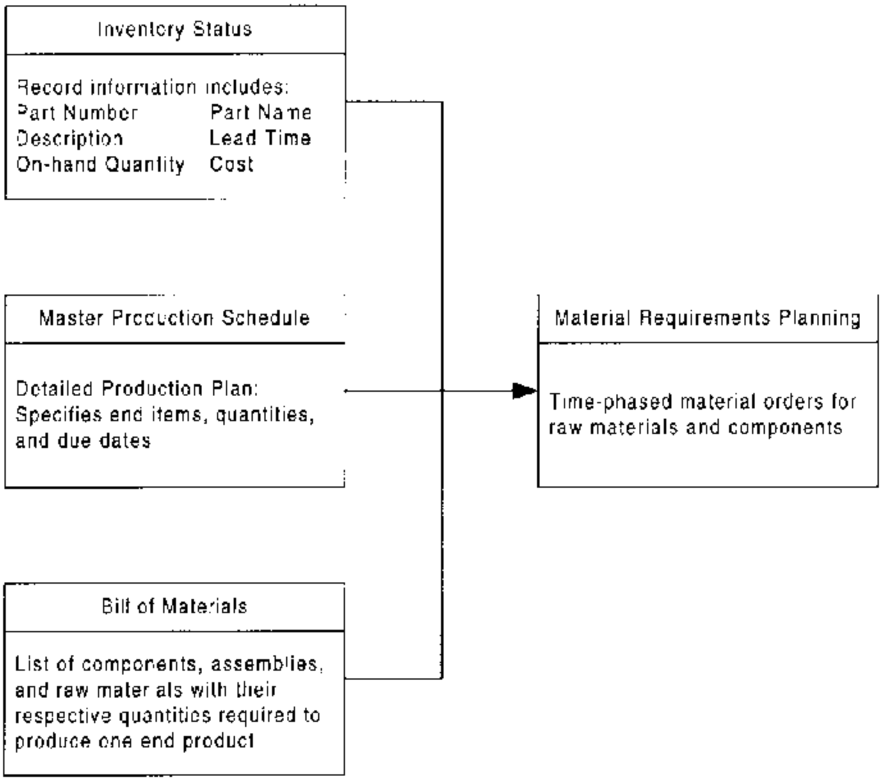

Inventory Status
Record information includes:
Part Number
Part Name
Description
Lead Time
On-hand Quantity
Cost
Master Production Schedule
Detailed Production Plan:
Specifies end items, quantities, and due dates
Bill of Materials
List of components, assemblies, and raw materials with their respective quantities required to produce one end product
Material Requirements Planning
Time-phased material orders for raw materials and components

Figure 6 MRP system inputs.

shown on this MRP record. The scheduled receipt for 30 items that is recorded in Week 24 was released prior to Week 22. The next row, "projected on hand," shows the inventory level anticipated at the end of the period. The quantity of 25 items that shows up prior to Week 22 is the inventory status at the end of Week 21. The "planned order release" is an MRP calculated value which recommends the quantity that will satisfy the demand for the item. This number is calculated from the projected on hand, gross requirements, and lead time for the item. In Week 25, there is a demand for 30 items, but the inventory status at the end of Week 24 shows an available balance of only 15 items. Since this item has a 3-week lead time, an order must be placed in Week 22 to satisfy demand for Week 25. The quantity of the order is determined by the lot size. The "planned order receipts" row shows the quantity planned to be received in the future based on the MRP suggested planned order releases.

An example of the mechanics of an MRP system is illustrated below. A multilevel bill of materials (BOM) for a table fan is shown in Fig. 7. An additional piece of information included in the bill of materials is the purchasing or production lead time. The final product, part number F6001, appears in Level 0 in the bill of materials. The table fan is assembled from a front guard and back guard assembly using three screws (Level 1). The back guard assembly is composed of the back guard, a variable speed switch, and a fan assembly (Level 2). The fan assembly is fabricated from a fan blade, motor, and electric cord (Level 3). The same information regarding the product structure can be presented in a different format on an indented bill of materials (see Table 6).

Assume that this company receives an order for one table fan and currently carries no inventory for this item. The company can determine the earliest promise date for the customer by considering the effects of the production and purchasing lead times on the total

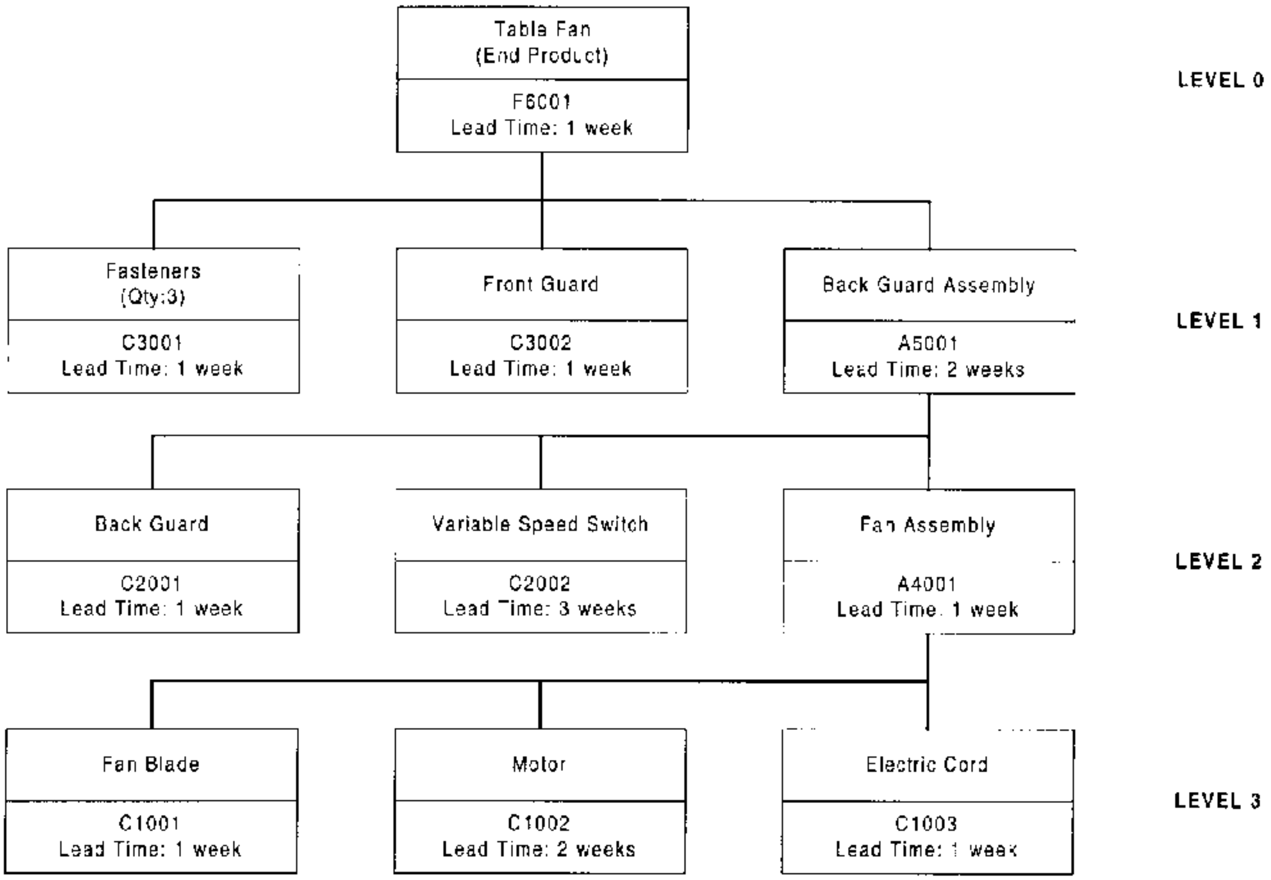


Figure 7 Multilevel BOM with lead times.

Table 6 Indented Bill of Materials

Part number	Quantity	Description
F6001	1	Table fan
.....C3001	3	Screws
.....C3002	1	Front guard
.....A5001	1	Back guard assembly
.........A4001	1	Fan assembly
..............C1001	1	Fan blade
..............C1002	1	Motor
..............C1003	1	Electri cord
.........C2001	1	Back guard
.........C2002	1	Variable speed switch

amount of time required to produce an end product. A Gantt chart is frequently created to graphically depict this lead time offset process (see Fig. 8).

2.2.2 Pull

2.2.2.1 Kanban

Kanban is a Japanese term which literally translated means "visible record." The term has been widely misinterpreted in the West and many industrialists use the term interchangeably with just-in-time production, stockless production, and numerous material handling strategies. The authors have visited many manufacturers which insist they have a kanban system in place; in reality, they generally have an inventory control system which has some visual pull aspects, but which varies widely from the original Japanese kanban system.

The most widely accepted version of kanban is that utilized as an integral part of the Toyota production system or just-in-time system. These systems employ a card (i.e., the visible record or signal) to send a straightforward message—either to deliver more parts to a production operation, or as a signal to produce more components. The primary difference in the true kanban approach and MRP is that in the kanban approach materials are pulled into the system based on downstream demand. In the traditional MRP

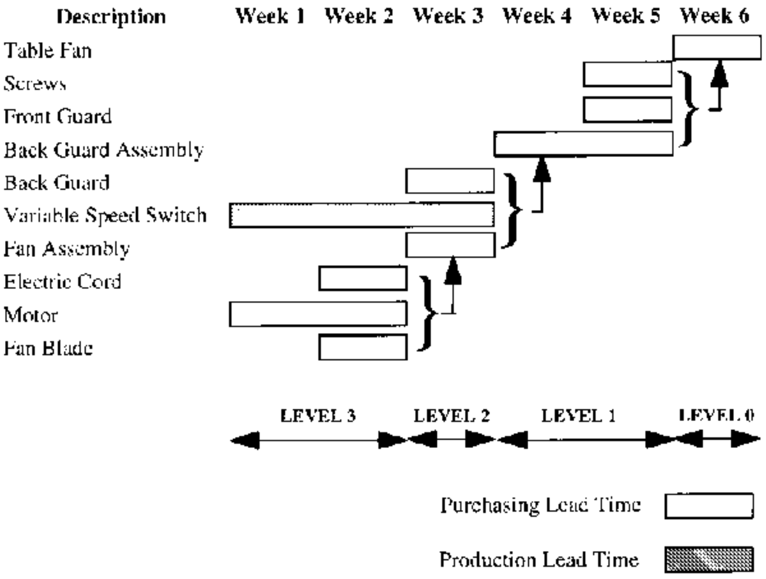


Figure 8 Gantt chart.

approach, predetermined schedules trigger the release of materials via a material router, job order or production ticket.

The kanban approach can be utilized for material control and movement throughout the manufacturing process, from raw material replenishment through production and distribution of finished goods. One of the most straightforward applications of kanban is for raw material replenishment. Under the MRP philosophy, purchase orders for raw materials are generated by the MRP algorithm based primarily on sales forecasts, firm orders, and supplier lead times for raw materials. There are three basic problems with this approach: (1) forecasts need to be accurate to within plus or minus 10–15%, (2) inventory accuracy needs to be maintained at 98%, and (3) generating and processing purchase orders is an expensive process which typically ranges from $100 to $400 per purchase order (which includes information systems support, accounting transactions, etc.).

Now consider a "two-bin" kanban approach for raw material replenishment. In its most basic form, every purchased component or raw material has two dedicated bins. A bin may consist of a tote pan, a series of tote pans, a pallet, or even multiple pallets of material. For this discussion, assume that a bin is a single tote pan. Each tote pan holds a specific quantity of a specific component. Calculation of the exact quantity is typically based on a formula such as the following:

Bin quantity = Leadtime (in days, including
supplier and internal)
× Maximum daily usage or
maximum daily production quantity
× Safety stock factor (for supplier
delivery or quality issues)

For example, if the screw in the table fan assembly has a lead time of 20 days and the daily usage is 500 screws per day, the bin quantity is 10,000 screws. The two-bin system for the screw is described in Fig. 9. In this scenario, the screws could be prebagged in daily quantities of 500. Generally, no inventory transaction record is necessary. When the production line requires more screws, a bag is removed from bin one for consumption. When the last bag is taken, this signals the material handler to send a preformatted fax to the supplier. This in turn signals the supplier to send exactly 10,000 screws to be delivered in 18 to 20 days based on an annual purchasing agreement or blanket purchase order. The average on-hand inventory in this scenario is 6000 screws (including a 2-day safety stock of 1000 screws) as depicted in Fig. 10.

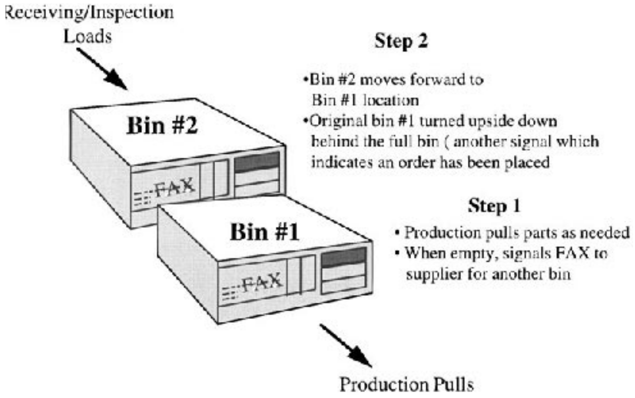

Receiving/Inspection Loads
Step 2
•Bin #2 moves forward to Bin #1 location
•Original bin #1 turned upside down behind the full bin (another signal which indicates an order has been placed
Bin #2
FAX
Step 1
• Production pulls parts as needed
• When empty, signals FAX to supplier for another bin
Bin #1
FAX
Production Pulls

Figure 9 Two-bin kanban system.

There are numerous advantages of this two-bin kanban material replenishment strategy, including:

Raw materials are ordered based on actual usage rather than forecast.

Does not require a purchase order for each replenishment cycle, just a fax to the supplier for a fixed amount.

Every component in the system has only one storage location, preferably near the point of use.

The system is straightforward and highly visual (inventory status can be visually determined).

Gives suppliers improved visibility and control (every time a fax is received for screw #C3001 the quantity and delivery schedule remain constant).

Guarantees first-in first-out inventory rotation.

A west-coast manufacturer of medical devices replaced their cumbersome MRP system with a two-

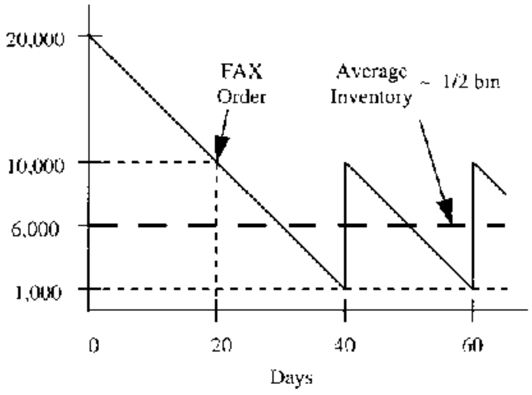


Figure 10 Average inventory.

bin kanban system in 1995–1996. Over 10,000 raw material components were involved. The results included:

An 82% reduction in raw material warehouse space.
Receiving cycle time reduced from 2-1/2 days to less than 1 day.
A dramatic reduction in labor involved in inventory management.
Fewer stockouts of raw materials.

Now we will examine the application of a kanban system on the shop floor. A fundamental aspect of this approach is that every part or subassembly has a special container which holds a fixed number of items. As indicated in the just-in-time discussion in Sec. 2.3.1, the general rule of thumb is the smaller the quantity, the better. Accompanying every container in the system are two cards which contain at least two vital pieces of information—the part or subassembly number and the quantity. One card is referred to as the *production* kanban card which serves as a signal to the operation which produces the part or subassembly. The second card is known as a *movement* or *conveyance* kanban which serves as a signal for the downstream operation. Associated with every operation is an in-process buffer or storage point, which may consist of an actual stock room or merely a space on the floor designated for the part container.

A second fundamental rule of the system is that the upstream operation *never* moves components until the downstream operation sends a signal (i.e., the production kanban card), thus denoting a true pull system. Every parts container in the system moves back and forth between its stock point and its point of use (the downstream operation) utilizing the cards as signals for action.

For the purpose of illustration, consider the table fan product described in Sec. 2.2.1.1. Assume that the end of the production line was set up as follows: (1) Workstation 1 assembles the back guard, variable speed switch, and fan assembly into the back guard assembly and supplies the assembly to Workstation 2; (2) Workstation 2 assembles the back guard assembly, front guard and a set of fasteners into the end product, the table fan, which is then moved to finished goods inventory; (3) finished goods inventory supplies the table fan directly to the customers (see Fig. 11).

The two-card system would work as follows:

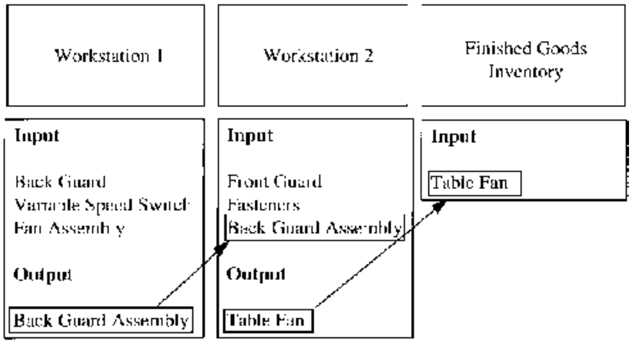


Figure 11 Table fan workstation layout.

1. Customer demand for the table fan would be satisfied from a "bin" in finished goods inventory. When the bin is emptied, the bin which has a C-kanban card attached to it is moved to the storage area at Workstation 2. The C-kanban card is then attached to a full bin and moved to finished goods inventory.
2. The P-kanban card that was attached to the full bin is detached and attached to the empty bin. The empty bin is then routed to the start of the production operations in Workstation 2 and signals a demand for production of table fans.
3. During the production of the table fans in Workstation 2, production line personnel work out of bins of raw materials which include back guard assemblies, front guards, and fasteners. When the bin of back guard assemblies is emptied, the empty bin with the C-kanban card is moved to the storage area of Workstation 1 to replenish the back guard assemblies. The C-kanban card is then attached to a full bin and moved to the appropriate area at Workstation 2.
4. The process is repeated as described in Step 2. The P-kanban card is attached to the empty bin and moved to the initial operation at Workstation 1, signaling a demand for production of back guard assemblies.

Figure 12 shows the conveyance and production kanban cards.

There are basic rules which simplify this sequence of events. First, no production occurs unless there is an empty container with a production card attached at a stock location. An operation remains idle until an actual demand is realized (this basic rule is often difficult for the Western production mentality which traditionally focuses on maximizing machine/operator utilization).

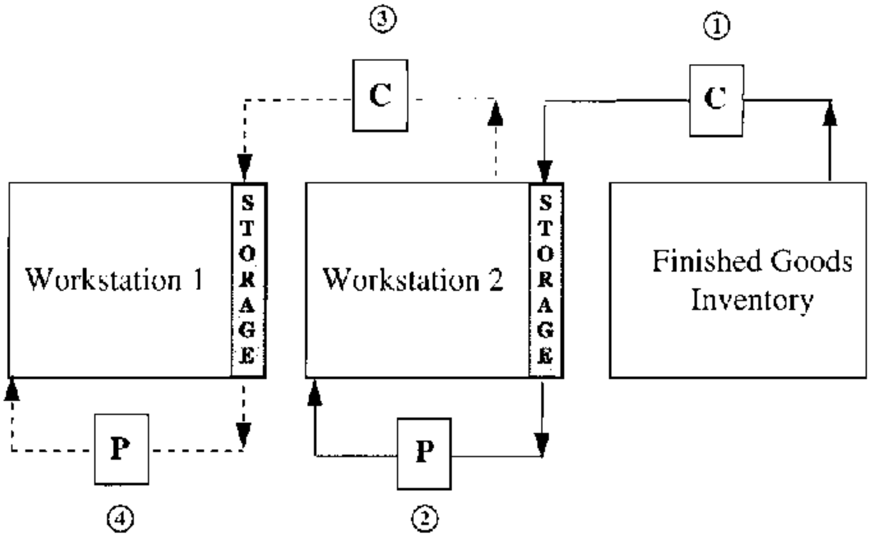


Figure 12 Conveyance and production kanban cards.

A second fundamental rule is that there is exactly one production and one conveyance kanban card per container. The number of containers for any given part number is determined by actual demand, the number of parts per container, setup times, etc. Finally, for any given part or subassembly number, there is a fixed quantity of parts as defined on the kanban card.

When this system and its fundamental rules are followed, it is simultaneously simple and precise, and sets the stage for continuous improvement. Furthermore, lot size reduction is a simple matter of reducing the number of kanban cards in the system. Throughput problems will arise, highlighting areas of opportunity which were previously hidden by excessive inventory. The above is an example of a two-card system which can be readily modified to meet individual company requirements.

2.2.2.2 CONWIP

CONWIP (CONstant Work In Process) is a pull philosophy whose strategy is to limit the total amount of in-process inventory that is allowed into the manufacturing process. The mechanism for release of materials or components into the process is signaled by the customer withdrawing or “pulling” a unit from finished goods inventory. Once the unit is removed from finished goods, a signal is sent to the initial workcenter to release additional materials into the process (see Fig. 13). Once materials or components are released into the system they will progress all the way through the system until reaching finished goods inventory. If finished goods inventory is filled, there will be no mechanism to release materials into the system.

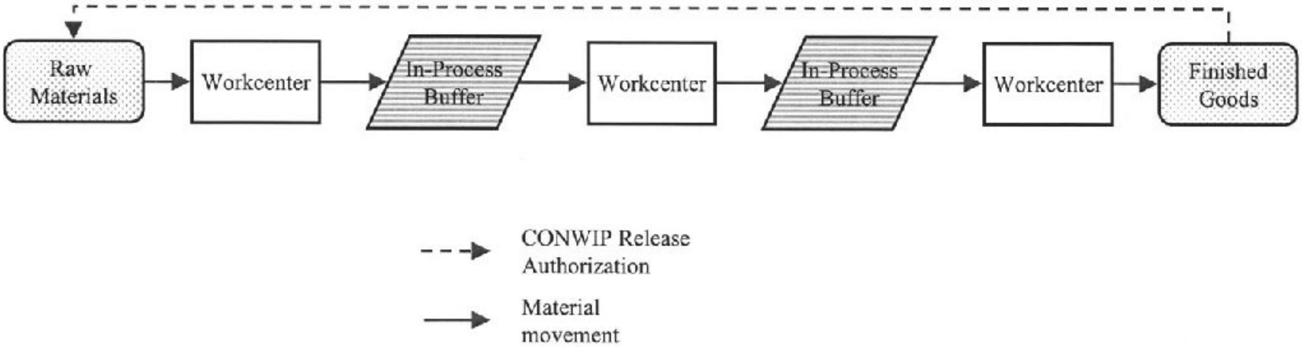


Figure 13 CONWIP control.

CONWIP can be considered as a specialized case of kanban; both systems use customer demand as the mechanism for material release. The major difference between CONWIP and kanban is their use of in-process buffers. Kanban systems route the materials through the line until all in-process buffers are full. Once materials have been released into the system, CONWIP systems allow the materials to progress all the way through the system until reaching finished goods inventory.

Another difference regarding these in-process buffers is their ability to protect upstream or downstream processes from work stoppages due to workcenter failures. Kanban buffers can protect downstream workcenters from failures in upstream workcenters. However, these buffers do not protect upstream workcenters from downstream failures. For instance, if a downstream workcenter fails or experiences significant downtime, an upstream workcenter will continue to operate only until its downstream buffer is filled. Meanwhile, demand for finished goods still grows at the end of the line. When the downstream workcenter becomes operational, an increased demand on the upstream workcenter occurs in order to fill the unsatisfied demand. This scenario occurs in cases where the demand rate exceeds the capacity of the system buffers.

CONWIP buffers help to decouple the upstream and downstream workcenters. In the case of a downstream workcenter failure, customer demand will be filled from finished goods inventory and new materials will continue to be released into the system. WIP will continue to build up in front of the down workcenter until that workcenter becomes operational. Once operational, the workcenter will have enough materials to satisfy the downstream demand and replenish finished goods inventory.

The ability to implement a CONWIP system is based on the following requirements:

1. All the parts in the production line flow through a single path.
2. The level of WIP in the system can be adequately measured.
3. The company's manufacturing strategy, in part, follows a make-to-stock philosophy.

In the event that a bottleneck exists in the system, a CONWIP pull system will have high levels of WIP towards the upstream workcenters. In this case, a hybrid system of CONWIP and kanban cells achieves the desired control (see Fig. 14). The kanban cells will limit the amount of parts released to the system if processing bottlenecks are severe enough.

2.2.3 System Comparisons

Although pull systems have many advantages over push systems (see Table 7), one of the biggest advantages is that the pull systems (CONWIP or kanban) limit the amount of in-process inventory in the system. This feature of a pull system is normally referred to as the "WIP cap" of the system. Pull systems will normally time the release of work closer to the point of when value will be added, as opposed to the push system, which generally pushes too much work into the system. Pushing more materials into the system increases the average WIP level but does not improve the amount of throughput. The WIP cap reduces the average level of WIP for a given throughput level while reducing the in-process inventory investment.

Additionally, pull systems have advantages over push systems in the following areas:

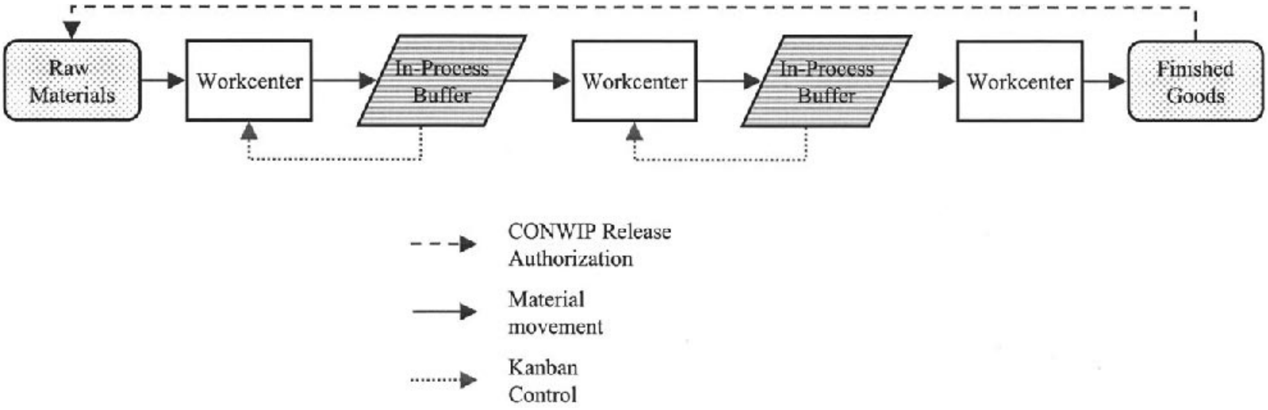


Figure 14 CONWIP/kanban hybrid control.

Table 7 Push vs. Pull Production

Push	Pull
Production scheduler or system is responsible for ensuring system performance	Production floor personnel oversee system performance
Production schedule generates build signals	Downstream customer demand authorizes build signals
"Push" begins at beginning of process	"Pull" begins at end of process
Materials "pushed" through the process, generally creating high WIP or bottlenecks	Materials "pulled" through the process
Production floor problems can be concealed through excessive WIP	Production floor problems are exposed creating necessity for attention
Intermittent communication between workcenters	Workcenters keep in constant communication
Production floor receives materials in batches	Materials released to production floor based on production rate
Production commonly decentralized	Production organized in cells
Product cycle times subject to increase	Product cycle times are reduced
WIP inventories can be high	WIP inventories are capped at low levels

1. *Manufacturing costs*. The WIP levels are capped not only from better timing of material releases than a push system, but when system disturbances do occur (e.g., machine downtime, product line changeovers, etc.), pull systems will not allow the WIP to exceed a certain level. Push systems cannot react in the same manner and generally WIP will run out of control before modifications to the system occur. Additionally, when engineering changes or job expediting is required, the presence of a WIP cap helps to reduce the manufacturing costs associated with these activities.
2. *Cycle time variability*. When there is a small variance in cycle times, there is a high degree of certainty regarding the length of time it takes a specific job to process through the system. Since production cycle times are directly associated with the WIP level (through Little's law), which is limited by the pull systems, these systems restrict significant increases in production cycle time.
3. *Production flexibility*. Push systems can often release an excessive amount of work into the production line causing severe congestion of the system. The high levels of WIP create a loss of flexibility due to the facts that: (1) engineering changes are not easily incorporated, (2) changes in scheduling priorities are hampered by the efforts required to move the WIP off the line to accommodate the expedited orders, and (3) release of materials to the floor is required earlier than scheduled, since the production cycle times would increase proportionally with the amount of WIP in the system.

In addition, pull systems will normally provide:

Immediate feedback if the product flow is stopped
Transfer of ownership of the process to members of the production line
Simplicity and visibility within the system
A sense of urgency to solve problems
Allocation of resources to the areas which ensure customer demand is satisfied.

Although the authors strongly support the use of pull systems, there are certain environments which favor MRP or MRPII systems over a pull (Kanban or CONWIP) system or vice versa. Typically in environments where the products are custom manufactured or are subject to low production volumes, MRP or MRPII systems are more appropriate. However, any environment which utilizes MRP for material planning is subject to problems in system performance if inventory record accuracy falls below 98%.

Pull systems are specifically targeted to manufacturing environments where production exhibits a continuous flow and production lead times are consistent (see Fig. 15). However, many production systems will fall between these two ends of the spectrum. It is quite common for these type of production systems to use a hybrid control system, which integrates aspects of both push and pull systems. For instance, using the MRP system as a top-level planning instrument for

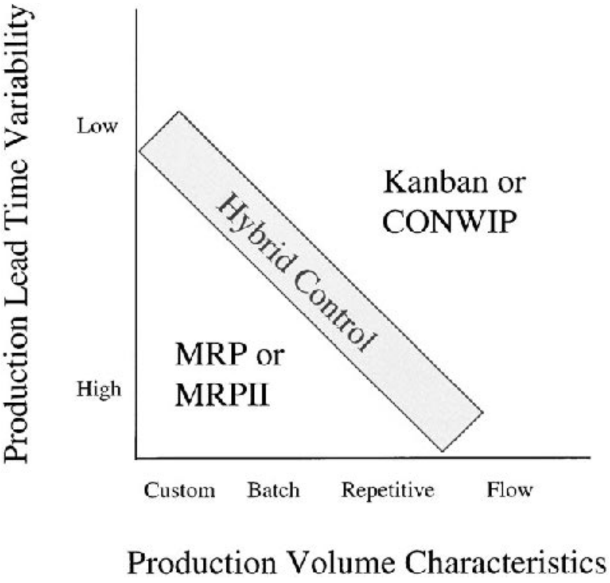


Figure 15 Production system controls.

annual material purchases while using pull mechanisms to control material movement on the shop floor is used quite commonly in repetitive manufacturing environments.

2.3 CONTEMPORARY MANUFACTURING PHILOSOPHIES

In this section the authors discuss four of the most widely accepted manufacturing philosophies including just in time, theory of constraints, and synchronous and flow manufacturing. There are numerous other approaches, but most are built on the fundamental premises of these four. These philosophies are presented to re-emphasize the need to streamline and optimize operations prior to automation. Also, many of the parameters of manufacturing systems, such as desired takt time, will dictate the level of automation required to support the system.

2.3.1 Just in Time

It is no secret that the Japanese have gained the dominant market share in numerous and diverse industries which were originally founded in the United States. Most informed analysts agree that there is only one common element to their success across all these diverse industries—the just-in-time (JIT) manufacturing system developed by Taiichi Ohno of Toyota.

Just in time is often misunderstood in Western culture as being solely an inventory reduction program. As will be shown, this is but one facet of a much larger manufacturing process. To understand JIT, it is first necessary to understand *manufacturing velocity*. Manufacturing velocity compares the current cycle time to the value-added time in any process. The difference between the two is the improvement opportunity zone. The formula for velocity is straightforward:

$$\text{Velocity} = \frac{\text{Current cycle time}}{\text{Value-added time}}$$

The average ratio among manufacturers in the United States is 120:1. This ratio implies that there are 120 hr of non-value-added time for every hour of value-added time! The manufacturer who achieves a ratio of 10:1 has a significant competitive advantage for numerous reasons, but primarily because the manufacturing pipeline is much shorter. At the beginning of the pipeline, suppliers are paid for raw materials and/or components. At the other end, customers pay for the products shipped. Higher velocities yield superior cash-flow positions and improve responsiveness to changes in the market.

There are numerous definitions of JIT. In the authors' opinion:

> Just in time is a pull-based manufacturing process focused on continuously increasing manufacturing velocity through the relentless elimination of waste, where waste is any activity that does not add value from the customer's perspective.

Waste is the use of any resource in excess of the absolute theoretical minimum required to meet customer demand. Waste most often takes the forms of excess inventory, material handling, queues, setup time, inspection and scrap. One of the founding fathers of JIT, Shigeo Shingo, became famous by popularizing the notion of the seven wastes. For the Toyota Corporation elimination of these seven wastes, described below, became the backbone of their JIT philosophy.

1. *Waste from overproduction.* One of the most difficult lessons U.S. manufacturers have learned from the Japanese is that premature production is highly undesirable. Finished goods are the most expensive form of inventory. In addition, if the goods are not required immediately, the factory has consumed resources (materials, labor, and process/machine capacity) which may have been used to increase

the manufacturing velocity of other goods which have an immediate demand. Premature production also conceals other wastes and therefore is one of the first that should be addressed.

2. *Waste of waiting time*. Any wait or queue time obviously decreases manufacturing velocity and does not add value to the end product or the customer. Waiting time for materials flowing through the manufacturing process is relatively straightforward to identify and systematically eliminate. Caution must be taken, however, in eliminating labor or machine waiting time because this may be more desirable than what appears to be value-added time. As described above, simply cranking out parts may contribute to premature production.
3. *Transportation waste*. Transportation time, whether in an automated or manual process, is nearly always non-value-added from the customer's perspective and is often viewed as a necessary evil. It is one of the most common wastes in manufacturing processes. Incoming materials, for example, as typically received, entered into the inventory tracking system, stored and subsequently pulled for production with yet another transaction in the tracking system. A central concept in the JIT system is to ensure that the minimum amount of material required to meet immediate customer demand is received nearest its point of use, "just in time" for production.
4. *Processing waste*. There are numerous categories of processing wastes ranging from removal of excess materials from components (e.g., removal of gates from a casting) and materials consumed in the manufacturing process (e.g., cutting fluids) to non-value-added machine setup time. Design for manufacturability (DFM), design for assembly (DFA), single minute exchange of dies (SMED) and line balancing are examples of methods aimed at the elimination of processing wastes.
5. *Inventory waste*. As mentioned above, many Western interpretations of JIT focused almost solely on reduction of inventory. In numerous cases this has had the net effect of merely pushing the burden of inventory carrying costs further upstream to the suppliers, who in turn incurred higher costs which were eventually passed back to the manufacturer. One of the most problematic aspects of excess inventories is that they obscure other areas of waste including poor scheduling, quality problems, line imbalances, excessive material handling/transportation (both within the factory walls and upstream and downstream of the factory).
6. *Waste of motion*. Somewhat analogous to transportation waste is the waste of motion. Motion wastes can take the form of reorienting a part from one operation to the next, reaching and/or searching for tools and any extra motions (automated or manual) required to perform a manufacturing operation.
7. *Waste from product defects*. In general, the further along the manufacturing process that a defect occurs, the more costly it becomes. Even quality inspections in the process to identify defects are a form of waste. The worst defect of all is one that reaches the customer because not only may the material and labor be lost, but the customer may be lost as well. Thus process control is clearly a central component of JIT.

Another approach to implementing JIT is by focusing on lot size reductions. This concept is portrayed in the JIT cause-and-effect diagram illustrated in Fig. 16. This approach is particularly effective because many of the benefits of JIT are realized by reducing lot sizes. For example, reducing lot sizes will decrease the amount of inventory in the system, which will yield lower inventory carrying costs and improve cash flow. Additionally, lower inventory will cause deficiencies in the system to surface which could otherwise go undetected because excess inventory tends to mask less than optimal conditions.

The relentless and continuous process of elimination of all seven wastes, or the alternative approach of reducing lot sizes, will increase manufacturing velocity, which is the essence of JIT. According to conservative estimates, the implementation of JIT should yield results, as shown in Table 8.

2.3.2 Theory of Constraints

The theory of constraints (TOC), popularized by Goldratt [1,2], is based on the premise that the key to continuous improvement is the systematic identification and exploitation of system constraints. A constraint is anything that limits a system from achieving higher performance with respect to its goal. This theory has application to any type of system, but has gained the most attention from its application to manufacturing systems.

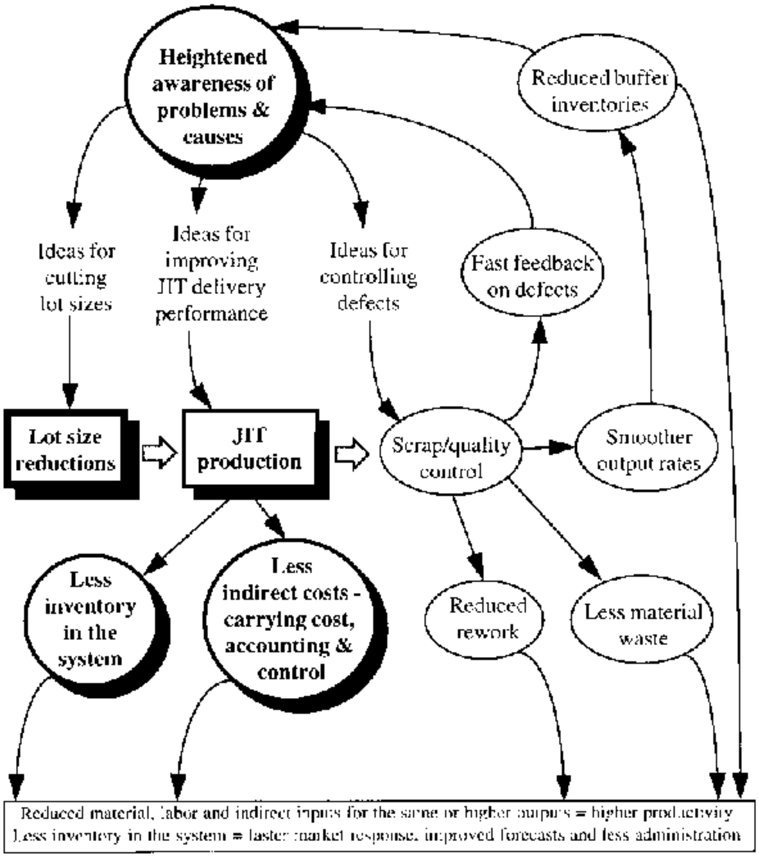

Heightened awareness of problems & causes
Reduced buffer inventories
Ideas for cutting lot sizes
Ideas for improving JIT delivery performance
Ideas for controlling defects
Fast feedback on defects
Lot size reductions
JIT production
Scrap/quality control
Smoother output rates
Less inventory in the system
Less indirect costs - carrying cost, accounting & control
Reduced rework
Less material waste
Reduced material, labor and indirect inputs for the same or higher outputs = higher productivity
Less inventory in the system = faster market response, improved forecasts and less administration

Figure 16 JIT Production—cause and effect.

Table 8 Typical JIT Implementation Results

Category	Typical results
1–3 years	
Cycle time	50–90% reduction
Work-in-process inventory	50–90% reduction
Scrap and rework	60–80% reduction
Setup time	50–90% reduction
Manufacturing floor space	30–60% reduction
Quality improvement metrics	10–1000 times improvement
3–7 years	
Overall quality	5–10 times improvement
Inventory turns	4–10 times improvement
Return on assets	Variable depending on industry

The application of TOC to the continuous improvement of manufacturing systems consists of five steps, as shown in Fig. 17. The first step is to identify and prioritize the system's constraints. Some constraints are easily identified, such as a machining center through which numerous products are routed. Indications may be excessive overtime, very high utilization compared to other operations, or numerous components in its queue waiting to be machined. Other constraints are more difficult to identify, such as poor scheduling practices or purchasing policy constraints. Once identified, the constraints need to be prioritized with respect to their negative impact on the goal.

The second step is to determine how to exploit the constraint. For example, if the constraint is a machining center, methods must be determined to increase its capacity. There may be numerous opportunities for

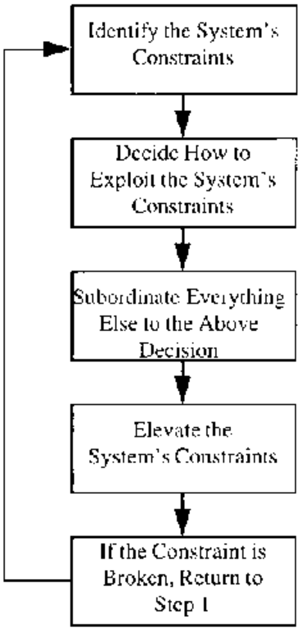

Identify the System's Constraints
Decide How to Exploit the System's Constraints
Subordinate Everything Else to the Above Decision
Elevate the System's Constraints
If the Constraint is Broken, Return to Step 1

Figure 17 TOC implementation.

improvement including setup reduction, improved scheduling, operator training, overtime, etc. The third step is to subordinate everything else to the above decision. This is a process of focusing the organizations attention on the constraint since it is the factor which limits the systems output. This is a major step because previously all operations received equal attention. The fourth step is to elevate the systems constraint which is similar to step three. It is intended to heighten awareness of the constraint through the organization and mobilize the organization's resources to tackle the constraint. Finally, if the constraint has been effectively eliminated, another constraint will surface and the process begins again. This is the TOC cycle of continuous improvement.

Another important concept of TOC is the drum–buffer–rope analogy. The drum is the desired pace of the production system which is typically determined by the capacity of the constrained resource. Thus the drum dictates the master production schedule. Since there will always be minor deviations from the planned schedule, actual material flow will differ from the plan. Therefore time and/or inventory buffers are built into the system at strategic points to increase the probability of attaining the desired throughput. Finally, the rope is the analogy for the mechanism which synchronizes material flow through all the nonconstraint resources without actually having to actively control each individual resource. The primary function of the rope is to pull materials to downstream operations at the right time and in the right quantity. For further information on the mechanics of the drum–buffer–rope philosophy see Goldratt [1,2], Srikanth and Cavallaro [3], and Umble and Srikanth, 1995.

There are numerous case studies of dramatic improvements attained by the application of TOC. For example, a custom manufacturer of cabinets made significant improvements through implementing a TOC manufacturing strategy. They were able to reduce their manufacturing lead time from an industry average of 4 weeks to only 2 days. They also increased sales from $6 to $10 million in 2 years while holding the number of employees constant. In another example, a Fortune 100 company pioneered the application of TOC to distribution and reported a $600 million reduction in inventory.

2.3.3 Synchronous Manufacturing

Synchronous manufacturing is not really a new technique and is based on the basic principles used by Henry Ford in the 1920s, the concepts of just-in-time manufacturing, and Goldratt's theory of constraints. Kanban (from JIT) and drum–buffer–rope (from TOC) both represent approaches to synchronized production control. The following definition by Srikanth and Cavallaro [3] states the underlying premise of synchronous manufacturing: "Synchronous manufacturing is an all-encompassing manufacturing management philosophy that includes a consistent set of principles, procedures and techniques *where every action is evaluated in terms of the common global goal of the organization.*"

In for-profit manufacturing organizations the global goal is generally straightforward—to make money. However, the concept of synchronous manufacturing can be applied to any manufacturing environment (e.g., where the global goal may be to produce on schedule with cost or profit being a secondary factor).

There are three fundamental elements of synchronous manufacturing. First, the manufacturing organization must explicitly define its global goal. The goal must be stated in terms that are readily understandable by the entire organization. If the goal is to make money, this can be further understood by the addition of commonly understood metrics such as throughput, inventory, and cost of goods sold. By focusing on these global metrics rather than on individual cost centers or other subsets of the manufacturing enterprise, organizations are more likely to achieve the global goal of making money.

The second element of synchronous manufacturing is to develop straightforward cause-and-effect relationships between individual actions and the global goal and its associated metrics. Here we can see the relationship of the theory of constraints to synchronous manufacturing. Actions which increase the throughput of nonbottleneck resources have zero impact on the common goal, whereas actions which increase the throughput of bottleneck resources have direct impact. Ongoing education of personnel throughout the enterprise as to how their actions, related to their spheres of influence, impact the global goal and its associated metrics is key to the success of synchronous manufacturing.

The third element is to manage the individual actions to ensure they are properly focused on the global goal. This also includes measuring the impact of actions against the metrics and refocusing where necessary. It is clear that all constraints, including market, capacity, material, logistical, managerial, and behavioral, must be managed.

The synchronous manufacturing philosophy enables the enterprise to focus its resources on the areas which have the greatest impact on the global goal. This process of focusing provides the basis for continuous improvement within the organization. Furthermore, synchronous manufacturing provides the basis for sound decision making. When considering automation, for example, strict adherence to this philosophy will ensure that only automation which makes money—the global goal—is implemented.

2.3.4 Flow Manufacturing

In a continually changing competitive marketplace which has encouraged global competition and a pronounced emphasis on customer satisfaction, more manufacturers are abandoning the traditional MRPII systems in favor of flow manufacturing. (Specific information presented in this section of the chapter is based on Constanza [4].) The catalyst for this change is the focus on the major inefficiencies caused by "push" systems which include growing inventory balances, increased manufacturing cycle times, decreased product quality, and reduced customer satisfaction. The benefits for converting to a flow manufacturer are numerous. Some typical results include:

Reduction in work-in-process inventories
Increased manufacturing output
Reduction of workspace requirements
Reduction in total material costs
Increased labor productivity ranging from 20 to 50%
Increased equipment capacity ranging from 20 to 40%
Reduction in manufacturing lead times ranging from 80 to 90%
Reductions in failure costs (scrap, rework, warranties) ranging from 40 to 50%.

The change to flow manufacturing requires changes that span both system and cultural boundaries. Some typical attributes of flow manufacturers include:

Production process based on customer order activity or demand (without standard production scheduling).
Product volume and mix adjusted daily.
Labor tracking and departmental absorption accounting is abandoned.
Streamlining production process through total employee involvement.
System driven towards zero in-process inventories.
Raw and in-process inventory turns greater than 20 per year.
Non-value-added activities are identified and minimized.
Focused on primary cost drivers—material and overhead rather than labor.
Use of total quality control techniques to eliminate external inspection stations and ensure product quality.
Utilization of takt times to drive the production floor layout.
Relieving inventories through backflushing the product's bill of materials once the product has exited the production floor, eliminating the need for numerous inventory transactions.
Use of concurrent engineering techniques to integrate engineering design changes into production.
Flex fences are used to help smooth demand and define allowable production rates. Flex fences allow for the variation of daily production demand (typically ±10%) without reducing the abilities to meet daily demand or increase levels of inventories.

The conversion to flow manufacturing requires an organizational commitment and significant re-engineering of the production process. The steps listed below indicate the major issues that must be addressed during that conversion. The list is not meant to be an all encompassing list but rather a general framework of how companies have approached the conversion to flow manufacturing. The basic implementation steps include:

1. Analyze market data.
2. Establish the line's takt time.
3. Develop sequence of events sheets.
3. Conduct time study analysis and brainstorm methods improvements.
5. Develop flow line design.
6. Implement multibin kanban system.

2.3.4.1 Key Implementation Elements

Analyze Market Data. One of the initial steps in the re-engineering process is to analyze the market data to determine the level of production capacity that is required to satisfy the market demands. The data that must be used to set the appropriate production capacity includes order arrival data for each product, projected market forecasts, booked orders, and intuition about product trends (growth and decline). The cross-functional team involved in selecting the future operating capacity must include representatives from production, purchasing, material control, engineering, quality, sales, and marketing.

It is critical to examine the order history on the basis of when the customers' orders were actually placed. By examining the data in this manner, the capacity can be designed to control production lead times. The historical data is often plotted (see Fig. 18) and the dashed line indicates the cross-functional team's selection of designed capacity to meet market requirements. The capacity of the cell, shown as the capacity bar drawn parallel to the x-axis, indicates the number of units that could be produced in a given day. All the "white space" below the designed capacity target line indicates the amount of excess capacity that could be used for handling spikes in the data (e.g., units that could not be built the previous day). The selected capacity drives many factors: the level of inventory required to support the production line, the level of automation changes required, the cell's takt time, etc.

Establish Line's Takt Time. Takt, a German word for rhythm or beat, indicates the rate a finished unit would be completed during the shift's effective hours. Once the production line's capacity is determined, the takt time is calculated by multiplying the number of effective hours expected per shift (6.5 was chosen to allow for work cell breaks, fatigue and delay, cleanup, inventory replenishment and ordering, etc.) times the number of shifts per day, all divided by the cells designed daily capacity.

$$\text{Takt} = \frac{\text{Effective work hours} \times \text{shifts/day}}{\text{Designed production rate}}$$

The takt time is based on the designed capacity and indicates the rate at which a finished unit would be produced by the production line during the effective hours of a given shift. It is necessary that the work content at any single workstation is targeted to equal the calculated takt time in order to satisfy the designed production rate. An example of a takt time calculation is given below.

Given:

Effective hours per employee = 6.0
Company runs single shift operation
Designed production rate = 25 units/day

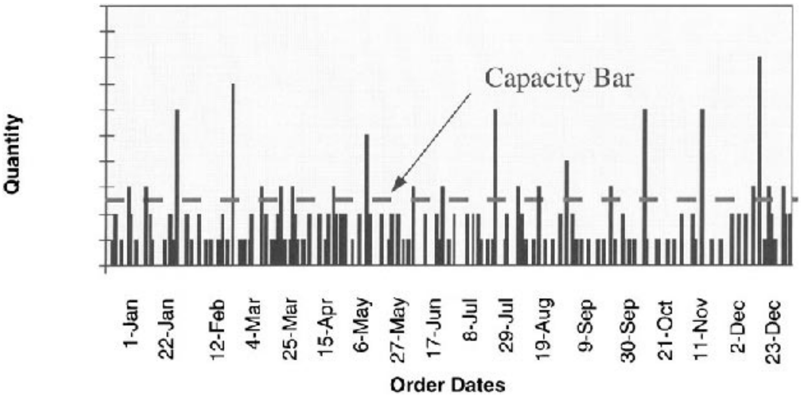


Figure 18 Daily order patterns.

$$\text{Takt} = \frac{6.0\,\text{hr} \times 1\,\text{shift/day}}{25\,\text{units/day}} = 0.24\,\text{hr/unit}$$
$$= 4.17\,\text{units/hr}$$

In this example approximately 4 units/hr will be produced by the production line. Each workstation in the production line would be balanced to be able to meet this production rate.

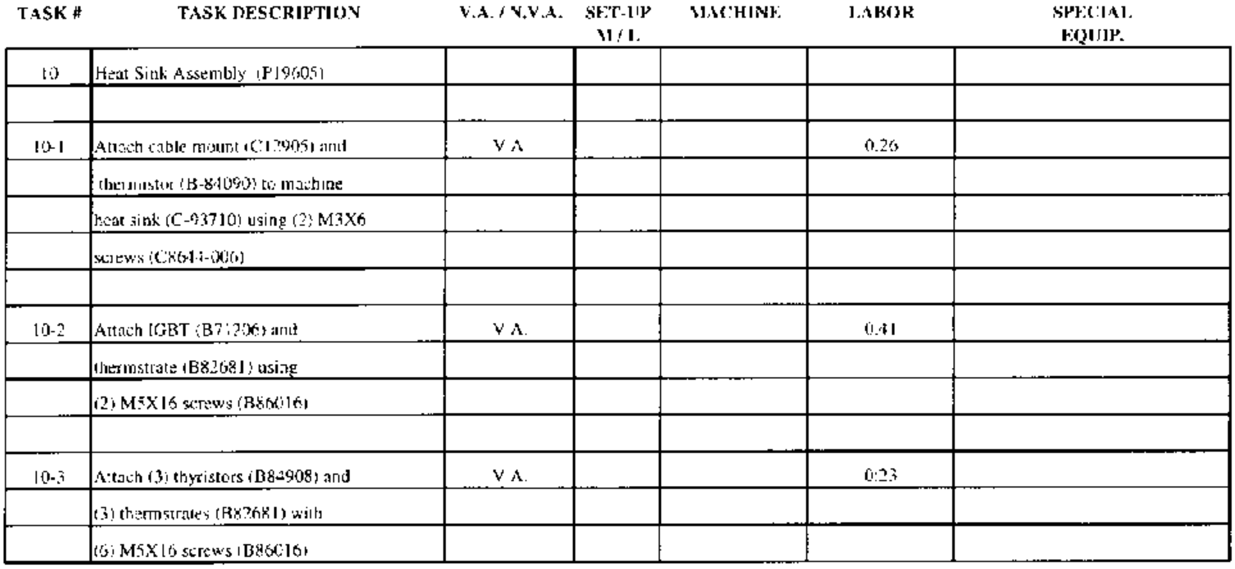
Develop Sequence of Events Sheets. The development of the sequence of events helps to outline the steps necessary to create the product or products. Sequence of events sheets determine a current performance benchmark and provide a measure of actual manufacturing cycle time. This documentation also aids in identifying areas where process improvements are required. It is important to note that sequence of events sheets are not the same as manufacturing routers in that they break the processes into discrete steps. An illustration of a typical sequence of events sheet is shown in Fig. 19.

An important feature of designing the sequence of events sheets is incorporating total quality control (TQC) method sheets into the documentation. Total quality control method sheets visually display workstation procedures and detail the required steps to produce products which meet the necessary quality standards. The quality of the products is enhanced through use of these TQC sheets and the ability to detect defects or rejects at early points in the process. Additionally, the sequence of events sheets are able to identify non-value-added steps which increase manufacturing costs and are not dictated by product specifications or customer demand. The identification of non-value-added steps allows for the calculation of the process design efficiency, below, which is used as an input to the continuous improvement process.

$$\text{Process design efficiency (\%)} = \frac{\text{Total value-added work}}{\text{Total work (including non-value-added activities)}}$$

Conduct Time Study Analysis and Brainstorm Methods Improvements. The purpose of the time study analysis is not to perform a micromotion study but to capture sufficient time study data to determine manufacturing cycle time and to aid in line balancing. During the time study data collection, any use of special equipment or the times required for searching for tools or waiting on equipment should be noted. In addition to recording time data, the purpose of this task is to identify potential process improvements. When possible, the production process should be videotaped to establish the current performance benchmark. Including line personnel and personnel not familiar with the production process, standard brainstorming techniques should be utilized to develop process improvements.

Develop Flow Line Design. The final flow line design is developed through the creation of "to-be" sequence of events sheets. Data from the time study is used to balance the line and adjust the work content at a



TASK #	TASK DESCRIPTION	V.A. / N.V.A.	SET-UP M / L	MACHINE	LABOR	SPECIAL EQUIP.
10	Heat Sink Assembly (P19605)					
10-1	Attach cable mount (C12905) and	V A			0.26	
	thermistor (B-84090) to machine					
	heat sink (C-93710) using (2) M3X6					
	screws (C8644-006)					
10-2	Attach IGBT (B71206) and	V A.			0.41	
	thermstrate (B82681) using					
	(2) M5X16 screws (B86016)					
10-3	Attach (3) thyristors (B84908) and	V A.			0:23	
	(3) thermstrates (B82681) with					
	(6) M5X16 screws (B86016)					

Figure 19 Typical sequence of events worksheet.

single station to be approximately equal to the calculated takt time. The goal of the line design is to eliminate WIP on the line and move towards one-piece flow. It is very common that process imbalances or bottlenecks occur that require alternate techniques that enable these processes to be integrated into the flow line.

Techniques to solve the process imbalances include: reducing cycle times at stations by removing non-value-added activities, acquiring additional resources to increase the capacity at the bottleneck stations, or creating WIP inventory by running bottleneck stations more hours than the rest of the line. One of the more common techniques is the use of in-process kanbans. In-process kanbans are placed on the downstream side of two imbalanced operations to balance the line. The calculation for the number of units in an in-process kanban is shown below:

$$\text{In-process kanban (\# of units)} = \frac{\text{Imbalance (min)} \times \text{daily capacity (units)}}{\text{Takt time (min)}}$$

An example of a situation where an in-process kanban is required is illustrated in Fig. 20. This flow line has three stations: drill press, mill, and assembly. The drill press and assembly operations require 30 min operations and the mill requires a 35 min operation. The daily capacity for this line is 42 units and the calculated takt time is 30 min. The 5 min imbalance between the mill and assembly requires an in-process kanban between these two stations. The calculation of the in-process kanban indicates that placing seven units between these two stations will allow this process to flow.

$$\begin{aligned}\text{In-process kanban (\# of units)} &= \frac{5\,\text{min} \times 42\,\text{units}}{30\,\text{min}}\\ &= 7\,\text{units}\end{aligned}$$

After the physical layout for the flow line is designed, the staffing requirements for the flow line are calculated. It is required that flow line personnel are able to adopt a "one-up one-down" philosophy. They must have the training and skills to staff adjacent workstations. The equation to calculate the number of required personnel is given below:

$$\text{\# personnel required} = \frac{\text{Designed production rate} \times \text{total labor time}}{\text{Effective work hours} \times \text{shifts/day}}$$

A sample calculation using the data provided below indicates that the appropriate staffing for the flow line would be three personnel.

Given:

Effective hours per employee = 6.0
Company runs single shift operation
Designed production rate = 25 units/day
Total labor time = 0.72 hr

$$\begin{aligned}\text{\# Personnel required} &= \frac{25\,\text{units/day} \times 0.72\,\text{hr/unit}}{6\text{hr/personnel} \times 1\,\text{shift/day}}\\ &= \frac{18\,\text{hr}}{6\,\text{hr/personnel}}\\ &= 3\,\text{personnel}\end{aligned}$$

Implement Multibin Kanban System. Developing a multibin kanban system requires significant data analysis and multifunctional team involvement. some of the major tasks involved include: identifying part, component, and subassembly usage; performing ABC analysis for all components and set weekly bin requirement quantities; determining production line packaging preferences; initiating vendor negotiations and/or training; determining company or vendor safety stocks; and establishing inventory control policies. Some discussion on each of these tasks, based on the authors' experiences, is included below.

Identify part, component, and subassembly usage. A worksheet is developed by exploding the bill of materials for the complete product. Typical

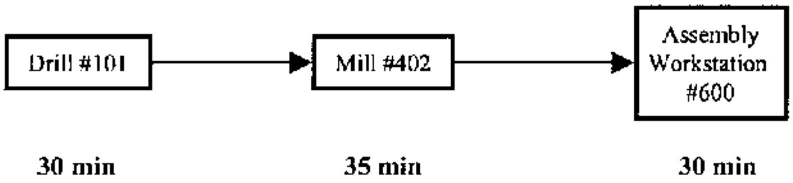


Figure 20 In-process kanban situation.

information recorded for each part includes: part number, description, quantity per assembly, quantity usage per week, unit cost, yield data, identification as external or internal part, and vendor information.

Perform ABC analysis and set weekly bin requirements. Calculate the annual costs of materials based on the designed capacity of the flow line. Items are segregated into ABC categories. Typical values for ABC categories are: "A" items represent 80%, "B" items represent 15%, and "C" items represent 5% of annual material costs. Bin sizes are sized to supply the flow line's designed weekly capacity plus compensation for rework, scrap, and overtime. Some typical bin quantities are sized to supply a week's quantity for "A" items, 2 weeks' quantity for "B" items, and 2 months' quantity for "C" items.

Determine flow line packaging preferences. Packaging quantities should be based on: part size in relation to handling, space requirements at workstations, and production line personnel preferences. The inventory storage should be decentralized at the workstation. When feasible, package quantities should be set equal to the flow line's designed daily capacity. Prepackaged parts will incur incremental costs but will ease inventory replenishment.

Initiate vendor negotiations and training. A key factor on identifying current or potential vendors is evaluate the vendor's ability to work under a multibin inventory system. Utilizing third party warehouses and consigned material have benefited many company's kanban systems.

Determine safety stocks. Safety stocks should be carried based on confidence in the vendor's ability to produce. Vendors normally will carry one bin at their facility ready for shipment. With short purchasing lead times, one bin may be sufficient. For longer lead times or problem vendors, an additional bin may be required to be held at the vendor's facility to ensure an uninterrupted supply of materials. Only extreme cases warrant safety stock held at the company's facility.

Establish inventory control policies. The responsibilities of buyer-planners, flow line or cell coordinators, and production line personnel must be identified and communicated. Identify procedures that are used for vendors with multiple parts for different flow lines. Many companies develop the responsibility for part replenishment to flow line personnel. Many times personnel will fax replenishment orders to vendors from the production floor.

2.4 WORLD-CLASS MANUFACTURING METRICS

World-class manufacturing (WCM) is a widely used but somewhat nebulous term. Numerous companies claim WCM status in their marketing promotions, but few have actually attained such status—and the bar is continuously on the rise. Since there is no definitive measure of world-class status, in this section the authors describe some of the attributes of world-class firms. These attributes span the entire manufacturing enterprise and include both qualitative and quantitative measures. Firms striving to attain WCM status can use these attributes to benchmark their progress towards achieving the WCM goal. These metrics are presented as a representative sample of current literature and should not be viewed as an all-encompassing list.

The foundation of WCM is the organizational culture, including leadership, strategic planning, employee empowerment, and human resources. Other key factors of WCM include customer focus, information technology, agility, quality, supplier management, and product development. Several key attributes in each of these areas are highlighted below. Some, such as invention to market of new products in less than 50 days, may appear as extremely ambitious goals but true world-class firms are achieving this goal. All the attributes are certainly not applicable across all industries, but as overall metrics the vast majority are relevant and appropriate.

2.4.1 Organizational Culture

2.4.1.1 Leadership

Top management is actively involved in creating a customer-oriented organization.

Management bases organizational values on published corporate strategy and vision statements.

Organizational values are communicated and reinforced through management actions.

Correspondence, newsletters, and internal meetings reflect organizational values.

CEO communicates quality values and organizational issues through internal and external publications (e.g. newsletter).

Employee recognition programs are spearheaded by top management.

Employees evaluate management's performance through annual leadership surveys.

2.4.1.2 Strategic Planning

Strategic planning process include customers and all levels of employees.

Core processes are reviewed annually for the purpose of improving customer focus and organizational performance.

Strategic objectives for departments or business units are developed and reviewed at least quarterly.

Each department/business unit maintains, communicates, and posts improvement goals and strategies.

Feedback on organizational performance is provided to all employees on a monthly basis.

2.4.1.3 Employee Empowerment

Organization actively invests in employees through training, educational reimbursement, etc. Typically 100% of employees are cross-trained.

Values are developed to ensure employees have opportunities to contribute to the organization.

High levels of participation are encouraged and solicited.

Employee involvement is encouraged, tracked, and measured.

Communication paths are open and available.

Suggestion systems and idea implementation systems are valued. Typical results of suggestion programs are one suggestion per employee per month with 98% implementation rate.

Employees are recognized and rewarded on a continual basis.

2.4.1.4 Human Resources

Team culture is supported through employee education programs.

Employee teams are involved in improving all core organizational processes that directly affect the workforce (e.g., personnel).

Employee and third-part satisfaction surveys are used to determine employee attitude and satisfaction levels.

Critical employee data is collected, analyzed, and used as an input into corporate continuous improvement programs (e.g., turnover, employee involvement, recognition, exit interviews).

Organizational recognition system fosters empowerment and innovation. Recognition, formal and informal, is given to both individual employees and teams.

Personal training requirements are identified through needs assessment surveys.

Training throughout the organization is aligned with the corporate strategy and measured against improved job performance.

A minimum of 20 annual training days per employee are provided.

Employee morale is measured and factored into improvement programs. Mean time between layoffs is 0 days.

Performance, recognition, and compensation system are integrated with strategic goals.

Organization is concerned with overall health and safety of employees and invests in wellness programs. Days since last accident approaches infinity.

2.4.2 Customer Focus

Customer orientation is a basic corporate value.

Customer values are integrated into corporate strategic plans.

Organization provides and encourages opportunities for customer contact in order to improve enhance customer relationships.

Customers are integrated into new product designs.

Semiannual customer focus groups and surveys benchmark organizational performance and define customer requirements.

Organizational advisory boards contain customer representation.

Employees responsible for customer contact are educated in customer interaction skills.

Customer complaint data is maintained, analyzed, and disseminated to the organization.

Written guarantees of organizational standards and performance are provided to customers.

Product or service quality in-process and after delivery to customer is tracked and indicates positive trend.

Customer turnover is measured and additional data gathered through exit interviews.

Market share has increased >10% due to customer focus (min. 3 years of data).

At least 98% of customer orders are delivered on-time.

Documented process exists for customer follow-up.

Customer orders are entered into manufacturing system within hours of customers placing orders rather than days.

At least 99% of customer orders are entered correctly into the enterprise management information system.

Less than 1% variability in order entry lead times.

Systems exist which focus on improving customer relationships

2.4.3 Information Technology

Introduction of new technology supports key business objectives and strategies (quality, agility, productivity, customers).

Information technologies integrate all business systems to provide real-time information to appropriate personnel.

Information systems are fully integrated and information is accessible throughout the organization.

Information collected is aligned with strategic goals (e.g., cycle time reduction).

Best practices are continuously sought to improve organization performance.

Benchmarks and competitive comparisons are utilized to improve critical processes.

Customer survey data provides input to improvement processes.

Financial data is summarized for employees on a regular basis.

Performance trends have been charted and posted throughout organization.

Key measures are collected on cycle times and costs for business and support processes. This data drives annual improvement objectives.

2.4.4 Agility

Manufacturing or production responsiveness is able to adapt to changing market conditions.

Flexible operating structures promote customer responsiveness.

Operations are run "lean."

Production designed around market demand and not economies of scale.

Process cycle times are continuously monitored and improved.

Daily production is satisfying market demand rather than min–max inventory levels.

Workforce is cross-trained.

Principles of JIT and other lean manufacturing techniques focus on reducing all classes of inventory.

Annual inventory turns >25.

Work-in-process inventory turns >100.

Production flow designed for lot sizes equal to 1.

On-hand inventory located at supplier's facility.

One hundred percent inventory accuracy.

Ratio of value added work to throughput cycle time $>50\%$.

Throughput time measured in hours rather than days or weeks.

Average setup times <10 min.

Utilized capacity exceeds 90%.

Lost production capacity due to breakdown losses $<1\%$.

Total productive maintenance program covers 100% of critical machinery.

Zero percent deviation from the weekly master production schedule.

2.4.5 Quality

Quality is built into the culture of the company through focused efforts to eliminate waste, customer returns, and non-value-added activities throughout the organization.

Integration of quality into the company culture as a method of operation as opposed to a program or slogan.

Quality is an organizational-wide responsibility and not the burden of one department or individual.

Thorough understanding and belief that quality improvement reduces overall costs.

Organizational awareness that quality is built into processes, not inspected in, and controlled processes produce defect-free products.

Detailed methods to map, measure and improve processes.

Total productive maintenance programs include predictive, preventive, and equipment improvement techniques.

Employees provided with training, tools, and information necessary to achieve high quality levels.

Less than 500 rejects per million parts.

Total cost of quality less than 5% of sales.

Control charts utilized throughout organization.

All business and support processes are documented.

Quality audits and supplier performance are tracked on an ongoing basis.

Quality-related data are posted and utilized throughout organization.

2.4.6 Product Development

Customers and key suppliers are integrated into cross-functional product design and development teams.

Investment in technologies and tools focus on reducing time to market, ensuring products meet customer needs and containing manufacturing costs.

Designs are reviewed, documented, and validated.

Tools and techniques include design for "X" (manufacturing, assembly/disassembly, environment, etc.), rapid prototyping (e.g., stereolithography), CAD/CAM/CIM, FEA, FMEA.

One hundred percent of product designs are evaluated based upon producibility.

Ninety-five percent of product designs meets cost targets.

Manufacturing process capability index (C_{pk}) > 1.33.

Less than 1% of first-year products requires engineering changes.

Engineering change response time < 1 day.

Product introduction index < 50 days (invention to market).

Active programs in place to reduce new product development time.

2.4.7 Supplier Management

Improvement of the quality and timeliness of raw materials and components as the primary performance metric.

Suppliers integrated into organization as an extended business unit or department. Programs to establish long-term partnerships developed.

Supply strategies are aligned with strategic objectives (customer service, quality, agility).

Total procurement cost is utilized as opposed to unit cost. Total costs include timeliness, accuracy, rejects, etc.

Education of suppliers to help improve supplier performance.

Suppliers receive regular performance feedback.

Procurement processes reevaluated on regular basis.

Supplier rating and certification program generate < 1% returns.

Certified supplier and supplier recognition system in place.

Delivered lot sizes < 22 days supply (< 37 days for international suppliers).

Manageable number of suppliers accomplished through using one supplier per item.

Number of alternative suppliers per item > 2.

Formal supplier certification process and published quality requirements exist.

Lead times controlled through JIT techniques.

REFERENCES

1. E Goldratt. The Goal. Great Barrington, MA: North River Press, 1989.
2. E Goldratt. Theory of Constraints. Great Barrington, MA: North River Press, 1990.
3. M Srikanth, H Cavallaro. Regaining Competitiveness—Putting the Goal to Work. Wallinford, CT: The Spectrum Publishing Company, 1987.
4. JR Costanza. The Quantum Leap in Speed to Market. Denver, CO: JIT Institute of Technology, 1994.

BIBLIOGRAPHY

Adair-Heeley C. The Human Side of Just-in-Time: How to Make the Techniques Really Work. New York: American Management Association, 1991.

Black JT. The Design of the Factory with a Future. New York: McGraw-Hill, 1991.

Chryssolouris G. Manufacturing Systems: Theory and Practice. New York: Springer-Verlag, 1992.

Cox III JF, Blackstone Jr JH, Spencer MS, eds. APICS Dictionary. American Production and Inventory Control Society, Falls Church, VA, 1995.

Diboon P. Flow manufacturing improved efficiency and customer responsiveness. IIE Solut (March): 25–29, 1997.

Fisher DC. Measuring Up to the Baldridge: A Quick and Easy Self-Assessment Guide for Organizations of All Sizes. New York: American Management Association, 1994.

Fogarty DW, Blackstone Jr JH, Hoffman TR. Production and Inventory Management. 2nd ed. Cincinnati, OH: Southwestern Publishing Co., 1991.

Gooch J, George ML, Montgomery DC. America Can Compete. Dallas, TX. George Group Incorporated, 1987.

Hall RW. Zero Inventories. Homewood, Ill: Dow-Jones Irwin, 1983.

Hall RW. Attaining Manufacturing Excellence: Just-in-time, Total Quality, Total People Involvement. Homewood, Ill: Dow-Jones Irwin, 1987.

Handfield RB. Re-Engineering for Time-Based Competition: Benchmarks and Best Practices for Production, R & D, and Purchasing. Westport, CT: Quorum Books, 1995.

Harding M. Manufacturing Velocity. Falls Church, VA: APICS, 1993.

Hopp WJ, Spearman ML. Factory Physics: Foundations in Manufacturing Management. Chicago, IL: Richard D. Irwin, 1996.

Kinni TB. America's best: industry week's guide to world-class manufacturing plants. New York: John Wiley & Sons, 1996.

Maskell BH. Performance measurement for world class manufacturing, part 1. Manuf Syst 7(7): 62–64, 1989.

Maskell BH. Performance measurement for world class manufacturing, part 2. Manuf Syst 7(8): 48–50, 1989.

Montgomery JC, Levin LO, eds. The Transition to Agile Manufacturing: Staying Flexible for Competitive Advantage. Milwaukee, WI: ASQC Quality Press, 1996.

Sandras W. Just-In-Time: Making It Happen. Essex Junction, VT: Oliver Wright Publications, 1987.

Schonberger R. Japanese Manufacturing Techniques. New York: The Free Press, 1982.

Sheridan JH. World-class manufacturing: more than just playing with the big boys. Industry Wk 239(13): 36–46, 1990.

Sheridan JH. How do you stack up? Industry Wk 234(4): 53–56, 1994.

Spearman ML, Woodruff DL, Hopp WJ. CONWIP: a pull alternative to kanban. Int J Prod Res 28(5): 879–894, 1990.

Steudel HJ, Desruelle, P. Manufacturing in the Nineties: How to Become a Mean, Lean, World-Class Competitor. New York: Van Nostrand Reinhold, 1992.

Suzaki, K. The New Manufacturing Challenge: Techniques for Continuous Improvement. New York: The Free Press, 1987.

Tompkins JA, White JA. Facilities Planning. New York: John Wiley, 1984.

Umble M, Srikanth M. Synchronous Manufacturing. Wallingford, CT: The Spectrum Publishing Company, 1995.

Urban PA. World class manufacturing and international competitiveness. Manuf Competit Frontiers 18(3/4): 1–5, 1994.

Wallace TF, Bennet SJ, eds. World Class Manufacturing. Essex Junction, VT: Oliver Wright Publications, 1994.

Chapter 6.3

Intelligent Manufacturing in Industrial Automation

George N. Saridis
Rensselaer Polytechnic Institute, Troy, New York

3.1 AUTOMATION

The evolution of the digital computer in the last 30 years has made it possible to develop fully automated systems that successfully perform human-dominated functions in industrial, space, energy, biotechnology, office, and home environments. Therefore, automation has been a major factor in modern technological developments. It is aimed at replacing human labor in

1. Hazardous environments
2. Tedious jobs
3. Inaccessible remote locations
4. Unfriendly environments.

It possesses the following merits in our technological society: reliability, reproducibility, precision, independence of human fatigue and labor laws, and reduced cost of high production.

Modern robotic systems are typical applications of automation to an industrial society [2]. They are equipped with means to sense the environment and execute tasks with minimal human supervision, leaving humans to perform higher-level jobs. Manufacturing on the other hand, is an integral part of the industrial process, and is defined as follows:

> Manufacturing is to make or process a finished product through a large-scale industrial operation.

In order to improve profitability, modern manufacturing, which is still a *disciplined art*, always involves some kind of automation. Going all the way and fully automating manufacturing is the dream of every industrial engineer. However, it has found several roadblocks in its realization: environmental pollution, acceptance by the management, loss of manual jobs, marketing vs. engineering. The National Research Council reacted to these problems by proposing a solution which involved among other items a new discipline called *intelligent manufacturing* [2].

Intelligent manufacturing is the process that utilizes intelligent control in order to accomplish its goal. It possesses several degrees of autonomy, by demonstrating (machine) intelligence to make crucial decisions during the process. Such decisions involve scheduling, prioritization, machine selection, product flow optimization, etc., in order to expedite production and improve profitability.

3.2 INTELLIGENT CONTROL

Intelligent control, has been defined as the combination of disciplines of artificial intelligence, operations research and control system theory (see Fig. 1), in order to perform tasks with minimal interaction with a human operator. One of its hierarchical applications, proposed by Saridis [3], is an architecture based on the *principle of increasing precision with decreasing intelligence (IPDI)*, which is the manifestation on a machine of the human organizational pyramid. The principle of IPDI is applicable at every level of the machine, reaffirming its universal validity. However, the coordina-

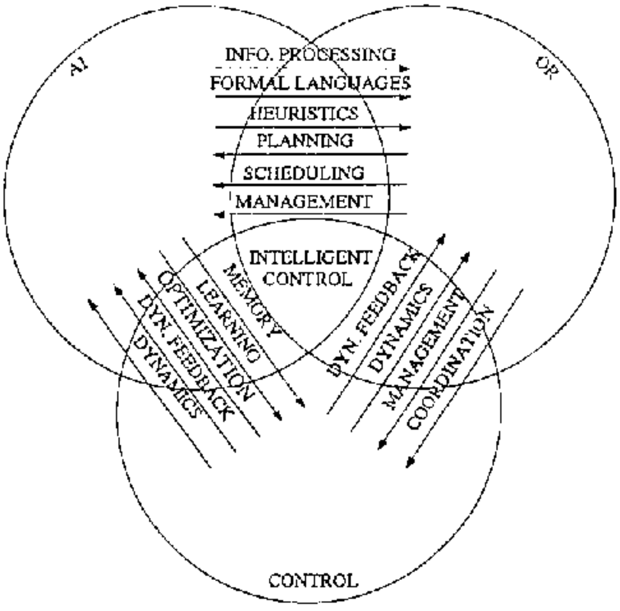


Figure 1 Definition of the intelligent control discipline.

tion may serve as a salient example of its application where the intelligence provided by the organization level as a set of rules is applied to the database provided by the execution level to produce flow of knowledge. The principle is realized by three structural levels (see Fig. 2):

1. Organization
2. Coordination
3. Execution.

Its implementation, according to Saridis, is based on the analytical realization of the various levels, with Boltzmann machines, Petri-net transducers, and control hardware respectively, using entropy as the common measure of evaluation. Robotics with sensor technology and data integration is a typical paradigm of such a procedure.

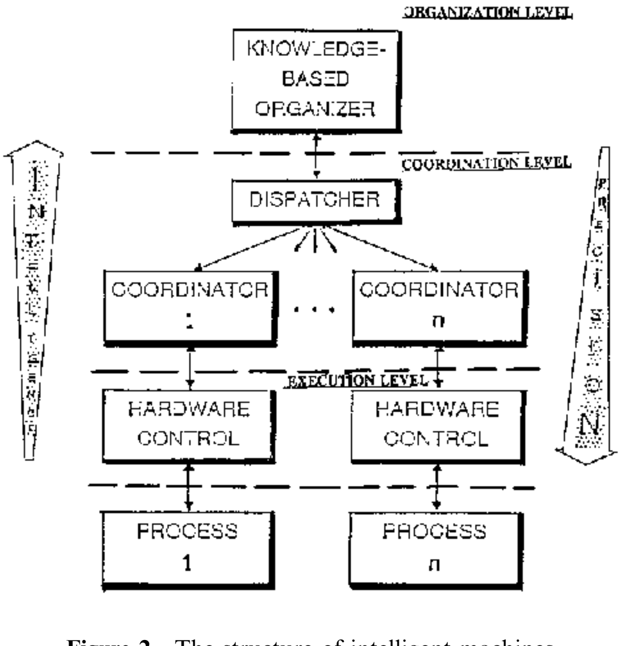


Figure 2 The structure of intelligent machines.

In order to implement an intelligent machine on analytical foundations, the theory of intelligent control has been developed by Saridis [4]. This theory assigns analytical models to the various levels of the machine and improves them through a generalized concept of selective feedback.

The intelligent control system is composed of three levels in decreasing order of intelligence and increasing order of precision as stipulated by the IPDI. However, with better understanding of the basics, new methodologies are proposed to analytically implement the various functions, without significantly changing the models at each level.

The *organization level* is designed to organize a sequence of abstract actions or rules from a set of primitives stored in a long-term memory regardless of the present world model. In other words, it serves as the generator of the rules of an inference engine by processing (intelligently) a high level of information, for reasoning, planning, and decision making. This can be accomplished by a two-level neural net, analytically derived as a Boltzmann machine by Saridis and Moed [5].

The *co-ordination level* is an intermediate structure serving as an interface between the organization and execution levels. It deals with real-time information of the world by generating a proper sequence of subtasks pertinent to the execution of the original command.

It involves co-ordination of decision making and learning on a short-term memory, e.g., a buffer. Originally, it utilized linguistic decision schemata with learning capabilities defined in Saridis and Graham [6], assigned subjective probabilities for each action. The respective entropies may be obtained directly from these subjective probabilities. Petri-net transducers have been investigated by Wang and Saridis [7], to implement such decision schemata. In addition, Petri nets provide the necessary protocols to communicate among the various co-ordinators, in order to integrate the activities of the machine.

Complexity functions may be used for real-time evaluation.

The *execution level* performs the appropriate control functions on the processes involved. Their performance measure can also be expressed as an entropy, thus unifying the functions of an intelligent machine.

Optimal control theory utilizes a nonnegative functional of the states of a system in the state space, which may be interpreted as entropy, and a specific control from the set of all admissible controls, to define the performance measure for some initial conditions, representing a generalized energy function. Minimization of the energy functional (entropy), yields the desired control law for the system.

In order to express the control problem in terms of an entropy function, one may assume that the performance measure is distributed over the space of admissible control according to a probability density function. *The differential entropy* corresponding to this density represents the uncertainty of selecting a control from all possible admissible feedback controls in that space. The optimal performance should correspond to the maximum value of the associated density. Equivalently, the optimal control should minimize the entropy function. This is satisfied if the density function is selected to satisfy *Jaynes' principle of maximum entropy* [3]. This implies that the average performance measure of a feedback control problem, corresponding to a specifically selected control, is an entropy function. The optimal control that minimizes the performance function maximizes the density function. The optimal control theory designed mainly for motion control, can be implemented for vision control, path planning and other sensory system pertinent to an intelligent machine by slightly modifying the system equations and cost functions. After all, one is dealing with real-time dynamic systems which may be modeled by a dynamic set of equations.

Hierarchically intelligent controls, as a theory, may be adapted to various applications that require reduced interaction with humans, from intelligent robotic to modern manufacturing systems. The heart of these operations is the specialized digital computer with variable programs associated with the specific tasks requested.

3.3 INTELLIGENT MANUFACTURING

Intelligent manufacturing is an immediate application of intelligent control. It can be implemented in the factory of the future by modularizing the various workstations and assigning hierarchically intelligent control to each one of them, the following tasks:

1. Product planning to the organization level
2. Product design and hardware assignment and scheduling to the co-ordination level
3. Product generation to the execution level.

The algorithms at the different levels may be modified according to the taste of the designer, and the type of the process. However, manufacturing can be thus streamlined and optimized by minimizing the total entropy of the process. Robotics may be thought as an integral part of intelligent manufacturing and be included as part of the workstations. This creates a versatile automated industrial environment where, every time, each unit may be assigned different tasks by just changing the specific algorithms at each level of the hierarchy (see Fig. 3). This approach is designed to reduce interruptions due to equipment failures, bottlenecks, rearrangement of orders, material delays, and other typical problems that deal with production, assembly, and product inspection. A case study dealing with a nuclear plant may be found in Valavanis and Saridis [1].

At the present time the application of such technology, even though cost-effective in competitive manufacturing, is faced with significant barriers due to [2]:

1. Inflexible organizations
2. Inadequate available technology
3. Lack of appreciation
4. Inappropriate performance measures.

However, international competition, and the need for more reliable, precisely reproducible products is direct-

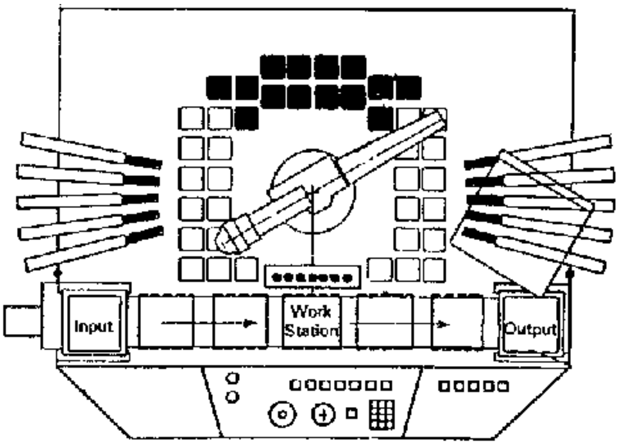


Figure 3 An intelligent automation workstation.

ing modern manufacturing towards more sophistication and the concept of an intelligent factory of the future.

REFERENCES

1. KP Valavanis, GN Saridis. Intelligent Robotic System Theory: Design and Applications. Boston, MA: Kluwer Academic Publishers, 1992, Boston, MA.
2. The Competitive Edge: Research Priorities for U.S. Manufacturing. Report of the National Research Council on U.S. Manufacturing. National Academy Press, 1989, Washington, DC.
3. GN Saridis. Architectures for intelligent controls. In: MM Gutta, NK Sinha, eds Intelligent Control Systems. IEEE Press, 1996, pp 127–148, Piscataway, U.
4. GN Saridis. "Toward the realization of intelligent controls." IEEE Proc 67(8): 1979.
5. GN Saridis, MC Moed. Analytic formulation of intelligent machines as neural nets. Symposium on Intelligent Control, Washington, DC, August 1988.
6. GN Saridis, JH Graham. Linguistic decision schemata for intelligent robots. Automatica IFAC J 20(1): 121–126, 1984.
7. F Wang, GN Saridis. A coordination theory for intelligent machines. Automatica IFAC J 35(5): 833–844, 1990.

Chapter 6.4

Measurements

John Mandel*
National Institute of Standards and Technology, Gaithersburg, Maryland

4.1 INTRODUCTION

The ancient Greeks believed that the mysteries of the universe could be elucidated by reasoning about them. Applying this philosophy to mathematics, they were very successful and developed the science of mathematics to a remarkable degree. But in the field of physics, chemistry, and biology, their philosophy did not allow them to make big advances. It was not until the Renaissance that scholars finally realized that, in these fields, it was necessary to perform experiments in order to discover the truth. The name of Galileo springs to mind, as one who performed experiments to uncover the laws of nature. Today we live in a period of experimentation in practically all fields of human endeavor. A fundamental aspect of modern experimentation is the making of measurements. Indeed, measurements transform the making of *qualitative* observations into the far more satisfactory establishment of *quantitative* facts.

Measurements can be discussed in many ways. A convenient way to look at them is to first classify them according to the field of scientific activity in which they fall. Thus we talk about physical and chemical measurements, biological measurements, economic measurements, demographic measurements, and many other types. In this chapter we will consider only physical and chemical measurements performed in laboratories. Also, because of limitations of space, we confine our discussion to one-way and two-way tables of measurements. Examples of such measurements include: the tensile strength of a steel bar, the heat of sublimation of gold, the Mooney viscosity of a sample of rubber, the amount of beta-carotene in a sample of human serum, and the amount of manganese in an ore. We are not concerned here with the way in which these measurements are carried out, but we are concerned with a close examination of the results of these measurements, with their *precision* and *accuracy*, and with the amount of confidence that we can have in them. These aspects of measurements are generally referred to as *statistical* properties. Indeed, the science of statistics can be of great usefulness in discussing the aspects of measurements with which we are concerned.

Let us discuss briefly the reasons for which we consider statistics in discussing measurements.

4.2 STATISTICS AND MEASUREMENT

The scientists who made measurements discovered early enough that the results of making repeated measurements of the same quantity seldom were identical to each other. Thus was born the concept that a measurement is the sum of two quantities: the ‘‘true’’ value of the quantity to be measured, and an ‘‘experimental error’’; in symbols,

$$y = \mu + \varepsilon \tag{1}$$

*Retired.

where y is the result of the measurement, μ is the true value, and ε is the experimental error.

Statisticians refined this idea by stating that ε is a member of a "statistical distribution" of experimental errors. Hence, to study measurements one would have to study statistical distributions. The names of Gauss and Laplace figure prominently in the establishment of the so-called "normal distribution" as the favored distribution for experimental errors. Today we know that this is not necessarily the case, and many nonnormal distributions are considered by scientists. However, the fundamental idea that there is a random, statistical element in experimental errors still prevails, and, therefore, statistics enters as a natural element in the study of experimental errors.

4.3 ONE-WAY CLASSIFICATIONS

Equation (1) is not always adequate to represent measurements. Often, a number of different laboratories are involved in an experiment, and some of the laboratories may decide to repeat their experiments several times. Table 1 presents data obtained by 10 laboratories in the determination of the heat of sublimation of gold [1]. Two methods were used, referred to as "second-law" and "third-law" procedures. The complete set of data, given in Table 1, shows that different laboratories made different numbers of replicate measurements.

A reasonable mathematical model for this experiment is given by the equation

$$y_{ij} = L_i + e_{ij} \tag{2}$$

where y_{ij} is the jth replicate obtained by laboratory i. L_i is the *systematic* error of laboratory i, and e_{ij} is the random error associated with the jth replicate in laboratory i.

Statistical textbooks present "analytical" methods of dealing with model equations. These are mathematical treatments based on a number of prior assumptions. The assumptions are seldom spelled out in detail and in reality many of them are often simply false. It is strange, but true, that this approach is essentially the one that the ancient Creeks used for the study of the universe, an approach that proved to be unproductive in all sciences except mathematics. We prefer to use the experimental method and study the data first by graphing them in an appropriate way.

Table 1 Heat of Sublimation of Gold

Lab							
Second-law data							
1	89,664	90,003					
2	87,176	91,206	92,638				
3	85,924	84,868	90,142				
4	89,133	85,876	89,815				
5	88,508	87,758					
6	88,464	85,015	85,687	89,500	88,911	86,480	90,212
7	86,862	86,300	87,041	86,927	86,931		
8	87,338	87,702	91,092	90,977			
9	81,981	98,741	77,563	84,593			
10	98,536	99,568	70,700	77,335	89,209		
Third-law data							
1	88,316	88,320					
2	88,425	87,626	87,747	87,975	88,120		
3	87,786	88,108	87,477				
4	88,142	88,566	87,514				
5							
6	87,912	87,917	87,882	87,791	87,638	87,845	87,489
7	86,653	86,517	86,570	86,691	86,531		
8	85,097	85,304	85,679	86,016			
9	85,984	89,821	85,191	84,624			
10	87,911	88,041	87,911	87,635	87,693		

4.4 A GRAPHICAL REPRESENTATION

For the time being, let us consider the data in Table 1 as originating in 19 laboratories. This signifies that we ignore temporarily the classification into second and third law. Thus the laboratory index i is considered to vary from 1 to 19.

Denoting by $\bar{x}$ the grand average of all 76 measurements in Table 1, and by s their overall standardized deviation, we calculate for each measurement the standardized value:

$$h_{ij} = \frac{y_{ij} - \bar{x}}{s} \tag{3}$$

where y_{ij} is the jth measurement obtained by laboratory i.

Equation (3) is simply a *linear* transformation of the measurement y_{ij}. A plot of h_{ij} is shown in Fig. 1. The h-values are portrayed as vertical bars in groups of the 19 laboratories.

Despite its simplicity, the graph is highly instructive. It reveals that:

1. There are noticeable differences between laboratories, even within each of the two methods (second and third law), in terms of agreement between replicates (precision).
2. The third-law data are visibly more precise than the second law data.
3. In the second-law data, laboratories 9 and 10 are appreciably less precise than the other laboratories.
4. In the third-law data, laboratories 7, 8, and 9 are less precise than the other laboratories.
5. The two methods are, on the average, very close to each other in terms of overall average value.

There is no way that a purely analytical approach would have revealed these facts. Any analytical

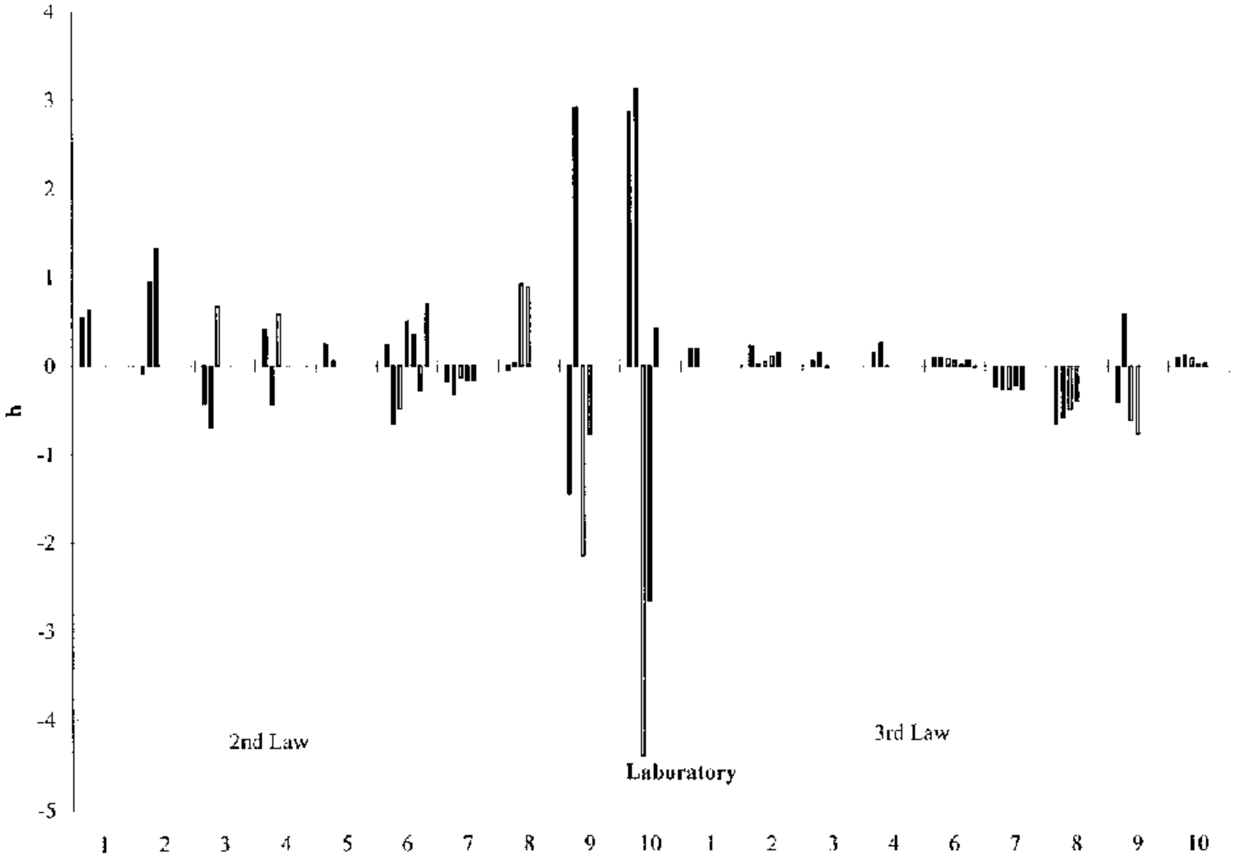


Figure 1 Heat of sublimation of gold.

approach would have to start with assumptions that can be shown not to be true for these data. Analytical approaches are based on pooled* estimates of the variability. We see from Fig. 1 that the pooling of within-laboratory variability, in which the results of laboratories 9 and 10 (second law) are pooled with the within-laboratory variability of the other laboratories is a distortion of the facts.

4.5 SEPARATE ANALYSES FOR SECOND- AND THIRD-LAW DATA

At this point it becomes reasonable to analyze the two sets of data separately. The main reason for this is the large difference in replication error (within-laboratory variability) between the second- and third-law data.

4.6 AN ITERATIVE ANALYSIS OF ONE-WAY DATA [2]

Let $\bar{y}_i$ be the average of all measurements made by laboratory i, and s_i the standard deviation of these measurements. Suppose there are N laboratories, and let n_i be the number of replicate measurements made by laboratory i. The statistic s_i is an estimate for the population parameter σ_i. The σ_i are, of course, *internal* (i.e., within-laboratory) variabilities. Let σ_B represent the standard deviation *between* laboratories, and Z represent the sample estimate of σ_B^2.

The mathematical model for our data is, again,

$$y_{ij} = \mu + L_i + \varepsilon_{ij} \tag{4}$$

where y_{ij} is the jth replicate made by laboratory i; L_i is the amount by which laboratory i differs, on the average, from the "true" value μ; and ε_{ij} is the replication error of y_{ij}; σ_B, the between-laboratory standard deviation is the standard deviation of L_i; and s_i is the estimated standard deviation of ε_{ij} for laboratory i.

For σ_i we have the estimates s_i obtained from the data of laboratory i[†], but for σ_B we have as yet no estimate. Our goal is to construct such an estimate, i.e., to find Z (an estimate of σ_B^2). We do this by an iterative process.

The total *weight* (reciprocal of variance), including within and between laboratory variability, of $\bar{y}_i$ is $1/(\sigma_B^2 + (\sigma_i^2/n_i))$. Represent this weight by w_i:

$$w_i = \frac{1}{\sigma_B^2 + \sigma_i^2/n_i} \tag{5}$$

Now, if all $\bar{y}_i$ are properly weighted, then the variance among them is unity. Hence

$$\frac{\sum w_i(\bar{y}_i - \bar{\bar{y}})}{N-1} = 1 \tag{6}$$

where the weighted average of all $\bar{y}_i$ (i.e., $\bar{\bar{y}}$) is given by

$$\bar{\bar{y}} = \frac{\sum w_i \bar{y}_i}{\sum w_i} \tag{7}$$

Let us start with an initial value, Z_0, for σ_B^2. Using this value, estimates of w_i, for all i, can be calculated using Eq. (5). Hence the quantity

$$F_0 = \sum w_i(\bar{y}_i - \bar{\bar{y}})^2 - (N-1) \tag{8}$$

can be calculated, since $\bar{\bar{y}}$ is given by Eq. (7). If $F_0 = 0$, then Z_0 is the correct value of Z. If not, then, Z_0 should be corrected, say by ΔZ:

$$Z = Z_0 + \Delta Z \tag{9}$$

If ΔZ is considered as a differential, then F becomes $F + \Delta F$, and

$$\Delta F = \frac{\partial F}{\partial Z} \Delta Z \tag{10}$$

It can be shown that

$$\frac{\partial F}{\partial Z} = -\sum_i w_i^2(\bar{y}_i - \bar{\bar{y}})^2 \tag{11}$$

On the other hand, if ΔZ is properly chosen, we should obtain a zero value for F, $F = 0$, hence

$$F_0 + \Delta F = 0 \tag{12}$$

or

$$\Delta F = -F_0 \tag{13}$$

As a result, Eq. (10) becomes

$$\Delta Z = \frac{-F_0}{-\sum w_i^2(\bar{y}_i - \bar{\bar{y}})^2} = \frac{F_0}{\sum w_i^2(\bar{i} - \bar{\bar{y}})^2} \tag{14}$$

We can continue this iterative process starting at every iteration with the previous value of F, instead of F_0, until ΔZ becomes so small that no further correction is required. The last value of Z is then accepted as an approximate estimate of σ_B^2.

The best estimate for μ can be shown to be $\bar{\bar{y}}$. Its standard error is $1/\sqrt{\sum_i w_i}$.

* A pooled estimate of a standard deviation is obtained by averaging the squares of all standard deviations and taking the square root of this average.

[†] We assume that n_i is at least 2, for all i.

A good starting value is $Z_0 = 0$. If, anywhere in the iterative process, the corrected value of Z is negative, then the variance between laboratories, σ_B^2, is considered to be zero.

4.7 THE GOLD DATA

Applying the above theory to the heat of sublimation of gold values, we obtain Tables 2 and 3 corresponding, respectively, to the second- and third-law data. For the second-law data, the estimate s_B is zero; for the third-law data s_B is 897.

Table 4 exhibits the results of the statistical analysis. For the second-law data, the estimate for μ is 88,156 with a standard error of 1934. For the third-law data, the estimate of μ is 87,439, with a standard error of 315. The difference between the two estimates of μ is 717, which is not significant, since the standard error of their difference is $\sqrt{1934^2 + 315^2} = 1959$. Thus, the two laws give consistent results, to within their uncertainties.

The above analysis, unlike the popular analysis of variance technique, makes no use of pooled estimates of variance. It treats the laboratories individually, allowing each one to have its own standard deviation of replication error. It allows for different numbers of replicates for the various laboratories. It has been shown [3] that it gives essentially the same results as an analysis based on the maximum likelihood principle for underlying normal distributions. Unlike this technique, however, it makes no assumptions about the nature of the underlying statistical distribution of errors, and of that of laboratory-to-laboratory variability.

Table 2 Heat of Sublimation of Gold: Second-Law Data Analysis

$N(I)$	$X(I)$	$S(I)$	$W(I) \times 1000$
2	89,833.500	239.709	0.034807
3	90,340.000	2,832.107	0.000374
3	86,978.000	2,790.512	0.000385
3	88,274.660	2,105.109	0.000677
2	88,133.000	530.330	0.007111
7	87,752.710	2,013.732	0.001726
5	86,812.210	293.458	0.058060
4	89,277.250	2,035.073	0.000966
4	85,719.500	9,153.021	0.000048
5	87,069.600	12,796.400	0.000031

Table 3 Heat of Sublimation of Gold: Third-Law Data Analysis

$N(I)$	$X(I)$	$S(I)$	$W(I) \times 1000$
2	88,318.000	2.828	0.001248
5	87,978.600	315.137	0.001218
3	87,790.340	315.522	0.001198
3	88,074.000	529.286	0.001118
7	87,782.000	161.121	0.001243
5	86,592.400	76.405	0.001247
4	85,524.000	406.947	0.001187
4	86,405.000	2,344.643	0.000460
5	87,838.210	168.894	0.001240

4.8 TWO-WAY TABLES

We will discuss the analysis of two-way tables by means of an example of real data, dealing with the density of aqueous solutions of ethyl alcohol [4]. The data are shown in Table 5. Since the values were given with six decimals, a rounding in the computer might cause a loss of accuracy. Therefore the data were coded by multiplying them by 1000. Table 5 also exhibits the seven column averages. The usual analysis of variance shown in Table 6 is highly uninformative. Indeed, a mere glance at the data convinces us that the rows are vastly different from each other, and so are the columns. Furthermore, the mean square for interaction is *not* a measure of any useful parameter, as we will show below.

However, if we make a graph of each row of the table against the corresponding column average, as seen in Fig. 2 [2] we obtain a bundle of essentially straight lines. It is apparent that the lines are *not* parallel. The fact that they are essentially straight lines allows us to postulate the following model:

$$y_{ij} = \alpha_i + \beta_i x_i + \varepsilon_{ij} \tag{15}$$

where y_{ij} is the value in row i and column j, x_j is the column-average of all values in column j, ε_{ij} is the amount by which the fit fails; and α_i and β_i are

Table 4 Heat of Sublimation of Gold: Parameters for Second- and Third-Law Data

Law	Estimate of μ	Std. error of estimate of μ	Std. dev. between laboratories
Second	88,156	1,934	0
Third	87,439	315	897

Table 5 Density of Aqueous Solutions of Alcohol: Coded Data*

991.108	990.468	989.534	988.312	986.853	985.167	983.266
983.962	983.074	981.894	980.460	978.784	976.886	974.780
973.490	971.765	969.812	967.642	965.290	962.750	960.034
969.190	967.014	964.664	962.129	959.440	956.589	953.586
959.652	956.724	953.692	950.528	947.259	943.874	940.390
942.415	938.851	935.219	931.502	927.727	923.876	919.946
921.704	917.847	913.922	909.938	905.880	901.784	897.588
899.323	895.289	891.202	887.049	882.842	878.570	874.233
875.989	871.848	867.640	863.378	859.060	854.683	850.240
851.882	847.642	843.363	839.030	834.646	830.202	825.694
826.442	822.174	817.866	813.515	809.120	804.672	800.171
798.116	793.882	789.618	785.334	781.027	776.682	772.297
Column averages						
916.1061	913.0482	909.8689	906.5682	903.1606	899.6446	896.0188

* Original values were multiplied by 1000. Thus the first value, in original units, is 0.991108.

the slope and intercept of the straight line corresponding to laboratory i. Equation 15 differs from the usual analysis of variance model, in that the lines corresponding to the various laboratories may have different slopes. In the usual analysis of variance model, all $\beta_i \equiv 1$ so that all lines have the same slope, i.e., are parallel to each other.

Equation (15) can be written in a more useful form,

$$y_{ij} = H_i + \beta_i(x_j - \bar{x}) + d_{ij} \tag{16}$$

where H_i is the "height" of the fitted line, and can be shown to be the ordinate of the line corresponding to the value $\bar{x}$, on the abscissa, $\bar{x}$ being the average of all x_j; the residuals, d_{ij}, are the quantities by which the estimate, $H_i + \beta_i(x_j - \bar{x})$, fails to reproduce exactly the measured value, y_{ij}. Comparing Eqs. (15) and (16) we see that

$$H_i = \alpha_i + \beta_i \bar{x} \tag{17}$$

It can be shown that for normal distributions, H_i and β_i are not correlated.

Table 8 lists for each laboratory, H_i, b_i (which is an estimate for β_i), and the standard deviation, $(s_d)_i$, of the residuals, i.e., the quantities by which the fitted values differ from the observed values of y_{ij}. The values of $(s_d)_i$ are seen to be very small compared to the observed values of y_{ij}.

Table 6 Density of Aqueous Solutions of Alcohol: Classical Analysis of Variance

Source	DF	SS	MS
Rows	11	360,394	32,763
Columns	6	3,772	629
$R \times C$	66	404.30	6.1258

It is, indeed, readily seen that the residuals, d_{ij}, are much smaller than the square root of the mean square of the interaction in the analysis of variance. In fact, the analysis of variance can be expanded to take into account the different slopes corresponding to different laboratories. To this effect, the sum of squares of interaction in Table 6 is partitioned into two parts: a systematic portion, denoted by "slopes" and a random part, which is simply the sum of squares of all residuals, d_{ij}.

This is done as follows. Let $\bar{y}_i$ represent the average of all the elements in row i. Then, b_i, the sample estimate for β_i, is given by the equation

$$b_i = \frac{\sum_j (y_{ij} - \bar{y}_i)(x_j - \bar{x})}{\sum_j (x_j - \bar{x})^2} \tag{18}$$

The sum of squares for the "slopes" is given by the equation

$$\mathrm{SS}_{\mathrm{slopes}} = \sum_i (b_i - 1)^2 \sum_j (x_j - \bar{x})^2 \tag{19}$$

Finally, the sum of squares of interaction in the analysis of variance table is partitioned as follows:

$$\mathrm{SS}_{R\times C} = \mathrm{SS}_{\mathrm{slopes}} + \mathrm{SS}_{\mathrm{residual}} \tag{20}$$

where $\mathrm{SS}_{\mathrm{residual}}$ can be calculated by subtracting $\mathrm{SS}_{\mathrm{slopes}}$ from $\mathrm{SS}_{R\times C}$. Alternatively, $\mathrm{SS}_{\mathrm{residual}}$ can be shown to be equal to the sum of squares of all d_{ij}:

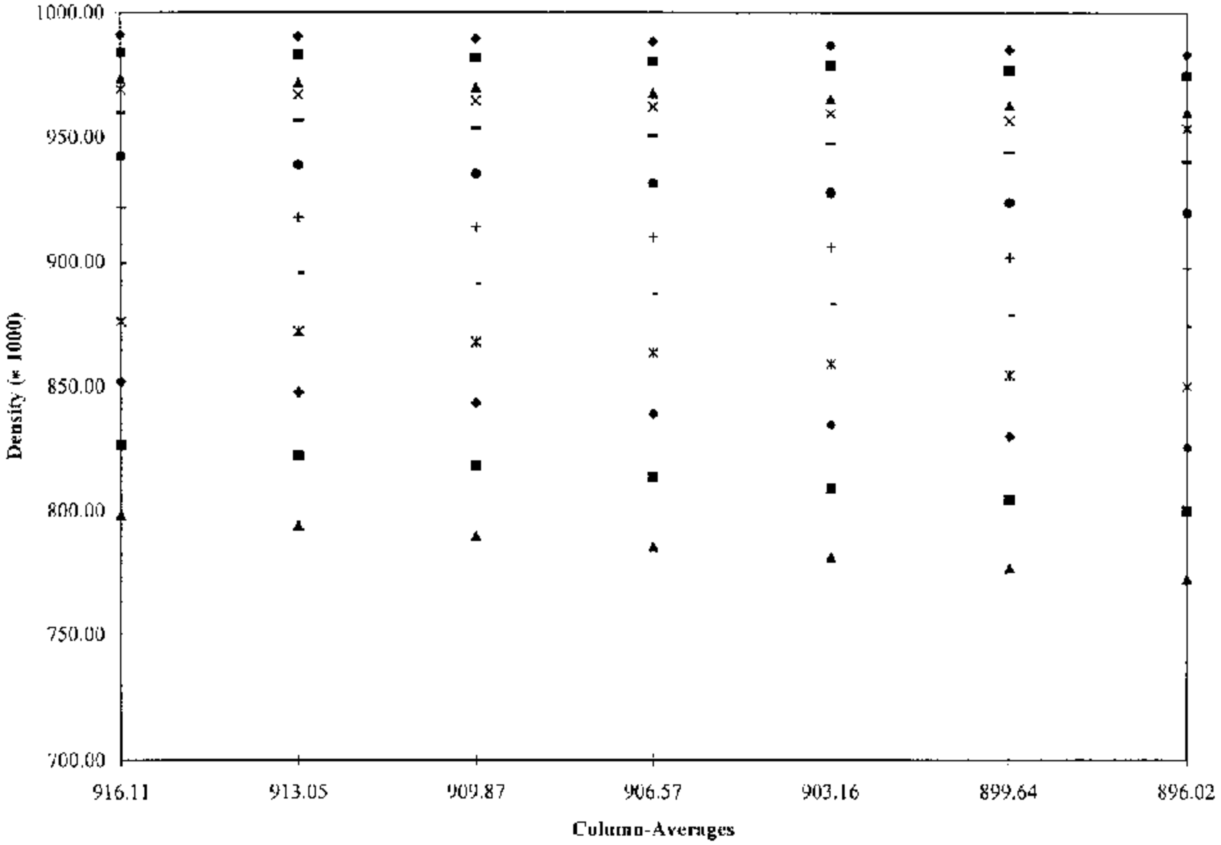


Figure 2 Density of aqueous solutions of alcohol: row-linear plot.

$$\mathrm{SS}_{\mathrm{residual}} = \sum_i \sum_j d_{ij}^2 \tag{21}$$

Applying this to our data, we obtain the expanded analysis of variance shown in Table 7. Note that the interaction mean square has dropped from 6.1258 to 0.0569. Obviously, the value 6.1258 has no physical interpretation. The square root of 0.0569 is 0.24, which is a measure of the magnitude of the residuals for the coded values [see Eq. (16)]. For the original values, the standard deviation would be 0.0024. Equation (16) is an expression of the "row-linear mode," because it is the result of considering all rows to be linear functions of each other, and, therefore, also linear functions of the column averages. Contrary to the classical analysis of variance, the straight lines representing rows of the table of original data have different slopes.

The expanded analysis of variance is still an unsatisfactory representation of the data because it is based on pooled estimates of variability. A more adequate representation is given by Table 8, which lists the parameters for each row of the table of data, namely, the average (H_i), the slope b_i (versus the column averages), and the standard deviation of fit (which measures the magnitude of the deviations of the fitted line from the observed values). The actual residuals in coded units are listed individually in Table 9.

Table 7 Density of Aqueous Solutions of Alcohol: Expanded Analysis of Variance

Source	DF	SS	MS
Rows	11	360,394	32,763
Columns	6	3,772	629
$R \times C$	66	404.30	6.1258
Slopes	11	401.17	36.47
Residual	55	3.13	0.0569

Table 8 Density of Aqueous Solutions of Alcohol: Row-Linear Parameters

Row	Height	Slope	Standard deviation of fit
1	987.8150	0.3941	0.4240
2	979.7770	0.4608	0.3902
3	967.2550	0.6708	0.2480
4	961.8017	0.7782	0.1594
5	950.3028	0.9587	0.0112
6	931.3623	1.1171	0.1102
7	909.8090	1.4988	0.1459
8	886.9297	1.2483	0.1644
9	863.2626	1.2805	0.1758
10	838.9228	1.3014	0.1908
11	813.4229	1.3065	0.2070
12	785.2795	1.2841	02382

4.9 FIT IN TERMS OF ALCOHOL CONCENTRATION AND TEMPERATURE [2]

So far we have established that the density data obey, to a good approximation, a row-linear model: each row is a linear function of the column averages. (Therefore the rows are linearly related to each other.) The two parameters describing the fitted straight line for each row are listed in Table 8; they are the height (ordinate corresponding to $\bar{x}$), the slope, and the standard deviation of fit. The slopes are definitely very different from each other. The fits to these lines are very good, though not perfect.

The authors of this study were, however, interested in the relation of the data to the concentration of the alcohol in the solutions, and to the temperature. They provide the alcohol concentrations as well as the temperatures. We can refer to these two variables as "marginal labels." Each row of the table of density values is associated with a specific alcohol concentration, and each column of the table with a specific temperature. The latter were 10, 15, 20, 25, 30, 35, and 40°C for the seven columns of the table. The alcohol concentrations, associated with the 12 rows of the table were (percent): 4.907, 9.984, 19.122, 22.918, 30.086, 39.988, 49.961, 59.976, 70.012, 80.036, 90.037, and 99.913. The remaining task, in the analysis of the data, is to express the two parameters, H_i and b_i, as functions of these concentrations, and the column variable, x_j, as a function of the temperature. We found that a slightly modified quadratic function generally provides a good fit. This function is

$$\hat{y} = A + Bz + Cz^2 \tag{22}$$

where z is a so-called Box–Cox transformation of the abscissa variable or to a quantity linearly related to this variable, and $\hat{y}$ is the fitted density. The variable z is given by the equation

$$z = \frac{x^\alpha - 1}{\alpha} \tag{23}$$

where x is the abscissa variable and α is a constant. Further details of this fitting process can be found in Mandel [2].

The fits for H_i, b_i, and x_j are shown in Figs. 3, 4, and 5 respectively. They are excellent for H_i and x_j, and good for b_i.

The formulas for our curve fit are as follows:

$$u = 2 + \frac{x - a}{b}$$

where x is the marginal label

Table 9 Density of Aqueous Solutions of Alcohol: Residuals from Row-Linear Fit

−0.5540	0.0111	0.3299	0.4086	0.2925	−0.0080	−0.4802
−0.5134	0.0078	0.2930	0.3800	0.2744	−0.0033	−0.4384
−0.3126	0.0137	0.1934	0.2376	0.1714	−0.0100	−0.2939
−0.2074	−0.0038	0.1202	0.1536	0.1163	0.0014	−0.1802
−0.0089	−0.0052	0.0109	0.0113	0.0092	−0.0048	−0.0128
0.1483	0.0004	−0.0799	−0.1096	−0.0780	−0.0010	0.1194
0.1938	0.0025	−0.1113	−0.1385	−0.1117	0.0072	0.1576
0.2086	−0.0082	−0.1265	−0.1592	−0.1127	0.0045	0.1935
0.2275	0.0021	−0.1348	−0.1703	−0.1250	0.0003	0.2000
0.2557	−0.0046	−0.1459	−0.1831	−0.1324	−0.0005	0.2103
0.2658	−0.0069	−0.1609	−0.1994	−0.1423	0.0037	0.2399
0.3028	−0.0046	−0.1862	−0.2320	−0.1635	0.0064	0.2770

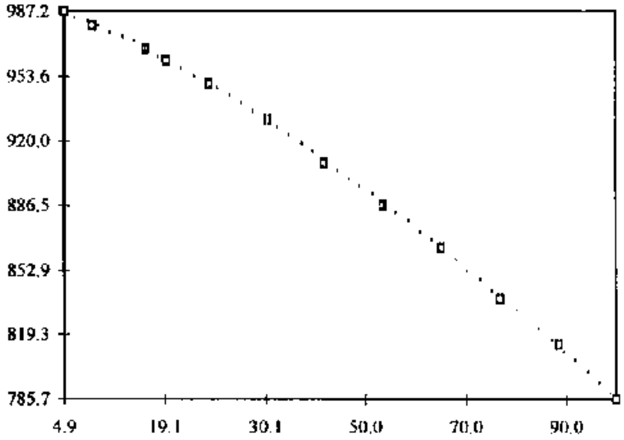


Figure 3 Curve 1: height versus percent alcohol.

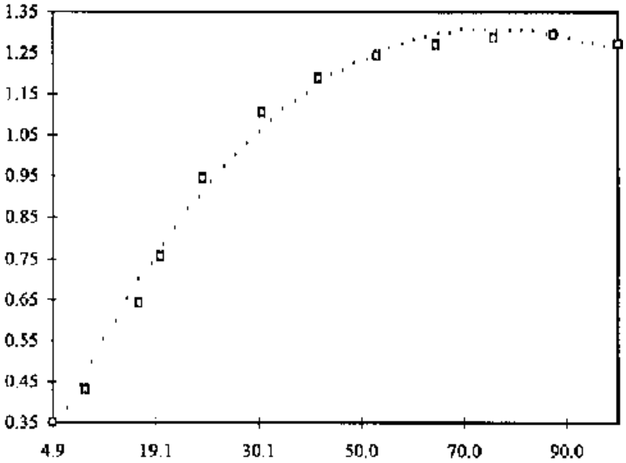


Figure 4 Curve 2: slope versus percent alcohol.

$$z = \frac{x^{\alpha} - 1}{\alpha}$$

The fitted value $= A + Bz + Cz^2$.

The parameters occurring in these equations are given in Table 10.

If the expressions of H_i, b_i, and x_j in terms of concentration of alcohol and temperature are substituted for the symbols H_i, b_i, and x_j in Eq. (16), there result estimated values of y_{ij} for all i and j. The differences between these estimates and the measured values of y_{ij} (Table 5) are listed in Table 11. These are in coded units. In the original units, the residuals are 1000 times smaller.

These residuals are larger than those in Table 9. This is due to the additional uncertainties introduced by the fitting of H_i, b_i, and x_j by quadratic equations. It should be remembered that the residuals are in coded units. In the original units they would be 1000 times smaller.

The current philosophy in the statistical profession is that the purpose of a statistical analysis is to "fit models to data." We have done this for the density

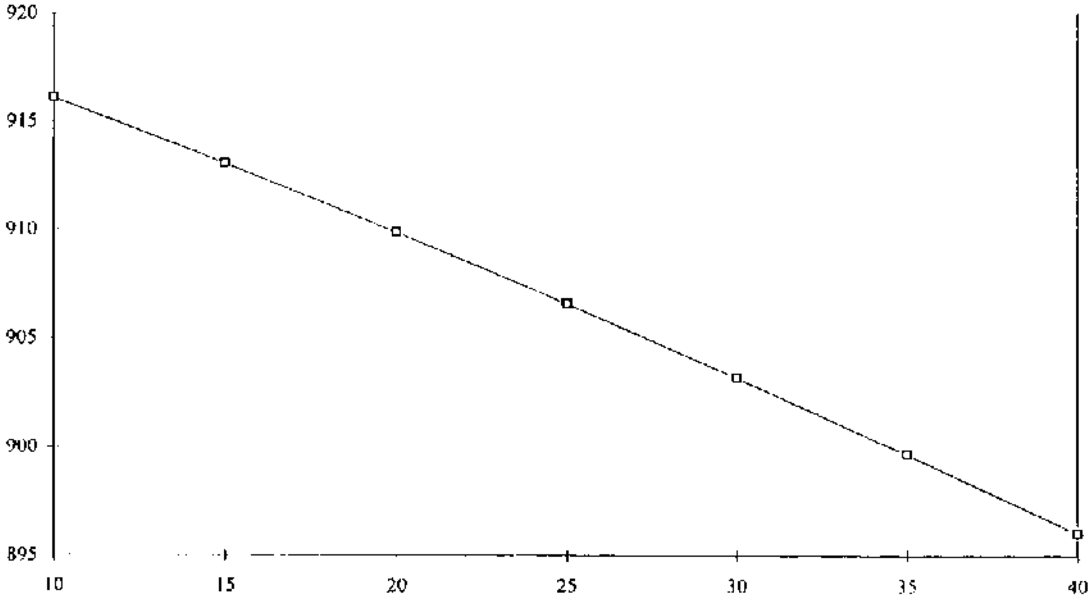


Figure 5 Curve 3: column averages versus temperature.

Table 10 Density of Aqueous Solutions of Alcohol: Parameters for Fits of Curves

Curve	*a*	*b*	*α*	*A*	*B*	*C*
H versus % alcohol	48.07834	106.1125	0.2	975.2964	228.3583	−417.9115
b versus % alcohol	48.07834	106.1125	0.2	−3.176193	9.874077	−5.42575
x versus temperature	25	26.45752	1.2	922.656	−14.31696	−0.5190612

Table 11 Density of Aqueous Solutions of Alcohol: Residuals from Fit in Terms of Alcohol Concentration and Temperature

0.508	0.945	1.128	1.064	0.802	0.352	−0.275
−1.742	−1.115	−0.722	−0.526	−0.517	−0.675	−0.987
−1.296	−0.827	−0.500	−0.309	−0.218	−0.237	−0.354
−0.330	−0.067	0.116	0.206	0.231	0.181	0.067
1.144	1.061	0.982	0.878	0.774	0.656	0.541
1.216	0.944	0.730	0.555	0.443	0.374	0.343
0.125	−0.107	−0.267	−0.351	−0.376	−0.308	−0.210
−0.603	−0.784	−0.870	−0.877	−0.797	−0.643	−0.417
−0.503	−0.658	−0.728	−0.702	−0.586	−0.385	−0.108
0.339	0.129	0.036	0.040	0.140	0.326	0.592
1.158	0.888	0.732	0.682	0.735	0.880	1.114
−0.037	0.375	−0.593	−0.685	−0.658	−0.527	−0.298

data, but unlike many published analyses, we have checked the model at every step. The *advantage* of our fit is that, for *every* concentration of alcohol between 5 and 100%, and for *every* temperature between 10 and 40°C, our formulas will give acceptable estimates of density. The disadvantage of our method of fitting is that the fit results in a slight loss of precision (see Table 11).

For example, for an alcohol concentration of 15% and a temperature of 32°C, we obtain the estimates

$$H_i = 973.9616$$

$$b_i = 0.62187$$

$$x_j = 901.7688$$

Then

Estimate of coded density

$$= 973.9616 + 0.62187 \times (901.7688 - 906.3451)$$

$$= 971.116$$

Estimate of density = 0.971116

Whoever uses our analysis should decide whether the analysis has shown itself to be useful to him or her.

REFERENCES

1. J Mandel, R Pauli. Interlaboratory evaluation of a material with unequal numbers of replicates. Anal Chem 42:1194–1197, 1970.
2. J Mandel. Analysis of Two Way Layouts. New York: Chapman and Hall, 1995.
3. MG Vangel, AL Rukhin. Maximum likelihood analysis for a series of similar experiments. To be published.
4. NS Osborne. Thermal expansion of mixtures of ethyl alcohol and water. Bull Bur Stand 9:387–398, 1913.

Chapter 6.5

Intelligent Industrial Robots

Wanek Golnazarian
General Dynamics Armament Systems, Burlington, Vermont

Ernest L. Hall
University of Cincinnati, Cincinnati, Ohio

5.1 INTRODUCTION

The Robot Industries Association (RIA) has defined an *industrial robot* as "a reprogrammable multifunctional manipulator designed to move material, parts, tools or specialized devices, through variable programmed motions for the performance of a variety of tasks." The most common types of manipulators may be modeled as an open kinematic chain of rigid bodies called *links*, interconnected by *joints*. Some have closed kinematic chains such as four-bar mechanisms for some links. The typical industrial robot is mounted on a fixed pedestal base which is connected to other links. The end effector attaches to the free end and enables the robot to manipulate objects and perform required tasks. Hard automation is differentiated because of the single function. Computer numerical control (CNC) machines have a smaller variety of tasks.

A more general definition of a robot is: a general-purpose, reprogrammable machine capable of processing certain humanlike characteristics such as judgment, reasoning, learning, and vision. Although industrial robots have been successfully used in a variety of manufacturing applications, most robots used are deaf, dumb, blind, and stationary [1]. In fact they have been used more like automated machines where the jobs are repetitive, dirty, dangerous, or very difficult. An industrial robot is limited in sensory capabilities (vision and tactile), flexibility, adaptability, learning, and creativity.

Current researchers are attempting to develop *intelligent robots*. Hall and Hall [1] define an intelligent robot as one that responds to changes to its environment through sensors connected to its controller. Much of the research in robotics has been concerned with vision (eyes) and tactile (fingers). *Artificial intelligence* (AI) programs using heuristic methods have somewhat solved the problem of adapting, reasoning, and responding to changes in the robot's environment. For example, one of the most important considerations in using a robot in a workplace is human safety. A robot equipped with sensory devices that detect the presence of an obstacle or a human worker within its work space could automatically shut itself down in order to prevent any harm to itself and/or the human worker.

Kohonen [2] suggests a higher degree of learning is possible with the use of *neural computers*. The intelligent robot is supposed to plan its action in the natural environment, while at the same time performing nonprogrammed tasks. For example, the learning of locomotion in an unknown environment is extremely difficult to achieve by formal logic programming. Typical robot applications in manufacturing assembly tasks may require locating components and placing them in random positions.

5.2 AUTOMATION AND ROBOTICS

In the manufacturing industry, the term *automation* is very common. It was introduced in the 1940s at the Ford Motor Company, where specialized machines helped manufacture high-volume production of mechanical and electrical parts. However, the high cost of tooling for new models limits production flexibility. This type of automation is referred to as *hard* or *fixed automation*. The advantage is high production rate. Hard automation is still being used for the production of light bulbs at General Electric at a rate of two billion light bulbs per year [3].

In the early 1950s numerically controlled (NC) machine tools were introduced. Later, NC machines evolved into computer numerical control (CNC) machines. Since CNCs can be easily reprogrammed through software to accommodate high product variety relative to hard automation, they are referred to as *soft* or *flexible automation* The reprogrammability of flexible automation equipment gives it a key advantage over hard automation. Robots and their development are a natural extension of the concepts of NC and CNC. The robot's reprogrammable feature has enhanced the flexibility of the automation systems and is referred to as an example of soft or flexible automation [3]

Industrial robots were first commercially marketed in 1956 by a firm named Unimation. In 1961, Ford Motor Company was the first to use a Unimate robot to unload a diecasting machine [4]. Since then, the automobile industry has been largely responsible for development of the flexible manufacturing system (FMS) with industrial robots. Introducing robot technology into the factories has improved productivity, quality, and flexibility which could not be realized on the basis of hard or fixed automation structure However, robots in the early 1980s were limited in capabilities and performance due to their drive mechanisms, controller systems, and programming environment. They were not well suited for most manufacturing tasks and were often too expensive. As a result, the increase of robot installation through the mid-1980s turned into a significant slump which lasted into the early 1990s [5]. Due to the steep decline in robot orders, many US manufacturers chose to pull out and ceded the market to foreign competitors. Adept Technology Inc. (San Jose, CA) is the only major U.S. robot manufacturer to survive in the $700 million market, with about $60 million in annual sales [6].

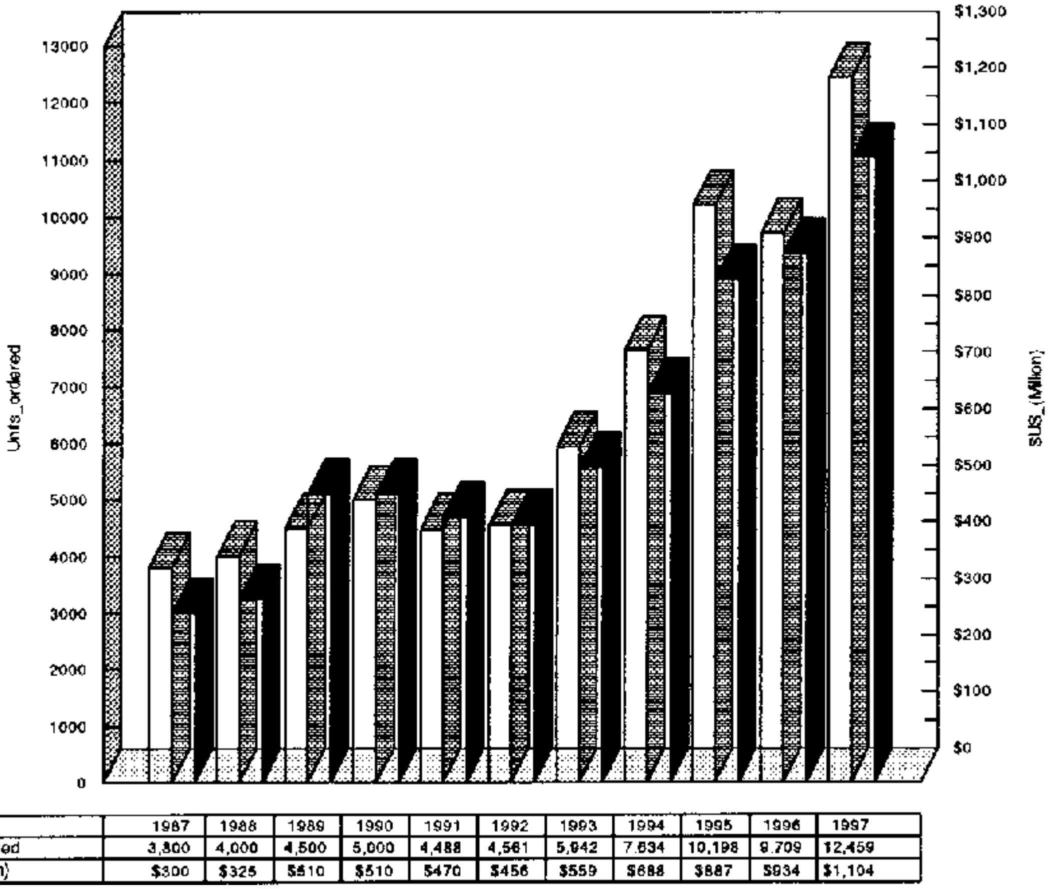
The introduction of robots is often justified on the basis that they perform consistently and productively. Often a few people suffer the loss of employment. Others believe robotic technology creates skilled jobs with greater creativity. However, the question of the social impact of robotics has yet to be adequately addressed [7].

Lower cost, greater reliability, and targeting tasks that are too difficult or dangerous for humans have led to a renewed interest in robotics during the early 1990s. Finally, U.S. manufacturers are realizing the significant impact robots can have in improving productivity, quality, flexibility, and time-to-market. Process repeatability and final product uniformity are more important than labor cost. And unlike dedicated machinery (fixed automation), which is designed to perform a specific task, today's robot can be used for multiple products with case. It has become the critical element in many applications such as welding, sealing, and painting. Other applications (material handling, assembly, and inspection) in nonautomotive industries such as electronics, consumer products, pharmaceutical, and service, are maturing rapidly [8]. According to the Robotic Industries Association (RIA), the robot industry is in a recovery mode, particularly in the United States, as manufacturers invest in robotics to stay competitive. Figure 1 illustrates the renewed strength in robotics since 1987 [5].

Record-breaking shipments from U.S. manufacturers have totaled 12,459 robots valued at $1.1 billion in 1997. This represents a 172% increase in robotic systems and a 136% increase in revenues since 1992. According to new statistics released by the Robotic Industries Association, the world's population of installed robots at the end of 1997 exceeded 500,000. The country that has the largest population of industrial robots is Japan (400,000); it is followed by the United States (80,000), and then the Western European nations combined (120,000).

Industrial robot applications are not limited to automotive industries. A summary of robot applications, along with the share of the robot market in the United States for the year 1995, is displayed in Table 1. Traditional applications of spot and arc welding, and spray painting continue to dominate. The market share for assembly robots has grown over the past decade. The discussion that follows, although not all-inclusive, offers an overview of such applications by type. In addition, excellent reviews of existing robot applications are given by Odery [4] and examples and case studies in the textbook by Asfahal [3].

Welding. This is the process of joining metals by fusing them together, commonly used in the fabrication



Year	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997
▯ Units_ordered	3,800	4,000	4,500	5,000	4,488	4,561	5,942	7.634	10,198	9.709	12,459
▯ $US_(Milion)	$300	$325	$510	$510	$470	$456	$559	$688	$887	$934	$1,104

Figure 1 Industrial robot market.

industry. Spot and arc welding were the top industrial robot applications of the early 1990s. Continuous arc welding is a more difficult process for a robot than spot welding. An arc weld has to be placed along a joint between two pieces of metal. Part misalignment or dimensional variations in parts am the major causes of problems often encountered in robot arc welding. Researchers are investigating the variety of sensors, that can provide feedback information for the purpose of guiding the weld path [10].

Table 1 United States Robot Market [9]

Application	Percent
Welding	53.0
Material handling	24.0
Assembly	10.0
Spray coating	8.5
Inspection	1.0
Other	3.5

Robotic arc welding cells provide the following advantages: higher productivity as measured by greater "arc-on" time, reduced worker fatigue, improved safety, and decreased idle time. Broshco, a division of Jay Plastic (Mansfield, OH), is a producer of robotic-welded assemblies for automobile seating units. It purchased a Motoman ArcWorld 2000, with six axes of motion, in order to improve productivity per person hour, weld quality, and work environment. The installation is part of the automation strategy for the new product line [11].

Material Handling. Applications of this type refer to grasping and movement of work parts from one point to another. Examples include machine loading and unloading, automated palletizing, and automated warehousing. Material-handling applications are typically simple with regard to degrees of freedom required. Specialized gripper design is an active area of research, where quick retooling enables the robot during the production cycle. Automated palletizing operations may require additional degrees of freedom with more sophisticated drive mechanisms, controllers, and expert programming features [12]. In GE's Electrical Distribution and Control plant (Morristown, TN) five CG-120 gantry robots (C&D Robotics) palletize a range of product sizes (cartons range from 8 × 8 in^2. to 12 × 40 in.^{2}). Since 1989, GE's robots have provided versatility over manual and conventional palletizing and eliminated injuries caused by manual lifting [13].

Spray Coating. This is the process of applying paint or a coating agent in thin layers to an object, resulting in a smooth finish. Industrial robots are suitable for such applications, where a human worker is in constant exposure to hazardous fumes and mist which can cause illness and fire. In addition, industrial robots provide a higher level of consistency than the human operator. Continuous-path control is required to emulate the motions of a human worker, with flexible multiple programming features for quick changeovers. Hydraulic drives are recommended to minimize electrical spark hazards. Chrysler Corp. has found an alternative process to fill the seams on its new LH vehicles to eliminate warping and inconsistent filling. In 1992, a four-robot station (Nachi Robotic Systems Inc.) at Chrysler's completely retooled plant (Ontario, Canada) successfully replaced the manual filling of the seams with silicon-bronze wire [14].

Assembly. Many products designed for human assembly cannot be assembled automatically by industrial robots. The integration of product design and assembly design belongs to the concept of *design for manufacturability* [15]. More recent research in design for assembly has been completed at the University of Cincinnati [16]. Design for manufacturability results in the design of factories for robots. Fewer parts, complex molding, and subassemblies which allow a hierarchical approach to assembly has lead to robotic applications. For example, the IBM Proprinter, which was designed for automatic assembly, uses 30 parts with mating capabilities (screwless) to assemble and test in less than 5 min. For part-mating applications, such as inserting a semiconductor chip into a circuit board (peg-in-hole), remote center compliance (RCC) devices have proved to be an excellent solution. More recently, Reichle (Wetzikon, Switzerland), a midsize manufacturer of telecommunications switching equipment, needed a system to automate the labor-intensive assembly of electronic connectors. Using hardware and software from Adept Technology Inc. (San Jose, CA), three AdeptOne robots reduce personpower requirements from 10 to 2. In addition, the system provides speed and a high degree of robotic accuracy [17].

Inspection and Measurement. With a growing interest in product quality, the focus has been on "zero defects." However, the human inspection system has somehow failed to achieve its objectives. Robot applications of vision systems have provided services in part location, completeness and correctness of assembly products, and collision detection during navigation. Current vision systems, typically two-dimensional systems, compare extracted information from objects to previously trained patterns for achieving their goals. Co-ordinate measuring machines (CMM) are probing machining centers used for measuring various part features such as concentricity, perpendicularity, flatness, and size in three-dimensional rectilinear or polar co-ordinate systems. As an integrated part of a flexible manufacturing system, the CMMs have reduced inspection time and cost considerably, when applied to complex part measurement.

Machine vision applications require the ability to control both position and appearance in order to become a productive component of an automated system. This may require a three-dimensional vision capability, which is an active research area [18]. At Loranger Manufacturing (Warren, PA), 100% inspection of the rim of an ignition part is required for completeness. Using back lighting and with the camera mounted in line, each rim is viewed using pixel connectivity. When a break in pixels is detected an automatic reject arm takes the part off the line [19].

Other processing applications for robot use include machining (grinding, deburring, drilling, and wire brushing) and water jet-cutting operations. These operations employ powerful spindles attached to the robot end effector, rotating against a stationary piece. For example, Hydro-Abrasive Machining (Los Angeles, CA) uses two gantry robots with abrasive water-jet machining heads. They cut and machine anything from thin sheet metal to composites several inches thick with tolerances of 0.005 in. for small parts to 0.01 in. for larger parts [19]. Flexible manufacturing systems combined with robotic assembly and inspection, on the one hand, and intelligent robots with improved functionality and adaptability, on the other, will initiate a structural change in the manufacturing industry for improved productivity for years to come.

5.3 ROBOT CHARACTERISTICS

In this section an industrial robot is considered to be an open kinematic chain of rigid bodies called links, interconnected by joints with actuators to drive them. A robot can also be viewed as a set of integrated subsystems [10]:

1. *Manipulator*: the mechanical structure that performs the actual work of the robot, consisting of links and joints with actuators.
2. *Feedback devices*: transducers that sense the position of various linkages and/or joints that transmit this information to the controller.
3. *Controller*: computer used to generate signals for the drive system so as to reduce response error in positioning and applying force during robot assignments.
4. *Power source*: electric, pneumatic, and hydraulic power systems used to provide and regulate the energy needed for the manipulator's actuators.

The manipulator configuration is an important consideration in the selection of a robot. It is based on the kinematic structure of the various joints and links and their relationships with each other. There are six basic motions or degrees of freedom to arbitrarily position and orient an object in a three-dimensional space (three arm and body motions and three wrist movements). The first three links, called the major links, carry the gross manipulation tasks (positioning). Examples are arc welding, spray painting, and waterjet cutting applications. The last three links, the minor links, carry the fine manipulation tasks (force/tactile). Robots with more than six axes of motion are called redundant degree-of-freedom robots. The redundant axes are used for greater flexibility, such as obstacle avoidance in the workplace. Examples are parts assembly, and machining applications. Typical joints are revolute (R) joints, which provide rotational motion about an axis, and prismatic (P) joints, which provide sliding (linear) motion along an axis. Using this notation, a robot with three revolute joints would be abbreviated as RRR, while one with two revolute joints followed by one prismatic joint would be denoted RRP.

There are five major mechanical configurations commonly used for robots: cartesian, cylindrical, spherical, articulated, and selective compliance articulated robot for assembly (SCARA). Workplace coverage, particular reach, and collision avoidance, are important considerations the selection of a robot for an application. Table 2 provides a comparative analysis of the most commonly used robot configurations along with their percent of use. Details for each configuration are documented by Ty and Tien [20]. Figure 2 shows the arm geometries for the most commonly used robot configuration: (a) cartesian (PPP), (b) cylindrical (RPP), (c) articulated (RRR), (d) spherical (RRP),

Table 2 Comparisons of Robot Configurations

Robot application	Configuration	Use (%)[a]	Advantage	Disadvantage
Cartesian assembly and machine loading	PPP	18	Linear motion in 3D; simple kinematics; rigid structure	Poor space utilization; limited reach; low speed
Cylindrical assembly and machine loading	RPP	15	Good reach; simple kinematics	Restricted work space; variable resolution
Spherical automotive manufacturing	RRP	10	Excellent reach; very powerful with hydraulic drive	Complex kinematics; variable resolution
Articulated spray coating	RRR	42	Maximum flexibility; large work envelope; high speed	Complex kinematics; rigid structure; difficult to control
SCARA assembly and insertion	RRP	15	Horizontal compliance; high speed; no gravity effect	Complex kinematics; variable resolution; limited vertical motion

[a] *Source*: VD Hunt. Robotics Sourcebook. New York: Elsevier, 1988.

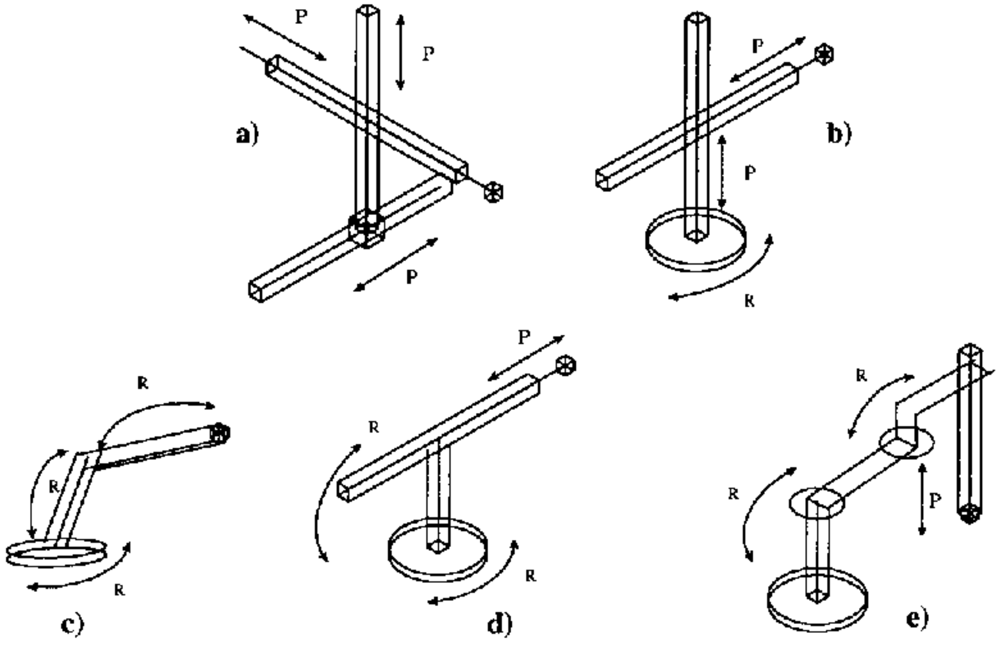


Figure 2 Common robot arm geometries.

and (e) SCARA (RRP). However, there are other configurations used in either research or specialized applications.

The gantry configuration is essentially a cartesian configuration with the robot mounted on an overhead track. A redundant robot configuration may be used when more than six degrees of freedom are needed to reach a particular orientation. All the above configurations are rigid serial links. A parallel robot configuration known as Steward platform, also exists. Lightweight flexible robot arms also exist for faster speed, and low energy consumption. Other classifications am possible based on transmission type. Industrial robots can be direct-driven arms (DDArm) and indirect driven arms. Most industrial robots used today are indirect-drive geared mechanisms. This drive mechanism may suffer from poor dynamical response under heavy mechanical load and gear friction, and backlash.

In DDArms no gears or other mechanical power conversion devices are used. High-torque motors are directly coupled to each joint, eliminating gear friction and increasing stiffness and speed. However, they are more difficult to control because the inertia changes and gravity effects are no longer suppressed by high gear ratios [21]. Experimental DDArms have been built at the MIT AI Laboratory and CMU Robotics Institute. The AdeptOne robot (four-axis SCARA) is the first commercially available direct-drive robot.

5.4 ROBOT CONTROL STRATEGIES

Robot manipulators typically perform a given task repeatedly. Yoshikawa [22] defines the fundamental elements of tasks performed by robots as:

1. *Gross manipulation*: to move the end effector, with or without a load, along a desired path (*position control*).
2. *Fine manipulation*: to exert a desired force on an object when in contact with it (*force control*).

Industrial manipulators require a high level of accuracy, speed, and stability in order to provide the mobility and flexibility needed to perform the range of tasks required in a manufacturing setting. An industrial robot is useful when it is capable of controlling its movement and the forces it applies to its environment. Accuracy, repeatability, and stability are important considerations when designing a suitable robot controller.

One of the major objectives of a robot is to position its tool from one point to another while following a planned trajectory. This is called controlled path motion or the *motion trajectory problem*. Motion trajectory control of the end effector applies to the tasks in the first category, gross manipulation, as defined by Yoshikawa. Robots, in general, use the first three axes for gross manipulation (position control) while the remaining axes orient the tool during the fine manipulation (force or tactile control). The dynamic equations of an industrial robot are a set of highly nonlinear differential equations. For an end effector to move in a particular trajectory at a particular velocity a complex set of torque (force) functions are to be applied by the joint actuators. Instantaneous feedback information on position, velocity, acceleration, and other physical variables can greatly enhance the performance of the robot.

In most systems, conventional single-loop controllers track the tasks, which are defined in terms of a joint space reference trajectory. In practice, the tracking error is compensated through an iterative process which adjusts the reference input so that the actual response (Y) of the manipulator is close to the desired trajectory (Y_d). When following a planned trajectory, control at time t will be more accurate if the controller can account for the end effector's position at an earlier time. Figure 3 represents the basic block diagram of a robot trajectory system interacting with its environment.

With increasing demands for faster, more accurate, and more reliable robots, the field of robotics has faced the challenges of reducing the required online computational power, calibration time, and engineering cost when developing new robot controllers. If the robot is to be controlled in real time the algorithms used must be efficient and robust. Otherwise, we will have to compromise the robot control strategies, such as reducing the frequency content of the velocity profile at which the manipulator moves.

The robot arm position control is a complex kinematic and dynamic problem and has received researchers' attention for quite some time. During the last several years, most research on robot control has resulted in effective but computationally expensive algorithms. A number of approaches have been proposed to develop controllers that are robust and adaptive to the nonlinearities and structural uncertainties. However, they are also computationally very difficult and expensive algorithms to solve. As of this day, most robot controllers use joint controllers that are based on traditional linear controllers and are ineffective in dealing with the nonlinear terms, such as friction and backlash.

One popular robot control scheme is called "computed-torque control" or "inverse-dynamics control." Most robot control schemes found in robust, adaptive, or learning control strategies can be considered as special cases of the computed-torque control. The computed-torque-like control technique involves the decomposition of the control design problem into two parts [23]:

1. Primary controller, a feedforward (inner-loop) design to track the desired trajectory under ideal conditions
2. Secondary controller, a feedback (outer-loop) design to compensate for undesirable deviations (disturbances) of the motion from the desired trajectory based on a linearized model.

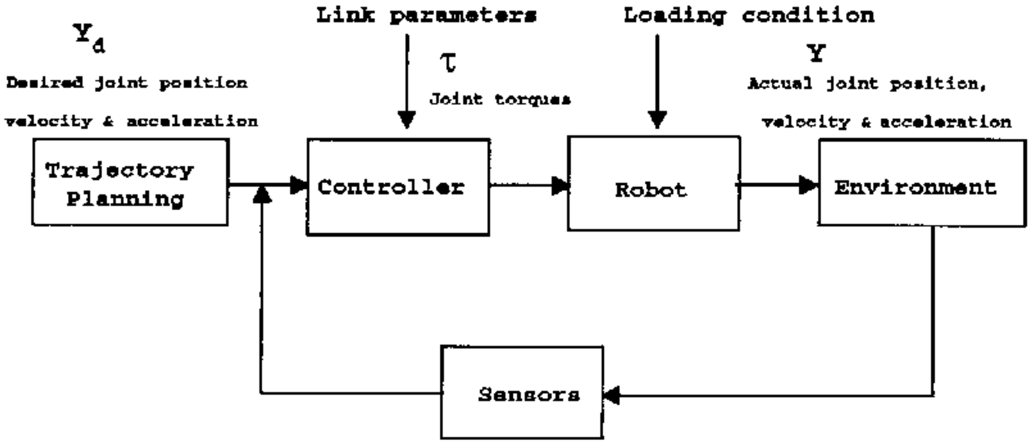


Figure 3 Basic control block for a robot trajectory system.

The primary controller compensates for the nonlinear dynamic effects, and attempts to cancel the nonlinear terms in the dynamic model. Since the parameters in the dynamic model of the robot are not usually exact, undesired motion errors are expected. These errors can be corrected by the secondary controller. Figure 4 represents the decomposition of the robot controller showing the primary and secondary controllers.

It is well known that humans perform control functions much better than the machinelike robots. In order to control voluntary movements, the central nervous system must determine the desired trajectory in the visual co-ordinates, transform its co-ordinate to the body co-ordinate, and finally generate the motor commands [24]. The human information processing device (brain) has been the motivation for many researchers in the design of intelligent computers often referred to as neural computers.

Psaltis et al. [25] describe the neural computer as a large interconnected mass of simple processing elements (artificial neurons). The functionality of this mass, called the *artificial neural network* (ANN), is determined by modifying the strengths of the connections during the learning phase. This basic generalization of the morphological and computational feature of the human brain has been the abstract model used in the design of the neural computers.

Researchers interested in neural computers have been successful in computationally intensive areas such as pattern recognition and image interpretation problems. These problems are generally static mapping of input vectors into corresponding output classes using a feedforward neural network. The feedforward neural network is specialized for the static mapping problems, whereas in the robot control problem, nonlinear dynamic properties need to be dealt with and a different type of neural network structure must be used. Recurrent neural networks have the dynamic properties, such as feedback architecture, needed for the appropriate design of such robot controllers.

5.5 ARTIFICIAL NEURAL NETWORKS

Artificial neural networks are highly parallel, adaptive, and fault-tolerant dynamic systems, modeled like their biological counterparts. The phrases "neural networks" or "neural nets" are also used interchangeably in the literature, which refer to neurophysiology, the study of how the brain and its nervous system work. Artificial neural networks are specified by the following definitions [26]:

1. *Topology*: this describes the networked architecture of a set of neurons. The set of neurons are organized into layers which are then classified as either *feedforward networks* or *recurrent networks*. In feedforward layers, each output in a layer is connected to each input in the next layer. In a recurrent ANN, each neuron can receive as its input a weighted output from other layers in the network, possibly including itself. Figure 5 illustrates these simple representations of the ANN topologies.
2. *Neuron*: a computational element that defines the characteristics of input/output relationships. A simple neuron is shown in Fig. 6, which sums *N* weighted inputs (called activation) and passes the result through a nonlinear transfer function to determine the neuron output. Some nonlinear functions that are often used to mimic biological neurons are: unit step function and linear transfer-function. A very common formula for determining a neuron's output is through the use of sigmoidal (squashing) functions:

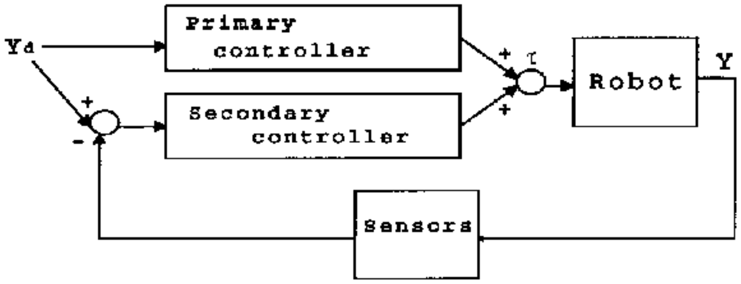


Figure 4 Controller decomposition in primary and secondary controllers.

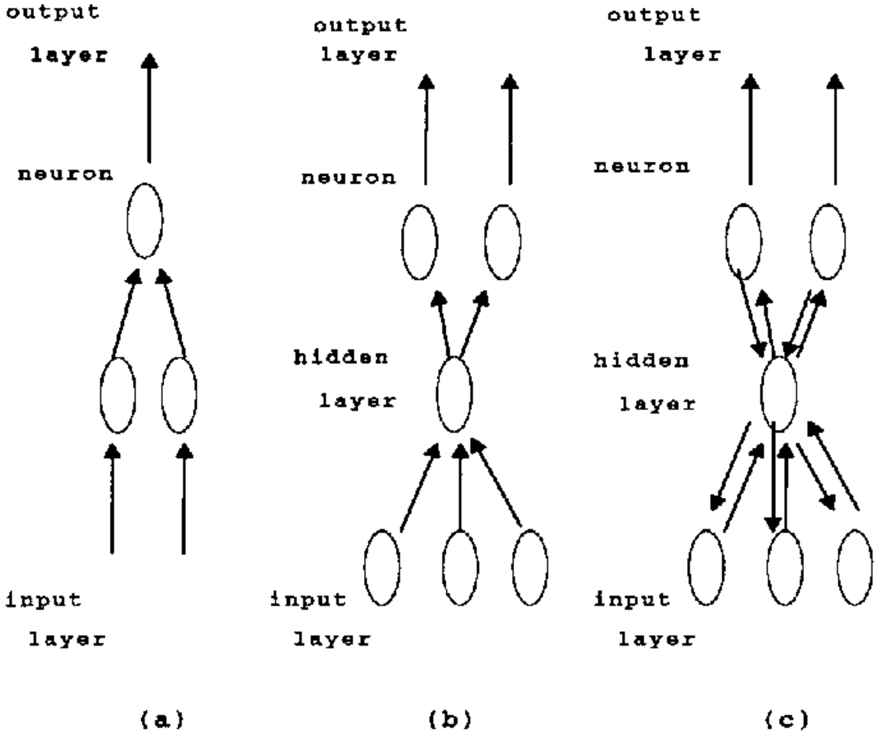


Figure 5 Artificial neural network topologies: (a) single-layer feedforward; (b) multilayer feedforward; (c) multilayer recurrent.

$$g(x) = (1 + e^{-kx}) \qquad \text{range of } (0, 1) \qquad (1a)$$

$$g(x) = \tanh(kx) \qquad \text{range of } (-1, 1) \qquad (1b)$$

For various values of the slope parameter, k, these functions are continuous and have derivatives at all points.

3. *Learning rules*: given a set of input/output patterns, ANNs can learn to classify these patterns by optimizing the weights connecting the nodes (neurons) of the networks. The learning algorithms for weight adaptation can be described as either supervised or unsupervised learning. In supervised learning, the desired output of the neuron is known, perhaps by providing training samples. During supervised training, the network compares its actual response, which is the result of the transfer function described above, with the training example and then adjusts its weight in order to minimize the error between the desired and its actual output. In unsupervised training, for which there are no teaching examples, built-in rules are used for self-modification, in order to adapt the synaptic weights in response to the inputs to extract features from the neuron Kohonen's self-organizing map is an example of unsupervised learning [27].

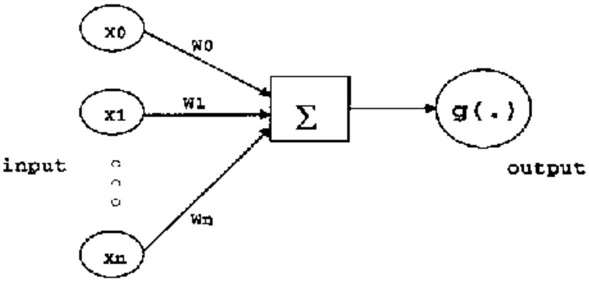


Figure 6 McCulloch and Pitts neuron.

One of the first models of an artificial neuron was introduced in 1943 by McCulloch and Pitts and is shown in Fig. 6. The model, known as the McCulloch–Pitts neuron, computes a weighted sum of inputs (x_i) and outputs (unit step function) a binary value (Y) according to whether this sum is above or below a certain threshold (θ) [28].

$$Y = g\left(\sum w_i x_i - \theta\right) \qquad (2)$$

where $g(p) = 1$ if $p \geq 0$; 0 otherwise.

McCulloch and Pitts proved that a synchronous network of neurons (M-P network), described above, is capable of performing simple logical tasks (computations) that are expected of a digital computer. In 1958, Rosenblatt introduced the "perceptron," in which he showed how an M-P network with adjustable weights can be trained to classify sets of patterns. His work was based on Hebb's model of adaptive learning rules in the human brain [29], which stated that the neuron's interconnecting weights change continuously as it learns [30].

In 1960, Bernard Widrow introduced ADALINE (ADAptive LINear Element), a single-layer perceptron and later extended it to what is known as MADALINE, multilayer ADALINE [31]. In MADALINE, Widrow introduced the steepest descend method to stimulate learning in the network. His variation of learning is referred to as the Widrow–Hoff rule or delta rule.

In 1969, Minsky and Papert [32] reported on the theoretical limitations of the single layer M-P network, by showing the inability of the network to classify the exclusive-or (XOR) logical problem. They left the impression that neural network research is a farce and went on to establish the "artificial intelligence" laboratory at MIT. Hence, the research activity related to ANNs went to sleep until the early 1980s when the work by Hopfield, an established physicist, on neural networks rekindled the enthusiasm for this field. Hopfield's autoassociative neural network (a form of recurrent neural network) solved the classic hard optimization problem (traveling salesman) [33].

Other contributors to the field, Steven Grossberg and Teuvo Kohonon, continued their research during the 1970s and early 1980s (referred to as the "quiet years" in the literature). During the "quiet years," Steven Grossberg [34,35] worked on the mathematical development necessary to overcome one of the limitations reported by Minsky and Papert [32]. Teuvo Kohonon [36] developed the unsupervised training method, the self-organizing map. Later, Bart Kosko [37] developed bidirectional associative memory (BAM) based on the works of Hopfield and Grossberg. Robert Hecht-Nielson [38] pioneered the work on neurocomputing.

It was not until 1986 that the two-volume book, edited by Rumelhart and McClleland, titled *Parallel Distributed Processing* (PDP), exploded the field of artificial neural networks [38]. In this book (PDP), a new training algorithm called *the backpropagation* method (BP), using the gradient search technique was used to train a multilayer perceptron to learn the XOR mapping problem described by Minsky and Papert [39]. Since then, ANNs have been studied for both design procedures and training rules (supervised and unsupervised), and are current research topics. An excellent collection of theoretical and conceptual papers on neural networks can be found in books edited by Vemuri [30], and Lau [40]. Interested readers can also refer to a survey of neural networks book by Chapnick [41] categorized by: theory, hardware and software, and how-to books.

The multilayer feedforward networks, using the BP method, represent a versatile nonlinear map of a set of input vectors to a set of desired output vectors on the spatial context (space). During the learning process, an input vector is presented to the network and propagates forward from input layers to output layers to determine the output signal. The output signal vector is then compared with the desired output vector, resulting in an error signal. This error signal is backpropagated through the network in order to adjust the network's connecting strengths (weights). Learning stops when the error vector has reached an acceptable minimum [26]. An example of feedforward network consisting of three layers is shown in Fig. 7.

Many studies have been undertaken in order to apply both the flexibility and the learning ability of backpropagation to robot control on an experimental scale [42–44]. In a recent study, an ANN utilizing an adaptive step-size algorithm, based on a random-search technique, improved the convergence speed of the BP method for solving the inverse kinematic problem for a two-link robot [45]. The robot control problem is a dynamic problem, where the BP method only provides a static mapping of the input vectors into output classes. Therefore, its benefits are limited. In addition, like any other numerical method, this novel learning method has limitations (slow convergence rate, local minimum). Attempts to improve the learning rate of BP have resulted in many novel approaches [46,47]. It is necessary to note that the most important behavior of the feedforward networks using the BP method is its classification ability or the generalization to fresh data rather than temporal utilization of past experiences.

A *recurrent network* is a multilayer network in which the activity of the neurons flows both from input layer to output layer (feedforward) and also from the output layer back to the input layer (feedback) in the course of learning [38]. In a recurrent network each activity of the training set (input

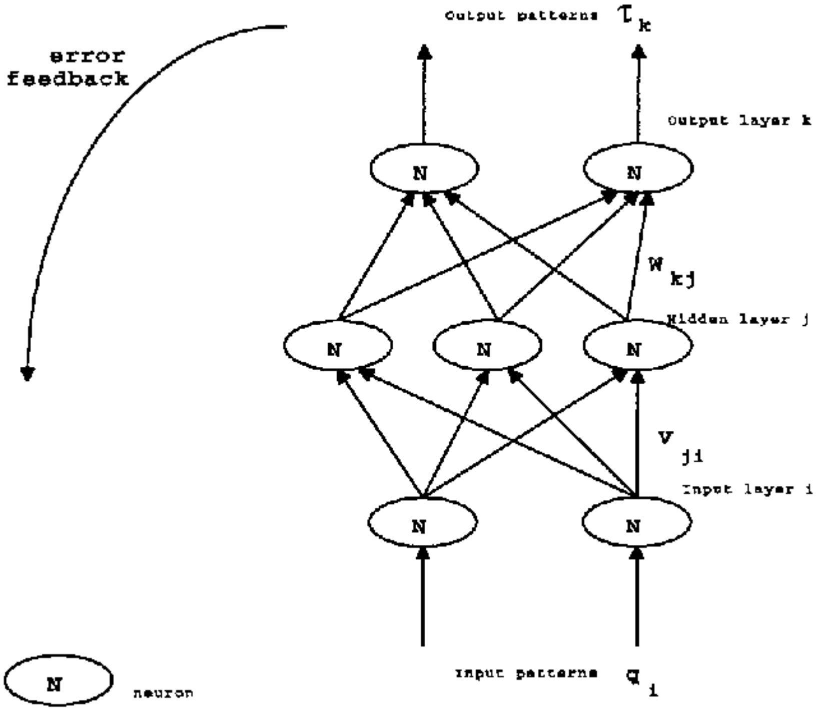


Figure 7 Artificial neural network—feedforward.

pattern) passes through the network more than once before it generates an output pattern, where in standard BP only error flows backward, not the activity. This network architecture can base its response to problems on both spatial (space) and temporal (time) contexts [48,49]. Therefore, it has a potential in modeling time-dependent processes such as in robotic applications. Figure 8 represents the recurrent ANN architecture.

It is clearly evident that a recurrent network will require a substantial memory in simulation (more connections) than a standard BP. Recurrent networks computing is a complex method, with a great deal of record keeping of errors and activities at each time phase. However, preliminary results indicate that they have the ability to learn extremely complex temporal patterns where data is unquantified with very little preprocessing, i.e., stock market prediction, and Fourier-transform relationships [50]. In feedforward networks where the training process is memoryless, each input is independent of the previous input. It is advantageous, especially in repetitive dynamic systems, to focus on the properties of the recurrent networks to design better robot controllers.

5.6 ROBOT ARM KINEMATICS

This section provides the mathematical formulations for the kinematic and dynamic analysis of robot manipulators. These formulations are then considered in design of the control algorithms for a four-axis SCARA industrial robot (AdeptOne). The treatment of robot manipulator kinematics and dynamics presented here is patterned primarily after discussions found in Lee [51], Spong and Vidyasagar [52], and Schilling [53]. Additional comprehensive sources of robot manipulator kinematics and dynamics can be found in investigations by Paul [54], Wolovich [55], Craig [56], Asada and Slotine [57], and Fu et al. [58].

The purpose of a robot manipulator is to position and interface its end effector with the working object. For example, a robot has to pick up a part from a certain location, put it down in another location, and so on. Robot arm kinematics deals with the geometry of robot arm motion as a function of time (position, velocity, and acceleration) without regard to the forces and moments that cause it. Specifically, one studies the functional relationship between the joint displacements and the position and orientation of the end effector of a robot as shown in Fig. 9. The problem of finding the

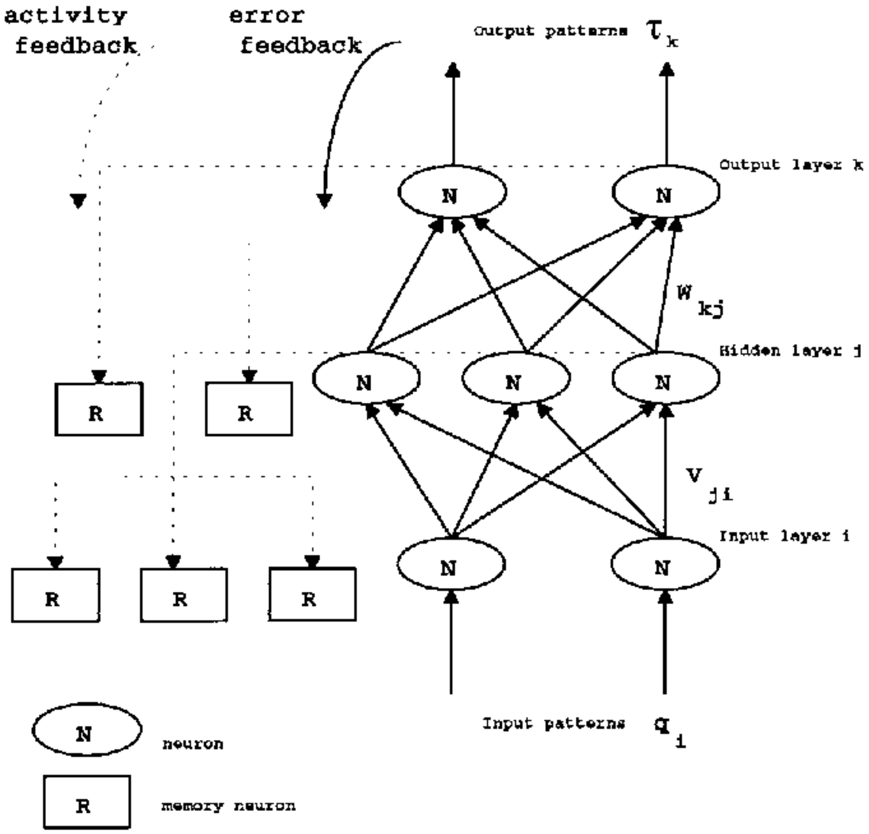


Figure 8 Artificial neural network—recurrent.

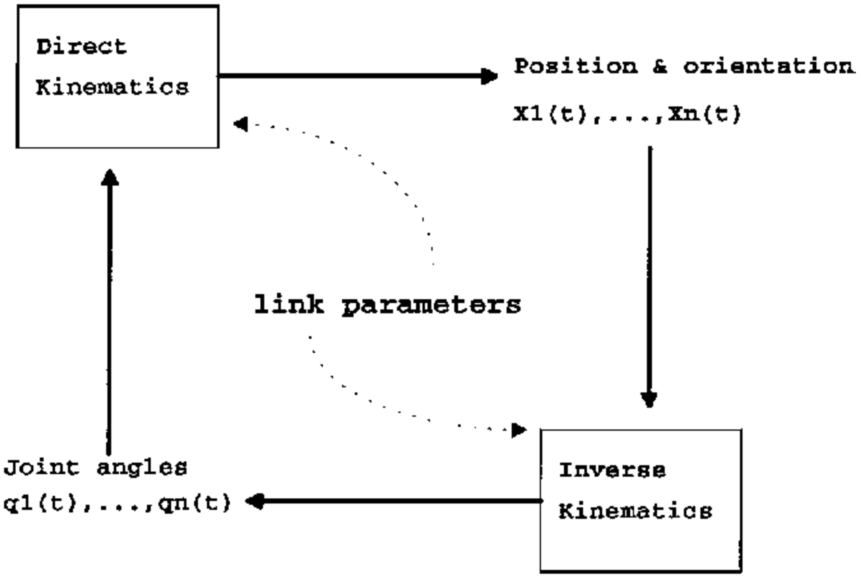


Figure 9 The direct and inverse kinematic problems.

end-effector position and orientation for a given set of joint displacements is referred to as the *direct kinematic problem*. Thus, for a given joint co-ordinate vector $\mathbf{q}$ and the global co-ordinate space $\mathbf{x}$, it is to solve

$$\mathbf{x} = \mathbf{f}(\mathbf{q}) \tag{3}$$

where $\mathbf{f}$ is a nonlinear, continuous, and differentiable function. This equation has a unique solution. On the other hand, given the end-effector position and orientation, the *inverse kinematic problem* calculates the corresponding joint variables to drive the joint servo controllers by solving

$$\mathbf{q} = \mathbf{f}^{-1}(\mathbf{x}) \tag{4}$$

The solution to this equation, also called the arm solution, is not unique. Since trajectory tasks are usually stated in terms of the reference co-ordinate frame, the inverse kinematics problem is used more frequently.

5.6.1 Homogeneous Transformation Matrix

Before further analyzing the robot kinematic problem, a brief review of matrix transformations is needed. Figure 10 illustrates a single vector defined in the $\{i\}$ co-ordinate frame $\mathbf{P}^i = (x', y', z')$. The task is to transform the vector defined with the $\{i\}$ co-ordinate frame to a vector with the $\{i-1\}$ co-ordinate frame, $\mathbf{P}^{i-1} = (x, y, z)$. Simply, this transformation is broken up into a rotational part and a translational part:

$$\mathbf{P}^{i-1} = \mathbf{R}_i\mathbf{P}^i + \mathbf{d}_i \tag{5}$$

Since rotation is a linear transformation, the rotation between the two co-ordinate frames is given by

$$\mathbf{P}^{i-1} = \mathbf{R}_i\mathbf{P}^i \tag{6}$$

Here, $\mathbf{R}_i$ is 3×3 matrix operation about the x, y, and z axes. These matrices are

$$\mathbf{R}_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix} \tag{7a}$$

$$\mathbf{R}_y(\theta) = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix} \tag{7b}$$

$$\mathbf{R}_z(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{7c}$$

The following general statements can be made:

1. A co-ordinate transformation $\mathbf{R}$ represents a rotation of a coordinate frame to a new position.
2. The columns of $\mathbf{R}$ give the direction cosines of the new frame axes expressed in the old frame.
3. It can be extended to a product of more than two transformation matrices.

To complete the transformation in Fig. 10, translation between frames $\{i\}$ and $\{i-1\}$ still needs to take place.

$$\mathbf{d}_i = \begin{bmatrix} dx \\ dy \\ dz \end{bmatrix} \tag{8}$$

However, translation is a nonlinear transformation, hence the matrix representation of eq. (5) can only be in 3×4 form:

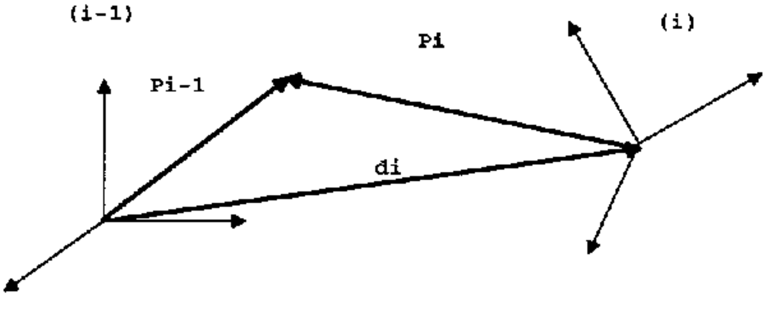


Figure 10 Transformation between two co-ordinate frames.

$$\mathbf{P}^{i-1} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \mathbf{R}_i \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} + \begin{bmatrix} dx \\ dy \\ dz \end{bmatrix} \tag{9}$$

$$\mathbf{P}^{i-1} = [R_i \quad d_i] \begin{bmatrix} \mathbf{P}^i \\ 1 \end{bmatrix}$$

Equation (9), where $[\mathbf{R}_i \; \mathbf{d}_i]$ is a (3×4) matrix and $[\mathbf{P}_i \; 1]^T$ is a (4×1) vector, can only be used to transform the components of $\mathbf{p}$ from frame $\{i\}$ to $\{i-1\}$. Due to singularity of the matrix above, the inverse of the transformation cannot be achieved. To incorporate the inverse transformation in Eq. (9), the concept of *homogeneous co-ordinate representation** replaces the (3×4) transformation matrix with a (4×4) transformation matrix by simply appending a final (1×4) row, defined as [0 0 0 1], to $[\mathbf{R}_i \; \mathbf{d}_i]$. Correspondingly, the $\mathbf{P}_{i-1}$ vector will be replaced by (4×1) vector of $\mathbf{P}_i = [\mathbf{P}_{i-1} \; 1]^T$.

$$\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \begin{bmatrix} & & & dx \\ & R_i & & dy \\ & & & dz \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} \tag{10}$$

It can be seen that the transformation equation (5) is equivalent to the matrix equation (10). The (4×4) transformation matrix, denoted $\mathbf{H}_i$, contains all the information about the final frame, expressed in terms of the original frame:

$$\mathbf{H}_i = \begin{bmatrix} R & d_i \\ 0 & 1 \end{bmatrix} \tag{11}$$

Using the fact that $\mathbf{R}_i$ is orthogonal it is easy to show that the inverse transformation $\mathbf{H}_i^{-1}$ is given by

$$\mathbf{H}_i^{-1} = \begin{bmatrix} R_i^T & -R_i^T d_i \\ 0 & 1 \end{bmatrix} \tag{12}$$

The matrix $\mathbf{H}_i$ has homogenized the representation of translation and rotations of a coordinate frame. Therefore, matrix $\mathbf{H}_i$ is called the *homogeneous transformation matrix*. The upper left (3×3) submatrix represents the rotation matrix; the upper right (3×1) submatrix represents the position vector of the origin of the rotated co-ordinate system with respect to the reference system; the lower left (1×3) submatrix represents perspective transformation for visual sensing with a camera; and the fourth diagonal element is the global scaling factor. In the robot manipulator, the perspective transformation is always a zero vector and the scale factor is 1. The frame transformation is now given by

$$\mathbf{P}^{i-1} = \mathbf{H}_i \mathbf{P}^i \tag{13}$$

This transformation, represented by the matrix $\mathbf{H}_i$, is obtained from simpler transformations representing the three basic translations along (three entries of $\mathbf{d}_i$), and three rotations (three independent entries of $\mathbf{R}_i$) about the frames axes of x, y, and z. They form the six degrees of freedom associated with the configuration of $\mathbf{P}$. These fundamental transforms, expressed in a 4×4 matrix notation, are shown as

$$\mathrm{Trans}(dx, dy, dz) = \begin{bmatrix} 1 & 0 & 0 & dx \\ 0 & 1 & 0 & dy \\ 0 & 0 & 1 & dz \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{14a}$$

$$\mathrm{Rot}(x, \theta) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta & 0 \\ 0 & \sin\theta & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{14b}$$

$$\mathrm{Rot}(y, \theta) = \begin{bmatrix} \cos\theta & 0 & \sin\theta & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\theta & 0 & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{14c}$$

$$\mathrm{Rot}(z, \theta) = \begin{bmatrix} \cos\theta & -\sin\theta & 0 & 0 \\ \sin\theta & \cos\theta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{14d}$$

The homogeneous transformation matrix is used frequently in manipulator arm kinematics.

5.6.2 The Denavit–Hartenberg Representation

The most commonly accepted method for specifying frame position and finding the desired transformation matrices is attributed to the Denavit–Hartenberg (D-H) representation [59]. In this method, an orthonormal Cartesian co-ordinate system is established on the basis of three rules [55]:

1. The z_{i-1}-axis lies along the axis motion the ith joint.
2. The x_i-axis is normal to the z_{i-1} axis, pointing away from it.

* The representation of an n-component position vector by an $(n+1)$-component vector is called homogeneous co-ordinate representation.

3. The y_i-axis complete the right-hand-rule co-ordinate system.

This is illustrated in Fig. 11. Note that joint i joins link $i-1$ with link i. Frame i, which is the body frame of link i, has its z-axis located at joint $i+1$. If the joint is revolute, then the rotation is about the z-axis. If the joint is prismatic, the joint translation is along the z-axis.

The D-H representation depends on four geometrical parameters associated with each link to completely describe the position of successive link coordinates:

$a_i =$ the shortest distance between z_i and z_{i-1} along the x_i.
$\alpha_i =$ the twist angle between z_i and z_{i-1} about the x_i.
$d_i =$ the shortest distance between x_i and x_{i-1} along the z_{i-1}.
$\theta_i =$ the angle between x_{i-1} and x_i about the z_{i-1}.

For a revolute joint, θ_i is the variable representing the joint displacement where the adjacent links rotate with respect to each other along the joint axis. In prismatic joints in which the adjacent links translate linearly to each other along the joint axis, d_i is the joint displacement variable, while θ_i is constant. In both cases, the parameters a_i and α_i are constant, determined by the geometry of the link.

In general we denote the joint displacement by $\mathbf{q}_i$, which is defined as

$\mathbf{q}_i = \theta_i$ for a revolute joint

$\mathbf{q}_i = d_i$ for a prismatic joint

Then, a (4×4) homogeneous transformation matrix can easily relate the ith co-ordinate frame to the $(i-1)$th co-ordinate frame by performing the following successive transformations:

1. Rotate about z_{i-1}-axis an angle of θ_i, $\mathrm{Rot}(z_{i-1}, \theta_i)$
2. Translate along the z_{i-1}-axis a distance of d_i, $\mathrm{Trans}(0, 0, d_i)$
3. Translate along the x_i-axis a distance of a_i, $\mathrm{Trans}(a_i, 0, d_i)$
4. Rotate about the x_i-axis an angle of α_i, $\mathrm{Rot}(x_i, \alpha_i)$

The operations above result in four basic homogeneous matrices. The product of these matrices yields a composite homogeneous transformation matrix ${}^i\mathbf{A}_{i-1}$. The ${}^i\mathbf{A}_{i-1}$ matrix is known as the D-H transformation matrix for adjacent co-ordinate frames, $\{i\}$ and $\{i-1\}$. Thus,

$$
\begin{aligned}
{}^i\mathbf{A}_{i-1} &= \mathrm{Trans}(0, 0, d_i)\ \mathrm{Rot}(z_{i-1}, \theta_i) \\
&\quad \mathrm{Trans}(a_i, 0, 0)\ \mathrm{Rot}(x_i, \alpha_i)
\end{aligned}
$$

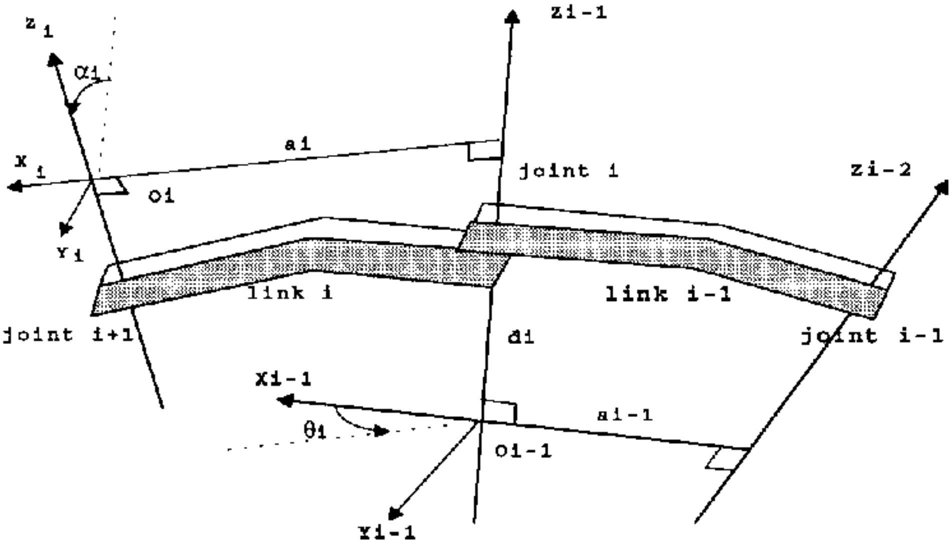


Figure 11 Denavit–Hartenberg frame assignment.

$$
{}^{i}\mathbf{A}_{i-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & d_i \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} \cos\theta_i & -\sin\theta_i & 0 & 0 \\ \sin\theta_i & \cos\theta_i & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}
$$

$$
\times \begin{bmatrix} 1 & 0 & 0 & a_i \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}
$$

$$
\times \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -\cos\alpha_i & \sin\alpha_i & 0 \\ 0 & \sin\alpha_i & \cos\alpha_i & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}
$$

$$
{}^{i}\mathbf{A}_{i-1} = \begin{bmatrix} \cos\theta_i & -\cos\alpha_i\sin\theta_i & \sin\alpha_i\sin\theta_i & \alpha_i\cos\theta_i \\ \sin\theta_i & \cos\alpha_i\cos\theta_i & -\sin\alpha_i\cos\theta_i & \alpha_i\sin\theta_i \\ 0 & \sin\alpha_i & \cos\alpha_i & d_i \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{15}
$$

Using the ${}^{i}\mathbf{A}_{i-1}$ matrix in Fig. 10, vector $\mathbf{P}_i$ expressed in homogeneous co-ordinates with respect to co-ordinate system $\{i\}$, relates to vector $\mathbf{P}_{i-1}$ in co-ordinate $\{i-1\}$ by

$$
\mathbf{P}_{i-1} = {}^{i}\mathbf{A}_{i-1}\mathbf{P}_i \tag{16}
$$

where $\mathbf{P}_{i-1} = (x, y, z, 1)^T$ and $\mathbf{P}_i = (x', y', z', 1)^T$.

5.6.3 Manipulator Arm Kinematic Equations

The transformation matrix of Eq. (16) relates points defined in frame $\{i\}$ to frame $\{i+1\}$. For a robot manipulator with six links, the position of the end effector (last link) with respect to the base is determined by successively multiplying together the single (D-H) transformation matrix that relates frame $\{6\}$ to frame $\{0\}$:

$$
\begin{aligned}
{}^{6}\mathbf{T}_5 &= {}^{6}\mathbf{A}_5 \\
{}^{6}\mathbf{T}_4 &= {}^{5}\mathbf{A}_4{}^{6}\mathbf{T}_5 = {}^{5}\mathbf{A}_4{}^{6}\mathbf{A}_5 \\
{}^{6}\mathbf{T}_3 &= {}^{4}\mathbf{A}_3{}^{6}\mathbf{T}_4 = {}^{4}\mathbf{A}_3{}^{5}\mathbf{A}_4{}^{6}\mathbf{A}_5 \\
{}^{6}\mathbf{T}_2 &= {}^{3}\mathbf{A}_2{}^{6}\mathbf{T}_3 = {}^{3}\mathbf{A}_2{}^{4}\mathbf{A}_3{}^{5}\mathbf{A}_4{}^{6}\mathbf{A}_5 \\
{}^{6}\mathbf{T}_1 &= {}^{2}\mathbf{A}_1{}^{6}\mathbf{T}_2 = {}^{2}\mathbf{A}_1{}^{3}\mathbf{A}_2{}^{4}\mathbf{A}_3{}^{5}\mathbf{A}_4{}^{6}\mathbf{A}_5 \\
{}^{6}\mathbf{T}_0 &= {}^{1}\mathbf{A}_0{}^{6}\mathbf{T}_1 = {}^{1}\mathbf{A}_0{}^{2}\mathbf{A}_1{}^{3}\mathbf{A}_2{}^{4}\mathbf{A}_3{}^{5}\mathbf{A}_4{}^{6}\mathbf{A}_5
\end{aligned} \tag{17}
$$

Generalized for n degrees of freedom, the base frame $\{0\}$ is assumed to be fixed. This is taken as the inertial frame with respect to which a task is specified. The body frame $\{n\}$ is the free moving end effector. The columns of the overall homogeneous transformation matrix, ${}^{n}\mathbf{T}_0$, corresponds to the position and orientation of the end effector $\mathbf{x}_n$, expressed in the base frame This transformation matrix, ${}^{n}\mathbf{T}_0$, will be a function of all n joint variables $(\mathbf{q}_n)$, with the remaining parameters constant:

$$
{}^{n}\mathbf{T}_0(x_n) = {}^{1}\mathbf{A}_0(q_1){}^{2}\mathbf{A}_1(q_2){}^{3}\mathbf{A}_2(q_3)\ldots{}^{n}\mathbf{A}_{n-1}(q_n) \tag{18}
$$

The final transformation matrix ${}^{n}\mathbf{T}_0$, also called the arm matrix, defines the final configuration of any end effector with respect to the inertia frame $\{0\}$, depicted in Fig. 12. The tool origin represents any appropriate point associated with the tool frame (or the transporting object). The origin (O_t) frame can be taken either at the wrist or at the tool tip, or placed symmetrically between the fingers of the end effector (gripper). The ${}^{n}\mathbf{T}_0$ matrix may be written as

$$
\begin{aligned}
{}^{n}\mathbf{T}_0 &= \begin{bmatrix} R_i & d_i \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} x_n & y_n & z_n & d_i \\ 0 & 0 & 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{n} & \mathbf{s} & \mathbf{a} & \mathbf{d} \\ 0 & 0 & 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} n_x & s_x & a_x & d_x \\ n_y & s_y & a_y & d_y \\ n_z & s_z & a_z & d_z \\ 0 & 0 & 0 & 1 \end{bmatrix}
\end{aligned} \tag{19}
$$

where three mutually perpendicular unit vectors, as shown in Fig. 12, represent the tool frame in a cartesian co-ordinate system. In the above equation:

$\mathbf{n}$ = normal unit vector, normal to the fingers of the robot arm following the right-hand rule.

$\mathbf{s}$ = unit sliding vector, pointing to the direction of the sideways motion of the fingers (open and close).

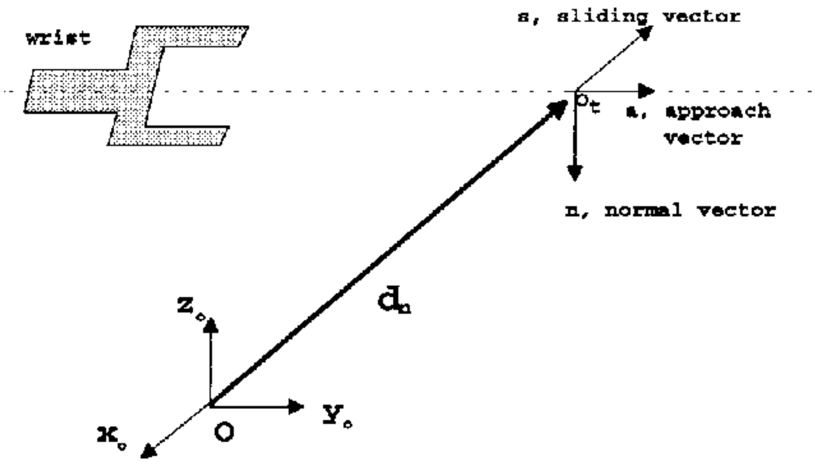


Figure 12 Tool co-ordinate system (frame).

a = unit approach vector, normal to the tool mounting plate of the arm.

Equation (20) represents the direct (forward) kinematic problem for a robot manipulator: given the joint displacements ($\boldsymbol{\theta}$) and the link parameters, find the position and the orientation of the end effector (**X**) in the base frame:

$$\mathbf{X} = \mathbf{f}(\boldsymbol{\theta}) \tag{20}$$

where **f** is a nonlinear, continuous, and differentiable function. This equation has a unique solution. This is best illustrated by an example.

Consider the four-axis horizontal-jointed SCARA robot, AdeptOne, in Fig. 13. This manipulator is unique because it is the first commercial robot to implement a *direct-drive** system for actuation. The robot consists of an RRP arm and a one degree-of-freedom wrist, whose motion is a roll about the fourth vertical axis. The (D-H) link kinematic parameters are given in Table 3 [53].

All the joint axes are parallel. The joint variable (vector-form) is $\mathbf{q} = [\theta_1, \theta_2, d_3, \theta_4]^T$. The first two joint variables, θ_1 and θ_2, are revolute variables which establish the horizontal component of the tool position. The third joint variable $\mathbf{d}_3$, a prismatic joint, determines the vertical component of tool origin. Finally, the last joint variable θ_4, which is of revolute kind, controls the tool orientation.

Using the values from Table 3 in Eq. (15), the ${}^i\mathbf{A}_{i-1}$ matrices are as follows:

$${}^1\mathbf{A}_0 = \begin{bmatrix} \cos\theta_1 & \sin\theta_1 & 0 & a_1\cos\theta_1 \\ \sin\theta_1 & -\cos\theta_1 & 0 & a_1\sin\theta_1 \\ 0 & 0 & -1 & d_1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{21a}$$

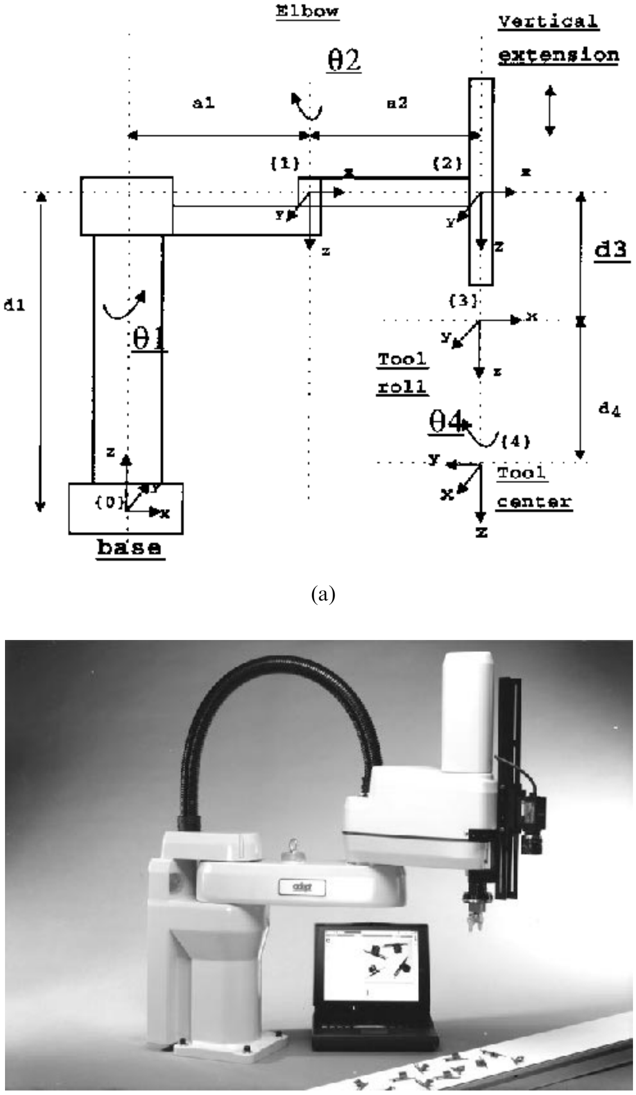
$${}^2\mathbf{A}_1 = \begin{bmatrix} \cos\theta_2 & -\sin\theta_2 & 0 & a_2\cos\theta_2 \\ \sin\theta_2 & \cos\theta_2 & 0 & a_2\sin\theta_2 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{21b}$$

$${}^3\mathbf{A}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \theta_3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{21c}$$

Table 3 Kinematic Parameters for a Four-Axis SCARA Robot

Axis	θ	d	a	α
1	q_1	$d_1 = 877$ mm	$a_1 = 425$ mm	π
2	q_2	0	$a_2 = 375$ mm	0
3	0	q_3	0	0
4	q_4	$d_4 = 200$ mm	0	0

* Direct drive is an electrical drive in which no gear reducer is used. Therefore the rotor of the electric motor is directly coupled to the load, hence the mechanical gears are not needed. This eliminates gear friction and backlash and allows for clean, precise, high-speed operation [60].



(a)

(b)

Figure 13 (a) A four-axis SCARA robot (AdeptOne). (b) The AdeptOne industrial robot.

$$
{}^{4}\mathbf{A}_3 = \begin{bmatrix} \cos\theta_4 & -\sin\theta_4 & 0 & 0 \\ \sin\theta_4 & \cos\theta_4 & 0 & 0_1 \\ 0 & 0 & 1 & d_4 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{21d}
$$

The forward kinematic solution, using Eq. (18) is therefore given by

$$
\begin{aligned}
{}^{\text{tool}}\mathbf{T}_{\text{base}(x_4)} &= {}^{4}\mathbf{T}_0(x_4) \\
&= {}^{1}\mathbf{A}_0(\theta_1){}^{2}\mathbf{A}_1(\theta_2){}^{3}\mathbf{A}_2(d_3){}^{4}\mathbf{A}_3(\theta_4) \qquad (22)
\end{aligned}
$$

$$
= \begin{bmatrix} \cos(\theta_1-\theta_2-\theta_4) & \sin(\theta_1-\theta_2-\theta_4) & 0 & 0 & a_1\cos\theta_1 + a_2\cos(\theta_1-\theta_2) \\ \sin(\theta_1-\theta_2-\theta_4) & -\cos(\theta_1-\theta_2-\theta_4) & 0 & 0 & a_1\sin\theta_1 + a_2\sin(\theta_1-\theta_2) \\ 0 & 0 & -1 & -1 & d_1-d_3-d_4 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}
$$

In Eq. (22) the rotation matrix ${}^{\text{tool}}\mathbf{R}_{\text{base}}$, the (3×3) upper left submatrix, expresses the orientation of tool frame {4} relative to the base frame {0} as

$$
{}^{\text{tool}}\mathbf{R}_{\text{base}} = [n\ s\ a] = \begin{bmatrix} n_x & s_x & a_x \\ n_y & s_y & a_y \\ n_z & s_z & a_z \end{bmatrix}
$$

$$
{}^{\text{tool}}\mathbf{R}_{\text{base}} =
$$

$$
\begin{bmatrix} \cos(\theta_1-\theta_2-\theta_4) & \sin(\theta_1-\theta_2-\theta_4) & 0 \\ \sin(\theta_1-\theta_2-\theta_4) & -\cos(\theta_1-\theta_2-\theta_4) & 0 \\ 0 & 0 & -1 \end{bmatrix} \tag{23}
$$

Note that the approach vector $\mathbf{a} = [0\ 0\ -1]$ is fixed, and independent of the joint variables. This is one of the characteristics of the AdeptOne robot, or even all SCARA robots, which are designed to manipulate objects directly from above. Industrial applications such as circuit board assembly, is the common area of use for this robot.

The vector $\mathbf{d}_i$, the right column vector in Eq. (22), represents position of the tool frame {4} relative to the base frame {0} as

$$
\mathbf{d}_i = \begin{bmatrix} dx \\ dy \\ dz \end{bmatrix} = \begin{bmatrix} a_1\cos\theta_1 + a_2\cos(\theta_1-\theta_2) \\ a_1\sin\theta_1 + a_2\sin(\theta_1-\theta_2) \\ d_1-d_3-d_4 \end{bmatrix} \tag{24}
$$

5.6.4 The Inverse Kinematic Solution

For the purpose of driving and controlling a robot, it is necessary to solve Eq. (20) for the joint variables since the actuator variables are the joint variables. This is the inverse (backward) kinematic problem associated with a robot manipulator: given the position and the orientation of the end effector (**X**) in the base frame, find the joint displacement (**$\boldsymbol{\theta}$**):

$$\boldsymbol{\theta} = \mathbf{f}^{-1}(\mathbf{X}) \tag{25}$$

The backward solution algorithm is generally more difficult than the forward solution. In the three-dimensional space, six co-ordinates are needed to specify a rigid body (three position co-ordinates and three angles of rotation). Since the six equations generated are nonlinear trigonometric functions and are coupled, a simple and unique solution for **q** may not even exist, For a high number of degrees of freedom the inverse kinematic problem can result in very complicated solution algorithms [58].

Some simplification is possible by properly designing the robot geometry. For example, when the axes for the three revolute joints of a six degree-of-freedom robot coincide at the wrist of the end effector, it is possible to decouple the six equations in Eq. (15) into two sets of three simpler equations [61]. The first set of equations decides the position of the first three joint variables. Once the first three joint variables are determined, the last three joint variables are obtained such that the end effector has the correct orientation.

Recall the four-axis SCARA (AdeptOne) example whose forward kinematic solution is defined by Eq. (22). Suppose that the position and orientation of the final frame (tool frame) is given as Eq. (19):

$$^{\text{tool}}\mathbf{T}_{\text{base}}(X_4) = \begin{bmatrix} n_x & s_x & a_x & d_x \\ n_y & s_y & a_y & d_y \\ n_z & s_z & a_z & d_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{26}$$

To find the corresponding joint variables $[\theta_1, \theta_2, d_3, \theta_4]$, we must solve the following simultaneous set of nonlinear trigonmetric equations:

$$\begin{aligned}
n_x &= \cos(\theta_1 - \theta_2 - \theta_4) \\
n_y &= \sin(\theta_1 - \theta_2 - \theta_4) \\
n_z &= 0 \\
s_x &= \sin(\theta_1 - \theta_2 - \theta_4) \\
s_y &= -\cos(\theta_1 - \theta_2 - \theta_4) \\
s_z &= 0 \\
a_x &= 0 \\
a_y &= 0 \\
a_z &= -1 \\
d_x &= a_1 \cos\theta_1 + a_2 \cos(\theta_1 - \theta_2) \\
d_y &= a_1 \sin\theta_1 + a_2 \sin(\theta_1 - \theta_2) \\
d_z &= d_1 - d_3 - d_4
\end{aligned}$$

Note that the SCARA robot has only four degrees of freedom. Therefore, not every element of the orientation matrix allows a solution for Eq. (22). The complete solution to the inverse kinematic problem of finding the joint variables $[\theta_1, \theta_2, d_3, \theta_4]$, in terms of the end-effector position and orientation for the SCARA manipulator is shown in Table 4.

This inverse kinematics solution is not unique due to multiplicity of q_2. Complete derivation for the solution for the four-axis SCARA (AdeptOne) robot manipulator can be found in textbooks by Spong and Vidyasagar [52] and Schilling [53]. In general, the inverse kinematic problem can be solved by various methods. The methods used usually are algebraic, geometrical, or iterative. The common approach is to use a closed form solution to the inverse kinematic problem. However, these solutions are manipulator-dependent and still often too difficult to solve in closed form. In many cases (nonclosed form) the iterative method is required with kinematically redundant robots [58].

One popular method is to use the N-dimensional Newton–Raphson algorithm or its modified version to solve the nonlinear equations [62]. Fast iterative techniques for the inverse kinematic problems have been reported based on nonlinear least-squares minimization with accurate and rapid convergence [63]. Other techniques based on the geometrical relationship between the various links and joints have been reported, depending on the points of reference chosen. In general, most researchers resort to numerical methods to obtain the inverse solution.

Table 4 Inverse Kinematics of a Four-Axis SCARA Robot

Base joint:

$$q_1 = \arctan 2[\pm\sqrt{1 - r^2}r], \text{ where } r^2 = \frac{d_x^2 + d_y^2 - a_1^2 - a_2^2}{2a_1 a_2}$$

Elbow joint:

$$q_2 = \arctan 2[d_x \; d_y] - \arctan 2[a_1 + a_2 \cos\theta_2 \; a_2 \sin\theta_2]$$

Vertical extension joint:

$$q_3 = d_1 - d_4 - d_z$$

Tool roll joint:

$$q_4 = q_1 - q_2 - \arctan 2[n_y \; n_x]$$

The use of iterative techniques raises the problem of accurate real-time implementation of manipulator kinematic control. The need to compute the inverse Jacobian at several points, importance of closeness of the initial solution to the exact solution (otherwise the algorithm diverges) and accumulation of the linearization error, reflects the computational complexity of the methods for online use.

Artificial neural network (ANN) theory has provided an alternative solution for solving the inverse kinematic problem. Artificial neural networks are highly parallel, adaptive and fault-tolerant dynamic systems modeled like their biological counterparts [39]. An application of ANNs to this problem is to train the network with the input data in the form of pairs of end-effector positions and orientations and the corresponding joint values. After the training is completed, the ANN can generalize and give good results (joint angles) at new data points (position and orientation).

Recently, ANNs have been augmented with an iterative procedure using the Newton–Raphson technique resulting in an increase in computational efficiency by twofold for the PUMA 560 robot [44]. An ANN-based scheme has also been proposed for computing manipulator inverse kinematics where no prior knowledge of the manipulator kinematics is required [64]. In the event that the physical structure of a robot is changed (or damaged) during an operation, ANN architecture can supplement the existing robot controller to learn the new transformations quickly without repairing the robot physically [42]. In a recent study, an ANN utilizing an adaptive step-size algorithm, based on a random-search technique, improved the convergence speed for solving the inverse kinematic problem for a two-link robot [45].

5.6.5 Manipulator Motion Kinematic Equations

Previously, we have described the forward and inverse arm solutions relating joint positions and end-effector position and orientation. However, the manipulator is stationary at a specific configuration. Typically, a robot is in motion to perform tasks such as spray painting, arc welding, and sealing. Continuous-path motion control is required in these applications where the tool must travel a specific paths a prescribed time (trajectory). One must then determine and control the velocity and acceleration of the end-effector between points on the path.

The forward kinematic, Eq. (20), relates the end-effector position to the joint displacements. When the end effector is in motion, an infinitesimal directional change in its position is determined by differentiating the kinematic equations (20) with respect to time, this yields

$$
\begin{aligned}
dx_m &= \frac{\partial x_m}{\partial q_1}\,\partial q_1 + \frac{\partial x_m}{\partial q_2}\,\partial q_2 + \frac{\partial x_m}{\partial q_3}\,\partial q_3 + \cdots + \frac{\partial x_m}{\partial q_n}\,\partial q_n \\
dx_m &= \mathbf{J}(\mathbf{q})d\mathbf{q}_n \qquad (27) \\
\dot{\mathbf{X}}_m &= \mathbf{J}(\mathbf{q})\dot{\boldsymbol{\theta}}_n
\end{aligned}
$$

The $\mathbf{J}(\mathbf{q})$ matrix, called the *manipulator Jacobian* or *Jacobian*, defines the linear transformation from joint co-ordinates to cartesian co-ordinates, n is the number of joints of the manipulator and m is the dimensionality of the cartesian co-ordinate of the tool under consideration. The Jacobian is one of the most important quantities in the analysis and control of robot motion. It is used for smooth trajectory planning and execution, in the derivation of the dynamic equations, and in transformation of forces applied by the end effector into the resultant torque generated at each joint. The generalized cartesian velocity vector, $\dot{\mathbf{X}}$ is defined as

$$
\dot{\mathbf{X}} = \begin{bmatrix} v \\ \omega \end{bmatrix}
$$

where $\mathbf{v} = [v_x, v_y, v_z]$ represents the linear velocity and the $\omega = [\omega\mathbf{x}, \omega_\mathbf{y}, \omega_\mathbf{z}]$ the angular velocity of the tool.

Consider the four-axis SCARA robot whose kinematic description has been developed. The Jacobian of the SCARA robot is

$$
\mathbf{J}(\mathbf{q}) = \left[\begin{array}{cccc}
-a_1\sin\theta_1 - a_2\sin(\theta_1-\theta_2) & a_2\sin(\theta_1-\theta_2) & 0 & 0 \\
a_1\cos\theta_1 + a_2\cos(\theta_1-\theta_2) & -a_2\cos(\theta_1-\theta_2) & 0 & 0 \\
0 & 0 & -1 & 0 \\
\hline
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
1 & -1 & 0 & -1
\end{array}\right] \qquad (28)
$$

The first three rows of $J(q)$ correspond to linear tool displacement, while the last three correspond to angular tool displacement. Then a joint space trajectory $q(t)$ corresponding to $x(t)$ can be obtained by inverting the Jacobian along with the inverse kinematic solution:

$$
\dot{\boldsymbol{\theta}} = \mathbf{J}(\boldsymbol{\theta})^{-1}\dot{\mathbf{X}} \qquad (29)
$$

By differentiating the Eq. (29) the desired joint accelerations are found as well,

$$\ddot{\boldsymbol{\theta}} = \dot{\mathbf{J}}(\boldsymbol{\theta})^{-1}\dot{\mathbf{X}} + \mathbf{J}(\boldsymbol{\theta})^{-1}\ddot{\mathbf{X}} \tag{30}$$

The problem with solving for the joint space differentials (velocity and acceleration) using the Jacobian is that at certain point in joint space, the tool Jacobian may lose its rank. That is, there is a reduction in the number of linearly independent rows and columns. The numerical solutions of Eqs. (29) and (30) produce very large differential values. The points at which $J(q)$ loses rank are called *joint-space singularities*. There are two types of joint-space singularities:

1. Boundary singularity, which occurs when the tool tip is on the surface of the work envelope. These are not particularly serious, because they can be avoided by mechanical constraints.
2. Interior singularity occurs inside the work envelope when two or more axes become collinear. These are more serious, because cancellation of counteracting rotations about an axis causes the tool tip position to remain constant.

From inspection of the Jacobian for the SCARA robot, Eq. (28), if and only if the upper-left 2×2 submatrix becomes singular ($|\mathbf{J}(q)| = 0$), the $\mathbf{J}(q)$ loses rank.

$$\begin{aligned}\Delta = |\mathbf{J}(q)| &= [-a_1 \sin\theta_1 - a_2 \sin(\theta_1 - \theta_2)][-a_3 \cos\theta_2] \\ &\quad - [a_2 \sin(\theta_1 - \theta_2)][a_1 \cos\theta_1 + a_2 \cos(\theta_1 - \theta_2)] \\ &= a_1 a_2 \sin\theta_1 \cos(\theta_1 - \theta_2) + a_2^2 \sin(\theta_1 - \theta_2) \\ &\quad \times \cos(\theta_1 - \theta_2) \\ &\quad - a_1 a_2 \sin(\theta_1 - \theta_2) \cos\theta_1 - a_2^2 \sin(\theta_1 - \theta_2) \\ &\quad \times \cos(\theta_1 - \theta_2) \\ &= a_1 a_2 [\sin\theta_1 \cos(\theta_1 - \theta_2) - \cos\theta_1 \sin(\theta_1 - \theta_2)] \\ &= a_1 a_2 \sin\theta_2 \end{aligned} \tag{31}$$

If $\sin\theta_2 = 0$, the Jacobian matrix is singular and has no inverse. This will occur when the elbow angle q_2 is an integer multiple of π. The tool tip is on the outer surface of the work envelope (arm is reaching straight out). whereas when $|q_2| = \pi$ the arm is folded inside the surface of the work envelope.

5.7 ROBOT ARM DYNAMICS

Similar to the robot kinematic problem, there are two types of robot dynamic problems, a direct (forward) dynamic problem and an inverse dynamic problem, as shown in Fig. 14. The problem of direct or forward dynamics is to calculate the joint trajectories (position, velocity, and acceleration), $q(t)$, given the force (torque) profile, $\tau(t)$, which causes the desired motion trajectory:

$$\mathbf{q} = \mathbf{f}(\boldsymbol{\tau}) \tag{32}$$

This problem is important in computer simulation studies of robot manipulators. However, the vast majority of robot manipulators are driven by actuators which supply a force for the prismatic joint and a torque for the revolute joint to generate motion in a prescribed trajectory. The controller must also call for torques to compensate for inertia when each link is accelerated. Therefore, the problem of inverse dynamics is to design efficient control algorithms to compute accurate force (torque) $\tau(t)$, which causes the desired motion trajectories $q(t)$:

$$\boldsymbol{\tau} = \mathbf{f}^{-1}(\mathbf{q}) \tag{33}$$

This section presents the inverse dynamic equation of motion for robot manipulators. The robot arm kinematic solution along with the velocity and the acceleration of the link coordinates will be used to obtain the inverse dynamic model for the four-axis SCARA robot (AdeptOne).

In order to control the robot manipulator in real time, at adequate sampling frequency, it is necessary to balance the generalized torques accurately and frequently from four sources [7]:

1. Dynamical source, arising from the motion of the robot:
   a. Inertia, mass property resisting change in motion, proportional to joint acceleration
   b. Coriolis, vertical forces derived from the link interactions, proportional to the product of the joint velocities
   c. Centripetal forces, constraining rotation about a point, proportional to the square of the joint velocity
2. Static source, arising from friction in the joint mechanism
3. Gravity source, arising from force of gravity on each link
4. External source, arising from external loads (tasks) on the end effector.

We can formulate the following expression for the inverse dynamics problem:

$$\boldsymbol{\tau} = \mathbf{M}(\boldsymbol{\theta})\ddot{\boldsymbol{\theta}} + \mathbf{C}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}}) + \mathbf{F}(\dot{\boldsymbol{\theta}}) + \mathbf{G}(\boldsymbol{\theta}) + \boldsymbol{\tau}_d \tag{34}$$

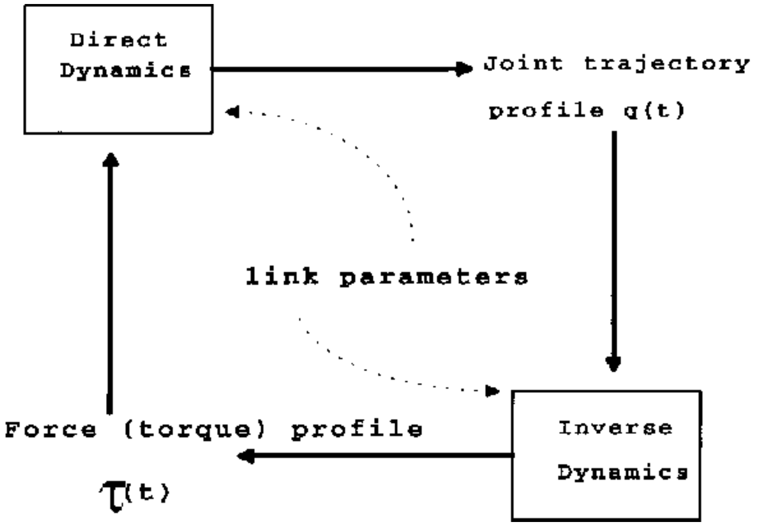


Figure 14 The direct and inverse dynamics problems.

where

$\mathbf{M}(\boldsymbol{\theta}) = (n \times n)$ inertia matrix of the robot.
$\mathbf{C}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}}) = (n \times 1)$ vector of centrifugal and Coriolis terms.
$\mathbf{F}(\dot{\boldsymbol{\theta}}) = (n \times 1)$ friction vector.
$\mathbf{G}(\boldsymbol{\theta}) = (n \times 1)$ gravity vector.
$\tau_d =$ disturbance due to unknown loading.

Detailed discussions on structure and dynamic properties of the robot equations of motion can be found in Lewis et al. [65] and Schilling [53]. An excellent collection of early kinematics and dynamics research articles, discussed in this section, by Lee et al. [51] is also available.

The dynamic characteristics of robot manipulators are highly nonlinear and therefore require a great number of mathematical operations. There are two forms of inverse dynamic solutions for a robot: (1) closed form, and (2) recursive form. If manipulators are kinematically and dynamically simple in design, an analytical expression for the closed-form dynamic equations can be derived [56]. Thus, the final analytical expression will have simple physical interpretation in terms of all the dynamic properties described above.

An analytical approach based on the Lagrange's energy function, known as Lagrange–Euler method (L-E), results in a dynamic solution that is simple and systematic. In this method, the kinetic energy (K) and the potential energy (P) are expressed in terms of joint motion trajectories. The resulting differential equations then provide the forces (torques) which drive the robot. Closed-form equations result in a structure that is very useful for robot control design and also guarantee a solution. However, its drawback is that it requires redundant calculations. For instance, when a force (torque) is applied to the end effector, of a serial link manipulator, the joint interaction results in considerable duplication of calculation in the subsequent link equations.

This duplication can be avoided in the recursive form and calculation can be made more efficient. In addition, the L-E method is inefficient, mainly because it uses the (4×4) homogeneous transformation matrices that are somewhat sparse due to combination of rotation and translation. Various attempts have been reported to simplify and improve the computational efficiency of the L-E formulation [66] and [67]. In general, these approximations, when used for control purposes, result in suboptimal dynamic performance (lower speed and position inaccuracy) [68].

One recursive approach called the Newton–Euler (N-E) formulation, has the advantage of speed and accuracy for online implementation [69]. The N-E method is based on the Newton's mass center theorem and the Euler's theory of kinetic momentum applied at each robot link. Each link is considered to be a free body and the equations of motion are obtained for each link in a recursive manner. It uses a set of forward and backward recursive equations. The forward iteration propagates kinematic information from base frame to the end effector. Once the forward iteration is complete, the backward iteration calculates and propagates the joint forces (torques) exerted on each link from end effector back to the base frame. The N-E formulation is simple and very fast. However, the derivation is messy (vector cross-product terms) and recursive equations destroy the structure of the dynamic model, which is very important for the design of robot controllers [51].

The advantage of using recursive form (numerical) over the closed form (analytical) is only the speed of computation, particularly as the number of axes increases. The L-E method has a computational complexity of order $O(n^4)$, n being the number of axes. This is in contrast to N-E formulation which has a computational complexity of order $O(n)$.

Lee et al. [68] also obtained an efficient set of closed-form solutions based on the generalized d'Alembert (G-D) principle that retain the structure of the problem with a moderate computational complexity of order $O(n^3)$. A symbolic program called algebraic robot modeler (AMR) has also been reported to generate efficient customized dynamic equations for a variety of robots [70].

In the following section, the complete dynamic model of the four-axis SCARA robot (AdeptOne) is derived based on the L-E formulation (see Fig. 13). The (AdeptOne) SCARA robot is kinematically simple, and its unique direct drive actuating system eliminates gear friction and backlash. It has closed-form dynamic solutions which will provide the simple physical interpretations (inertia, centrifugal and Coriolis forces, friction, and gravity) necessary to design the robot controller.

5.7.1 The Lagrange–Euler Formulation

The Lagrange–Euler method describes the dynamic behavior of a robot in terms of work and energy stored in the system. The constraining forces are eliminated during the formulation and the closed-form solution is derived independent of any co-ordinate system. This method is based on the utilization of [52]:

1. The (4×4) Denavit–Hartenberg matrix representation, ${}^{i}\mathbf{A}_{i-1}$, which describes the spatial relationship between the ith and $(i-1)$th link co-ordinate frames.
2. Fundamental properties of kinetic energy (K) and potential energy (P).
3. Lagrange's equation of motion:

$$\tau_i = \frac{d}{dt}\left[\frac{\partial L}{\partial \dot{q}_i}\right] - \frac{\partial L}{\partial q_i} \qquad i = 1, 2, \ldots, n \quad \text{links} \tag{35}$$

where

$L =$ lagrangian function $= (K - P)$
$q_i =$ generalized co-ordinates of the robot arm
$\dot{q}_i =$ first derivative of q_i
$\tau_i =$ generalized torques corresponding to q_i.

The generalized torques act on link i in the direction of the q_i co-ordinate frame. In the case of a revolute joint, it is composed of a torque vector, or when prismatic it is a force vector.

Since potential energy is only position dependent, Eq. (35) can be further defined as

$$\tau_i = \frac{d}{dt}\left[\frac{\partial K}{\partial \dot{q}_i}\right] - \frac{\partial K}{\partial q_i} + \frac{\partial P}{\partial q_i} \tag{36}$$

Let us begin by deriving the kinetic energy stored in a moving robot manipulator link (i) with both translation and rotation, in three-dimensional space:

$${}^{i}K = \frac{1}{2}{}^{i}m{}^{i}v_0^{T\,i}v_0 + \frac{1}{2}{}^{i}\omega_0^{T\,i}\omega_0 \tag{37}$$

where

${}^{i}m =$ mass of link i.
${}^{i}v_0 = (3 \times 1)$ linear velocity vector of the center mass with respect to reference frame.
${}^{i}\omega_0 = (3 \times 1)$ angular velocity vector of the link i with respect to reference frame.
${}^{i}I_0 = (3 \times 3)$ inertia tensor matrix of link i.

Since energy is additive, the total kinetic energy stored in the whole arm linkage is then given by

$$K = \sum_{i=1}^{n} {}^{i}K \tag{38}$$

Recall the homogeneous transformation matrix, ${}^{i}\mathbf{T}_0$, that gives the position and orientation of any link, ${}^{i}\mathbf{r}$, with respect to base frame as

$$r = {}^{i}T_0{}^{i}r \tag{39}$$

Differentiation with respect to time gives the velocity of the link position, ${}^{i}v_0$, with respect to base frame as

$${}^{i}v_0 = \frac{dr}{dt} = \frac{d}{dt}({}^{i}T_0){}^{i}r = \sum_{j=1}^{i} \frac{\partial{}^{i}T_0}{\partial\theta_j}\dot{\theta}_j{}^{i}r \tag{40}$$

Substituting Eq. (40) in Eq. (37) and subsequently in Eq. (38) yields

$$K = \sum_{i=1}^{n} \frac{1}{2}\,{}^{i}m \left(\sum_{j=1}^{i} \frac{\partial\,{}^{i}T_0}{\partial\theta_j}\,\dot{\theta}_j\,{}^{i}r \right)^{T} \sum_{j=1}^{i} \frac{\partial\,{}^{i}T_0}{\partial\theta_j}\,\dot{\theta}_j\,{}^{i}r$$

$$+\frac{1}{2} \left(\sum_{j=1}^{i} \frac{\partial\,{}^{i}T_0}{\partial\theta_j}\,\dot{\theta}_j\,{}^{i}r \right)^{T} {}^{i}I_0 \sum_{j=1}^{i} \frac{\partial\,{}^{i}T_0}{\partial\theta_j}\,\dot{\theta}_j\,{}^{i}r$$

$$K = \sum_{i=1}^{n} \frac{1}{2}\,{}^{i}m \sum_{j=1}^{i} \sum_{k=1}^{i} \frac{\partial\,{}^{i}T_0^{T}}{\partial\theta_j}\,{}^{i}r^{T}\,{}^{i}r\,\frac{\partial\,{}^{i}T_0}{\partial\theta_k}\,\dot{\theta}_j\dot{\theta}_k$$

$$+\frac{1}{2} \sum_{j=1}^{i} \sum_{k=1}^{i} \frac{\partial\,{}^{i}T_0^{T}}{\partial\theta_j}\,{}^{i}r^{T}\,{}^{i}I_0\,{}^{i}r\,\frac{\partial\,{}^{i}T_0}{\partial\theta_k}\,\dot{\theta}_j\dot{\theta}_k \tag{41}$$

Asada and Slotine [57] suggest to rewrite the expressions in Eq. (41) by using the $(n \times n)$ *manipulator inertia tensor matrix*, $\mathbf{M}$,

$$\mathbf{M} = \sum_{i-1}^{n} \left[\begin{array}{l} {}^{i}m \displaystyle\sum_{j=1}^{i} \sum_{k=1}^{i} \frac{\partial\,{}^{i}T_0^{T}}{\partial\theta_j}\,{}^{i}r^{T}\,{}^{i}r\,\frac{\partial\,{}^{i}T_0}{\partial\theta_k} \\ \quad + \displaystyle\sum_{j=1}^{i} \sum_{k=1}^{i} \frac{\partial\,{}^{i}T_0^{T}}{\partial\theta_j}\,{}^{i}r^{T}\,{}^{i}I_0\,{}^{i}r\,\frac{\partial\,{}^{i}T_0}{\partial\theta_k} \end{array} \right] \tag{42}$$

The matrix $\mathbf{M}$ incorporates all the mass properties of the entire arm linkage. The manipulator inertia matrix, also called mass matrix, is symmetrical and in quadratic form. Since the kinetic energy is positive, unless the robot is at rest, the inertia matrix is *positive definite*.

$$K = \frac{1}{2} \sum_{j=1}^{n} \sum_{j=1}^{n} M_{ij}\theta_i\theta_j \tag{43}$$

Note that M_{ij}, component of inertia matrix $\mathbf{M}$, is a function of joint variables $\mathbf{q}$ and represents coupling between the i and j links. The diagonal term M_{ii} represents the self-inertial coefficients. Since the mass $\mathbf{M}$ is positive definite, the coefficient of coupling falls between 0 (no inertial interaction) and 1 (tightly coupled).

Since the mass matrix $\mathbf{M}$ involves Jacobian matrices, which are configuration dependent and can vary, two links can be highly coupled in one configuration and completely decoupled in another. Desirably, the manipulator inertia matrix would be diagonal with constant self-inertial terms. This will allow the dynamic properties to stay the same for all configurations and, the control algorithms simple [71].

The kinetic energy depends on q and dq/dt. Therefore,

$$K(q, \dot{q}) = \frac{1}{2}\,\dot{q}^{T} M(q) \dot{q} \tag{44}$$

The potential energy of position vector defined in Eq. (36) in a gravity field $\mathbf{g} = [g_x\ g_y\ g_z\ 0]$ is

$${}^{i}P = -\,{}^{i}mgr = -\,{}^{i}mg\,{}^{i}T_0\,{}^{i}r \tag{45}$$

Then the total arm potential energy is

$$P = -\sum_{i=1}^{n} m_i g\,{}^{i}T_0\,{}^{i}r \tag{46}$$

Note that the potential energy depends only on the joint variable q.

The terms required in the Lagrangian equation (36) are now given by

$$\text{1st term:} \quad \frac{d}{dt}\left[\frac{\partial K}{\partial \dot{q}_i}\right] = M(q)\ddot{q} + \dot{M}(q)\dot{q}$$

$$\text{2nd term:} \quad \frac{\partial K}{\partial q_i} = \frac{1}{2}\frac{\partial}{\partial q}(\dot{q}^{T} M(q)\dot{q})$$

$$\text{3rd term:} \quad \frac{\partial P}{\partial q_i} = -\sum_{i=1}^{n} m_i g\,{}^{i}T_0\,{}^{i}r$$

Therefore, the arm dynamic equation is

$$\tau = M(q)\ddot{q} + \dot{M}(q)\dot{q} - \frac{1}{2}\frac{\partial}{\partial q}(\dot{q}^{T} M(q)\dot{q})$$

$$+ \sum_{i=1}^{n} m_i g\,{}^{i}T_0\,{}^{i}r \tag{47}$$

Defining the Coriolis/centripetal vector as

$$C(q, \dot{q}) = \dot{M}(q)\dot{q} - \frac{1}{2}\frac{\partial}{\partial q}(\dot{q}^{T} M(q)\dot{q}) \tag{48}$$

The final dynamic form defined in Eq. (34) is derived. The friction and disturbance terms can be added to complete the dynamic equation. The robot dynamic equation represents nonlinearities due to co-ordinate transformations which are trigonometric. Additional nonlinearities appear in both kinetic and potential energy terms due to Coriolis and centrifugal terms. The Coriolis terms represent the velocity coupling of links j and k felt at joint i, in torque or force form, whereas the centrifugal terms reflect the velocity coupling of only link j at joint i. The Coriolis and centrifugal coefficients are small and become important when the robot is moving at high speed.

The gravity coefficients arise from the potential energy stored within individual links. These coefficients also vary with configuration. Equation (47) provides the dynamic properties based on the assumption that no loads are attached to the arm. Effect of a load on the dynamic coefficients is additive and therefore can be considered as a point mass and an extension of the last link (tool). Robot manipulators use actuators (electric or hydraulic) to move. Since the actuator and motor inertias are decoupled values acting only at the joints, they can be treated as additive terms to the mass matrix [82]. Also, note that each torque (force) computation involves the summation over a possible range of n joints. This creates a computation complexity in the order of $O(n^4)$. This high order of calculations is very slow for online implementation. In practice, the manipulator dynamic must be modeled accurately for precise and high speed motion control.

5.7.2 Dynamic Model of the SCARA Robot

Tasks performed by robots are defined previously as gross manipulation and fine manipulation tasks. Motion trajectory control of the end effector applies to the tasks in the first category. Robots, in general, use the first three axes for gross manipulation (position control), while the remaining axis orients the tool during the fine manipulation (force or tactile control).

Robots, in parts assembly applications, are to pick a component up with a vertical movement, move it horizontally and then downwards vertically for insertion. These can be achieved by the four-axis SCARA robots. Examples of robots which belong to this class include the AdeptOne robot, the Intelledex 440 robot, and the IBM 7545 robot. The first three axes of a SCARA robot position the end effector, while the fourth axis orients the tool through a roll motion [53].

This section provides the dynamic equations of AdeptOne SCARA robot, based on the L-E formulation discussed above. AdeptOne robot is a direct-drive robot. As there is no gearbox, the friction at each joint is very small. The vector of joint variables is

$$[\theta_1 \quad \theta_2 \quad d_3 \quad \theta_4]$$

where the third joint is prismatic, while the remaining three joints are revolute.

From reviewing the results of Table 4, it is clear that the tool roll angle θ_4 has no effect on the end-effector position, therefore θ_4 can be ignored in our investigation of the motion trajectory control. The mass of the fourth link and the attached end effector is also assumed to be very small in comparison with the masses of the other links.

The link-co-ordinate diagram for this robot is shown in Fig. 13 for the first three axes. The three links are assumed to be thin cylinders of mass m_1, m_2, and m_3. The vertical column height d_1 is stationary, and when joint 1 is activated, only rod of length a_1 (mass of m_1) rotates. The dynamic relationships are governed by Eq. (47):

$$\tau = M(q)\ddot{q} + \dot{M}(q)\dot{q} - \frac{1}{2}\frac{\partial}{\partial q}(\dot{q}^T M(q)\dot{q}) + \sum_{i=1}^{n} m_i g\, {}^iT_0\, {}^ir$$

where the matrix $M(q)$ is $(n \times n)$ symmetrical and positive definite with elements $m_{ij}(q)$, and n is the number of joints. We will adopt the following notation:

l_i, length of link i
m_i, mass of link i
$I = (m_i a_i^2)/12$, mass moment of inertia of link i about its center of mass of thin cylinder
g, gravity constant.

Then the dynamic equation for the first joint of the manipulator would be

$$\begin{aligned}\tau_1 = {} & \left[\left(\frac{m_1}{3} + m_2 + m_3\right)a_1^2 + (m_2 + 2m_3)a_1 a_2 \cos\theta_2 \right.\\ & \left. + \left(\frac{m_2}{3} + m_3\right)a_2^2\right]\ddot{\theta}_1 - \left[\left(\frac{m_2}{2} + m_3\right)a_1 a_2 \cos\theta_2 \right.\\ & \left. + \left(\frac{m_2}{3} + m_3\right)a_2^2\right]\ddot{\theta}_2 - (m_2 + 2m_3)a_1 a_2 \sin\theta_2 \dot{\theta}_1 \dot{\theta}_2 \\ & + \left(\frac{m_2}{2} + m_3\right)a_1 a_2 \sin\theta_2 \dot{\theta}_2^2 \qquad (49)\end{aligned}$$

Since axis 1 is aligned with the gravitational field, the gravity term on joint 1 is zero.

Second joint:

$$\begin{aligned}\tau_2 = {} & -\left[\left(\frac{m_2}{2} + m_3\right)a_1 a_2 \cos\theta_2 + \left(\frac{m_2}{3} + m_3\right)a_2^2\right]\ddot{\theta}_1 \\ & + \left(\frac{m_2}{3} + m_3\right)a_2^2 \ddot{\theta}_2 + \left(\frac{m_2}{2} + m_3\right)a_1 a_2 \sin\theta_2 \dot{\theta}_2^2 \\ & \qquad (50)\end{aligned}$$

Again, there is no gravitational loading on joint 2.

Third joint:

$$\tau_3 = m_3 \ddot{\theta}_3 - g m_3 \qquad (51)$$

Joint 3 is a prismatic joint, used to establish the vertical tool position, and is completely independent of the first two joints. Thus, there are no Coriolis and centrifugal forces on this joint. However, The mass of the third joint effects the motion of the first two joints by acting as a load.

Equations (49)–(51) are referred to as the inverse dynamic equations of the three-axis SCARA robot.

One needs to know the resulting motion trajectories: $\theta, \dot{\theta}, \ddot{\theta}$ before one can calculate the joint torques, τ. These equations are time-varying, nonlinear, and coupled differential equations. The trajectory problem of this robot manipulator revolves around finding the joint torques $\tau_i(t)$, such that the joint angles $\theta_i(t)$ track the desired trajectories $\theta_{id}(t)$. Torque-based control techniques are built directly on the dynamic models of the robot manipulators.

5.8 ROBOT NEURAL CONTROLLER

In order to design intelligent robot controllers, one must also provide the robot with a means of responding to problems on the temporal context (time) as well as spatial (space). It is the goal of the intelligent robot researchers to design a neural learning controller to utilize the available data from the repetition in the robot operation. The neural learning controller based on the recurrent network architecture has the time-varying feature that once a trajectory is learned, it should learn a second one in a shorter time.

In Fig. 15, the time-varying recurrent network will provide the learning block (primary controller) for the inverse dynamic equations discussed above. The network compares the desired trajectories: $\theta_d, \dot{\theta}_d, \ddot{\theta}_d$, with continuous paired values for the three-axis robot $\theta, \dot{\theta}, \ddot{\theta} : \tau$, at every instant in a sampling time period. The new trajectory parameters are then combined with the error signal from the secondary controller (feedback controller) for actuating the robot manipulator arm.

Neural networks can be applied in two ways in the design of the robot controller described in Fig. 4: (1) system identification model and (2) control. ANNs can be used to obtain the system model identification which can be used to design the appropriate controller. They can also be used directly in design of the controller [72] once the real system model is available. Neural network approaches to robot control are discussed in general by Psaltis et al. [25], and Yabuta and Yamada [73]. These approaches can be classified into:

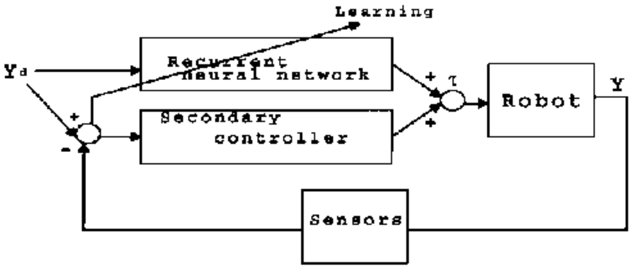


Figure 15 Recurrent neural learning controller.

1. *Supervised control*, a trainable controller, unlike the old teaching pendant, allows responsiveness to sensory inputs. A trainable neuromorphic controller reported by Guez and Selinsky [74] provides an example of a fast, real-time, and robust controller.
2. *Direct inverse control* is trained for the inverse dynamics of the robot. Kung and Hwang [75] used two networks online in their design of the controller.
3. *Neural adaptive control*, neural nets combined with adaptive controllers, results in greater robustness and ability to handle nonlinearity. Chen [76] reported the use of the BP method for a nonlinear self-tuning adaptive controller.
4. *Backpropagation of utility*, involves information flowing backward through time. Werbos's backpropagation through time is an example of such a technique [77].
5. *Adaptive critic method*, uses a critic to evaluate robot performance during training. Very complex method; requires more testing [78].

In the direct inverse control approach, the recurrent neural network will learn the inverse dynamics of the robot in order to improve the controller performance. The neural network model replaces the primary controller (see Fig. 4). In this approach a feedback controller (secondary controller) will be used to teach the network initially. As learning takes place, the neural network takes full control of the system.

Kawato and his research group were successful in using this approach in trajectory control of a three-degree-of-freedom robot [24, 79]. Their approach is known as feedback-error-learning control. However, their neural network structure was simply the linear collection of all nonlinear dynamic terms (they called them subsystems) in the dynamic motion equation. Learning was purely for the estimates of the subsystems. As the degrees of freedom increase the network size needs to increase (order of n^4). For example, for six degrees of freedom 942 subsystems are needed, compared with 43 for the three degrees of freedom. However, due to the parallel processing capability of the neural network, the implementation of Kawato's method is still an attractive method.

Goldberg and Pearlmutter [80] have demonstrated the utility of the feedback-error-learning approach for the motion control of the first two joints of the CMU

DDArm II, using temporal windows of measured positions as input to the network. The output of the network is the torque vector. Newton and Xu [81] used this approach to control a flexible space robot manipulator (SM^2) in real time. The trajectory tracking error was reduced by 85% when compared to conventional PID control scheme. More recently, Lewis et al. [82] developed an online neural controller, based on the robot passivity properties (system cannot go unstable if the robot cannot create energy), using a similar approach with good tracking results. The feasibility and performance of the feedback-error-learning control with global asymptotic stability has also been reported [83, 84]. The design of a compact and generic *recurrent network* has also shown promising results in replacing the need for custom subsystems-type design, such as the one by Kawato's group [85]. The proposed controller performs based on the systematic design approach and the recurrent network's time-varying feature.

REFERENCES

1. EL Hall, BC Hall. Robotics: A User-Friendly Introduction, Orlando, FL: Saunders College Publishing, Holt, Rienhart and Wilson, 1985, pp 1–8.
2. T Kohonen. Introduction to neural computing. Neural Net 1: 3–16, 1988.
3. CR Asfahl. Robots and Manufacturing Automation, New York: John Wiley & Sons, 1992, pp 1–10.
4. NG Odrey. Robotics: applications. In: RC Dorf, ed.-in-chief. The Electrical Engineering Handbook, Boca Ratio, FL: CRC Press, 1993, pp 2175–2182.
5. J Holusha. Industrial robots make the grade. New York Times, Wed Sept 7, 1994, pp c1, d5.
6. P Sinton. Faster, flexible robots tackle the job. San Fran Chron May 15, 1995, sec B, p 1, col 6.
7. PJ McKerrow. Introduction to Robotics. Reading MA: Addison-Wesley, 1991, pp. 14–23.
8. M Weil. New competitiveness spurs record robot sales (part 2 of 2). Manag Autom 9(6): 5–8, 1994.
9. RIA. Robotics Industry Has Best Year Ever in 1997. Ann Arbor, MI: Robotics Industries Association, 1998 (http://www.robotics.org).
10. RD Klafter, TA Chmielewski, M Negin. Robotic Engineering: An Integrated Approach. Englewood Cliffs, NJ: Prentice-Hall, 1989, pp 244–247.
11. J Rottenbach. Quality takes a seat via welding (part 2 of 2). Manag Autom, 7(6): 16, 1992.
12. EL Hall, GD Slutzky, RL Shell. Intelligent robots for automated packaging and processing. Qual Use Computer Comput Mech Artif Intell Robot Acoust Sens 177: 141–146, 1989.
13. D Labrenz. Robots lend muscle to palletizing (part 2 of 2). Manag Autom 7(6): 16–20, 1992.
14. K Sauer. Robotic arc spray eliminates warping (part 2 of 2). Manag Autom 7(6): 10, 1992.
15. G Boothroyd. C Poli, LE Murch. Automatic Assembly. New York: Marcel Dekker, 1982.
16. RL Hokestra. Design for assembly. PhD dissertation, University of Cincinnati, Cincinnati, OH, 1992.
17. G Farnum. Robots figure in flexible assembly (part 2 of 2). Manag Autom 7(6): 9, 1992.
18. JH Nurre, EL Hall. Three dimensional vision for automated inspection. Proceedings of Robots 13 Conference, Gaithersburg, MD, May 7–11, 1989, pp 16–1 to 16–11.
19. O Davies. Vision fires up parts inspection (part 2 of 2). Manag Autom 7(6): 14, 16, 36, 1992.
20. AJ Ty, CH Tien. Robotics: robot configuration. In: RC Dorf, ed.-in-chief. The Electrical Engineering Handbook. Boca Raton FL: CRC Press, 1993, pp 2154–2162.
21. D Kaiser. Valves, servos, motors, and robots. In: DM Considine, ed.-in-chief. Process/Industrial Instruments and Controls Handbook, 4th ed. New York: McGraw-Hill, 1993, pp 9.86–9.87.
22. T Yoshikawa. Foundations of Robotics: Analysis and Control. Cambridge MA: The MIT Press, 1990, pp 1–12.
23. AJ Koivo. Fundamentals for Control of Robotic Manipulators. New York: John Wiley & Sons, 1989. pp 306–338.
24. H Miyamoto, M Kawato, T Setoyama, R Suzuki. Feedback error learning neural network model for trajectory control of a robotic manipulator. Neural Net 1: 251–265, 1988.
25. D Psaltis, A Sideris, A Yamamura. AA A multilayered neural network controller. IEEE Control Syst Mag 8(2): 17–21, 1988.
26. RP Lippman. An introduction to computing with neural nets. IEEE ASSP Mag 4(2): 4–22, 1987.
27. M Chester. Neural Networks: A Tutorial. Englewood Cliffs, NJ: Prentice-Hall, 1993.
28. J Hertz, A Krogh, RG Palmer. Introduction to the Theory of Neural Computation. Reading MA: Addison-Wesley, 1991.
29. DO Hebb. The Organization of Behavior: A Neuropsychological Theory. New York: Wiley, 1949.
30. V Vemuri. Artificial neural networks: an introduction. In: V Vemuri, ed. Artificial Neural Networks: Theoretical Concepts. IEEE Computer Society Press, 1988, pp 1–12.
31. B Widrow, MA Lehr. 30 years of adaptive neural networks: Perceptron, madaline, and backpropagation. Proc IEEE 78: 1415–1442, 1990.
32. ML Minsky, SA Papert. Perceptrons. MIT Press, Cambridge, MA 1969.
33. JJ Hopfield, TW Tank. 'Neural' computation of decisions in optimization problems. Biol Cybern 52: 141–152, 1985.

34. S Grossberg. Studies of Mind and Brain. Boston Studies in the Philosophy of Science, vol 70, Boston, MA: Reidel Publishing Company, 1982.
35. S Grossberg, ed. Neural Networks and Natural Intelligence. Cambridge MA: MIT Press, 1988.
36. T Kohonen. Self-organized formation of topologically correct feature maps. Biol Cybern 43: 59–69, 1982.
37. B Kosko. Bi-directional associative memories. IEEE Trans Syst Man Cybern 18(1): 49–60, 1988.
38. R Hecht-Nielsen. Neurocomputing. Reading MA: Addison-Wesley, 1990, pp. 182–190.
39. DE Rumelhart, GE Hinton, RJ Williams. Learning internal representation by error propagation. In: DE Rumelhart, JL McClelland, eds. Parallel Distributed Processing: Exploration in the Microstructure of Cognition, vol 1. Cambridge, MA: MIT Press, 1986, pp 318–362.
40. C Lau, ed. Neural Networks: Theoretical Foundations and Analysis. New York: IEEE Press, 1992.
41. P. Chapnick. Lots of neural nets books. AI Expert June: 21–23, 1992.
42. G Josin, D Charney, D White. Robot control using neural networks. IEEE International Conference on Neural Networks, vol 2, 1988, pp 625–631.
43. M Kuperstein, J Wang. Neural controller for adaptive movements with unforeseen payload. IEEE Trans Neural Net 1(1): 137–142, 1990.
44. A Guez, Z Ahmad, J Selinsky. The application of neural networks to robotics. In: PGJ Lisboa, ed. Neural Networks: Current Applications. London: Chapman & Hall, 1992, pp 111–122.
45. W Golnazarian, EL Hall, RL Shell. Robot control using neural networks with adaptive learning steps. SPIE Conference Proceedings, Intelligent Robots and Computer Vision XI: Biological, Neural Net, and 3-D Methods, vol 1826, pp 122–129, 1992.
46. RA Jacobs. Increased rates of convergence through learning rate adaptation. Neural Net, 1: 295–307, 1988.
47. N Baba. A new approach for finding the global minimum of error function of neural networks. Neural Net 2: 267–373, 1989.
48. BA Pearlmutter. Learning state-space trajectories in recurrent neural networks. Neural Comput. 1: 263–269, 1989.
49. FJ Pineda. Recurrent backpropagation and the dynamical approach to adaptive neural computation. Neural Comput 1: 161–172, 1989.
50. M. Caudell, C Butler. Understanding Neural Networks: Computer Exploration, Advanced Networks, vol 2. Cambridge, MA: MIT Press, 1992, pp 79–112.
51. CSG Lee. Robots arm kinematics, dynamics, and control. Computer 15(12): 62–80, 1982.
52. MW Spong, M Vidyasagar. Robot Dynamics and Control. New York: Wiley, 1989.
53. RJ Schilling. Fundamentals of Robotics: Analysis and Control. Englewood Cliffs, NJ: Prentice-Hall, 1990.
54. RP Paul. Robot Manipulators. Cambridge, MA: MIT Press, 1981.
55. WA Wolovich. Robotics: Basic Analysis and Design. New York: Holt, Rinehart, and Winston, 1986.
56. JJ Craig. Introduction to Robotics: Mechanics and Control. Reading, MA: Addison-Wesley, 1986.
57. H Asada, JE Slotine. Robot Analysis and Control. New York: John Wiley & Sons, 1986, pp 133–183.
58. KS Fu, RC Gonzalez, CSG Lee. Robotics: Control, Sensing, Vision and Intelligence. New York: McGraw-Hill, 1987, pp 201–223.
59. J Denavit, RS Hartenberg. A kinematic notation for lower-pair mechanisms based on matrices. J Appl Mech 22:215–221, 1955.
60. H Asada and K Youcef-Toumi. Direct Drive Robots. Cambridge, MA: MIT Press, 1987.
61. D Pieper. The kinematics of manipulators under computer control. PhD dissertation, Stanford University, CA, 1968.
62. S Oh, D Orin, M Bach. An inverse kinematic solution for kinematically redundant robot Manipulators. J Robot Syst 1(3): 235–249, 1984.
63. AA Goldenberg, DL Lawrence. A generalized solution to the inverse kinematics of robotic manipulator. J Dyn Syst Meas Control 107: 103–106, 1985.
64. L Nguyen, RV Patel. A neural network based strategy for the inverse kinematics problem in robotics. Proceedings of International Symposia on Robotics and Manufacturing, vol 3, 1990, pp 995–1000.
65. FL Lewis, CT Abdallah, DM Dawson. Control of Robot Manipulators. New York: Macmillan Publishing Company, 1993, pp 140–158.
66. J Hollerbach. A recursive lagrangian formulation of manipulator dynamics and a comparative study of dynamics formulation complexity. IEEE Trans Syst Man Cybern SMC-10 (11): 730–736, 1980.
67. M Renaud. An efficient iterative analytical procedure for obtaining a robot manipulator dynamic model. In M Brady, ed. Robot Res, MIT Press, 1984, pp 749–768.
68. CSG Lee, BH Lee, R Nigham. Development of the generalized d'Alembert equations of motion for mechanical manipulator. Proceedings of the 22nd Conference on Decision and Control, December 14–16, 1983.
69. LYS Luh, M Walker, R Paul. Resolved-acceleration control of mechanical manipulators. IEEE Trans Autom Control AC-25: 468–474, 1980.
70. CP Neuman, JJ Murray. Customized computational robot dynamics. J Robot Syst 4(4): 503–526, 1987.
71. VD Tourassis, CP Neuman. Properties and structure of dynamic robot models for control engineering applications. Mechanism Mach Theory, 20(1): 27–40, 1985.
72. KS Narendra, K Parthasarathy. Identification and control of dynamical systems using neural networks. IEEE Trans Neural Net 1(1): 4–27, 1990.

73. T Yabuta, T Yamada. Neural network controller characteristics with regard to adaptive control. IEEE Trans Syst Man Cybern 22(1): 170–176, 1992.

74. A Guez, J Selinsky. A trainable neuromorphic controller. J Robot Syst 5(4): 363–388, 1988.

75. S Kung, J Hwang. Neural network architectures for robotic applications. IEEE Trans Robot Autom 5(5): 641–657, 1989.

76. F Chen. Back-propagation neural networks for nonlinear self-tuning adaptive control. IEEE Control Syst Mag April: 44–48, 1990.

77. P Werbos. Backpropagation through time: what it does and how it does it. Proc IEEE 78: 1550–1560, 1990.

78. P Werbos. An overview of neural networks for control. IEEE Control Syst Mag 11(1): 40–42, 1991.

79. M Kawato, K Furukawa, R Suzuki. A hierarchical neural-network model for control and learning of voluntary movement. Biol Cybern 57: 169–185, 1987.

80. KY Goldberg, BA Pearlmutter. Using backpropagation with temporal windows to learn the dynamics of the CMU direct-drive arm II. In: DS Touretzky, ed. Advances in Neural Information Processing Systems I. Palo Alto, CA: Morgan Haufmann Publishers, 1989, pp 356–363.

81. RT Newton, Y Xu. Real-time implementation of neural network learning control of a flexible space manipulator. IEEE International Conference on Robotics and Automation, Atlanta, GA, May 2–6, 1993, pp 135–141.

82. FL Lewis, K Liu, A Yesildirek. Neural net robot controller with guaranteed tracking performance. IEEE Trans Neural Net 6(3): 703–715, 1995.

83. M Kawato. Feedback-error-learning neural network for supervised motor learning. In: R Eckmiller, ed. Advanced Neural Computers, Elsevier Science Publishers BV (North-Holland), 1990, pp 365–372.

84. D Patino, R Carelli, B Kuchen. Stability analysis of neural networks based adaptive controllers for robot manipulators. Proceedings of the American Control Conference, Baltimore, MD, June 1994, pp 609–613.

85. W Golnazarian. Time-varying neural networks for robot trajectory control. PhD dissertation, University of Cincinnati, OH, 1995.

Chapter 6.6

Industrial Materials Science and Engineering

Lawrence E. Murr
The University of Texas at El Paso, El Paso, Texas

6.1 INTRODUCTION AND BRIEF HISTORICAL PERSPECTIVE

We often refer to our current, wide use of materials—metals, ceramics, plastics/polymers, semiconductors, superconductors, composites, etc.—as representing the "Age of Materials." The reality, however, is that the age of materials began with the Stone Age, dating well before 10,000 BC, when prehistoric peoples fashioned crude tools, weapons, and other commodities from stone. But stone continues in use today, not as a basis for tools and weapons but as a basis for building materials and other construction applications. The Great Pyramids and Hanging Gardens of Babylon, among other notable developments in antiquity, are of stone construction, and they exemplify not only the variations in stone as a material, but also the selective use of these variations in both the engineering and art of construction. In addition, a variety of clays and wood-augmented stone materials remain a significant part of our materials technology.

The earliest metal to be used by people in a significant way was copper, and the world's earliest known man-made copper objects—beads, pins, and awls—were fabricated about 8000 BC in the region which is near Turkey and Iran, and often designated as the start of the Copper (or Chalcolithic) Age. Copper was probably first discovered in its native state in various locations worldwide in prehistoric times, and there is evidence of copper mining in the Balkans circa 5000 BC. Copper came to prominence in Europe as early as the fourth millennium BC when artisans—the first extractive metallurgists—learned to extract it by smelting ore. In this early ore smelting or copper extraction process, copper oxide ore was heated in open and closed rock and clay crucibles or hearths buried in the ground, using wood, sometimes mixed with the ore, as the fuel. Somehow it was discovered that blowing air through hollow sticks, embedded in the fuel–ore mass, fanned the fire to create heat intense enough to free the copper as a molten mass. Actually the molten mass consisted of slag which contained impurities from the ore which separated upon cooling and could be chipped away, leaving relatively pure copper metal blocks. These blocks, or ingots, were later reheated, and the liquid copper poured into a cast or mold, or fashioned into tools, such as ax blades, and weapons, sharper and more workable than stone. Indeed it is the quality of workability or *malleability* which continued to set copper apart as a unique material over a period of at least 10,000 years, including the current technologies of the world. Even during the period of the earliest uses of native copper, native gold had also been discovered and was being utilized. Somehow it was observed that hammering the two metals caused copper to harden (or *work harden*), whereas gold remained malleable. Those who recognized and utilized these performance qualities were probably the first adaptive metallurgists. This intuition about the physical and mechanical properties of native metals probably dates prior to the Copper Age. The demand for copper increased with agricultural inventions like the plow and wheeled carts, and traders, prospectors, and early metallurgists exchanged ideas and toolmaking skills.

While metallurgy began with a realization that copper could be hammered into shapes which held sharp edges far better than gold (or even silver which was also discovered as native metal in prehistoric times), the utility of metallurgy in the shaping of civilizations began with the development of systematic processes to extract metal from its ore. Copper smelting—or ore reduction using a hearth—was the basis for metal recovery well into the iron age.

During the thousands of years native copper was worked and copper smelting was discovered and developed as a technological process, tin, a white metal, was also somehow smelted. Tin smelting was easier than copper because tin melts at only 232°C. Somehow, it is believed, a metallurgist in antiquity discovered that tin could be mixed with copper not only to produce a wide range of working options but even alterations in the appearance or luster of copper. The mixing of tin and copper to produce an alloy we call bronze ushered in the Bronze Age around 3000 BC, and created a versatility in the use of copper by lowering the processing temperatures, which made it more castable and more tractable in a wide range of related applications. This phenomenon was expanded with the addition of zinc to tin bronze in the Middle Ages to form castable gun metal (88% Cu, 10% Sn, 2% Zn). In reality, this was the foundation of the "age of alloys" which began a technological phenomenon we today call "materials by design." Metal alloys first discovered in the Bronze Age influenced the development of an enormous range of metal combinations which attest to essentially all of the engineering achievements of modern times, including atomic energy, space flight, air travel, communications systems and microelectronic devices, and every modern structural and stainless-steel building, vessel, or commodity items, including eating utensils.

It is believed that early man found iron meteorite fragments which were shaped by hammering into tools, weapons, and ornaments, because iron is rarely found in the native state, in contrast to copper, gold, or silver. In addition, chemical analysis of archeological specimens often shows 7–15% nickel, and natural iron–nickel alloys (awaruite, $FeNi_2$, and josephinite, Fe_3Ni_5) are extremely rare and almost exclusively confined to a geological region in northwestern Greenland. Ancient writings from India and China suggest that ferrous metals may have been extracted by the reduction of iron ore in hearth processes similar to copper ore reduction as early as 2000 BC, and this extraction process was well established over a wide area of the ancient world between about 1400 and 1100 BC, establishing an era referred to as the Iron Age. As time went by, variations in hearth or furnace design emerged along with discoveries of variations in the properties of iron brought on by *alloying* due to the absorption of carbon These included a reduction in the temperature required for the production of molten *(pig) iron* and the production of melted iron used to make castings *(cast iron)* when the carbon content was high (between 3 and 4 wt%). Low-carbon metal, in contrast, was relatively soft, ductile, and malleable, and could be hammered or hammer welded at forging temperatures (in an open-hearth furnace), and corresponded to what became generally known as *wrought iron*. Somewhere near 1 wt% carbon (or somewhere between a very low-carbon- and a very high-carbon-containing iron) the iron–carbon alloy produced could be made to exhibit a range of hammer-related characteristics, not the least of which involved extreme hardness when it was variously cooled by immersion in water or some other liquid from a high temperature. The process of *quenching* was the forerunner of modern steel production. The quenched metal could be reheated for short periods at lower temperature to reduce the hardness and concomitant brittleness in what is now called *tempering*. The metal could also be variously worked in tempering cycles which provided a range of hardness and strength to the final products, forerunners of specialty steels.

6.2 MATERIALS FUNDAMENTALS: STRUCTURE, PROPERTIES, AND PROCESSING RELATIONSHIPS IN METALS AND ALLOYS

We now recognize metals as a unique class of materials whose structure, even at the atomic level, provides distinctive properties and performance features. For example, it is the atomic structure, or more accurately the electronic structure of metal atoms, which gives rise to magnetic properties and magnetism. The atomic structure also controls both thermal and electrical conductivity in metals because of the special way in which metal atoms are bound together to form solid crystals. Furthermore, metals can be mixed to form alloys having properties which can either mimic the constituents or produce completely different properties. It is this diversity of properties and the ability to alter or manipulate properties in a continuous or abrupt manner which have made metals so tractable in commerce and industry for thousands of years. In this section, we will illustrate some of the more important features of metals and alloys—their structures, properties, processing, and performance issues—by following a case history concept based on chronology of the ages of metals. That is, we will first emphasize and a few exemplary copper alloys and then illustrate similar metallurgical and materials features using steel and other examples. There will be some redundancy in these illustrations, especially in regard to the role of crystal structures and defects in these structures. The emphasis will involve physical and mechanical metallurgy whose themes are variants on the interrelationships between structure, properties, processing, and ultimately, performance.

6.2.1 The Physical and Mechanical Metallurgy of Copper and Copper Alloys

Following the extraction of metals, such as copper and their refining to produce useful forms for industrial utilization, metallurgy becomes an adaptive process involving physical, mechanical, and chemical issues, such as electroplating and corrosion, for example.

To illustrate a wide range of these metallurgical processes and related fundamental issues, we will begin with the example in Fig. 1a which shows tiny (75 μm diameter) copper wires that provide access to

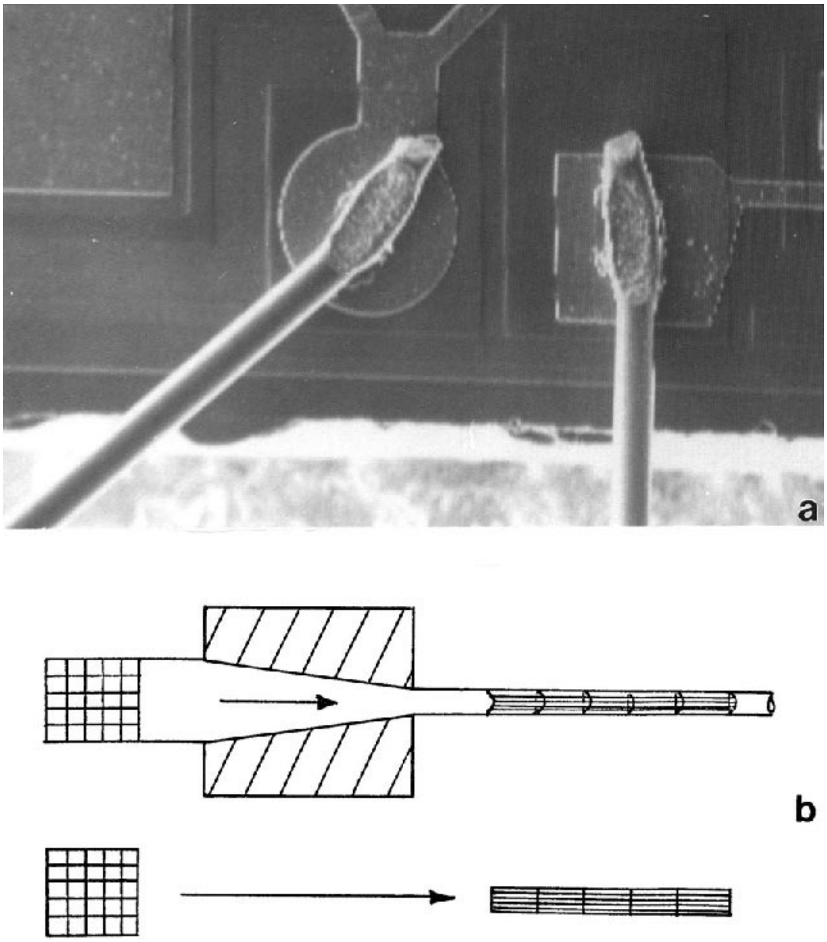


Figure 1 (a) Small section of integrated circuit showing 75 μm diameter copper, bonded connector wires. (b) Schematic representation for mechanical drawing of these fine connector wires through a hardened die. The drawing operation shown represents inhomogeneous deformation which is contrasted with homogeneous deformation of a simple structural lattice.

a microchip in an electronic circuit. Such microchips can be found in a great variety of devices from hand-held calculators to ignition systems in automobiles. These tiny copper or other suitable wires must be drawn in a series of wire-drawing operations which begin with the production of 99.99% copper rod (nominally 5/16 in. (cm) diameter in 5500 lb spools) in a rod mill. In this wire-drawing process or series of processes, illustrated schematically in Fig. 1b, a copper wire is drawn through a die to reduce its diameter. This process produces a deformation or displacement in characteristic "units" in the initial rod. To facilitate this deformation and the drawing operation, high temperatures are often used and this differentiates "hot" drawing from "cold" drawing operations. The heat in such operations facilitates the relaxation of atomic-level distortions (or defects) in the copper wire. A simple analogy might be the difference in spreading cold butter and warmed butter on breakfast toast. The warm butter "flows" better under the pressure of the knife. This notion of flow during deformation is a very important issue because if the flow becomes interrupted during drawing (Fig. 1b), cracks may form and the wire could either break during the drawing operation, or cracks present could compromise the wire in Fig. 1a during operation—for example, vibrations in an automobile ignition control system could cause a cracked wire to break.

The flow of copper during processing such as wire drawing in Fig. 1b is a fundamental issue which involves an understanding of the structure of copper and how this structure accommodates deformation. To some extent, fundamental issues of metal structure, like copper, begin with the atomic structure, although the atomic structure itself is not altered during wire drawing. It is the arrangement of the atoms in a crystal structure or structural units composed of atoms, which is altered and it is these atoms or atomic (structural) units which must "flow."

6.2.1.1 Electronic Structure of Atoms, Unit Cells, and Crystals

To examine this process in detail, let us look at the series of schematic diagrams depicting the formation of various atomic structural units in copper (and other metals) in Fig. 2. Here the copper atoms are characterized by 29 electrons (atomic number, Z) in a specific orbital arrangement (atomic shells designated 1, 2, 3, etc. corresponding to electron shells K, L, M, etc, and subshells designated s, p, d, etc.). Electrons occupying subshells have the same energy, while all electrons in each shell have total energies which are ideally proportional to $2n^2$; consequently, the electrons in the K-shell ($n = 1$) are closest to the nucleus and have the highest or ground-state energy, while electrons farthest from the nucleus ($n = 4$ in copper) are less tightly bound and can be more easily knocked off or otherwise donated or shared in chemical reactions which on[y depend upon this electronic structure. It is in fact this electronic structure which is unique to each atom or chemical element and allows the elements to be ranked and grouped in a periodic arrangement or chart which often shows similarities or systematic differences which can be used to understand or predict chemical, physical, and even solid structural behaviors. For example, when the elements are arranged as shown in Fig. 3, those elements in similar rows and columns

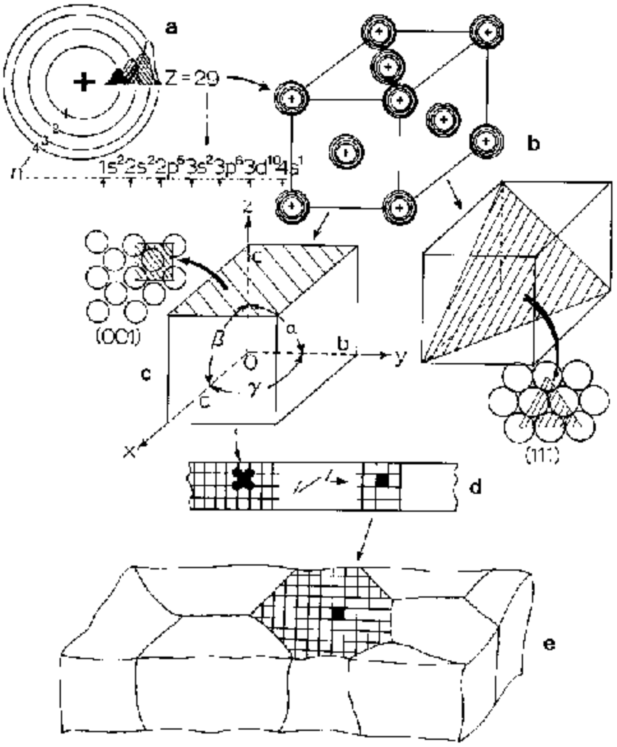
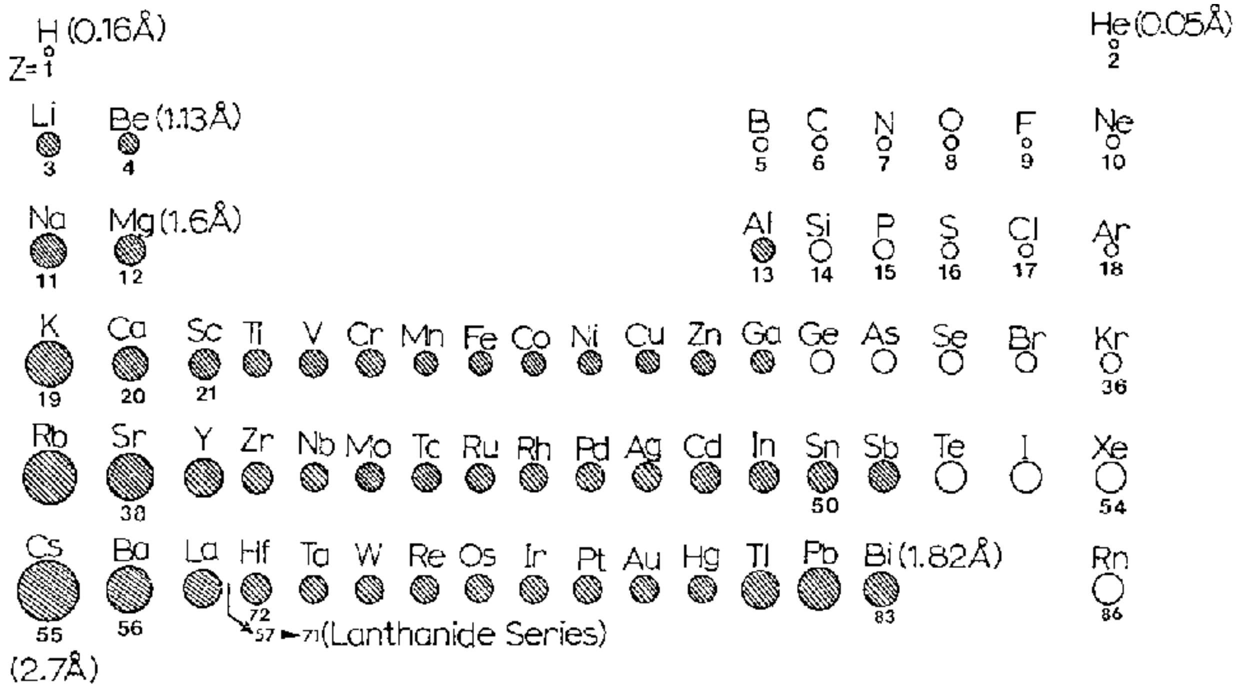


Figure 2 Schematic representation of the evolution of metal (copper) crystal unit cells and polycrystalline grain structures. (a) Copper atom representation. (b) Copper FCC unit cell arrangement showing (1 1 1) crystal plane representation (shaded) and close-packed atomic arrangement. (c) Crystal unit cell conventions showing (0 0 1) plane (shaded) and atomic arrangement. (d) Simple plane view strip along the wire axis in Fig. 1a. (e) Three-dimensional section view through copper wire in Fig. 1a showing crystal grains made up of unit cells.


H (0.16Å)
Z= 1
He (0.05Å)
2
Li
3
Be (1.13Å)
4
B C N O F Ne
5 6 7 8 9 10
Na
11
Mg (1.6Å)
12
Al Si P S Cl Ar
13 14 15 16 17 18
K Ca Sc Ti V Cr Mn Fe Co Ni Cu Zn Ga Ge As Se Br Kr
19 20 21 36
Rb Sr Y Zr Nb Mo Tc Ru Rh Pd Ag Cd In Sn Sb Te I Xe
38 50 54
Cs Ba La Hf Ta W Re Os Ir Pt Au Hg Tl Pb Bi (1.82Å) Rn
55 56 72 83 86
(2.7Å)
57 ⟶ 71 (Lanthanide Series)

ATOMIC NUMBER - ELEMENT SYMBOL - ELECTRONIC STRUCTURE SEQUENCE

1	H	$1s$	34	Se	$[Ar]3d^{10}4s^24p^4$	67	Ho	$[Xe]4f^{11}6s^2$
2	He	$1s^2$	35	Br	$[Ar]3d^{10}4s^24p^5$	68	Er	$[Xe]4f^{12}6s^2$
3	Li	$[He]2s$	36	Kr	$[Ar]3d^{10}4s^24p^6$	69	Tm	$[Xe]4f^{13}6s^2$
4	Be	$[He]2s^2$	37	Rb	$[Kr]5s$	70	Yb	$[Xe]4f^{14}6s^2$
5	B	$[He]2s^22p$	38	Sr	$[Kr]5s^2$	71	Lu	$[Xe]4f^{14}5d6s^2$
6	C	$[He]2s^22p^2$	39	Y	$[Kr]4d5s^2$	72	Hf	$[Xe]4f^{14}5d^26s^2$
7	N	$[He]2s^22p^3$	40	Zr	$[Kr]4d^25s^2$	73	Ta	$[Xe]4f^{14}5d^36s^2$
8	O	$[He]2s^22p^4$	41	Nb	$[Kr]4d^45s$	74	W	$[Xe]4f^{14}5d^46s^2$
9	F	$[He]2s^22p^5$	42	Mo	$[Kr]4d^55s$	75	Re	$[Xe]4f^{14}5d^56s^2$
10	Ne	$[He]2s^22p^6$	43	Tc	$[Kr]4d^55s^2$	76	Os	$[Xe]4f^{14}5d^66s^2$
11	Na	$[Ne]3s$	44	Ru	$[Kr]4d^75s$	77	Ir	$[Xe]4f^{14}5d^76s^2$
12	Mg	$[Ne]3s^2$	45	Rh	$[Kr]4d^85s$	78	Pt	$[Xe]4f^{14}5d^96s$
13	Al	$[Ne]3s^23p$	46	Pd	$[Kr]4d^{10}$	79	Au	$[Xe]4f^{14}5d^{10}6s$
14	Si	$[Ne]3s^23p^2$	47	Ag	$[Kr]4d^{10}5s$	80	Hg	$[Xe]4f^{14}5d^{10}6s^2$
15	P	$[Ne]3s^23p^3$	48	Cd	$[Kr]4d^{10}5s^2$	81	Tl	$[Xe]4f^{14}5d^{10}6s^26p$
16	S	$[Ne]3s^23p^4$	49	In	$[Kr]4d^{10}5s^25p$	82	Pb	$[Xe]4f^{14}5d^{10}6s^26p^2$
17	Cl	$[Ne]3s^23p^5$	50	Sn	$[Kr]4d^{10}5s^25p^2$	83	Bi	$[Xe]4f^{14}5d^{10}6s^26p^3$
18	Ar	$[Ne]3s^23p^6$	51	Sb	$[Kr]4d^{10}5s^25p^3$	84	Po	$[Xe]4f^{14}5d^{10}6s^26p^4$
19	K	$[Ar]4s$	52	Te	$[Kr]4d^{10}5s^25p^4$	85	At	$[Xe]4f^{14}5d^{10}6s^26p^5$
20	Ca	$[Ar]4s^2$	53	I	$[Kr]4d^{10}5s^25p^5$	86	Rn	$[Xe]4f^{14}5d^{10}6s^26p^6$
21	Sc	$[Ar]3d4s^2$	54	Xe	$[Kr]4d^{10}5s^25p^6$	87	Fr	$[Rn]7s$
22	Ti	$[Ar]3d^24s^2$	55	Cs	$[Xe]6s$	88	Ra	$[Rn]7s^2$
23	V	$[Ar]3d^34s^2$	56	Ba	$[Xe]6s^2$	89	Ac	$[Rn]6d7s^2$
24	Cr	$[Ar]3d^54s$	57	La	$[Xe]5d6s^2$	90	Th	$[Rn]6d^27s^2$
25	Mn	$[Ar]3d^54s^2$	58	Ce	$[Xe]4f5d6s^2$	91	Pa	$[Rn]5f^26d7s^2$
26	Fe	$[Ar]3d^64s^2$	59	Pr	$[Xe]4f^36s^2$	92	U	$[Rn]5f^36d7s^2$
27	Co	$[Ar]3d^74s^2$	60	Nd	$[Xe]4f^46s^2$	93	Np	$[Rn]5f^46d7s^2$
28	Ni	$[Ar]3d^84s^2$	61	Pm	$[Xe]4f^56s^2$	94	Pu	$[Rn]5f^67s^2$
29	Cu	$[Ar]3d^{10}4s$	62	Sm	$[Xe]4f^66s^2$	95	Am	$[Rn]5f^77s^2$
30	Zn	$[Ar]3d^{10}4s^2$	63	Eu	$[Xe]4f^76s^2$	96	Cm	$[Rn]5f^76d7s^2$
31	Ga	$[Ar]3d^{10}4s^24p$	64	Gd	$[Xe]4f^75d6s^2$	97	Bk	$[Rn]5f^97s^2$
32	Ge	$[Ar]3d^{10}4s^24p^2$	65	Tb	$[Xe]4f^96s^2$	98	Cf	$[Rn]5f^{10}7s^2$
33	As	$[Ar]3d^{10}4s^24p^3$	66	Dy	$[Xe]4f^{10}6s^2$	99	Es	$[Rn]5f^{11}7s^2$

Figure 3 Partial periodic system of the elements showing the metals (shaded). The elements are ranked sequentially by atomic number Z (or the number of electrons) which are illustrated in an electronic structure notation shown below for $Z = 1$ (hydrogen) to $Z = 99$ (einsteinium). Relative sizes of the atoms of these elements are included. Atomic radii are shown in parentheses (Å).

possess characteristic valences and ionic or atomic sizes. Consequently, this chart becomes a quick guide to predicting specific combinations of, or substitutions of, elements in a structure which may be sensitive to a specific charge compensation (or balance) and size.

Normally, when atoms combine or are bound to form solids, they also first create units or unit cells that have specific geometries (Fig. 2) which can be described in a conventional cartesian co-ordinate system by a series of so-called Bravais lattices. These form distinct and unique *crystal systems* or crystal structural units. There are seven unique systems which are composed of a total of 14 Bravais lattices shown in Fig. 4. These unit cells have atomic dimensions denoted *a*, *b*, and *c* illustrated schematically in Fig. 2, and these unit dimensions and the corresponding co-ordinate angles (α, β, γ) delineate the unit cell geometries, as implicit in Fig. 4. It is not really understood exactly why different metal atoms will arrange themselves in specific solid crystal units, but it has something to do with the electronic configuration and the establishment of some kind of energy minimization within the coordinated, atomic unit. Consequently, metals such as copper, silver, gold, palladium, and iridium normally form a

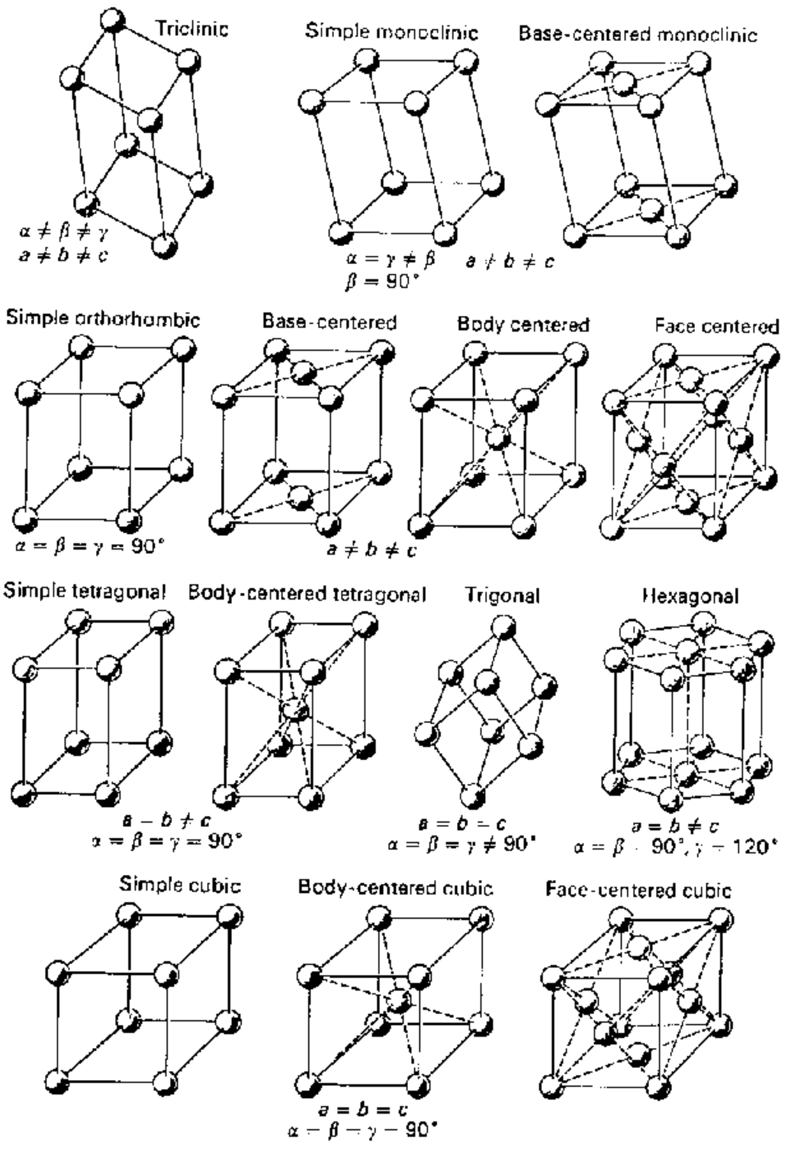


Figure 4 Atomic unit cells characterizing crystal systems composed of a total of 14 Bravais lattices shown.

face-centered cubic (FCC) unit cell, while metals such as iron, tantalum, and tungsten, for example, will be characterized by unit cells having a body-centered cubic (BCC) lattice (or crystal) structure.

When one examines the atomic arrangements in the unit cells as shown for the shaded planes in Fig. 2, these planes also exhibit unique spatial arrangements of the atoms, and are designated by a specific index notation called the *Miller* or *Miller–Bravais* (in the special case of hexagonal cells) *index notation*. This notation is simply derived as the reciprocals of the intercepts of each plane with the corresponding axes (x, y, z in Fig. 2), the reduction to a least common denominator, and the elimination of the denominator; e.g. $1/p$, $1/q$, $1/r$ where p, q, r represent intercepts along x, y, z respectively referenced to the unit cell. Consequently for the plane denoted (1 1 1) in Fig. 2, $p = 1, q = 1, r = 1$, and for the plane denoted (0 0 1), $p = \infty, q = \infty, r = 1$. Correspondingly, if we consider the plane formed when $p = 1, q = 1$, and $r = 1/2$, the plane would be designated (1 1 2) while a plane with $p = 1, q = 1$, and $r = 2$ (extending one unit outside the unit cell) would be designated a (2 2 1) plane which, while illustrated in a geometrical construction outside the unit cell, could just as easily be shown in a parallel construction within the unit cell with $p = 1/2, q = 1/2$, and $r = 1$. This notation can often be a little confusing, especially when the opposite or a parallel plane is denoted. For example, the opposite face of the FCC cell for copper in Fig. 2 [plane opposite (0 0 1) shown shaded] coincides with the zero axis (z) and must therefore be referenced to a unit cell in the opposite direction, namely $p = \infty, q = \infty$, and $r = -1$. We designate such a plane $(0\,0\,\bar{1})$ (the negative index is denoted with a bar above for convenience and convention). You can easily show from your analytical geometry training that for the cubic unit cells, directions perpendicular to specific crystal planes have identical indices. That is, a direction [1 1 1] is perpendicular to the plane (1 1 1). Note that directions are denoted with brackets. We should also note that planes of a form, for example all the faces (or face planes) in the copper unit cell (in Fig. 2) can be designated as {0 0 1} and directions which are perpendicular to these planes (or faces) as ⟨0 0 1⟩.

We have taken the time here to briefly discuss the notations for crystal planes and directions because these are very important notations. It should be apparent, for example, that looking at the atomic arrangements for (0 0 1) and (1 1 1) in Fig. 2 is tantamount to viewing in the corresponding [0 0 1] and [1 1 1] directions. Imagine that you are half the size of the copper atom and walking in one or the other of these directions in a copper crystal. Imagine how much tighter you would fit in the [1 1 1] direction than the [0 0 1] direction. These directional and geometrical features are correspondingly important for many physical and mechanical properties of metals, including copper for analogous reasons. Electrical conductivity, which can be visualized in a very simple way by following a single electron traversing a finite length along some specific crystal direction like ⟨0 0 1⟩ is a case in point which we will describe a little later.

6.2.1.2 Polycrystals, Crystal Defects, and Metal Deformation Phenomena: Introduction to Structure—Property Relationships

Now let us return to Fig. 2 and Fig. 1. We are now ready to examine the "test strip" in Fig. 1 which illustrates a small linear segment along the tiny copper wire. A piece of this strip can be imagined in Fig. 2 to represent first a continuous arrangement of unit cells like a repeating arrangement of crystal blocks. You should note that in the context of our original discussion and illustration in Fig. 1 of the wire drawing process, that like individual atoms, the FCC unit cell for copper is also not distorted or displaced by the deformation. That is, the unit cell dimension for copper ($a = b = c = 3.6\,\text{Å}$) is not altered. What is altered, is the longer-range periodicities of these cells through the creation of *defects* which are characteristic of these atomic alterations. Such alterations can also include the creation of crystal domains, which are illustrated in Fig. 2. Such domains, called crystal grains, can be visualized as irregular polyhedra in three dimensions, and solid metals are normally composed of space-filling collections of such polyhedra which produce polycrystalline solids in contrast to single-crystalline solids. The *interfaces* or common boundaries where these grains or individual crystal polyhedra meet also represent defects in the context of a perfect, long-range (continuous) crystal structure. It is easy to demonstrate in fact how *grain boundaries* as defects can have an important influence on both the strength and the electrical conductivity of the copper wires in Fig. 1.

Figure 5 provides a graphical illustration of a test section through a polycrystalline copper wire and a single-crystal section of wire (which contains no grain boundaries). The grain size or average grain diameter is shown in this section as D in reference to the thickness view in Fig. 2e. Note that the grain boundaries constitute a discontinuity in a particular crystal direction or orientation. If a wire sample as shown in Fig. 1

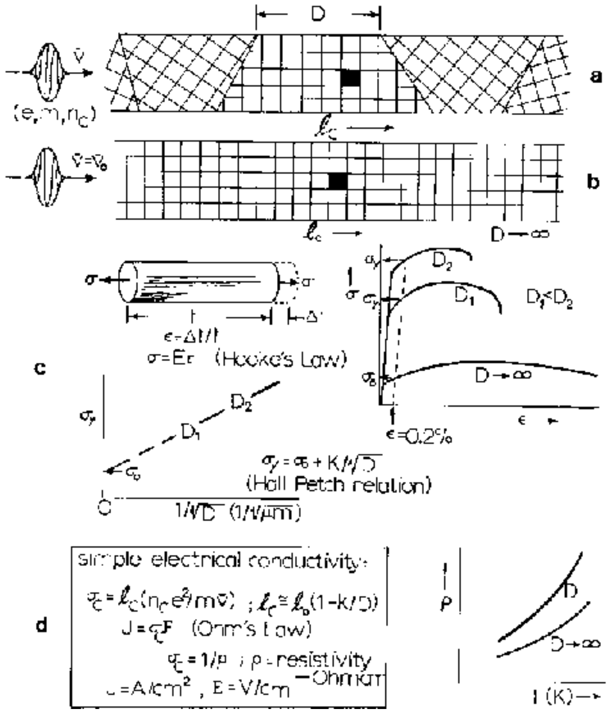


Figure 5 Fundamental features distinguishing polycrystalline (a) and single crystal metal sections (b) and their role in defining simple mechanical (tensile) (c) and electrical (conductivity/resistivity) (d) properties. Note in (c) that ideally as the rod extends it also "necks" or reduces its cross-sectional area. This radial strain (ε') is contrasted with the tensile strain (ε) by *Poisson's ratio*, $\nu = \varepsilon'/\varepsilon$; where $\varepsilon' = \Delta d/d_o$, and d_o is the original rod diameter: $\Delta d = d - d_o$ (a negative value).

were pulled in tension and the stress applied (stress = force/cross-section area) plotted as a function of the elongation or strain (ε in %), a simple engineering stress–strain diagram will result, as shown in Fig. 5. Note that there is a linear, elastic portion of these curves where the stress is proportional to the strain: $\sigma = E\varepsilon$; called *Hooke's law*. Furthermore, if a straight line parallel to this linear elastic portion is extended from a strain of 0.2%, the intercept with the curve denotes an engineering yield point or *yield stress* (σ_y) at which point the metal is no longer deforming elastically, and the deformation becomes plastic and nonrecoverable. The maximum stress attained by these curves is called the ultimate tensile stress (UTS), while σ_F denotes the stress at fracture. Note that the polycrystalline wire has a much higher UTS and σ_F than the single crystalline wire (where $D = \infty$). Note also that the single crystal wire has a much larger strain (or elongation) and therefore is much more *ductile* than the polycrystalline wire. Note also that the plastic regime dominates these curves—there is only a very small elastic or recoverable deformation. To understand this, we need to examine the concept of plastic deformation. This excursion will also help explain the difference between the polycrystalline and single crystal wire behaviors, i.e., the effect of grain size, D illustrated in Fig. 5.

Figure 6 depicts a simple reference square in the single crystal section of Fig 4. This section does not necessarily have any simple geometrical relationship to the crystal lattice geometry depicted which has been related to the FCC unit cell for copper. However, it must be noted that, during plastic deformation, the unit cell is not itself distorted, and the only sensible way this can occur is for a co-ordinated slip between groups of unit cells shown to the right in Fig. 6. That is, the "average" elongation (Δl) experienced by the

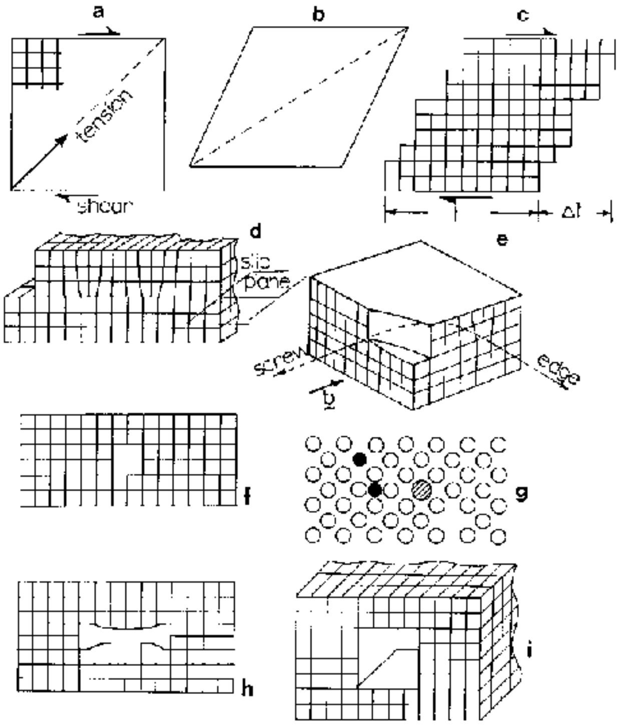


Figure 6 Simple crystal defects. (a) to (d) illustrate the creation of dislocations by systematic shear deformation resolved as slip [(a) to (c)]. The concept of a total dislocation line defect varying from edge to screw is shown in (e). (f) shows a vacancy and (g) shows substitutional impurities on atomic sites and an interstitial defect not on a site. (h) is also a vacancy disc which could appear as a dislocation loop. (h) and (i) show volume defects.

plastically deforming wire is the combined slip of cooperating segments. As it turns out, slip in FCC crystals occurs exclusively along {1 1 1} planes at room temperature. This is because, as illustrated in Fig. 6, these are the close-packed planes (refer also to Fig. 2(b) and the atoms require only a modest "force" resolved in the slip plane to translate the atoms into equilibrium positions. This process can occur without breaking bonds between atoms or without stretching bonds. In fact, unless at least half the unit translation for two reference (1 1 1) planes is achieved, the planes can relax back to their original positions—elastic deformation (Fig. 6). Here the reader must be cautioned about the confusion in the crystal lattice notations of Figs. 2, 5, and 6. For simplicity, we have utilized the FCC (0 0 1) plane as a reference "block" as a convenience, but slip in FCC does not occur in the {0 0 1} planes as shown, but rather in the {1 1 1} planes. This requires some close examination of the geometries and crystal structure (unit cell) notations.

Figure 6a–d illustrates a rather simple scheme for slip to account for strain (elongation) in a crystal deformed in tension. And while we do not allow for a unit cell distortion in response to a stretching and rotation about the tension line (dotted in Fig. 6a and b), or a unit shear, combinations of tension and shear over millions of lattice or unit cell dimensions can rationalize the displacements necessary to accommodate huge plastic deformations, such as those which may characterize a wire drawing operation depicted schematically in Fig 1b.

If the unit displacements imposed on the lattice by slip are not translated through the lattice, these units of deformation accumulate in the lattice as defects characterized by an extra half plane for each such displacement unit on the slip plane (Fig. 6d). These slip-related defects are called *dislocations*, and the unit displacements which characterize them are called a *Burgers vector*, and denoted **b**.

As shown in the solid section view of Fig. 6e, a total dislocation in a metal crystal (or any other crystalline material) can be characterized as *edge* or *screw*. A dislocation is in effect a *line defect* in a crystal, and its *character* (edge, screw, or a mixture) is determined by the angle (α) between its Burgers vector (**b**) and the dislocation line vector (ξ). Note in Fig. 6e that for a pure edge dislocation in the {0 0 1} face the Burgers vector, **b**, is perpendicular to the dislocation line ($n = 90^\circ$), while for the emergent screw dislocation line to the left it is parallel ($\alpha = 0^\circ$). In the FCC structure, as we noted, the slip plane is {1 1 1} and the direction of the Burgers vector is $\langle 1\,1\,0 \rangle$. Note that the screw dislocation emerging on the surface creates a *spiral* step or unit ramp. Atoms added systematically to this step can allow the face to grow by a spiral growth mechanism. Finally, note that the total dislocation in Fig. 6e can be created by imagining the solid section to consist of a block of Jell-O which is sliced by a knife from face to face and the (1 0 0) face displaced by a unit amount (/**b**/). The character of the line creating a demarcation for this cut and displacement is the dislocation line whose character changes from edge to screw ($\alpha = 90^\circ$ to 0°) in a continuous way. At the midpoint of this quarter circle the character is said to be *mixed* ($\alpha = 45^\circ$).

Having now defined grain boundaries or crystal interfaces in a polycrystalline metal as defects (actually these boundaries are often referred to as *planar defects* in a crystalline (or polycrystalline) solid, as well as *dislocation line defects*, it is probably instructive to illustrate a few other common crystal defects as shown in Fig. 6f–i. Utilizing the simple unit cell schematic, Fig. 6f shows a missing atom in a crystal lattice. Such missing atoms are called vacancies. In contrast to grain boundary planar defects and dislocation line defects, a vacancy is ideally a *point defect*. Other point defects can be illustrated by *substitutional impurities* which sit on the lattice (atom) sites, or *interstitial atoms* which can be the same lattice atoms displaced to nonlattice sites, or other impurity atoms which occupy nonlattice sites (Fig. 6g). In Fig. 6h a series of vacancies form a vacancy disc which, if large enough, will allow the lattice to essentially collapse upon this disc. This creates a loop whose perimeter is a dislocation: a *dislocation loop*. This points up an interesting property of a dislocation line. It can end on itself (as in forming a loop), on a "surface," such as an internal grain boundary, or a free surface, or on another dislocation line. Dislocations never simply end in a crystal.

Finally, in Fig. 6i, we illustrate a void—*a volume defect*—in a crystal portion formed by creating a large number of vacancies. Such defects can be formed by neutron damage in a nuclear reactor where atoms are displaced creating vacancies which can *diffuse* or migrate through the lattice to create an aggregate which can grow. This process is facilitated by high temperature and when helium is created as an interstitial by-product of the nuclear reaction, it can also migrate by diffusion to these void aggregates to form crystallographic "bubbles." This process causes the original metal in the reactor environment to swell to accommodate these bubbles, and this nuclear reactor swelling (by more than 6%) can pose unique engineering problems in reactor design and construction.

Figure 7 illustrates some of the crystal defects discussed above and illustrated schematically in Fig. 6. Note especially the dislocations emanating from the grain boundary in Fig. 7b which can be reconciled with dislocations formed by the creation of a surface step in Fig. 6d. This is a simple example of the properties of grain boundaries, which can also block dislocation slip and create *pile-ups* against the grain boundary barrier. It is these features of dislocation generation and barriers to slip which account for the effects of

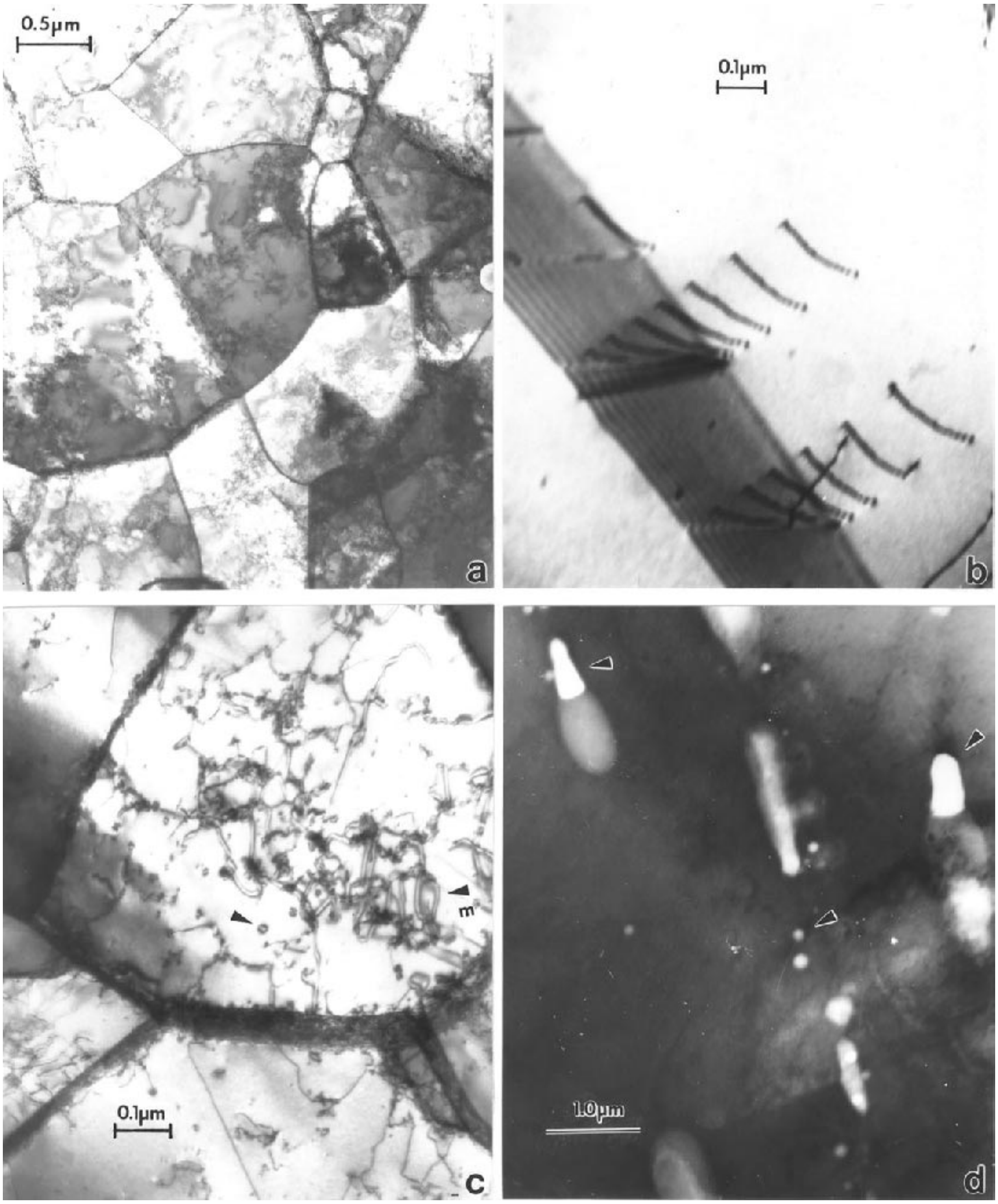


Figure 7 Examples of crystal defects in a number of metals and alloys observed in thin-film specimens in the transmission electron microscope. (a) Grain boundaries in BCC tantalum. (b) Dislocation emission profiles from a grain boundary in FCC stainless steel (type 304). (c) Dislocation loops in tantalum (arrows). (d) Voids (arrows) in solidified copper rod, some associated with small copper inclusions which are introduced during the solidification process in the melt.

grain boundaries and grains of varying size, D, on the mechanical properties of metals.

If we now return to Fig. 5, but continue to hold Fig. 6 in reference, the differences in the stress–strain curves for single-crystal and polycrystalline copper should become apparent. During tensile straining, slip is activated at some requisite stress. Processes similar to those illustrated in Fig. 6d occur, and many dislocations are created during straining which are not translated through the crystal but begin to interact with one another, creating obstacles to slip and eventual thinning or *necking* of the crystal, and fracture. For polycrystalline copper, slip is quickly blocked by the grain boundaries which rapidly produce a host of dislocations which create additional barriers to slip, causing necking, and failure. As the grain size, D, is reduced, the stress necessary to achieve deformation rises and the elongation will often correspondingly decrease. That is, as the strength (UTS) increases, the ductility decreases, and the ability to form or draw copper (its *malleability*) is also reduced.

We might argue that other detects such as vacancies, interstitials, and substitutional elements will also influence the mechanical (stress–strain) properties of metals by creating barriers to the motion (slip) of dislocations. In fact it is the variance to slip or dislocation motion created by substitutional elements which accounts in part for the variations in a metal's behavior (such as strength, ductility, and malleability) when it is alloyed (or mixed) with another metal. We will discuss these features in more detail later.

Now let us finally consider electrical conductivity in the depictions of Fig. 5. Consider that ℓ_c denotes a mean free path (average) for all the conduction electrons, n_c, which move at some mean velocity, $\bar{v}$. If n_c, e (the charge), m (the mass), and $\bar{v}$ are unchanged, only ℓ_c will significantly influence conductivity. For a single crystal, this mean free path would depend upon the crystal structure and the direction of conduction of the crystal orientation of the wire if it were a single crystal. That is, the kinds of atoms and their organization relative to electrons moving as mass points or particles would cause collisions creating impedance to electron flow. However, as soon as crystal defects were added to the wire structure, the mean free path (propensity for electron collisions) would be altered, invariably shortened; thereby reducing the conductivity. Obviously some defects would have more dramatic influences than others. Grain boundaries would be expected to have a major effect, but including other crystal defects in a polycrystalline metal would compound the reduction of conductivity. And while reducing the grain size might strengthen a wire, it might also correspondingly reduce the conductivity.

Temperature would also have a very important effect on metal conductivity because increasing temperature would create significant vibrations of lattice atoms. In fact at very high temperatures the effects could become so large that atoms would be dislodged from their normal lattice sites, creating a vacancy upon leaving the site and an interstitial upon squeezing into a nonlattice site. Not only would the vibrating lattice atoms provide impedance to electron flow by creating additional collisions, but the "defects" created would have a similar effect. Consequently, resistivity ($\rho = 1/\sigma_c$) is usually linearly related to temperature in metals, decreasing with decreasing temperature. Crystal defects will alter the slope or the shape of this trend. When a metal becomes a superconductor, the resistivity abruptly drops to zero. When this occurs, electrons no longer behave as single mass points colliding with the lattice, and defects, rather than impeding the process, contribute by promoting increased current density, J.

Utilizing a few simple schematics, we have examined some fundamental issues which have illustrated some equally simple mechanisms to explain physical and mechanical properties of metals like copper. But these illustrations only begin to provide a metallurgical basis for such behavior under specific conditions. For example, the drawing of very thin copper wires from bar stock or cast rod involves more than simple tensile deformation. On forcing the metal through the die in Fig. 1b, the stresses are not simple uniaxial (along the wire direction), but involve a more complex stress system which simultaneously reduces the wire cross-section. Consequently, understanding practical metal-forming operations such as wire drawing involves far more than a simple tensile test. Correspondingly, tensile data, such as that reproduced in Fig. 8 for a number of metals and alloys, must be carefully evaluated and cautiously applied in any particular engineering design or manufacturing process. This is because many practical engineering considerations will involve multiaxial stress or strain, at elevated temperatures, and at strain rates which are far different from the conditions used in developing the diagrams shown in Fig. 8 where tensile samples were strained at a rate of 10^{-3}s^{-1}. Also, the microstructure conditions are not known or unrecorded for the particular materials, and the effects of simple defects such as grain boundaries (and grain size, D) are therefore unknown. Furthermore, as shown in Fig. 9 the production of dislocations and their ability

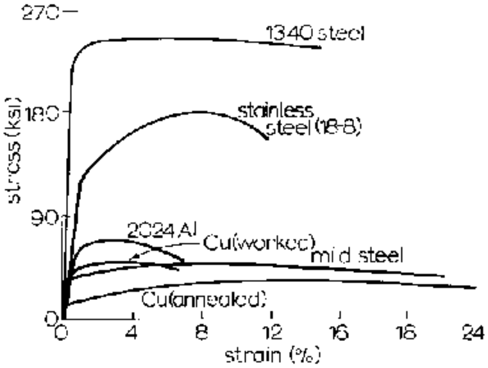


Figure 8 Some comparative tensile stress-strain diagrams for a number of metals and alloys at room temperature (~ 20°C). Crystal structures are noted in parentheses. (After data in Ref. 1.)

to slip on specific crystal planes and crystal directions is different for different crystal structures, and often significantly altered by alloying (adding other metals). This can account in part for differences in metals and alloys of the same crystal structure (such as Cu, Ni, 70Cu–30Zn, and stainless steel in Fig. 8), and different structures, such as FCC Cu, BCC mild steel in Fig. 8. In addition, prior deformation (or prior working) plays an important role in the shape of a stress–strain diagram and in the utilization of metals and alloys in any commercial or industrial process.

6.2.1.3 Metalworking and Microstructure

To examine these issues in more depth, let us consider the major metalworking processes: wire drawing, rolling and shaping, upset (compression) forging, and sheet-forming operations—die forming, deep drawing, and punching. These are illustrated in simple sketches reproduced in Fig. 10. None of these processes involves only a uniaxial (tensile) strain state, and many such operations are not performed at room temperature or at very low strain rates. For example, many wire-drawing operations correspond to drawing strain rates of 10^4 sec^{-1} in contrast to the straining of a tensile sample in Fig. 8 which is nominally only 10^{-3} sec^{-1}—a factor of 10 million! To examine these features let us look at the concept of stress or strain state illustrated in the schematic views of Fig. 11. Figure 11 also defines the strain notation elements which ideally illustrate the metalworking fundamentals for each process in Fig. 10. In addition to the stress or strain components, Fig. 11 illustrates the concept of stress or strain tensors as a notational scheme for directional properties in crystals. In fact, essentially all crystal properties are directional and can be treated as tensors. Figure 11 also shows other deformation modes which, while not necessarily characteristic of metal processing, represent some practical performance issues: twisting or *torsion, fatigue* due to cyclic stress, and very slow, cumulative straining, or *creep*.

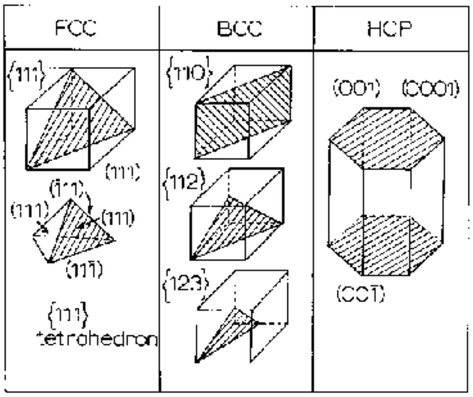


Figure 9 Slip systems in metal crystals.

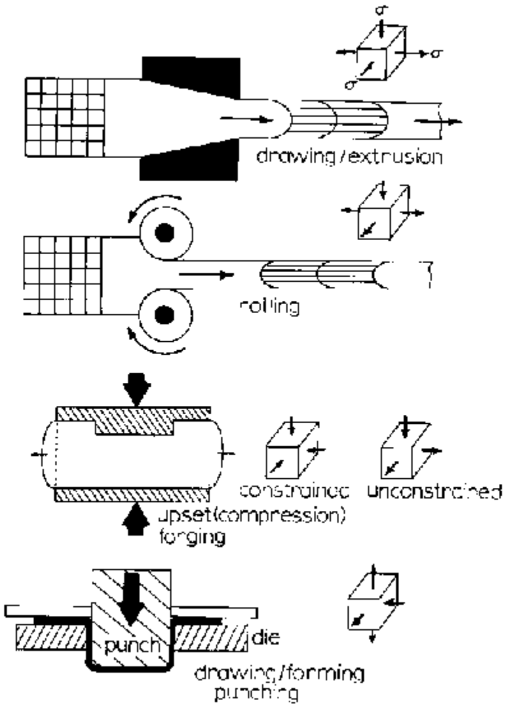


Figure 10 Principal metal forming processes involving various states of strain or stress (σ in cube notations).

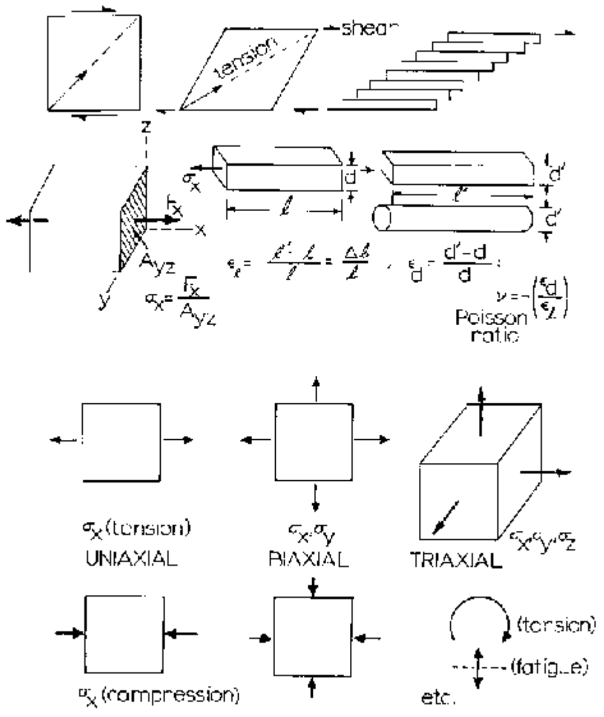

shear
tension
z
x
y
σ_x
F_x
A_{yz}
$\sigma_x = \frac{F_x}{A_{yz}}$
ℓ
ℓ'
d
d'
$\epsilon_\ell = \frac{\ell' - \ell}{\ell} = \frac{\Delta\ell}{\ell}$
$\epsilon_d = \frac{d' - d}{d}$
$\nu = -\left(\frac{\epsilon_d}{\epsilon_\ell}\right)$
Poisson ratio
σ_x (tension)
UNIAXIAL
σ_x, σ_y
BIAXIAL
TRIAXIAL
$\sigma_x, \sigma_y, \sigma_z$
σ_x (compression)
(torsion)
(fatigue)
etc.

Figure 11 Illustrations of stress, strain, and strain (or stress) states as modes of deformation.

While strain, ε, and its mode of imposition during the processing and service of metals is illustrated in simple schematics in Fig. 11, these can all correspond to different temperatures or rates of straining ($\dot{\varepsilon} = d\varepsilon/dt$). Consequently, all these processes are multifunctional and parametric interactions implicit in subjecting a metal to different stress, (σ), strain (ε), strain rate ($\dot{\varepsilon}$), and temperature, (T) can be described in a mechanical equation of state:

$$d\sigma = \left(\frac{d\sigma}{d\varepsilon}\right)d\varepsilon + \left(\frac{d\sigma}{d\dot{\varepsilon}}\right)d\dot{\varepsilon} + \left(\frac{d\sigma}{dT}\right)dT \qquad (1)$$

or in a functional system of strain state equations:

$$\begin{array}{lll} [\sigma = f(\varepsilon, \dot{\varepsilon}, T)]_{\mathrm{I}} & \text{(uniaxial)} & \\ [\sigma = f(\varepsilon, \dot{\varepsilon}, T)]_{\mathrm{II}} & \text{(biaxial)} & (2) \\ [\sigma = f(\varepsilon, \dot{\varepsilon}, T)]_{\mathrm{III}} & \text{(triaxial or multiaxial)} & \end{array}$$

Such equations can be written as so-called constitutive relationships which can be iterated on a computer for a range of process and/or performance data to model a manufacturing (forming) process. At low temperatures (near room temperature) forming processes are very dependent upon microstructures (such as grain structure and size, dislocations, and other defect features) and their alteration (or evolution) with deformation processing. At high temperatures, these microstructures are altered significantly by annihilation, recovery, or growth processes. Correspondingly, the performance data is changed accordingly.

Figure 12 and 13 will serve to illustrate some of these microstructural features and act as a basis to understand other elements of the mechanical equation of state. For example, Fig. 12 not only illustrates the concept of thermal recovery (or grain growth), but also recrystallization, which is depicted by reverse arrows indicating a grain size reduction or recrystallization to the right or grain growth to the left. Actually grain growth can follow recrystallization as well. Recrystallization will occur when grains are "deformed" significantly creating a high degree of internal (stored) energy which drives nucleation and growth of new grains. This is a kind of grain refinement, and can either occur by heavy deformation (at large strains) such as in wire drawing followed by annealing at elevated temperature (called *static recrystallization*) or by *dynamic recrystallization*, where the process occurs essentially simultaneously with the deformation. Such deformation is said to be *adiabatic* because it creates local temperature increases which induce recrystallization. Such adiabatic deformation requires high strain at high strain rate:

$$\Delta T = K\varepsilon\dot{\varepsilon} \qquad (3)$$

where K is a constant. Often extremely high straining creates regions of intense local shearing where high dislocation densities nucleate very small grains which are themselves deformed. These regions are called *adiabatic shear bands.*

Figure 12 also illustrates some rather obvious microstructural differences between BCC tantalum and FCC copper and stainless steel. In addition, the *annealing twins* in copper, which contribute to its characteristic straight and parallel boundaries, are also characteristic of the FCC stainless steel and especially apparent for the larger-grain stainless steel. These twin boundaries are special boundaries having low energy in contrast to the regular grain boundaries (20 mJ/m^2, compared to about 800 mJ/m^2 in stainless steel) and are coincident with the {1 1 1} close-packed slip planes in FCC metals and alloys. These features, of course, also contribute to fundamental differences in the mechanical properties of BCC and FCC metals and alloys (see Fig. 8). Finally, with regard to the microstructural features revealed in Fig. 12, we should point out that the surface etching to reveal these microstructures constitutes a very old and powerful technique

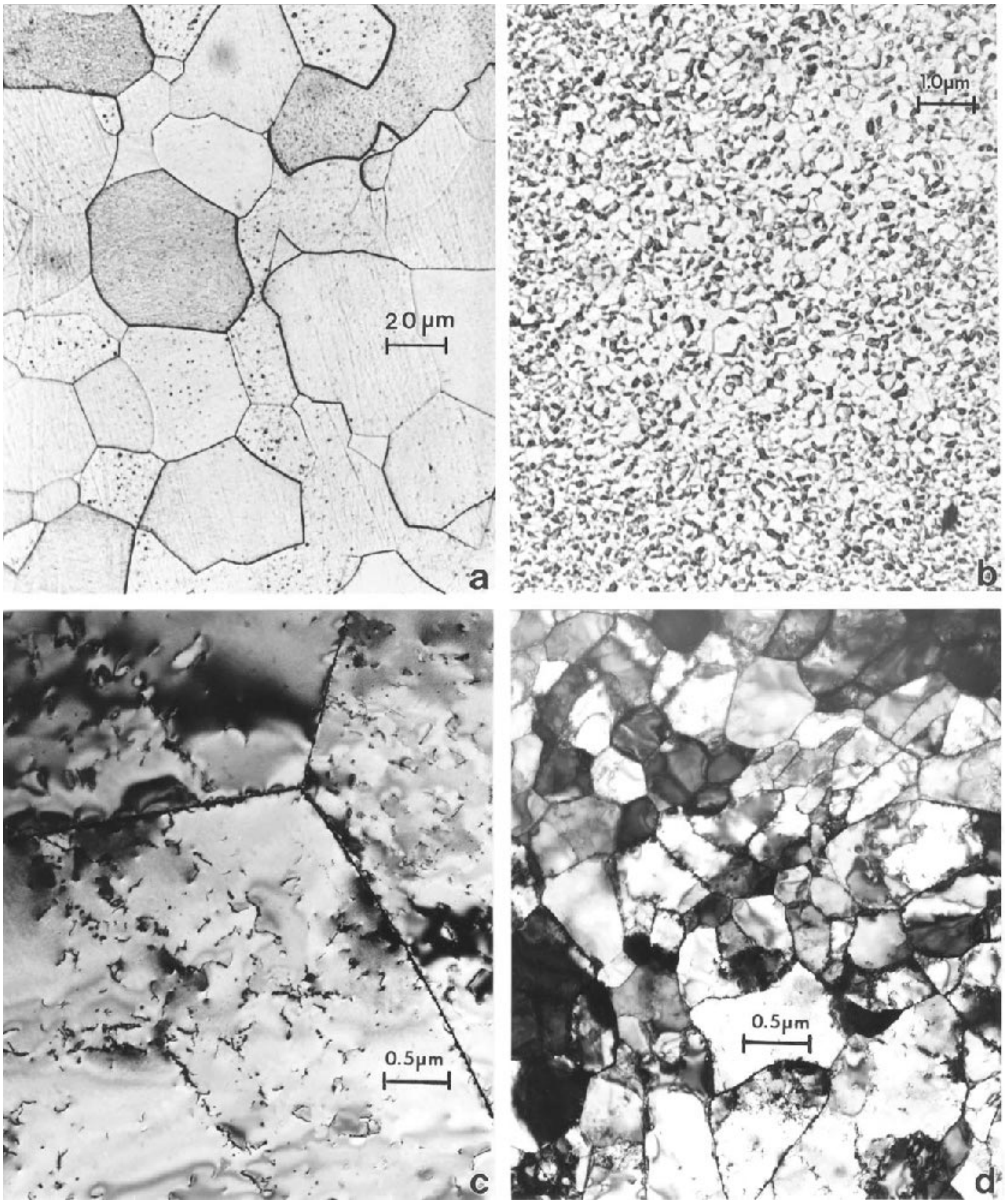


Figure 12 Grain size changes illustrating recrystallization (right) or grain growth (left). (a) and (b) show optical micrographs of tantalum, while (c) and (d) show transmission electron micrographs corresponding to grain structures in (a) and (b) respectively.

called light metallography. Selected etchants create surface relief which can reveal grain boundary and related microstructure, including different phases within a metal.

In contrast to Fig. 12, Fig. 13 shows various dislocation microstructures within individual grains in response to various modes of deformation, including strain. The creation of high densities of such dislocations creates a subgrain cell-like microstructure which can also be regarded as a kind of grain refinement which can strengthen or harden the original, undeformed or unprocessed metal:

$$\sigma = \sigma_0 + KD^{-1/2} + K'd^{-1} \tag{4}$$

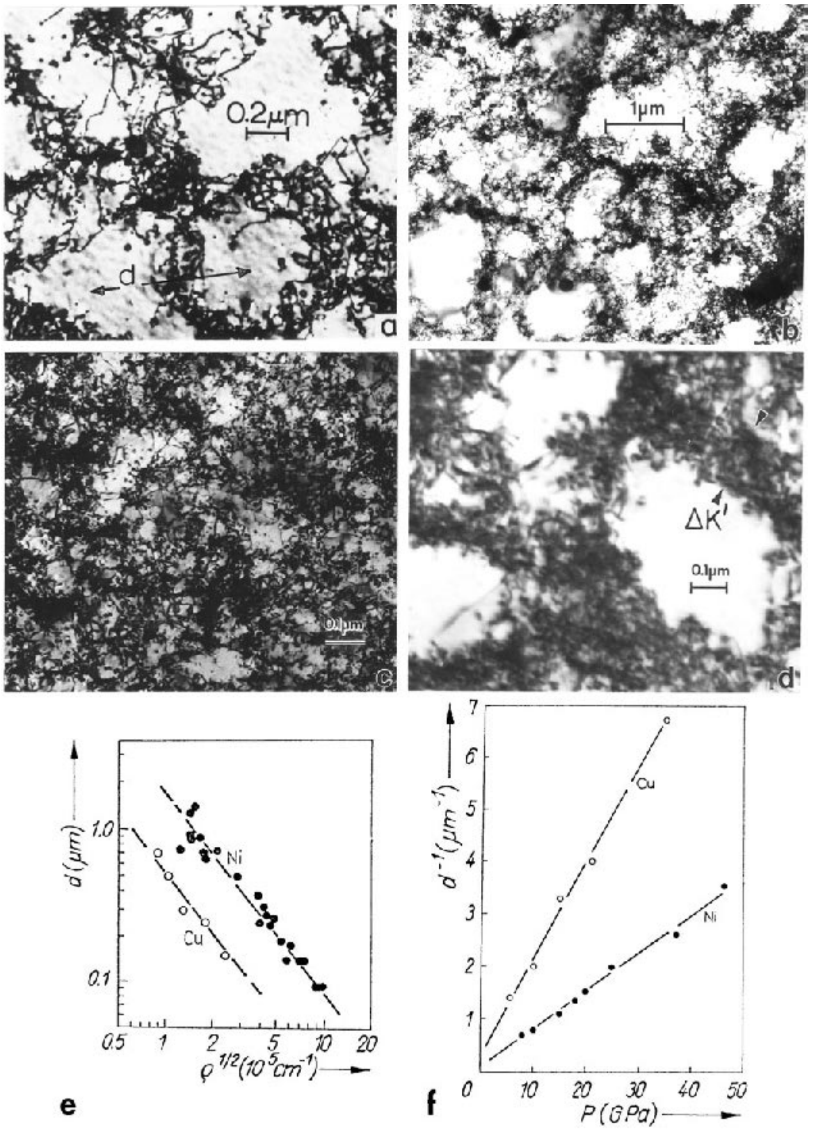


Figure 13 Dislocation cell structure in copper and nickel influenced by stress and strain. (a) Copper shock loaded at a peak pressure of 15 GPa (plane stress). (b) Cast copper rod (multiaxial stress) (c) Copper shock loaded at 30 GPa for comparison with (a). (d) Copper shocked repeatedly at 15 GPa. Note the wide (and thick) dislocation cell walls ($\Delta K'$) in contrast to (a). (e) Dislocation cell size d versus dislocation density ρ for drawn and shocked Cu and Ni. (f) Reciprocal dislocation cell size versus plane-wave shock pressure (stress) for Cu and Ni. [Data in (e) and (f) after Ref. 2.]

where d is the *mean cell size* illustrated in Fig. 13. This equation is a simple modification of the Hall–Petch equation shown in Fig. 5. The constant, K', in Eq. (4) can vary in response to the thickness of the *cell wall*, denoted $\Delta K'$ in Fig. 13. The formation of such *dislocation cells* in deformed metals and alloys is a consequence of multiple slip (or slip on more than one plane or parallel system of planes such as on all four of the {1 1 1} planes shown as a Thompson tetrahedron in Fig. 9) and the formation of a kind of equilibrium arrangement both within the cell wall, and in establishing the cell itself. These arrangements depend upon the attraction and repulsion of dislocations and minimizing their total internal energy. Consequently, increasing the stress or strain increases the dislocation density and this reduces the size of the dislocation cells, d. In fact, in Eq. (4), $K'/d = K''\sqrt{\rho}$ where K'' is a new constant and ρ is the dislocation density. An increase in temperature during deformation will cause the cell wall thickness to decline and such cells, when sufficiently small, and at sufficiently high temperature, can become new grains which can grow. Figure 13e and f illustrate this variation of dislocation cell size with shock pressure (stress) and strain (ε) in wire drawing for both copper and nickel. This data shows a common feature of microstructure evolution for several stress/strain states as well as stress or strain. In this regard it is worth noting that Fig. 13a and c illustrate the notion of *microstructure evolution* for copper shock loaded at different stress levels or copper rod drawn at two different strain values. The decrease of dislocation cell sizes with increasing stress or strain for different strain or stress states (and even strain rates) is often referred to as the principle of similitude. This kind of microstructure evolution also applies to the tensile stress-strain diagrams in Fig. 8. However, not all metals form dislocation cells, which are easily formed when dislocations can slip or cross-slip from one slip plane to another or where, as in BCC, multiple slip systems can be activated (Fig. 9).

The ability of dislocations to cross-slip in FCC metals such as copper, and other metals and alloys, depends upon the intrinsic energetics of dislocations as line defects. For example, we mentioned earlier that grain boundaries differ from annealing twin boundaries in FCC metals like copper and alloys such as stainless steel (see Fig. 12) because their specific surface (interface) energies are very different. Dislocations have been shown to have line energies which are proportional to the Burgers vector squared:

$$\varepsilon\ (\text{dislocation line}) = aG(b)^2 \tag{5}$$

where a is a constant which will vary for an edge or screw dislocation (Fig. 6f), G is the shear modulus equal to $E/2\ (1-\nu)$ in Fig. 5 (where ν is Poisson's ratio), and b is the scalar value of the Burgers vector. As it turns out, this dislocation energy in FCC metals can be lowered by the dislocation splitting into two *partial dislocations* whose separation, δ, will be large if the energy required to separate the partials is small. In FCC, this is tantamount to forming a planar region (planar interface) lying in the {1 1 1} plane (the slip plane), which creates a fault in the stacking sequence of the {1 1 1} planes called a *stacking fault*. These stacking faults therefore are characterized by a specific interfacial free energy of formation of unit area of faulted interface. Consequently, in metals with a high stacking fault free energy, total dislocations do not dissociate widely, while in very low stacking fault energy metals, the dislocations form very extended partials separated by long stacking faults. When this wide dissociation occurs during deformation, cross-slip is reduced and the microstructure is characterized by planar arrays of dislocations or long stacking faults. Stacking faults are another type of crystal defect which can refine a particular grain size just like dislocation cells. Of course in FCC, since these faults are coincident with {1 1 1} planes, and annealing twins are also coincident with {1 1 1} planes, there is a simple, functional relationship between twin boundaries and stacking faults. This relationship is implicit in Fig. 14 which not only illustrates the simple, periodic geometry for {111} plane stacking and stacking fault/twin boundary relationships, but also depicts the simplest (schematic) features of a total dislocation dissociation (splitting) into two partial dislocations. A few exam-

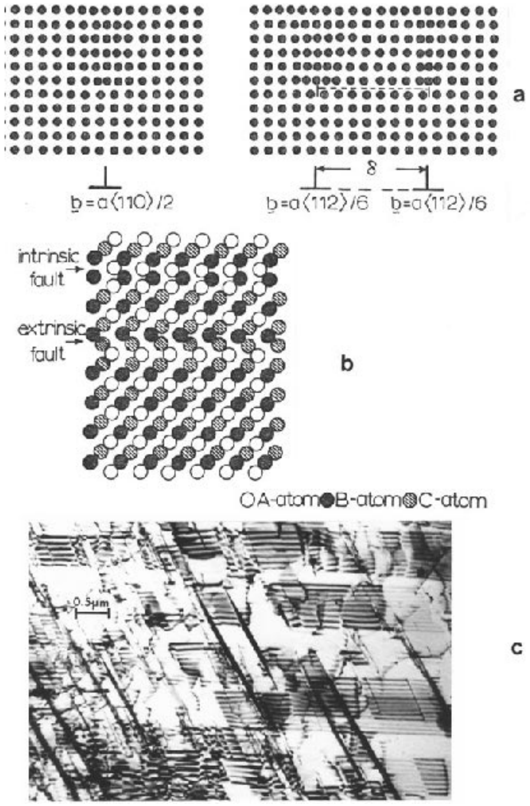
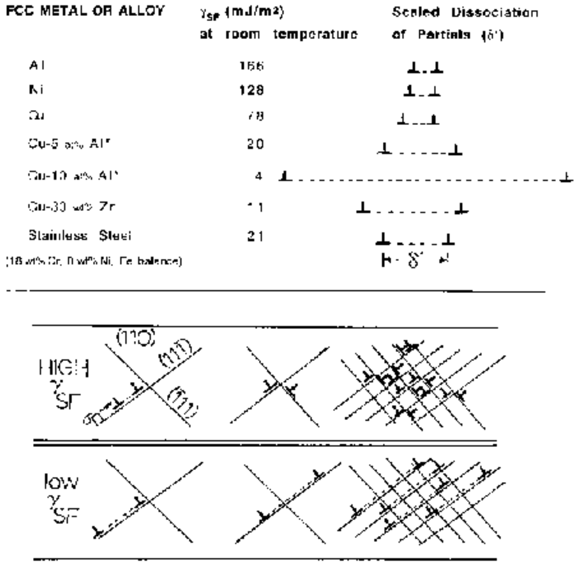


Figure 14 Schematic diagrams depicting the creation of partial dislocations and stacking faults in metals and alloys. (a) Total dislocation splitting into two partial dislocations separated by a region of stacking fault (δ). (b) Periodic atomic stacking model for FCC showing intrinsic and extrinsic stacking faults. (c) Transmission electron microscope image of stacking faults, in 304 stainless steel strained 6% in tension. Many stacking faults overlap.

ples of extended and overlapping stacking faults commonly observed in deformed stainless steel are also shown in Fig. 14c.

Stainless steel (18% Cr, 8% Ni, balance Fe) provides a good example for stacking faults in Fig. 14c because it has a very low stacking-fault free energy (γ_{SF}) in contrast to copper. In fact FCC alloys invariably have lower stacking fault free energies than pure metals, and this feature is illustrated in Fig. 15 which also illustrates a very important aspect of adding or mixing metals to form alloys. Even dilute alloys such as Cu–Al where only small amounts of aluminum are added to copper (the aluminum substituting for copper atoms) can produce a wide range of stacking-fault free energies and, correspondingly, a wide range of mechanical behaviors and a correspondingly wide range of characteristic microstructures and microstructure evolution. A simple analogy for the effects of stacking-fault free energy on deformation behavior of metals and alloys can be seen from the cross-slip diagram in Fig. 15. Extended stacking faults can be imagined to resemble a hook-and-ladder truck in contrast to a small sports car for higher stacking-fault energy which, when driven by a resolved stress (σ_n) in the primary slip plane during deformation, can more easily negotiate cross-slip. This phenomenon, when considered in the context of the mechanical equation of state [Eq. (1)], can reasonably account for the most prominent aspects of the mechanical, as well as the physical properties of crystalline or polycrystalline metals and alloys.



FCC METAL OR ALLOY	γ_{SF} (mJ/m²) at room temperature	Scaled Dissociation of Partials (δ')
Al	166	⊥.⊥
Ni	128	⊥_⊥
Cu	78	⊥__⊥
Cu-5 at% Al*	20	⊥-----⊥
Cu-10 at% Al*	4	⊥--------------------------⊥
Cu-30 wt% Zn	~1	⊥--------⊥
Stainless Steel (18 wt% Cr, 8 wt% Ni, Fe balance)	21	⊥----⊥ ├ δ' ┤

Figure 15 The concept of stacking-fault free energy (γ_{SF}) and its influence on dislocation dissociation and cross-slip in FCC metals and alloys. $\gamma_{SF} \propto 1/\delta$, where δ' is the separation of partials. For high stacking-fault free energy the partials are very close and cross-slip easily to form "cells." For low stacking-fault free energy the partials split widely and there is little or no cross-slip resulting in planar (straight) arrays of dislocations. *Aluminum atoms substitute for copper on a corresponding basis. (Data from Ref. 3.)

6.2.2 Alloying Effects and Phase Equilibria

Industrial metals are more often mixtures of elements—or alloys—than pure elements such as copper, silver, gold, nickel, iron, zinc, etc. Exceptions of course involve metals such as copper, as discussed previously, which is required in a pure form for a wide range of commercial wire applications. Recycled metals, which now constitute a huge portion of the industrial metals market, are also largely elemental mixtures. Combinations of elements create a wide range of new or improved properties or performance qualities, and even small elemental additions to a host metal can create a wide range of properties and performance features. Examples which come readily to mind include additions of carbon to iron to create a wide range of steels, and small additions of copper, magnesium, and/or silicon to aluminum to create a wide range of strong, lightweight alloys. In order to form an alloy it is necessary that there is some solubility of one element in another to create a *solid solution*. Some combinations of two elements forming binary alloys are completely soluble over their entire range of mixing (0–100%) forming complete solid solutions; for example, copper and nickel or copper and gold. Such solid solutions can be either *ordered* or *disordered* depending upon where the atoms are located in the unit cells. Figure 16 illustrates this concept as well as the profound effect order or disorder can have on properties such as electrical resistance.

Solid solutions such as those represented in Fig. 16 are referred to as *substitutional solid solutions* because the two atoms substitute freely for one another on the lattice sites. Of course in these ideal forms of alloying, as in many other circumstances, the relative sizes of the constituent atoms are important. In addition, the electronic structures of the elements are important in determining compound-forming tendencies as well as the actual structural features as illustrated in specific unit cell arrangements (e.g., Figs 3 and 4).

Phase equilibria (equilibrium states) as phase relations describing the mixing of two or more elements (or components which behave as single constituents) are

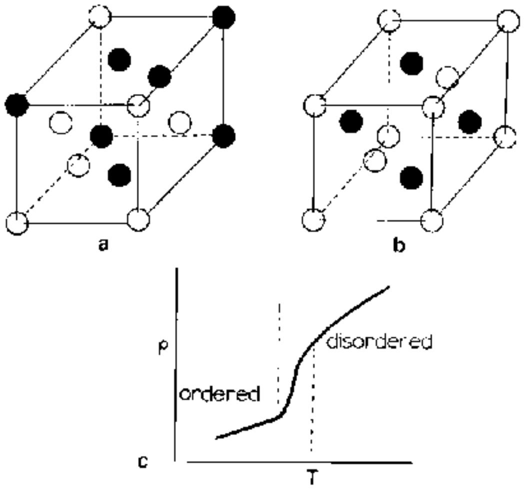


Figure 16 Order–Disorder phenomena in a simple binary, solid–solution alloy. (a) Face-centered cubic, disordered 50 Cu: 50 Au. (b) Face-centered cubic, ordered 50 Cu: 50 Au. (c) Resistivity versus temperature for ordered and disordered 50 Cu: 50 Au solid–solution alloy.

controlled by equilibrium thermodynamics, and can be displayed in a variety of *phase diagrams* as illustrated in Fig. 17 for mixtures of two elements: *binary phase diagrams*. Equilibrium thermodynamics implies variations of temperature (T), pressure (P), and corresponding volumes (V); or concentrations (X) of components in the system expressed as either weight percent or atomic percent. In Fig. 17 variations of temperature and concentration are shown at constant (or standard) pressure (760 mm Hg). Equilibrium boundaries as phase boundaries separate regions of liquid (L), solid (S) and solid/liquid mixtures (L + S). For binary combinations where the metals are miscible in the liquid state and immiscible in the solid state, a *eutectic* forms, as illustrated in Fig. 17b for complete immiscibility and in Fig. 17c for partial solubility. Compounds also form as a consequence of the differences in valences. The more electropositive the one element and the more electronegative the other the greater is the tendency toward compound formation. Simple systems may be joined by compounds, as shown in Fig. 17d. Since, as shown in the periodic table in Fig. 3, elements in the upper left portion of the table are the most electropositive in contrast to those electronegative elements in the upper right, the relative behavior of phase diagrams can often be estimated when taking into account the relative sizes of alloying elements. Generally, alloys of elements having the same valence, crystal structure, and size ratio tend to form continuous solid solutions. A metal (or element) of lower valence tends to dissolve one of higher valence. Alloys of elements having the same valence but different sizes have limited solubilities and usually do not form compounds while elements having differences in valence form stable compounds (Fig. 17c–e).

Phase equilibria is an important issue in manufacturing processes which involve bonding to thin layers, etc., and solder connections. Thin layers and solders in bonded automotive parts are often prone to diffusion at elevated engine temperatures which can create new phases susceptible to cracking and fracture.

There are, of course, more complex binary phase phenomena as shown for example in Fig. 17c which involves different crystal structures, solubilities, valences, and ionic/atomic sizes. In addition, phase equilibria involving three (ternary phase diagrams), four (quaternary phase diagrams) or more elements are also common among more advanced industrial materials or materials systems. There are also nonequilibrium materials systems, and *mechanical alloying* is a contemporary example. In this process, partially soluble or insoluble (immiscible) components can be mixed or milled together in a high-speed attritor to produce mechanically bonded, fine-powder mixtures which can be *sintered* into fully dense, bulk systems at high temperature and pressure. Figure 18 shows an example of such a system involving tungsten and hafnium carbide compared with an equilibrium aluminum alloy forming compounds as complex precipitates of (Fe, Cu, Al) Si.

6.2.2.1 Precipitation and Phase Transformations

Figure 18b illustrates the formation of second-phase precipitates in a dilute aluminum alloy (2024), and these features are also implicit in the phase diagrams shown in Fig. 17d and e in particular. To some extent, precipitation and phase transformation exhibit fundamental similarities: both can have chemical or compositional as well as crystallographic differences in contrast to the matrix in which they form. Both may involve *diffusion* of one or more elements through the matrix to the site of *nucleation*, or they may involve the mass transport of atoms into new coincidences by shear (stresses) or other invariant straining phenomena. These are generally referred to as diffusionless transformations. The formation of a critical nucleus for precipitation or transformation is often illustrated ideally as shown in Fig. 19, which involves the formation of a new (spherical) phase within a matrix as a

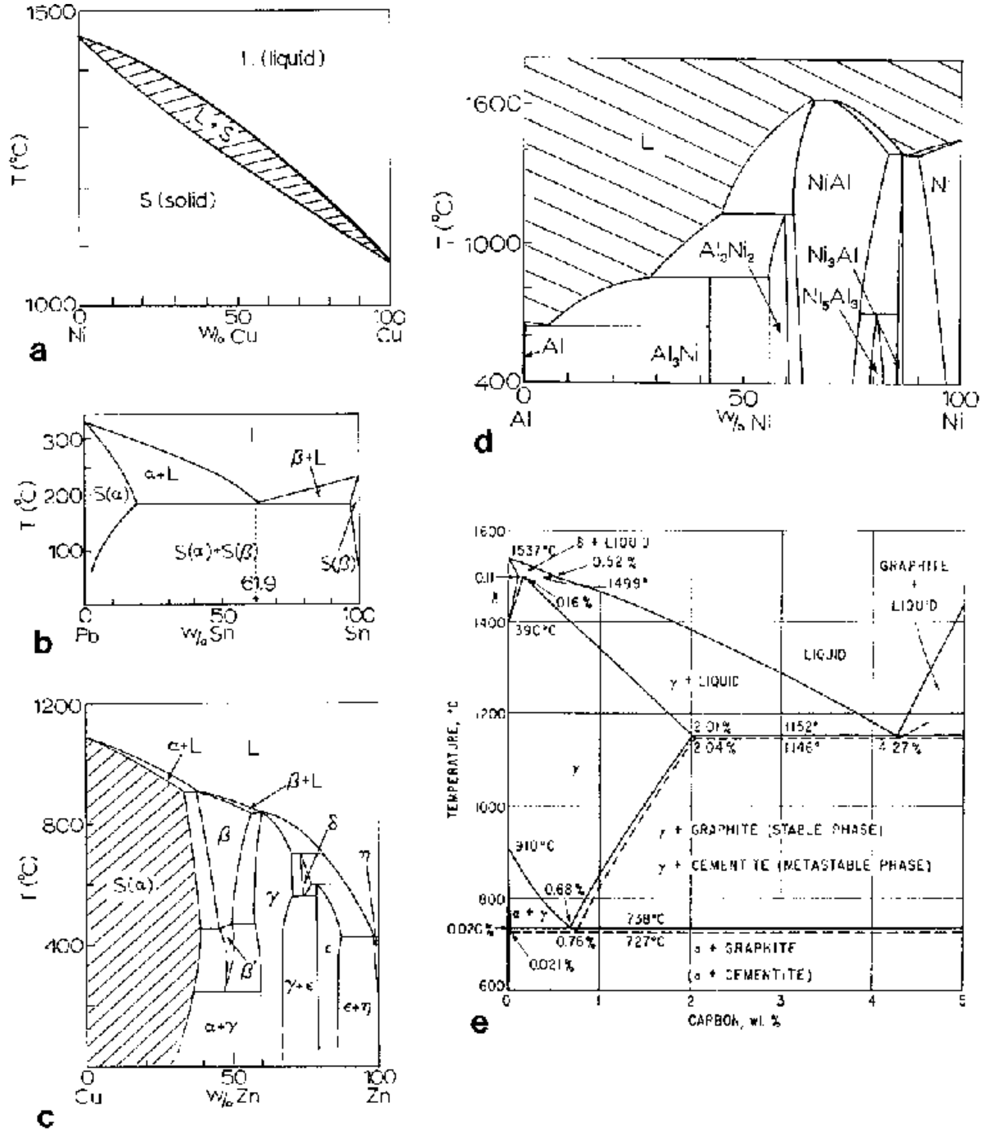


Figure 17 Binary phase diagrams. (a) Solid–solution diagram for Ni-Cu. (b) Simple eutectic diagram for Pb–Sn. (c) Copper–zinc phase diagram. Single phases are denoted α, β etc. respectively. $\alpha + \gamma$ is a two-phase region. (d) Systems joined by compound formation in the Al–Ni phase diagram. (e) Iron–carbon phase diagram portion. (Based on data in Ref. 4.)

consequence of either *homogeneous* or *heterogeneous nucleation*. In homogeneous nucleation, a spontaneous precipitate or new phase forms while in heterogeneous nucleation a surface or substrate provides some additional energy to initiate the process, and as a consequence the overall energy to drive the process is considerably reduced from homogeneous nucleation. The volume of new phase to create a critical nucleus is also reduced in heterogeneous nucleation, as illustrated in Fig. 19.

As illustrated in Fig. 17, phase transformations can occur in moving through a temperature gradient at some fixed composition, and this is especially notable in the case of the iron-carbon diagram (Fig. 17e) where heating and cooling alter the constitution of steel and, ultimately, the properties. The iron–carbon diagram represents the limiting conditions of equilibrium and is basic to an understanding of *heat-treatment* principles as applied to steel manufacture and behavior. The conventional iron phase is BCC ferrite (α). But iron

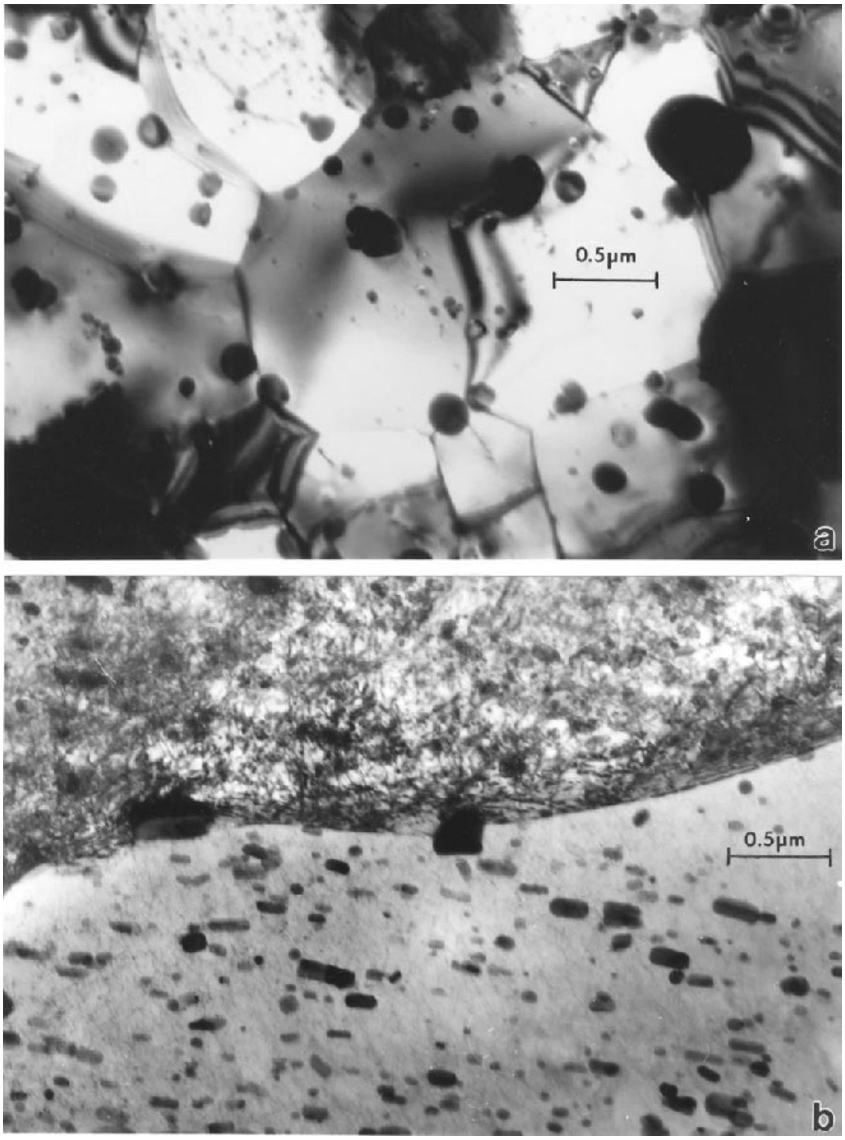


Figure 18 Second phase particles in a matrix. (a) Hafnium carbide particles in a tungsten matrix (courtesy of Christine Kennedy, UTEP). (b) (Fe, Cu, Al) Si precipitates in a dilute aluminum alloy (2024). (Courtesy of Maria Posada, UTEP.)

also exists as FCC austenite (γ). These are called *allotropic* forms of iron. When a plain carbon steel of approximately 0.8% carbon is cooled slowly from the temperature range at which austenite is stable, all of the ferrite and a precipitate of Fe_3C called cementite form a new precipitate phase called pearlite, which is a lamellar structure. Cementite is metastable, and can decompose into iron and hexagonal close-packed (HCP) carbon (or graphite). Consequently in most slowly cooled cast irons, graphite is an equilibrium phase constituent at room temperature. Like the graphite in a pencil, sheets of (0 0 1) carbon [actually (0 0 0 1) in the HCP structure] slide one over the other to produce a solid lubricant effect. It is this lubricating quality of cast iron which makes it particularly useful in engine block and related, extreme wear applications. Of course pure graphite powder is also a common solid lubricant.

Figure 20 illustrates some of the more common microstructures which characterize precipitation and other phase phenomena in iron–carbon alloys described above. Figure 20f also shows another example of iron–

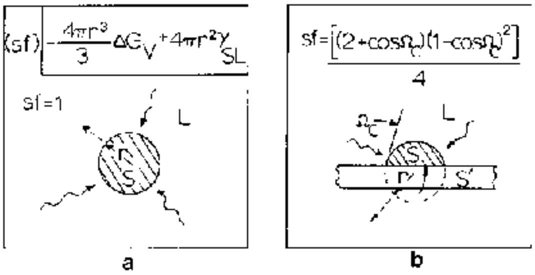

(sf) $-\frac{4\pi r^3}{3}\Delta G_V + 4\pi r^2\gamma_{SL}$
sf = 1
L
r
S
a
sf = $\frac{[(2+\cos\Omega_c)(1-\cos\Omega_c)^2]}{4}$
Ω_c
L
S
r'
S'
b

Figure 19 Simple, idealized features of homogeneous (a) and heterogeneous (b) nucleation. ΔG_v is the volume free energy, γ_{SC} is the solid (*S*) precipitate-liquid (*L*) interfacial free energy, *r* is the radius, and Ω_c is the contact angle for the nucleus forming on a solid substrate in (b).

carbon precipitates which can form in ferrite. Common precipitate stoichiometries include FeC_x, $Fe_{23}C_6$, and Fe_6C. Heat treatment can also produce other crystal structures. Martensite is a common example and is characteristic of transforming FCC austenite to a body-centered tetragonal phase or a body-centered cubic phase. Martensite, like cementite and other iron–carbon phases is considerably harder than either ferrite or austenite. As illustrated in Fig. 20, distributions of these precipitates or phases in the iron matrix provide for wide variations in mechanical properties–hardness, brittleness, ductility, toughness, yield strength, etc Adjustments, or the inability to accommodate adjustments, in volume changes when phases and precipitates form create internal stresses which can manifest themselves in dislocations or dislocation arrangements. These can be adjusted by *tempering* the steel at some temperatures which can provide for stress relaxation through dislocation rearrangement and annihilation, or by the formation of new or additional carbide precipitates by carbon diffusion in the solid phase matrix. This also creates additional variations in toughness, ductility, and related mechanical responses.

The alloying of chromium, nickel, and other elements with iron produces more corrosion-resistant steels or stainless steels. These alloys also generally contain reduced carbon ($< 0.08\%$C), but can be sensitized to corrosion by depletion of the chromium through carbide formation ($Cr_{23}C_6$, for example). This depletion is often preferential in grain boundaries where the diffusion of Cr and C are very rapid, and $Cr_{23}C_6$ precipitates can form, allowing preferential corrosion at the grain boundaries. Sensitization occurs when these alloys are *aged* at temperatures between about 600 and 800°C for long times. Figure 21 illustrates this sensitization behavior and the associated carbide precipitation on grain boundaries in 304 stainless steel (18% Cr, 9% Ni, 0.07% C, balance Fe).

Figure 21(c) illustrates a notable effect of even simple tensile deformation on the sensitization and precipitation behavior of 304 stainless steel. Actually, deformation (straining) plays a unique role in the development of nucleation sites for precipitation as well as the creation of other transformation products, the most notable being strain-induced BCC (α') martensite. Invariant strains create martensite nuclei as illustrated in Fig. 22a and b. This strain-induced phase transformation is also observed to be strain-state dependent as shown in Fig. 22c. The significance of this feature is that it becomes very difficult to process stainless steel in a biaxial mode, and even more difficult to draw wire or cable because of the very rapid production of α'-martensite. Figure 22d–f also illustrates additional deformation microstructures which can result in 304 stainless steel as a consequence of systematic movements of dislocations and partial dislocations as discussed previously. The specific shifts or displacements on the {1 1 1} slip planes are indicated in Fig. 22d and e.

6.2.3 Strengthening Mechanisms in Materials

Figure 12 and 13 have already demonstrated the ability to partition a crystal by creating grain boundaries or dislocation cell boundaries with the grains. The resulting strengthening has been demonstrated in Eq. (4) as an increased yield strength (stress σ) which, in the case of metals usually results in a systematic hardness increase as well, since hardness is linearly related to the yield stress: $H = K\sigma$, where K is often about 3.

This strengthening of a matrix is also apparent in Fig. 18 where dispersed particles (HfC) and precipitates provide some microstructural partitioning, while Fig. 20 shows a few additional examples of new phases and transformation products acting as microstructural partitioning phenomena; including precipitation strengthening. Precipitation occurs through heat treatment which creates a terminal solid solution which has decreasing solid solubility with decreasing temperature. Dense, fine precipitates (spaced a mean distance, *L*) act primarily as obstacles to dislocation motion, preventing plastic flow and deformation of the host or matrix. Similar to Eq. (4) precipitates or noncoherent particles within a grain will add an increment of engineering stress:

$$\sigma = \sigma_0 + K/\sqrt{D} + K'/L \qquad (6)$$

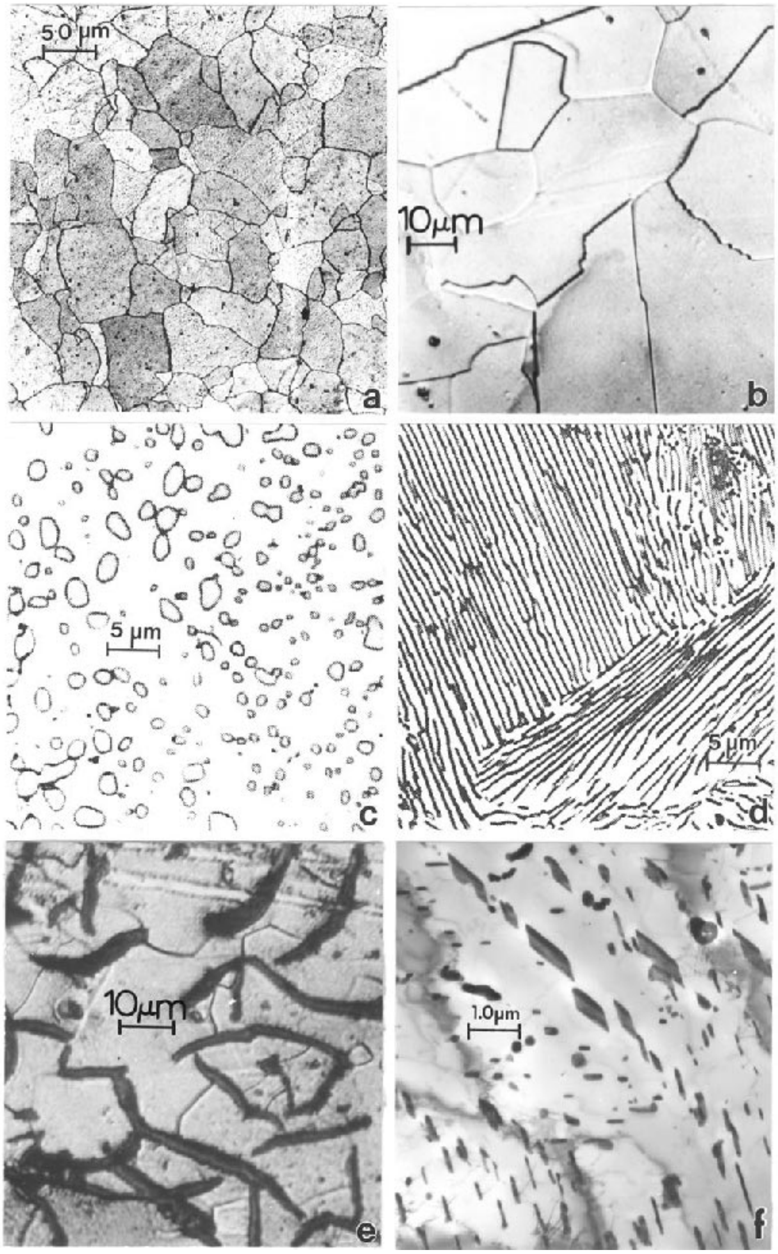


Figure 20 Microstructural variations in iron–carbon alloys. (a) BCC ferrite with carbon in solution. (b) FCC austenite with carbon in solution. The straight boundaries are annealing twin boundaries. (c) Cementite (Fe_3C) precipitates in ferrite. (d) Alternating lamellae of ferrite and cementite creating pearlite. (e) Graphite flakes (dark) in ferrite. (f) Iron carbides ($M_{23}C_6$ and FeC_x) in ferrite.

Deformation can itself add incremental strength or hardness to a material by creating dislocations or dislocation cells as shown in Fig. 13, by creating stacking faults or twin faults (deformation twins), or even a deformation-induced phase, as shown in Fig. 22. These features work harden the material by partitioning the grain structure: the more dense the deformation-induced microstructure the greater the residual stress or hardness (work hardening). These features are incremental and can be illustrated in a slightly different form of Eqs. (4) and (6):

$$\sigma = \sigma_0 + \sum_i K_i \lambda_i - m \tag{7}$$

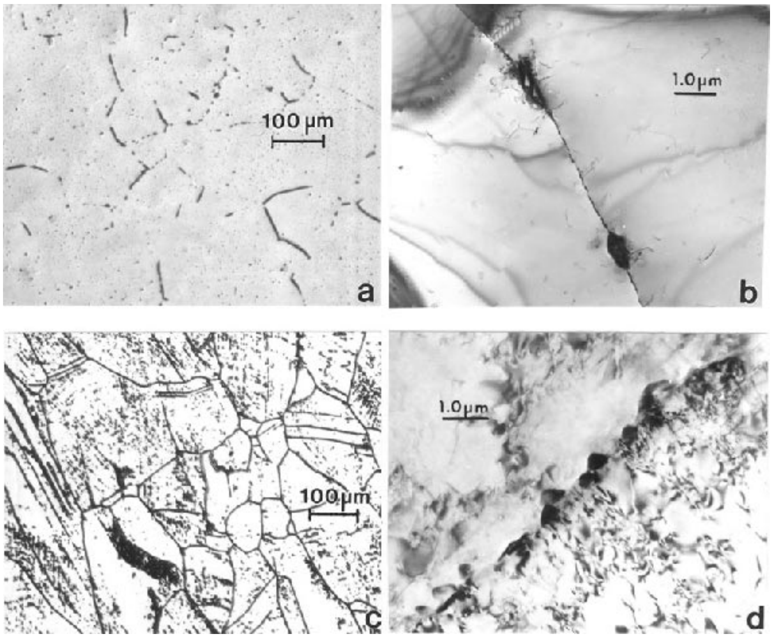


Figure 21 Sensitization and carbide precipitation at grain boundaries in 304 stainless steel aged 50 h at 25°C. (a) Mill processed prior to aging. (b) Details of carbides at grain boundaries in (a) (TEM image). (c) Mill processed and deformed in tension by straining 20% prior to aging. Note enhanced intergranular corrosion in contrast to (a). (d) Details of carbides at grain boundaries. The corrosion exhibited in (a) and (c) is indicative of the degree of sensitization. (Courtesy of Julio Maldonado, UTEP.)

where K_i is the associated material constant for a partitioning microstructure, i, having a mean spacing, λ_i, and m can vary from about 0.5 to 1. Figure 23 illustrates the general concept of strengthening mechanisms in classical materials science, consistent with Eq. (7). Figure 24 shows some additional examples of these strengthening mechanisms as strengthening microstructures which partition the existing grain structure. Each of the transmission-electron micrograph (TEM) images in Fig. 24 corresponds to a region within a polycrystalline grain. They correspond to partitioning parameters d, Δ, and L in Fig. 23 (Figs. 24a, b, and c and d respectively).

It should be noted that microstructural partitioning in materials systems also usually promotes hardening because for many metals (and some alloys) the hardness and yield stress are linearly related: $\sigma_y = AH$, where for pure metals like copper, $A \cong 1/3$. This is an important consequence since hardness, as a manufactured product quality can often be improved by the same kinds of microstructural adjustments performed to improve the overall strength of the material or material system. As a case in point, the W–HfC system shown in Fig. 18a goes from a Vickers hardness of around 600 with a 1% addition of HfC to about 1200 with a 5% addition of HfC to the W matrix.

The addition of a dispersed phase such as HfC to a matrix as shown in Fig. 18a produces a composite, often called a metal–matrix composite (MMC). Such composites form a unique class of materials which involve glass-reinforced plastics, ceramic-reinforced metals and alloys, and other variations. Such composites can involve particles having a wide range of sizes and size distributions as well as fibers; either continuous, discontinuous, unidirectional, bidirectional, etc. These variations in materials composites produce a wide range of new materials and materials systems and applications as well as a range of strengthening mechanisms. Not only do fibers obstruct dislocation motion in the so-called host material (or matrix) through the kind of partitioning as shown in Fig. 24, but intrinsic strength is provided through the fiber itself in the case of a fiber composite. Very fine, single-crystal fibers are especially useful in tailoring unique and very strong composites (SiC in aluminum, BC in nickel, etc). For example, where continuous fibers (f) are oriented in the direction of an applied stress in a host (matrix) material (m) the modulus of elasticity of the composite (E_c) is given ideally by

$$E_c = E_f V_f + E_m V_m \tag{8}$$

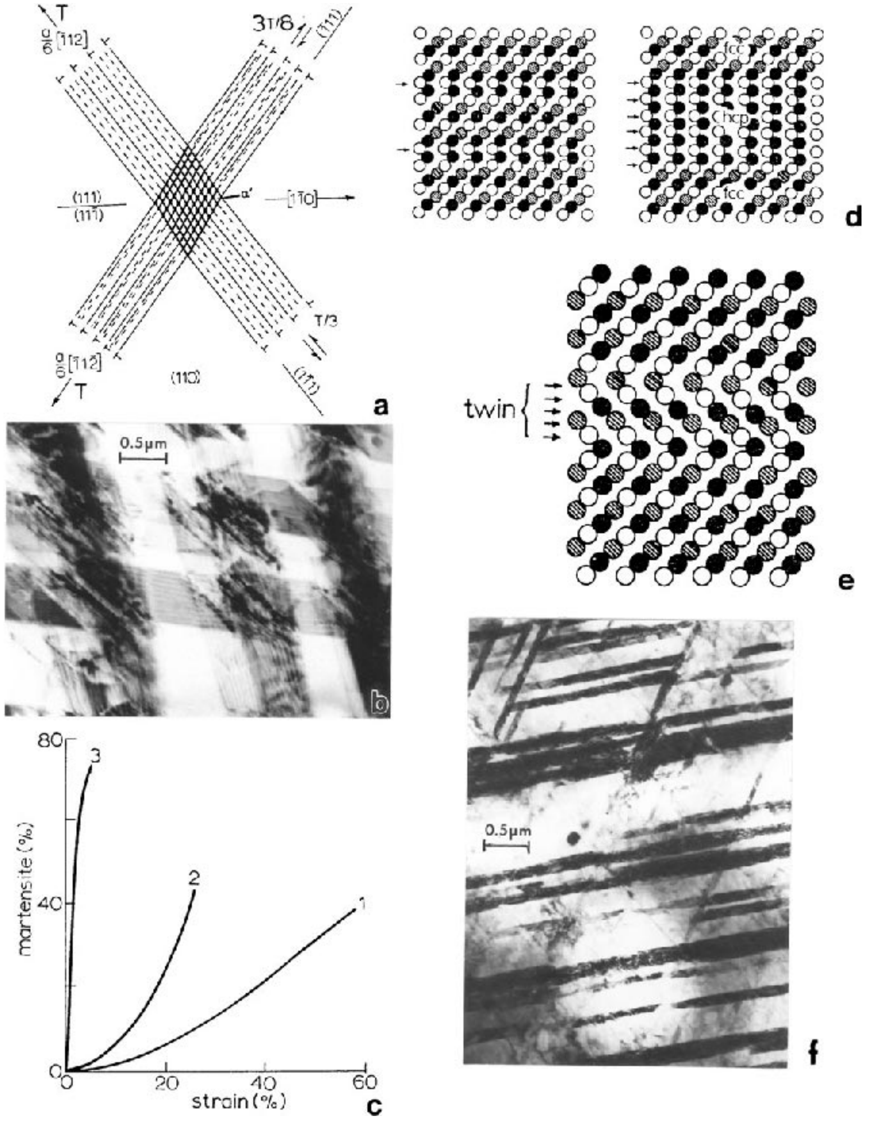


Figure 22 (a) Invariant strain producing specific, intersecting faults which produce a martensite nucleus. (b) TEM image of α'-martensite created at strain-induced intersections shown in (a) in 304 stainless steel. (c) α'-Martensite volume fraction at equivalent strain for different strain states in deformed 304 stainless steel denoted 1 (uniaxial), 2 (biaxial), and 3 (triaxial). (d) Stacking-fault and epsilon-martensite models in FCC materials. (e) Twin formation in {1 1 1} planes in FCC. (f) TEM images of deformation twins in shock-loaded 304 stainless steel.

which is often called the rule of mixtures for a binary composite where E_f and E_m are the corresponding fiber and matrix moduli of elasticity, and V_f and V_m are the corresponding fiber and matrix volume fractions respectively.

From Eq (8) it can be observed that the yield strength of a simple, oriented, fiber-reinforced composite material is given by

$$\sigma_c = \sigma_f V_f + \sigma_m V_m \tag{9}$$

where σ_f and σ_m are the corresponding fiber and matrix yield strengths (stresses).

In contrast to fibers which are oriented in the direction of an applied stress and are correspondingly strained identically to the matrix (isostrain condition), the fibers can be oriented (perpendicular) so that they are not strained identically to the matrix. This is ideally an isostress situation where

$$E_c \cong E_f E_m / (V_f E_m + V_m E_f) \tag{10}$$

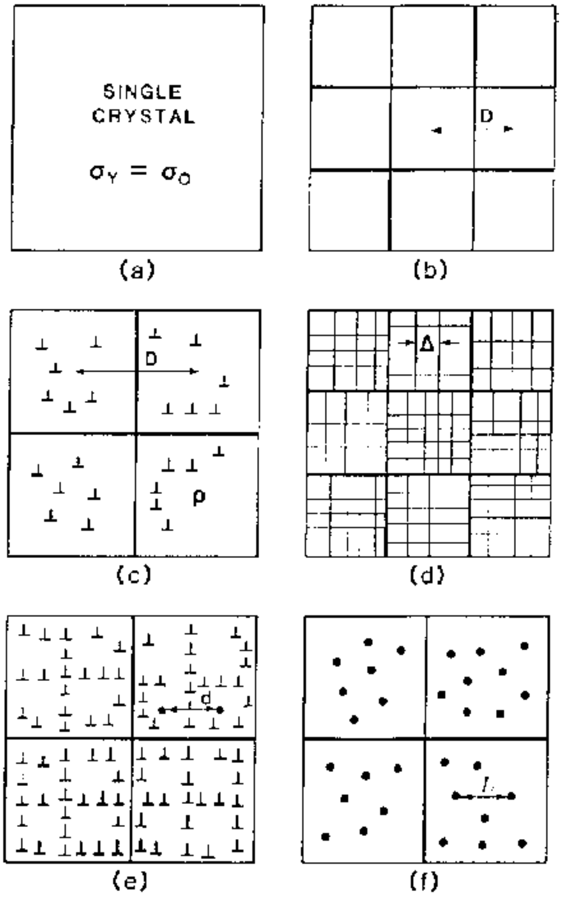


Figure 23 Schematic views of microstructure partitioning/strengthening. (a) Single crystal. (b) Polycrystalline material with grain size D. (c) Polycrystalline material with dislocations. (d) Polycrystalline material with twins or faults spaced Δ. (e) Dislocation cells (of spacing d) in polycrystalline material. (f) Polycrystalline material with precipitates or dispersoids with wear spacing L.

This kind of relationship can also approximate the case of dispersed particles or discontinuous fibers in a matrix as shown in Fig. 18a, which is clearly an isostress situation.

6.3 CERAMIC AND POLYMERIC MATERIALS

Up to this point, we have treated industrial materials science and engineering in the context of primarily metals and metallurgy because they are somewhat historically at the heart of the industrial revolution (and evolution) worldwide. In addition, metals continue to dominate industrial materials utilization, with metal cast parts accounting for a significant fraction of all manufactured products. Ceramic materials are also a significant materials component because they include glass products which constitute very old and continuous manufacturing and artistic commodity businesses. In contrast to polymeric or plastic materials, ceramics represent the extreme in brittle behavior, while polymers represent the extreme in "ductile" or plastic behavior. Ceramics are characterized by traditional crystal unit cells, while polymers are unique in their structure, constituted primarily by complex and often continuous molecular chains. Ceramics are characterized by ionic (and covalent) bonds between their atoms while polymers are characterized by covalent bonds. Both ceramics and polymers are normally poor electrical conductors; some are excellent insulators. On the other hand, the so-called high-temperature *superconductors* such as $YBa_2Cu_3O_7$ (*yttrium barium copper oxide*) which have superconducting transition temperatures above 90 K, are also oxides. Some oxides are *semiconductors* and simple structural differences can have a significant effect. For example, the creation of an additional oxygen vacancy in the $YBa_2Cu_3O_7$ structure (to $YBa_2Cu_3O_6$) or the addition of oxygen to vacant sites in $YBa_2Cu_3O_7$ to produce $YBa_2Cu_3O_8$ can have a dramatic effect on the electronic behavior, as illustrated in Fig. 25. Figure 25 also shows the fundamental features which differentiate (metallic) conductors, superconductors, semiconductors, and *insulators* both in terms of simple electronic or band structure considerations, and simple resistance (R) (or resistivity, ρ) versus temperature, T (R–T signature) curves. Of course the more popular and classical semiconductor materials consist of *silicon* or *gallium arsenide* (GaAs), and devices or chips which are fabricated from them, constituting a huge microelectronics industry sector worldwide.

6.3.1 Structure and Properties of Ceramics

Ceramic materials are inorganic compounds or similar materials mixtures often characterized by metal oxides or other elements bonded by ionic or covalent bonds, or combinations of ionic and covalent bonds. Their properties are often linked to the bonding which can include layer structures or molecular sheets with a strong covalent bonding in the sheet direction, and weak *van der Weals* or so-called *hydrogen bonding* between the sheets. Micacious materials are common examples of silicate sheet structures. Talc is another sheet structure in which the silicate sheets slide one over the other as a kind of solid-state lubricant. This

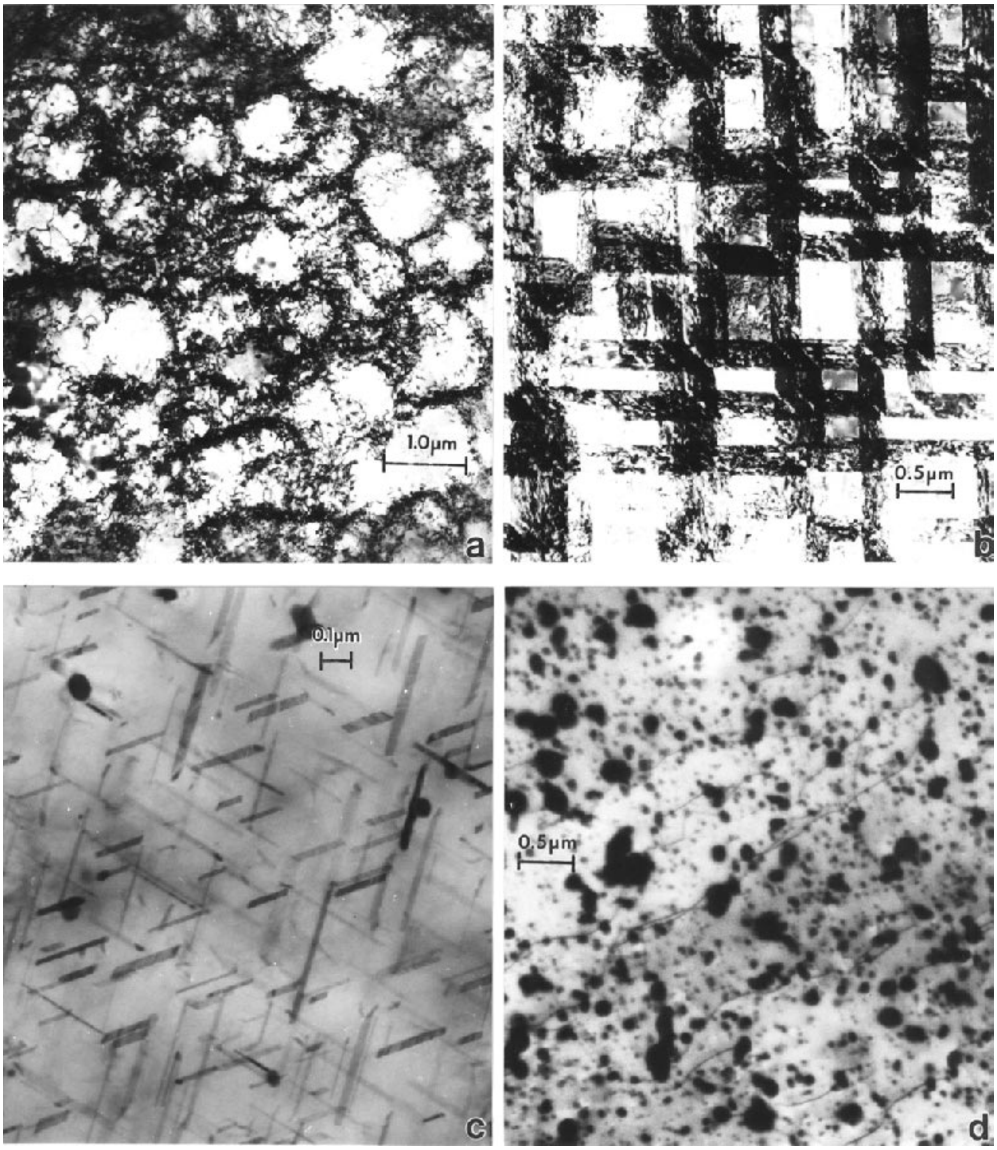


Figure 24 Examples of microstructural partitioning contributing to strengthening of materials. (a) Dislocation cells in copper. (b) Deformation twins in shock-loaded nickel. (c) Crystallographic {1 1 1} precipitates in aged 6061-T6 aluminum. (d) Thorium dioxide dispersed particles (2 vol%) in an 80 Ni–20 Cr alloy. Note dislocations anchored on particles.

is what makes baby powders ‘‘soft’’—talc sheets sliding one over the other. This is similar to the effect of graphite lubrication where the covalently bonded carbon layers slide one over the other. This is what makes cast irons machine so easily, etc., as noted previously. Many ceramic materials are crystalline and have structures (or unit cells) which are either in the form of the Bravais lattices shown in Fig. 4, or structural modifications of these fundamental lattice arrangements—such as those shown in Fig. 26. However, some ceramics are not crystalline, and have random structures. Glass in all of its common varieties is the best example. Sometimes glass is referred to as a frozen liquid or a liquid of very high viscosity. Figure 27 illustrates some examples of ceramic crystal structure and glass structure in various perspectives and applications.

Many ceramic materials are very hard and brittle and have very high melting points. In fact, HfC, a ceramic material illustrated in Figs. 18 and 26, is currently the material with the highest known melting

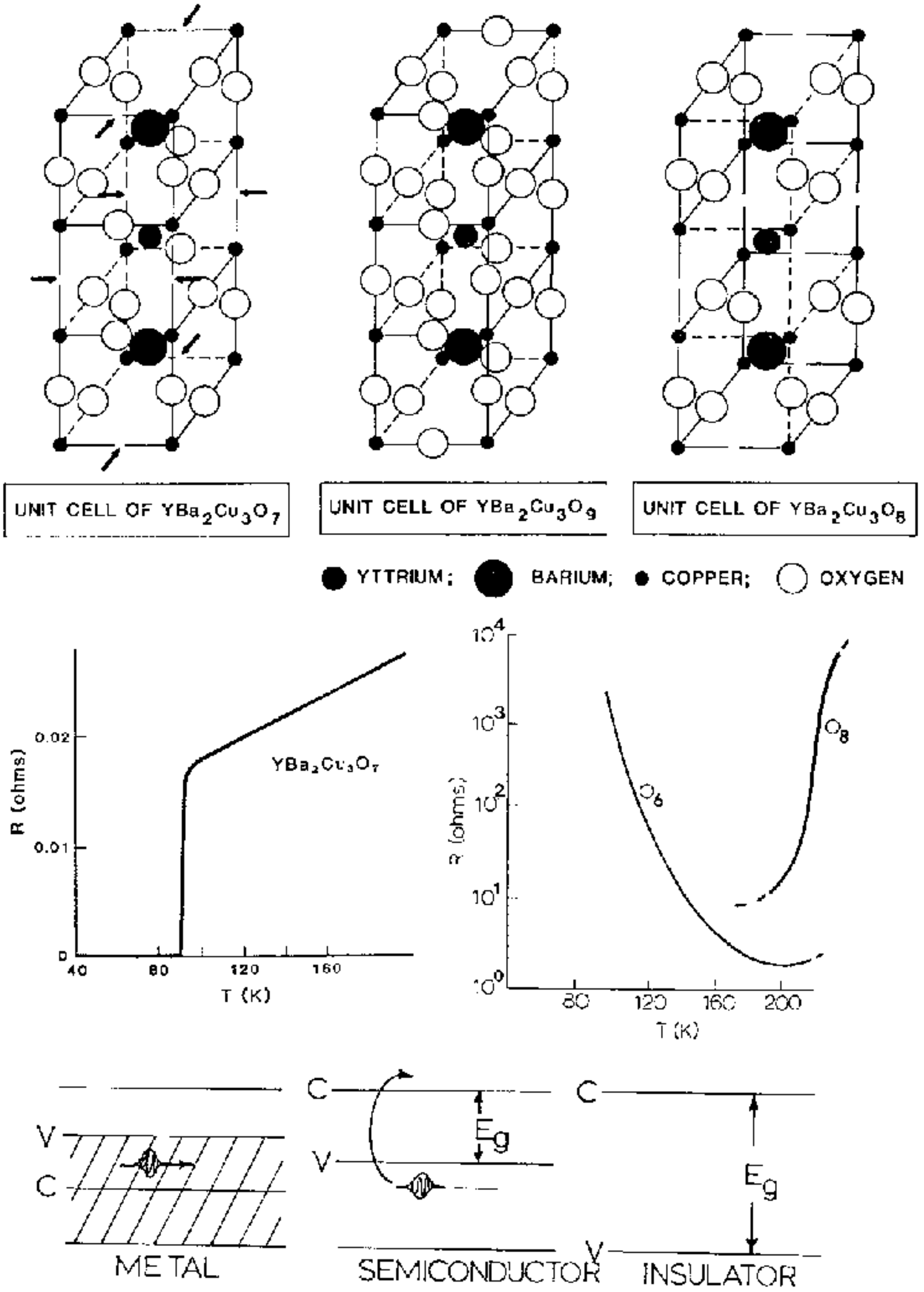


Figure 25 High-temperature ceramic superconductor and related structures illustrate examples of metallic conduction, semiconduction, and insulating or very high-resistance behavior as well as superconductivity. Resistance–temperature diagrams illustrate corresponding signatures for these kinds of behavior. The simple valence-conduction band sketches below show fundamental electronic considerations. In semiconductors and insulators an energy gap (E_g) shown must be overcome. Temperature provides the driving energy in semiconductors. The (O_9) structure shown top center is "ideal." It cannot exist. Only the (O_8) structure can exist. Resistance–temperature signatures are shown below the structures which are ideally stacked perovskite crystal cubes (see Fig. 26). *C* and *V* in bottom sequence denote the bottom of the conduction band and top of valence band, respectively. In a metal, these bands overlap.

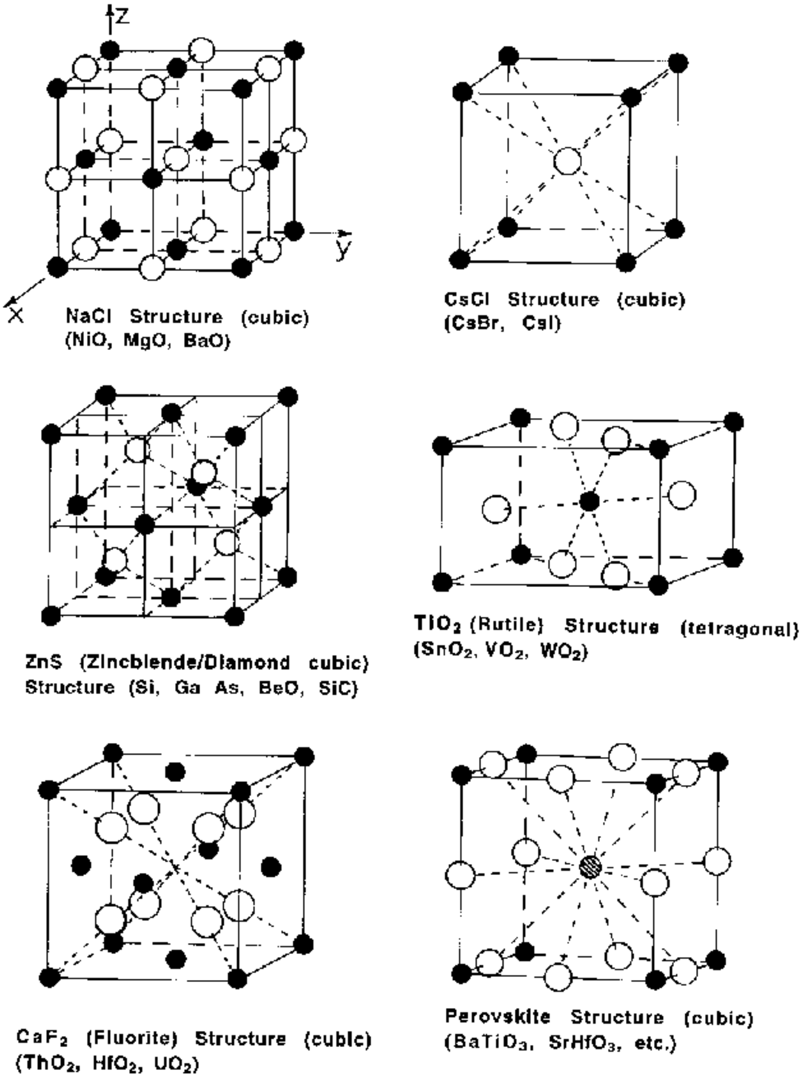

z
y
x
NaCl Structure (cubic)
(NiO, MgO, BaO)
CsCl Structure (cubic)
(CsBr, CsI)
ZnS (Zincblende/Diamond cubic)
Structure (Si, Ga As, BeO, SiC)
TiO2 (Rutile) Structure (tetragonal)
(SnO2, VO2, WO2)
CaF2 (Fluorite) Structure (cubic)
(ThO2, HfO2, UO2)
Perovskite Structure (cubic)
(BaTiO3, SrHfO3, etc.)

Figure 26 Examples of ceramic crystal structures.

point, 4600°C. It is in fact these very high melting points which make ceramics attractive for a host of very high-temperature operations and processes. Ceramic engine materials are a current focus not only in the automotive industry, but aerospace and other areas involving materials and devices required to operate at very high temperature (1000–3000°C).

In addition to being hard, most ceramics, including glass, are very brittle. This means that there is no plasticity and very limited elastic behavior. Many ceramics have very low *fracture toughness* as a consequence of these features: they shatter with impact. In the context of crystal defects discussed in earlier sections, ceramic crystals often have very low densities of dislocations and therefore cannot slip; hence, their brittle behavior. On the other hand, the creation of systematic vacancies in atomic sites—especially the cation (positive ion) sites in the crystal lattice can significantly alter other properties of ceramics. Some ceramics materials with good semiconducting behavior become insulators and conduct essentially no current. Other transparent crystals become colored or opaque as the stoichiometry and cation vacancy concentration change. This is a simple indication of optical property changes, and many crystalline ceramics exhibit a wide range of optical properties having commercial applications. Substitutional elements in glasses account for their colors as well as optical properties such as *photochromism*, where the degree of transparency or opacity changes with light intensity (see Fig. 27).

Many ceramic crystals, including common quartz, exhibit a connection between polarizability and

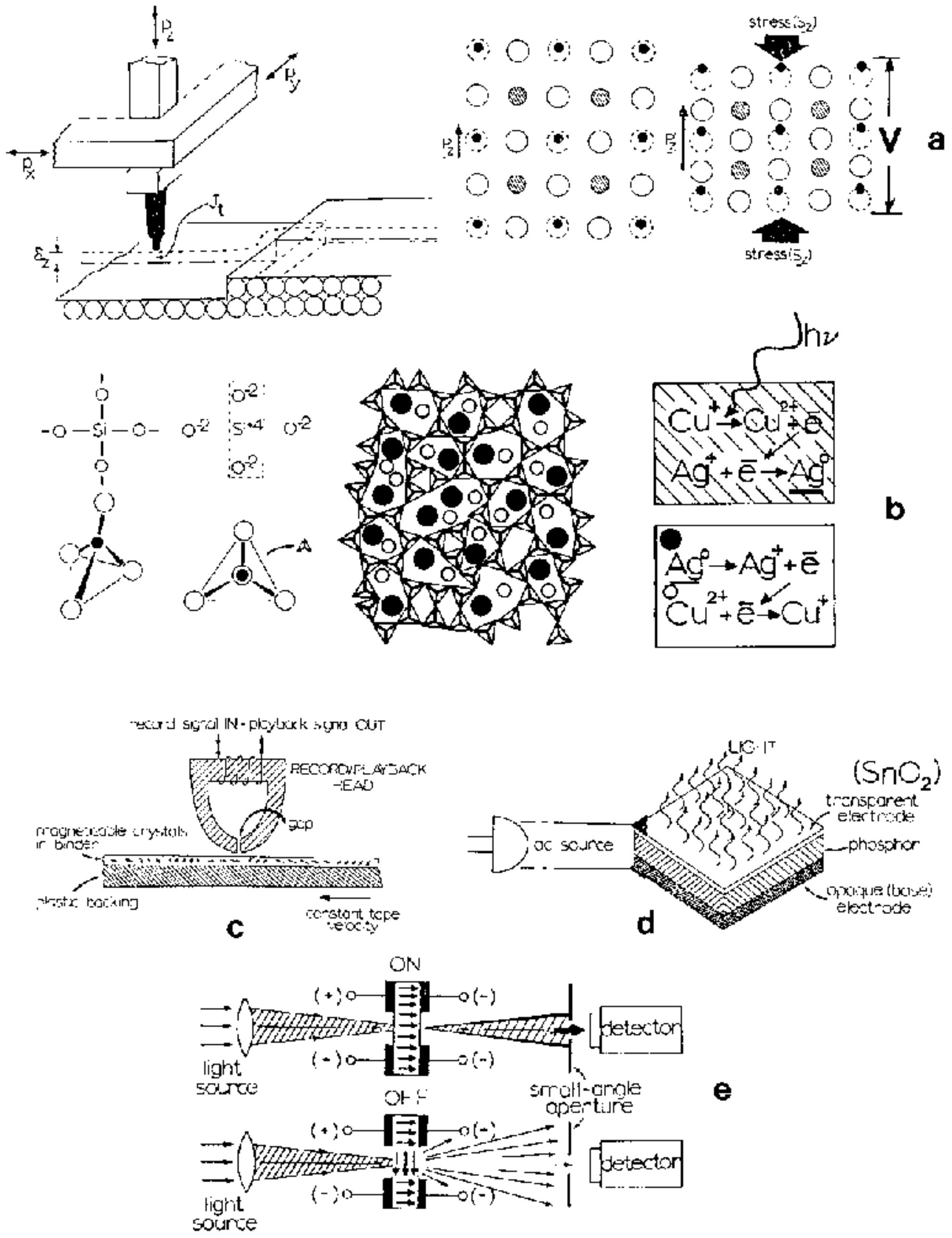


Figure 27 Examples of structure–property–performance features of ceramic materials. (a) Piezoelectric drive system for scanning tunneling microscope (left) operates by applying a voltage (*V*) to a perovskite structure to cause corresponding stress (or displacement) (right). (b) Photochromic effect in glass caused by light-induced (*hν*) intervalence transfer. The shading corresponds to darkened glass. (c) Polarization (magnetic) of ferrite crystals in magnetic tape recording. (d) Light emitting, electroluminescent device. (e) Ferroelectric light valve (lead–zirconate–titanate ceramic) [(c) and (d) are from Ref. 5.]

mechanical forces or stress. That is, a stress applied along some specific crystal axis (direction) can create structural asymmetries which polarize the structure. This creates a corresponding electrical voltage—a phenomenon called the *piezoelectric effect*. Materials exhibiting this property are also termed *ferroelectric materials*. This electromechanical effect is reversible. That is, the application of a voltage along specific crystal directions in ferroelectric materials can create a corresponding displacement or a mechanical response, as shown schematically in Fig. 27a. Micropositioning devices utilize this principal while accelerometers, transducers etc. measure mechanical force through electrical signals created, i.e., they convert a mechanical force or sound pressure into an electrical response. Ferroelectric (transparent) materials can also be used as optical switches or filters by applying an electric field to polarize the structure, thereby deflecting or absorbing light (Fig. 27e).

Figure 27 illustrates a variety of properties and applications of ceramic materials. In addition, ceramics in a variety of forms—solids, porous materials, fibers, etc. are especially effective thermal insulators in addition to being superior electrical insulators. Space shuttle tiles are a common example of outstanding thermal insulation qualities of pure silica fibers. Ceramic processing normally involves powder processing, which includes hot isostatic pressing (HIP), and slip casting, which involves ceramic particles in a water suspension, as well as sol–gel processing of powders in appropriate slurry gels. Ceramics are produced in molds or cbs other shape forms.

6.3.2 Structure and Properties of Polymers

There are two general classes of polymeric materials: *elastomers* (or rubbers) and *plastics*. Plastics are further classified as *thermosetting plastics* (such as epoxy resin-based materials) and *thermoplastics*. Elastomers can be stretched elastically over extended elongations and when relaxed, essentially return to their original shape. Consequently there may be no permanent (or plastic) deformation. Plastics, on the other hand, can have a large elastic range or even an elastic-like range followed by an extended range of permanent deformation prior to failure—hence the term plastics. Plastics have a structure similar to glass, but are dominated by chainlike molecular arrangements. These molecular chains can be folded or branched, and this can account in part for the deformation behavior of plastics. These molecular chains are bonded covalently—a process called chain polymerization. Common polymers rarely have a corresponding crystal structure. The most general features of polymer structure are illustrated in Fig. 28.

Like ceramics, many thermoplastics are molded (by *injection molding*) from powders to form products. Thermosetting plastics often provide a host or matrix for composites—for example, epoxy–graphite fiber composites or even epoxy-glass fiber composites. Plastics containing soft or rubber components also produce plastic composites having a wide range of energy absorbance, etc. Plastic sheet containing hard ceramic particles is also utilized in novet wear applications.

The automotive industry is one of the larger users of plastic materials, both thermosetting and thermoplastic. A modern automobile offers an interesting example not only of essentially every kind of basic material–metals, alloys semiconductors, ceramics, plastics, and composites—but also a wide variety of each of these materials systems.

Figure 29 shows for comparison the tensile stress–strain behavior of a common plastic material (polymethyl methacrylate, PMMA) over a range of temperatures compared with some common ceramic and metallic materials. This comparison not only illustrates the unique mechanical properties of these materials, but also their complementarity over a wide range of temperature and stress environments.

6.4 CREEP, FRACTURE, FATIGUE AND RELATED INDUSTRIAL MATERIALS PERFORMANCE ISSUES

6.4.1 Creep in Metals and Plastics

Creep is a property of metals (and alloys) and plastics where progressive plastic (or permanent) deformation takes place over a long period of time at some relatively constant load or stress. In effect, this is a time-dependent strain (or elongation), and is particularly important for materials operating at elevated temperature where the creep rate (or stain rate, $\dot{\varepsilon}$) is accelerated. Creep rates actually vary with time, and a typical creep curve for a metal as shown in Fig. 30a usually demonstrates three, well-defined stages leading to fracture: primary, secondary, and tertiary creep. Primary creep follows an instantaneous elongation (or strain, ε_0) and represents a stage where the creep rate decreases with time. Secondary creep represents a steady-state creep with a constant creep rate, while in the third stage of creep the creep rate rapidly accelerates to fracture. As in other deformation processes in crystalline or polycrystalline metals or alloys, dislocations slip to allow for straining.

Creep in plastics also occurs at elevated temperatures but does not occur in discrete stages because the mechanism is different from crystalline metals or alloys, as shown on comparing Fig. 30a and b. Creep of polymeric materials is often measured by a so-called creep modulus defined as the ratio of the initial applied stress, σ_0, to the creep strain after a particular time, $\varepsilon(t)$, and at a constant temperature. Plastics with a high

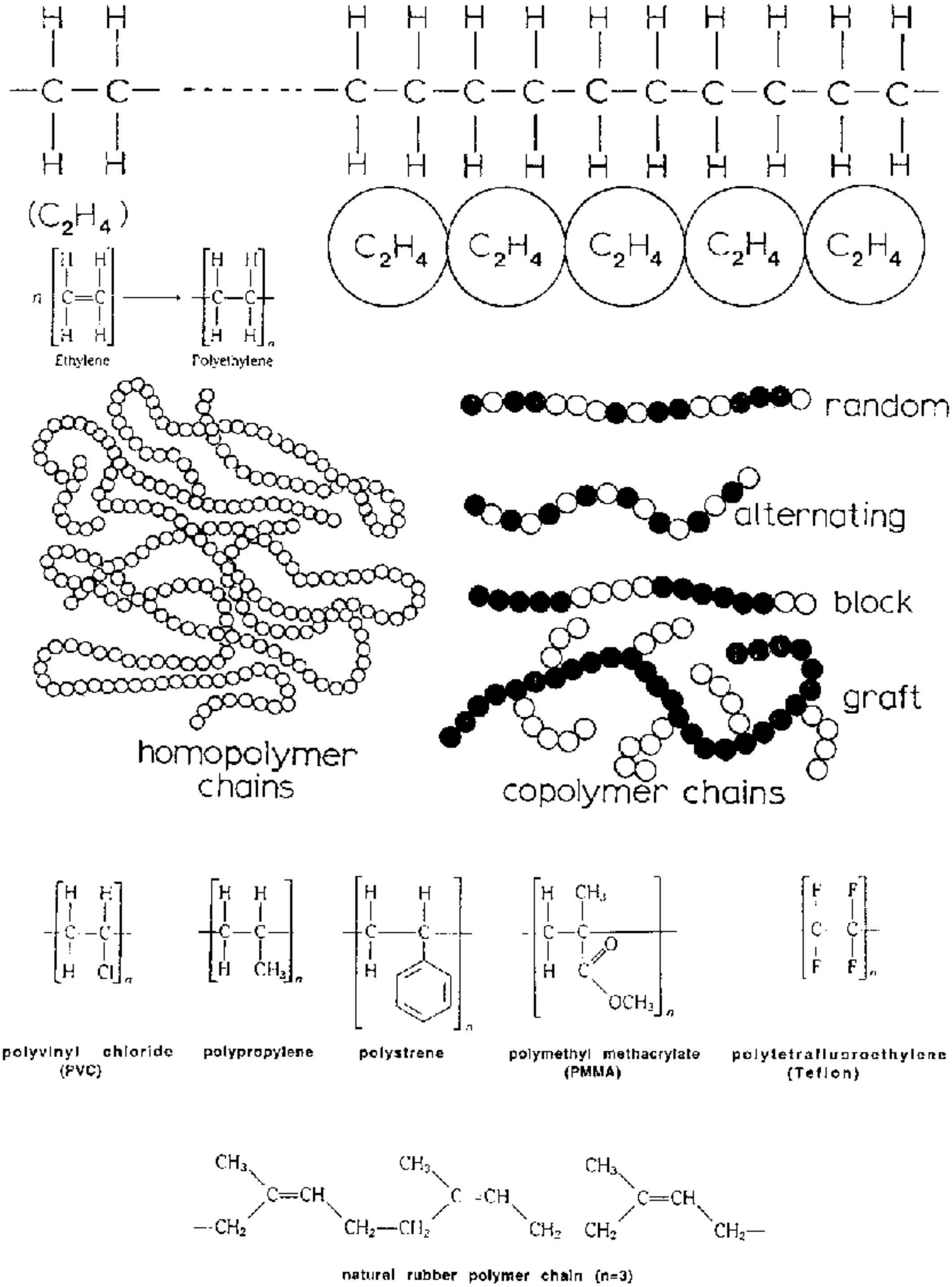


Figure 28 Structural formulas illustrating covalent bonding in polymeric materials along with simple schematic representations of polymers and polymer chain structures.

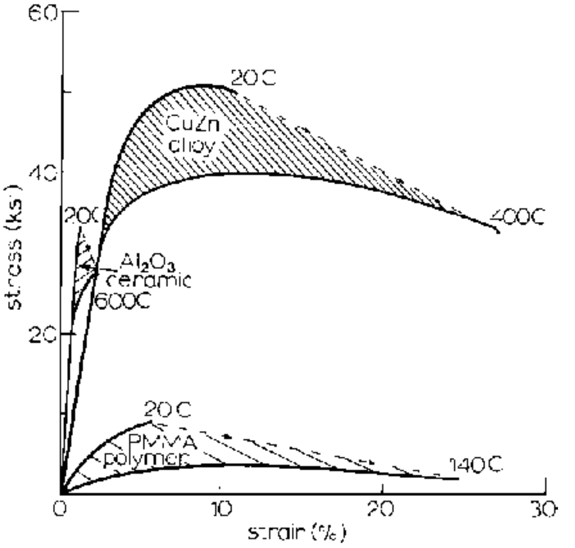


Figure 29 Tensile stress–strain curves for several different engineering materials at various temperatures. Temperature regimes are shown variously shaded. PMMA polymers are commonly used in a variety of tail-light reflectors.

creep modulus have low creep rates. When plastics are reinforced with other fibers or dispersed phases, the creep rate declines, i.e., the creep modulus is increased. For example, nylon has a creep modulus of around 120 ksi (1 ksi $\equiv 10^3$ psi $= 70.3$ kg/cm^2 $= 6.9$ MPa) at 20°C and 10 hr test time while the creep modulus for the same nylon, containing about 30% glass fibers (in the loading direction), increases to 700 ksi for the same test conditions.

Creep in metals and alloys can also involve diffusional phenomena and "sliding" of subgrain structures on even sliding of especially small grains, that is, slip along grain boundaries as opposed to slip in specific crystal planes in the matrix. At sufficiently high temperatures, some or all of these mechanisms may be very prominent At temperatures in excess of half the melting point ($T > 0.5T_m$), extremely high strains can be achieved in ordinary tensile testing. This phenomenon, commonly referred to as *superplasticity*, is dominated by sliding of equiaxed, small grain microstructures or often concurrently by rotations of small grains. Superplasticity is also strain-rate dependent: $\sigma_T = K'\dot{\varepsilon}_T^m$, where m is called the strain-rate sensitivity exponent, and the subscript T refers to the temperature specific to the straining.

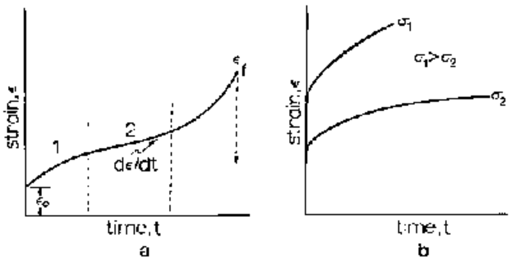


Figure 30 Idealized creep curve for a metal (a) showing primary, secondary, and tertiary creep stages marked 1, 2, 3, respectively. ε_f is the fracture strain. Note that the creep rates ($\dot{\varepsilon} = d\varepsilon/dt$) are different in the three creep stages. In (b), an idealized polymeric material like polystyrene strains with time in a somewhat different fashion.

It is probably instructive to look back retrospectively at the mechanical equation of state, Eq. (1). This equation illustrates [along with Eq. (2)] that for any particular state of straining (or stress), deformation is very specifically dependent upon not only the strain (ε) and strain rate ($\dot{\varepsilon}$), but the temperature at which the straining occurs.

6.4.2 Fracture of Materials

Tensile fracture or breakage involving the separation of solid materials occurs in all kinds of materials. We noted in Fig. 30 that fracture terminates creep deformation. Fracture can be characterized in many materials by two extremes: *ductile fracture* involving extensive plastic deformation, and *brittle fracture* involving essentially no plastic deformation. Fracture involves crack development and propagation and in this context ductile fracture is characterized by slow crack propagation, while brittle fracture is characterized by very rapid or catastrophic crack growth. Brittle fracture in metals includes cleavage or separation on specific crystal planes. However, fracture in polycrystalline metals can also include intergranular brittle fracture on grain boundaries. This often occurs when impurities segregate to the grain boundaries and lower the adhesive energies or the corresponding interfacial free energies. Low temperatures and very rapid rates of straining such as impact tend to favor brittle fracture, and as we mentioned previously, some steels and other materials exhibit a ductile to brittle fracture transition as the temperature is decreased. This is also true of plastics which almost universally become very brittle as the temperature is lowered. Ceramics, because of their general lack of plastic behavior, also characteristically exhibit brittle fracture over a wide range of temperatures Modern design strategies for ceramics in fact include efforts to improve the ductility of ceramics.

Fracture modes (such as ductile or brittle fracture, etc.) are implicit in the appearance of corresponding fracture surface morphologies (or fractographs) and these features along with simple fracture mode schematics are illustrated in Fig. 31. Note in Fig. 31a that ductile fracture is a cup and cone feature originating from void development and coalescence as the material necks. Shear fracture is a geometrical modification of this process as implicit in Fig. 31b. Figure 31c is illustrated by intergranular brittle fracture of iridium observed in the scanning electron microscope.

6.4.2.1 Fracture Toughness

Toughness is measured as the amount of energy a material can absorb without fracturing. Fracture toughness testing consists of impacting a notched specimen from the rear of the notch with a hammer which swings on an arm (a pendulum) from various heights (h) above the specimen. If the energy is considered to be the simple potential energy of the hammer (mgh; m is the mass and g is the gravitational constant), then testing over a range of temperatures can indicate when a material is brittle (low fracture energy) or ductile (high fracture energy); or the appearance of a brittle–ductile transition.

6.4.2.2 Fatigue Failure of Metals and Plastics

Repeated or cyclic stressing of materials can cause *fatigue fracture* at stresses often well below the static stress required for fracture. The reader can relate to this phenomenon by taking a steel wire which cannot be fractured by tensile straining, but which can be broken by bending back and forth. In addition, if the cycles of bending back and forth are rapid, the material will also become hot, and break sooner. This heating occurs as a consequence of adiabatic heating characterized as the product of the strain and the strain (or cyclic stress) rate.

For metals and plastics, fatigue failure or correspondingly fatigue life can be characterized by cyclic stress amplitude (S or $\sigma_a = (\sigma_{\max} - \sigma_{\min})/2$ versus number of cycles to failure (N or N_f), commonly called SN curves. Figure 32 illustrates several examples of these curves. It can be noted in Fig. 32a that for the steel curve, there is a limiting minimum stress amplitude below which there will be no failure. This is illustrated by the dotted line and is sometimes called the fatigue or *endurance limit* (σ_{fat}). Both the critical stress amplitude and the fatigue limit are altered by the magnitude of mean stress ($\sigma_{\text{mean}} = (\sigma_{\max} + \sigma_{\min})/2$). A simple relationship is often used to relate the stress amplitude, σ_a, the mean stress, σ_{mean}, and the fatigue limit (or critical fatigue stress amplitude, σ_{fat}:

$$\sigma_a = \sigma_{\text{fat}}(1 - \sigma_{\text{mean}}/UTS) \tag{11}$$

where UTS is the ultimate tensile strength (or stress).

Fatigue fracture of both metals and polymers is often characterized by striations or so-called clamshell markings on a clean fracture surface. The spacings of these striations reflect the random application of the cyclic stress that causes slow fatigue crack growth, and become very small for low stress or high cycle (higher frequency) fatigue. Figure 33 illustrates these features for metal fatigue fracture surfaces observed in the scanning electron microscope.

6.4.3 Corrosion and Wear Phenomena

Corrosion and wear of metals and alloys involve material deterioration or degradation by chemical and physical attack respectively. Corrosion and wear are also interactive, that is corrosive wear involving the simultaneous effects of both chemical attack and physical removal of surface material may also occur. Electrochemical attack involving electron flow through characteristic anodic and cathodic cells or volume elements on the metal surface is the most general mechanism driving corrosion phenomena: oxidation–reduction half-cell reactions. Zinc dissolution in acid is a prime example:

$$\text{Zn} + \text{HCl} \rightarrow \text{ZnCl}_2 + \text{H}_2 \uparrow \tag{12}$$

or

$$\text{Zn} + \text{H}^+ \rightarrow \text{Zn}^{2+} + \text{H}_2 \uparrow \tag{13}$$

or

$$\text{Zn} \rightarrow \text{Zn}^{2+} + 2\bar{e}\ (\textit{oxidation half-cell reaction})$$

$$2\text{H}^+ + 2\bar{e} \rightarrow \text{H}_2 \uparrow \text{(reduction half-cell reaction)} \\ \Rightarrow \text{cathodic reaction} \tag{15}$$

where ē represents an electron.

Table 1 lists the so-called electromotive series or the standard half-cell potentials for some selected metals referenced to hydrogen (standard hydrogen electrode). Metals which are more reactive than hydrogen are assigned negative potentials and are considered to be anodic to hydrogen, i.e.,

$$\text{M} \rightarrow \text{M}^{n+} + n\bar{e} \qquad \text{(metal oxidized to ion)} \tag{16}$$

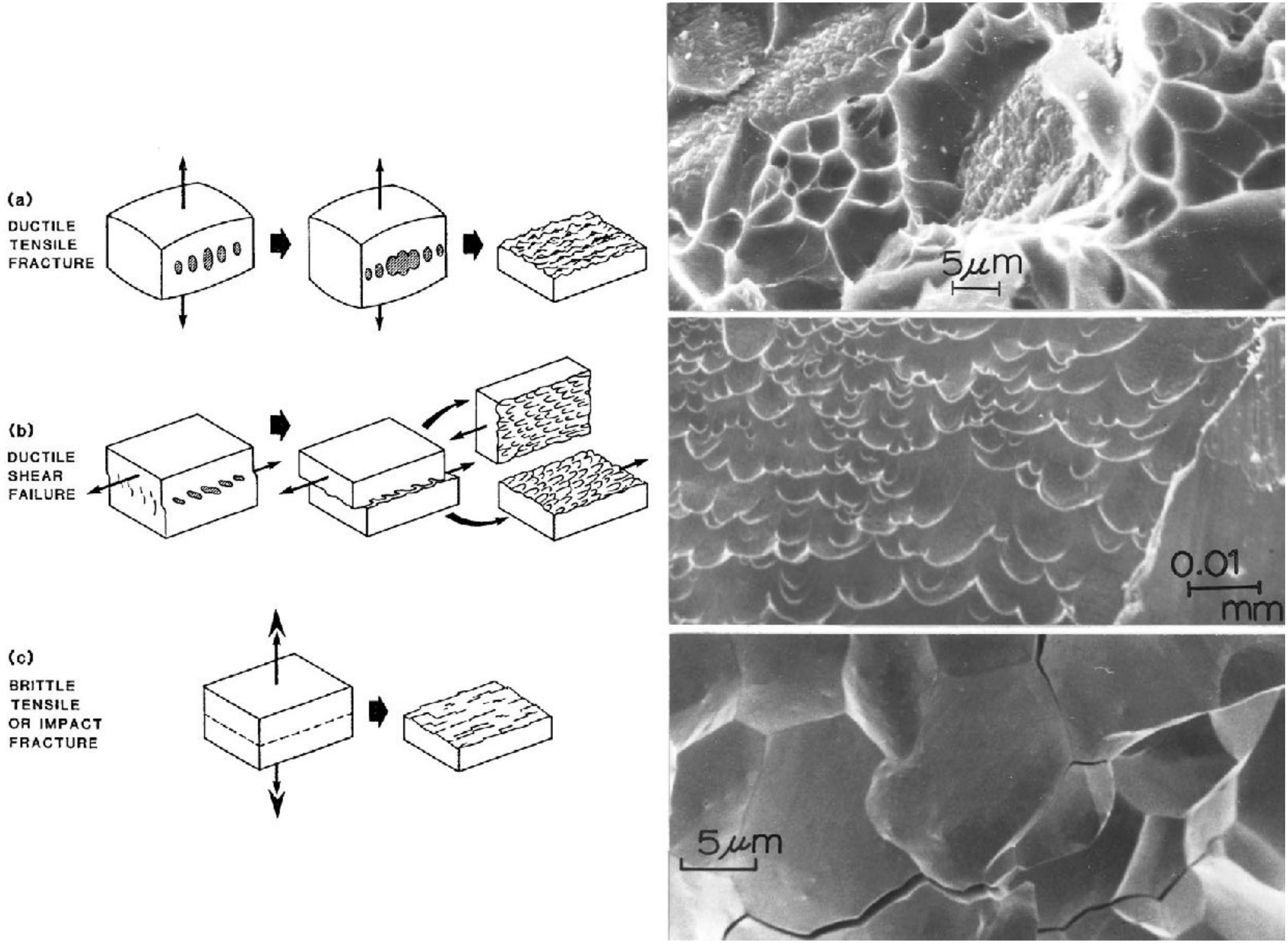


Figure 31 Schematic representation of common fracture modes along with corresponding fractography images of fracture surfaces observed in the scanning electron microscope (a) Ductile fracture in stainless steel. (b) Shear of Ti–Mo wire (c) Intergranular fracture in Ir. (From Ref. 6.)

Table 1 Standard Electrode Potentials (Electromotive Series) at 25°C

Element		Electrode potential [in volts versus the standard hydrogen electrode (SHE)]
Au		+ 1.498
Pt		+ 1.200
Ag	(Less corrosive)	+ 0.799
Hg	Cathodic	+ 0.788
Cu	↑	+ 0.337
H_2		0.000
Pb	↓	−0.126
Sn	Anodic	−0.136
Ni	(More corrosive)	−0.250
Co		−0.277
Cd		−0.403
Fe		−0.440
Cr		−0.744
Zn		−0.763
Al		−1.662
Mg		−2.363

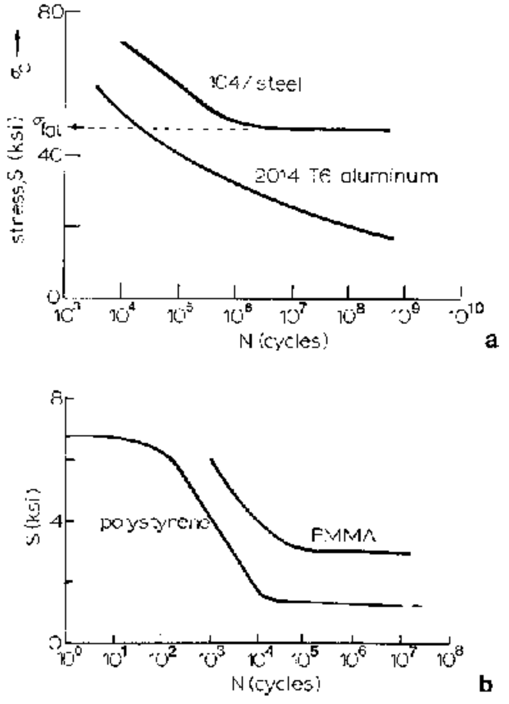


Figure 32 Stress amplitude (S) versus number of cycles to failure (N) for some common metals (a) and plastics (b). (After Refs 7–9.)

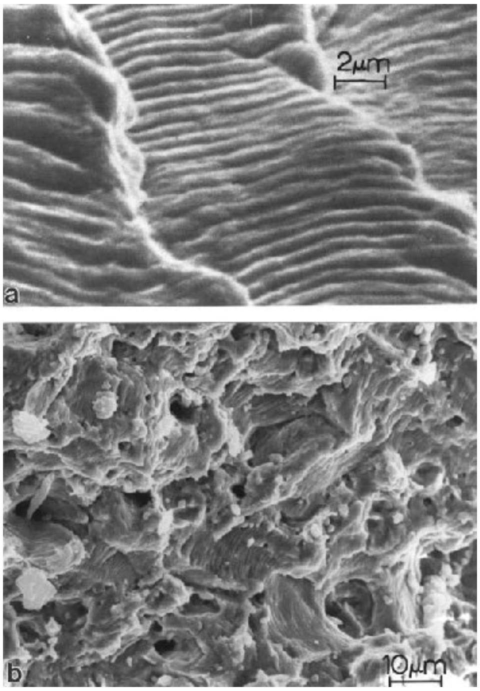


Figure 33 SEM fractographs showing fatigue striations on the fracture surface for copper (a) and a stainless–steel screw (b). The stainless–steel screw surface also shows intermixed ductile features as a consequence of failure in a prosthetic leg bone implant. This represents relatively low cycle fatigue. (From Ref. 6)

$$2H^+ + 2\bar{e} \rightarrow H_2 \qquad \text{(hydrogen ions reduced to hydrogen gas)} \qquad (17)$$

Correspondingly, metals which are less reactive than hydrogen are assigned positive potentials and are considered to be cathodic to hydrogen, i.e.,

$$M^{n+} + n\bar{e} \rightarrow M \qquad \text{(metal ions reduced to atoms)} \qquad (18)$$

$$H_2 \rightarrow 2H^+ + 2\bar{e} \qquad \text{(hydrogen gas oxidized to hydrogen ions)} \qquad (19)$$

As a consequence of these potential differences, it can be observed in Table 1 that if two dissimilar metals are coupled electrically or immersed in an electrolyte, the more negative metal halt-cell reaction will be oxidized, while the more electropositive metal will be reduced, and the overall electrochemical potential for the cell will be the sum of the half-cell potentials: a galvanic cell of Zn (−0.763 V) and Cu (+0.337 V) will produce an overall potential of −1.1 V (because the Cu half-cell reduction requires a sign change, i.e., −0.763–0.337). In a galvanic couple, the electrode which is oxidized is effectively the anode while the reduced electrode is correspondingly the cathode.

This proximity of electrochemically different metals, for example, can be utilized in fundamental ways to extract or deposit more electropositive metals from ionic solutions, or to protect one metal as another is sacrificed. For example, copper is far more electropositive than metals such as iron or aluminum. Consequently, copper in solution will deposit on Fe or Al: autodeposition. Prior to about 1980, this process, called *cementation*, utilized iron scrap to extract copper from copper sulfate solutions in many mining operations in the United States. Since aluminum has an intrinsic, and very tenacious, thin oxide on its surface, copper sulfate solutions require a small amount of chloride ion ($\sim$ 75 ppm) to react with the oxide layer and expose the aluminum surface for autodeposition of copper from solution. The coating of steel sheet with zinc to produce galvanized steel is another common example.

Corrosion control can also be achieved by similar methods called *cathodic protection* where electrons are supplied to a metal structure to be protected: galvanic coupling with a more anodic metal, i.e., more anodic than the one being protected. Underground steel tanks and especially buried pipeline steels are cathodically protected by connecting them to a sacrificial anode of zinc or magnesium, for example (see Table 1). Correspondingly, tanks and pipes can also be connected to a d.c. source through an anode to form a simple circuit where electrons are supplied to the steel to suppress its corrosion. This process can also involve so-called anodic protection where passive films are formed on metals by externally impressed anodic currents. Corrosion control can also be achieved by various coatings, such as paints or plastics of course. However, if these crack or peel, the exposed surface may then corrode.

The electrochemical differences characteristic of coupled metals is also a feature of even more fundamental differences in local stoichiometries in a huge array of materials systems and compositions. For example, precipitates within a material or segregation or depletion of selective elements in or near grain boundaries can produce local anodic/cathodic couples which promote accelerated corrosion reactions (refer to Fig. 21, for example). In addition, the creation of reaction products within a corroding and separating grain boundary can create enormous stresses which can accelerate boundary separation. Figure 34 illustrates an example of a related corrosion phenomenon called *exfoliation corrosion*, very common in a variety of rolled aluminum alloys where very elongated grains occur in the rolling direction. Similar phenomena can occur in general when (tensile) stresses applied to a corroding system accelerate crack advance or crack opening, thereby accelerating the corrosion process overall. This latter phenomenon is referred to as *stress corrosion cracking* (SCC).

Deformation can often enhance corrosion in general. For example, sensitization of grain boundaries can often be accelerated by stress (or strain) because dislocations produced as a consequence enhance the diffusion of elements to or from the grain boundary. Figure 21 illustrates this phenomenon. This is an important issue because manufactured products which retain a large amount of stored energy, even stainless steel, may be far more prone to corrosion in some environments than similar products which have been annealed or otherwise treated to reduce dislocation density, etc. Welded materials can also be prone to various corrosion phenomena because of stoichiometric variations in the heat-affected zone.

Of course when oxide layers or other corrosion products on a material surface are eroded away or removed by mechanical contact with the surface, there is an acceleration of the corrosion reactions. *Erosion corrosion* can occur in systems transporting fine particle slurries or in extremely fast fluid flow. This is especially troublesome at elbows or bends in pipes transporting slurries, etc.

In addition, erosive wear without corrosion can also compromise material performance along with other material surface removal through sliding contact wear. Wear in a general sense accounts for an enormous amount of material degradation, even in the absence of any significant corrosive phenomena, including oxidation. Usually very hard materials wear softer materials, and of course lubricants separating surfaces will retard contact wear phenomena. The application of very hard coatings to softer materials, is also often utilized to retard wear. TiN and TiC sputter-deposited coatings to cylinder walls and other engine parts is a typical example.

6.5 CLOSURE

Materials science and engineering is concerned with the search for basic materials knowledge and its application in the production of commodity products and manufactured goods. The three main types of materials—metals and alloys, plastics (polymers), and

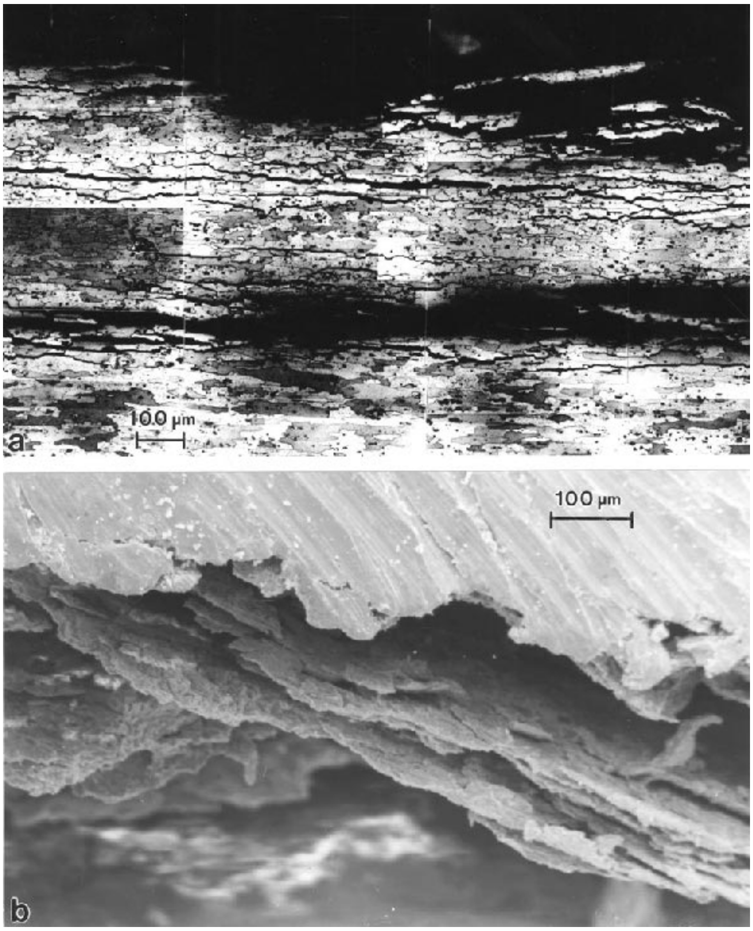


Figure 34 Example of exfoliation corrosion in a 2024 aluminum alloy sheet sample from a section of a KC-135 air cargo plane body skin. (a) Optical microscope image of section showing elongated grain structure and the intergranular corrosion which characterizes this phenomenon. (b) Scanning electron microscope image. (Photographs courtesy of Maria Posada, UTEP.)

ceramics have been the thrust of this brief overview, while composites and other materials issues have been discussed briefly. Lacking from this overview have been topics involving magnetic materials—ferromagnetism, paramagnetism, etc., elaboration of optical properties of materials, and electronic materials, particularly details relating to device or integrated circuit fabrication. Moreover, this introduction to materials science and engineering has included variations of fundamental issues which were thought to be more important in an industrial or manufacturing context than in a broader materials context. While detailed materials selection strategies and materials design criteria have not been covered in any specific way, it is hoped that this presentation will have provided sufficient substance to guide the industrial or manufacturing engineer toward appropriate selections and adequate understanding of contemporary and new materials. In addition, it is hoped that the presentation has evoked some sensitivities of the interrelationships between materials, manufacturing strategies, energy requirements, and environmental issues either relating to the materials themselves, or the consequences of their manufacture.

The reader might test his or her basic appreciation of the role that materials science and engineering play in the industrial context by referring in retrospect and in summary to Figs. 2 and 5. While at the risk of sounding fictitious, these two schematics represent the essence of the structure–properties–processing–performance “tetrahedron” which broadly characterizes materials science and engineering This model can be applied to essentially all manufactured materials components and products.

The Further Reading list will provide some additional insight into areas described in this chapter and in the pursuit of more specific materials issues which have not been addressed. The reader is encouraged to peruse these recommended readings where necessary and appropriate.

ACKNOWLEDGMENT

This work was supported by a Mr. and Mrs. MacIntosh Murchison Endowed Chair at The University of Texas at El Paso.

REFERENCES

1. J Marin. Mechanical Behaviour of Engineering Materials. New York: Prentice-Hall, 1962, p 24.
2. LE Murr, CS Hiou, S Pappu, JM Rivas and SA Quinones. Physica Status Solidi (a), 253: 49, 1995.
3. LE Murr. Interfacial Phenomena in Metals and Alloys. Reading, MA: Addison-Wesley, 1975.
4. Metals Handbook. 8th ed. vol 8. Materials Park, OH: ASM International, 1973.
5. LE Murr. Solid-State Electronics. New York: Marcel Dekker, 1978.
6. LE Murr. What Every Engineer Should Know About: Material and Component Failure, Failure Analysis and Litigation. New York: Marcel Dekker, 1987.
7. HW Hayden, WG Moffat anf J Wolff. In: The Structure and Properties of Materials, vol III. New York: Wiley, 1965, p 15.
8. P Beardmore, S Rabinowitz. Treat Mater Sci Technol 6: 267, 1975.
9. MN Riddell, GP Koo, and JL O'Toole. Polymer Engng Sci 6, 363, 1966.

FURTHER READING

Barsoum MW. Fundamentals of Ceramics. New York: McGraw-Hill, 1997.

Berkowitz AE, Kneller E, eds. Magnetism and Metallurgy. 2 vol. New York: Academic Press, 1969.

Ceramics and Glasses, vol 4. Engineered Materials Handbook. Materials Park, OH: ASM International, 1991.

Chawla KK. Composite Materials. New York: Springer-Verlag, 1987.

Courtney TH. Mechanical Behavior of Materials. New York: McGraw-Hill, 1990.

Deboo GJ, Burrows CH. Integrated Circuits and Semiconductor Devices: Theory and Applications. 2nd ed. New York: McGraw-Hill, 1977.

Fontana MG. Corrosion Engineering. 3rd ed. New York: McGraw-Hill, 1986.

Hatfield MH, Miller JH. High Temperature Superconducting Materials. New York: Marcel Dekker, 1988.

Heck C. Magnetic Materials and Their Applications. New York: Crane-Russak, 1974.

Hirth JP, Lothe J. Theory of Dislocations. 2nd ed. New York: Wiley, 1982.

Massalski TB. Binary Alloy Phase Diagrams. Materials Park, OH: ASM International, 1986.

Meyers MA, Chawla KK. Mechanical Metallurgy: Principles and Applications, Englewood Cliffs, NJ: Prentice-Hall, 1984.

Moore GT, Kline DE. Properties and Processing of Polymers for Engineers. New York: Prentice-Hall, 1984.

Murr LE, ed. Shock Waves for Industrial Applications. Park Ridge, NJ: Noyes Publications, 1988.

Murr LE. Electron and Ion Microscopy and Microanalysis: Principles and Applications. 2nd ed. New York: Marcel Dekker, 1991.

Rose-Innes C, Rhoderick EG. Introduction to Superconductivity. New York: Pergamon Press, 1976.

Smith WF. Structure and Properties of Engineering Alloys. 2nd ed. New York: McGraw-Hill, 1993.

Smith WF. Principles of Materials Science and Engineering. 3rd ed. New York: McGraw-Hill, 1996.

Sze SM. VLSI Technology. 2nd ed. New York: McGraw-Hill, 1988.

Chapter 6.7

Forming and Shaping Processes

Shivakumar Raman
University of Oklahoma, Norman, Oklahoma

7.1 INTRODUCTION

Finished products can be fabricated from a raw material or unfinished state using combinations of various individual manufacturing processes. These processes include casting, forming, machining, and joining. Forming processes are regarded as mass-conserving processes and serve as commercial alternatives to material removal processes. They differ from casting and foundry processes in that they do not undergo a state transformation from liquid to solid. Forming processes have existed for several years, evolving from a traditional blacksmith's art into a sophisticated computer-controlled precision process. Many of the forming operations such as coining and flashless forging can be considered to be ''near-net-shape'' or even ''net-shape'' operations, whereby little postfinishing operations are required after the application of the principal process.

From a processing standpoint, the majority of forming processes employ a combination of temperatures and pressures to process a raw material into a finished product. Hence, the materials are subjected to varying levels of pressures and various states of stresses during processing. Material processing design considerations in forming are significantly guided by standard mechanics formulations and handbook-based data that document material testing results. The replication and simplification of the material deformation characteristics in forming has been the subject of much work in the last 50 years.

7.2 PLASTIC DEFORMATION FUNDAMENTALS

It is important to consider the terminology associated with forming and hence, a simple revision of fundamentals is in order. When a material is subjected to tension in a tension test, it undergoes deformation. This deformation is elastic in the initial stages of load application, in that the relationship between the load and elongation is linear. The material recovers to its initial size upon release of the load. Beyond the yield point, this elongation becomes plastic. Deformation is however, uniform until the ultimate tensile strength (UTS) is reached, after which the material begins to neck. Nonuniform deformation extends until the ultimate fracture of the specimen. The extent of each region as well as the stress levels experienced are dependent on the material being tested. The material's ability to undergo significant deformation prior to fracture describes its toughness and ductility. While ductility is measured in terms of the engineering strain at failure, toughness can be described as the area under the true-stress–true-strain curve. The material's behavior in the elastic region impacts its stiffness.

Hardness of many materials can be related to their tensile yield stress or ultimate stress and reflect on their toughness and ductility. It should hence come as no surprise that most hard and brittle materials cannot be formed at room temperature. Temperature has a softening effect on materials and improves the ductility of most metals and alloys. Hence, softer ductile materials readily lend themselves to forming and shaping. Most brittle materials are considered to behave better in compression than in tension. Compression tests can be used to determine properties during compression, while the torsion test is common for determining properties in shear. Although these tests highlight particular properties, the prediction of the three-dimensional stress state, stress concentrations, etc., in forming is nontrivial. This is further compounded by impact and shock conditions (experienced during the use of hammers in forging) as well as cyclic loads that lead to mechanical and thermal fatigue.

Plastic deformation on a microscopic scale can be caused by slip or twinning. Slip in single crystals can be caused by the action of a shear stress and twinning is a complex mechanism whereby twins within certain materials tend to form abruptly during processing. Defects, such as vacancies, impurities, grain boundaries, and edge and screw dislocations, cause imperfections in the crystal lattice, causing the material's strength to deviate from theoretical calculations. A plane containing a dislocation requires a lower shear stress to initiate slip. On the other hand, the tangling of dislocations moving in different directions increase the shear stress to cause deformation. This higher stress state increases a metal's strength and is called work hardening. Plastic deformation at room temperature for most metals can lead to work hardening with the exception of materials that recrystallize close to room temperature. It is to be also noted that the strain hardening index of a material reduces with increasing temperature.

7.3 HOT AND COLD WORKING

As mentioned earlier, a combination of pressures and temperatures are applied to materials to change their shape in forming operations. The processing of materials heated beyond their recrystallization temperature (about 60% of their melting point) is termed as hot working. Cold working is typically conducted close to the room temperature (up to 30% of the melting point of the material). Warm working is a term applied to the processing conducted at intermediate temperatures. Cold-working processes are characterized by the application of high mechanical stresses for shape change, while hot working utilizes higher temperatures to reduce the mechanical loads required.

The grain structure is significantly impacted by the temperatures and pressures experienced during forming and hence, for the accurate control of properties the behavior of the material must be fully understood. Hot working is used in many instances to improve the mechanical properties and the grain structure of a cast material, by converting it into a wrought structure. Cold working tends to promote work hardening in many metals and is used as such for improving the strength of the metal in many applications. Depending on whether a bulk deformation or surface deformation is applied to the metal, the relevant areas may be hardened using cold working. Another advantage of the cold-working processes is their adherence to tighter tolerances and surface finish requirements. In fact, the tolerances produced in processes such as precision forging matches the tolerances produced by many machining processes. In such applications forming may be the preferred alternative, owing to their lower costs and their ability to produce uniform properties. Hot working, on the other hand, affords poorer dimensional stability and tolerances. Depending on the temperature of heating, various degrees of scaling, flaking, and flash are also experienced in many applications.

7.4 ROLLING OPERATIONS

The basic rolling process; termed flat rolling or simply rolling, relies on the application of compressive stresses on a strip to reduce its thickness. Because of this processes' ability to prepare sheet metal for subsequent sheet metal operations, rolling is considered an extremely important process in manufacturing. The rolling process is also regarded as a semicontinuous process whereby a finite-sized continuous sheet or rod is produced that may be cut into an appropriate number of pieces. Aside from the standard flat-rolling process, several shape-rolling operations are available that have the ability to produce a wide variety of shapes, sizes and forms. For example, thread rolling is used as a commercial alternative to machining in the production of threads. Shape rolling is used to produce a wide range of cross-sections, such as I-sections and C-sections.

A typical rolling operation is illustrated in Fig. 1. As is illustrated, two rolls are used in a two-high mill configuration, separated by a small gap. This gap provides for the insertion of the strip, which is ‘‘scooped’’ in by friction with the moving rolls. The thickness reduction achieved depends on the roll diameter and the coefficient of friction between roll and strip. In addition, the material and the geometry of the strip and the temperature during processing can affect the

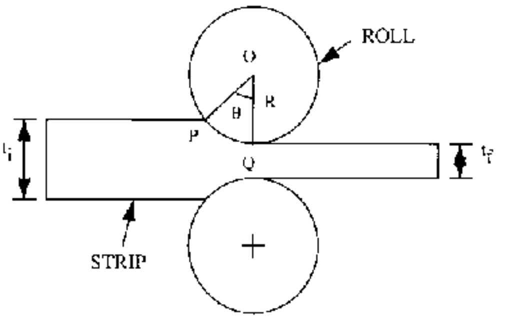


Figure 1 Schematic of the rolling process.

quality of the thickness reduction. For instance, wider strips of small thickness experience little to no spreading (in the width direction) and all the thickness reduction is reflected as a length change. In square cross-sections and cross-sections involving lower width-to-thickness ratios, significant spreading may be experienced.

The elastic deformation of the rolls and their drives can lead to flattening of rolls during rolling while the temperatures in hot rolling can cause nonuniform expansion of the rolls leading to a thermal camber. Thermal camber makes the roll diameter larger at the middle than at the corners. Other roll defects include zipper cracks, and edge cracks.

Friction is essential for rolling, but, excessive friction is undesirable. Lubricants are hence used to reduce friction and roll wear. Common roll materials are forged steels and cast steels and rolls are often ground and polished for cold-working applications. The typical strip materials include aluminum alloys, copper alloys, and ferrous metals. Lubricants commonly used in cold rolling are fatty acids and mineral oils and hot rolling employs graphite and compounded oils among others. The speed of the strip at entry into the roll gap is lower than the speed at exit, analogous to flow of fluids in converging channels. The roll speed typically lies in between the strip speeds, creating a no-slip or neutral point approximately midway in the roll gap.

The roll forces depend on the area of contact between the roll and strip and the average flow stress of the strip in the roll gap. The power consumed in the rolling process depends on the speed of rolling and the roll forces. Hence, the accurate determination of roll forces is basic to the calculation of power. Moreover, the roll forces affect the design of the drive system and the rolls themselves. Small rolls have a lower contact area and lower forces, while larger rolls result in larger forces. Larger forces also tend to deform the rolls more and for this reason, small rolls may be backed by larger rolls, thus reducing the contact areas. This is common in cluster mill configurations. Large thickness reductions may hence be achieved without increasing the forces in this configuration. Tandem rolling, where a series of rolls with successively reducing roll gaps are placed in tandem, is also common for achieving large thickness reductions, stepwise. Three-high and four-high mills are some other rolling configurations.

7.5 FORGING OPERATIONS

Forging is an old but versatile operation that has been used to shape materials for several centuries. The modern practice of forging is, however, sophisticated in large-scale production, utilizing computer-controlled automated machinery for the accurate production of fine details on parts. In forging, complicated shapes may be produced by the application of pressure, gradually or suddenly, within or without dies.

Hammers are used for sudden application of loads, while presses are employed for gradual pressure application. The type of energy supplied such as pneumatic, hydraulic, or electromechanical power also influences the pressures generated in forging. Hydraulic equipment are used for larger power applications and smoother operations, while mechanical units are employed for higher production rate and ease of automation.

The die material selection is affected by the choice of loading. It is common to employ die (alloy) steels that contain nickel, chromium, molybdenum, and vanadium, for the forging of many metals and alloys. Aluminum, magnesium, and copper alloys are more easily forged (and at lower temperatures) than steels, stainless steels, and titanium alloys. Lubrication, as in rolling, is applied to reduce friction and die wear. Moreover, in forging, it also serves as a parting agent between the dies and the workmaterial. In hot forging. solid-state lubricants such as glass and graphite are preferred, while in cold forging, soaps and mineral oils are commonly used.

In hot forging, it is also advantageous to preheat the dies to avoid uneven thermal contraction and consequent thermal stresses. Scales caused by oxidation in hot forging can also create nonuniform stresses upon deformation as well as flaking. Hence, hot forging is seldom a precision process and tight tolerances and tight finishes are more common in cold forging.

In open-die forging a workpiece is typically placed in an anvil (with optional die cavities) and the shape changed with pressure applied through the action of a plunger (hammer or press). Rapid changes in shape is not typical in open-die forging. Examples include small runs of sharp-edge break-up in the change of square cross-sections to round ones.

Closed-die forging (Fig. 2) is performed within a die cavity through the application of load by a punch. Wherever possible, large changes in shapes are not achieved in a single step. Instead, the processing is broken down into intermediate activities, such as edging, where material is drawn into a localized area, and fullering, where material is distributed away from an area. Blocking is another intermediate operation that is essentially a prefinish-forging stage. Large changes in shapes require significant pressures that can cause breakage in the work material and the dies. Insufficient pressures lead to the improper filling of the die cavities. In each case the determination of the forces and the areas of deformation is very critical. The determination of forces is also important to the calculation of the power consumed. In flashless forging, little to no squeezing of material between the die halves or flash is experienced. Precision forging is an advanced type of net-shape closed-die forging operation whereby high pressures are used to replicate very fine details on parts that require little postprocessing. Operations such as heading and coining are also basic variations of the common closed-die forging process. The coining operation is characterized by the fine detailing and excellent surface finish obtained. Heading is one of very few manufacturing operations used for increasing the cross-section of a workpart in solid state. Bolt and screw heads are often processed this way, without material wastage. Roll forging combines the rolling and forging operations in a simple, but, unique way for part processing.

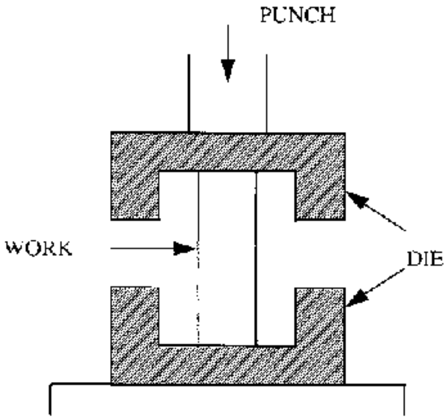


Figure 2 Schematic of the closed-die forging process.

7.6 EXTRUSION

Extrusion is the squeezing of metal parts using a plunger, through die cavities, to reduce the cross-section of parts. The extrusion process is conducted in an extrusion chamber that houses the work material. The shape of the part extruded depends on the shape of the extrusion die. A wide variety of cross-sections can be extruded using this basic process. Extrusion is termed direct if the directions of the plunger action and the flow of the extruded product are the same. It is called indirect when the plunger acts in an opposite direction to the direction of flow of the extruded product. The extrusion ratio is the ratio of the old cross-section to the extruded cross-section of the part. The ratio of the perimeter to cross-sectional area of the extruded product is another variable in extrusion and is termed the shape factor.

As can be imagined, significant friction is experienced by the work material at the walls and the corners of the extrusion chamber. This friction tends to slow the operation and higher forces must be applied to overcome this friction. One way of reducing this friction is to apply lubrication. Another method is to perform hydrostatic extrusion, where an incompressible fluid medium is used to transmit the pressures applied by the plunger to the work material.

In direct extrusion (Fig. 3), the improper design of the chamber and die can result in poor extrusions through the formation of a dead zone. Providing for a gradual flow of the metal into the die cavity by supplying a die angle reduces the tendency to form the dead zone. The dead zone prevents proper flow of metal through the die and also alters the finish of the extruded product. The normal method of scribing horizontal and vertical lines in specimens is used to understand the deformation and flow of metal through the die experienced in extrusion. Aluminum and its alloys are extruded easily into a wide variety of shapes for applications that include door handles and decorative items. Extrusion of ferrous metals is accomplished using harder dies.

Drawing is very similar to extrusion in that a material is pulled through a die to reduce its cross-section. The angles and lands are somewhat different as are the

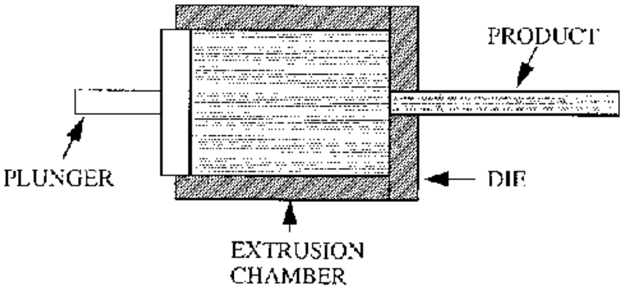


Figure 3 Schematic of the extrusion process.

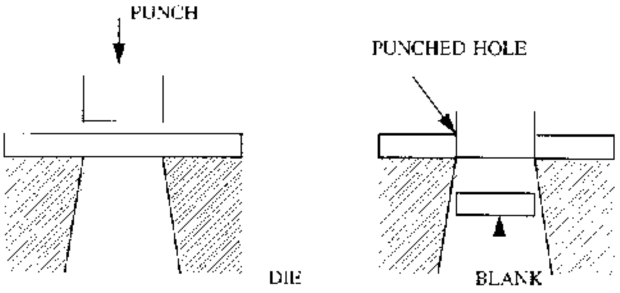


Figure 4 Schematic of the shearing process.

pressures applied. Fine drawing of wire is often accomplished through diamond dies. However, many applications also employ coated tool steels and carbides for the die material. Lubrication is used in many cases to reduce friction and power consumption.

7.7 SHEET-METAL OPERATIONS

Sheet-metal operations employ very similar principles of pressure application as do bulk material operations discussed so far. However, temperature application is not relevant in most cases. Since small thicknesses and cross-sections are handled, extreme caution is required in the application of forces.

Shearing is an initial sheet-metal operation used to separate sheet-metal blanks from semicontinuous sheet. Subsequent operations include bending, spinning, stretch forming, and deep drawing. Shearing operations (Fig 4) may be subclassified into blanking and punching. The punched hole is the workpart in punching and the discarded blank is wastage. In blanking, the blank is the part and the hole is discarded. The blanking and punching operations are performed in shearing presses by the action of a plunger and die. The plunger descends on the blank, separating it from the strip through the propagation of fracture in the gap between the die and plunger. The clearance between the die and the punch is critical to the quality of the sheared product. The larger this clearance and the more ductile the material in the gap, the greater the tendency for significant deformation prior to fracture. Smaller clearances and less ductile materials may in some situations lead to cleaner separation. However, generalized rules-of-thumb regarding blank and hole quality in terms of clearance are difficult to formulate.

Deep drawing (Fig. 5) is another process that utilizes a die and punch to pull the material into a cavity. However, this process tends to rely on the deformation prior to fracture rather than the fracture of material desired in shearing. Accordingly, smoother punches are employed with an edge radius that is replicated on the product. A sheared blank material stretches, conforming to the punch in its descent, and is eventually stripped by a stripper plate, preventing it from remaining attached to the punch in its ascent. The strip is subjected to a complex stress state during deep drawing. Often, the inner fibers are subjected to compression, while the outer fibers are in tension. The determination of the neutral axis is required for stress–strain calculations. Experimental techniques for determining strain at various points of the drawn cup include the scribing method discussed elsewhere in this chapter. Using these strains and data obtained from handbooks and material tests, the stresses at various points can be determined. Such calculations assist in the design of punches and dies as well as the drawn material's dimensional specifications.

Deep drawing has several applications and is most common in the manufacture of cooking utensils. A large range of ductile materials are deep drawn, including aluminum, copper, zinc, and steel. The prevention of cracks and wrinkles along bends is a chief concern of deep-drawing designers.

Bending is a common sheet-metal process achieved using a press brake. The ability of a material to undergo bending is characterized by its bend radius. The bend radius depends on the ductility of a material

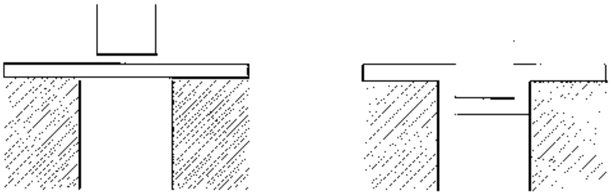


Figure 5 Schematic of the deep-drawing process.

and the sheet thickness. The outer fibers in bending are typically subjected to tension and the inner fibers are put in compression. This makes the determination of the neutral axis very important. Compensation must also be provided for springback or the elastic recovery after bending. Bent products abound in applications, automobile and aircraft bodies serving as principal examples. Casings and bodies for home airconditioners and refrigerators, and boxing for packaging and storage containers all illustrate bending at various angles and varying sheet thicknesses. Bending also serves in certain applications where the principal purpose is to increase the moment of inertia and stiffness of parts. Material is typically clamped and bent at the desired places with a die.

Stretching is another process that uses a form block to achieve material elongation along critical places. The material is typically held in clamps at ends and the die is moved on the material to stretch it. Spinning is used for making symmetrical parts using a process that resembles the making of clay pots. A rotating mandrel serves as the support for a sheet and a die is used to conform the sheet over the mandrel. The advantage of the process is that rather complicated shapes may be formed relatively easily.

Tube bending is another operation that employs bending principles to bend pipes and tubes along various angles. The compression-type tube bender uses fixed and movable rollers and a lever arm to achieve bending. The fixed roller is clamped to the base, while the movable roller is carried by a lever arm hinged at the fixed roller. When the lever arm is moved along an arc (its natural path), the movable roller pushes the tube being bent against the fixed roller, thus bending it. Provision is made on the two rollers to accommodate the tube. In order to prevent the wrinkling and buckling of tubes, tubes are often filled with sand.

Explosive forming is a process that uses a controlled explosion to form sheets into intricate shapes. A radioactive charge is detonated causing a shock wave that deforms the metal at high strain rates. The shock wave is typically carried by water that acts as the coupling drive, pushing the material against the die.

BIBLIOGRAPHY

Degarmo EP, Black JT, Kohser RA. Materials and Manufacturing Processes. 7th ed. Englewood Cliffs, NJ: Prentice-Hall, 1990.

Kalpakjian S. Manufacturing Engineering and Technology. 3rd ed. Reading, MA: Addison-Wesley, 1995.

Metals Handbook. 9th ed. Vol 14: Forming and Forging. Metals Park, OH: ASM International, 1988.

Wick C, Benedict JT and Veilleux RF, eds. Tool and Manufacturing Engineers' Handbook. 4th Ed. Vol 2: Forming. Dearborn, MI: Society of Manufacturing Engineers, 1984.

Chapter 6.8

Molding Processes

Avraam I. Isayev
The University of Akron, Akron, Ohio

8.1 INTRODUCTION

Molding processes are the major periodic manufacturing operations in the polymer industry. Enormous amounts of plastic, rubber, and thermoset parts ranging from automobile bumpers and personal computer and refrigerator housings to bottles and tires are produced by molding processes. The automotive, appliance, computing, beverage, and tire industries are deeply associated with molding. The molding of plastics alone is an industry of enormous volume and scope. There are many variations of molding technology. These processes include compression, injection, injection–compression, coinjection, transfer, resin transfer, blow, rotational molding, and thermoforming. Several books of a general nature are available describing various polymer processing operations including molding processes [1–12]. The present chapter briefly describes these various processes. It also gives an up-to-date overview of theoretical and experimental investigations conducted in the molding area by different research groups.

8.2 COMPRESSION MOLDING

8.2.1 Technology

Compression molding is one of the oldest techniques for manufacturing rubber, thermoset, and plastic products. Compression molding dates back to the origin of the rubber industry [13]. For many years, this has been the standard technique for molding, but recently it has been replaced by injection molding to some extent. By comparison, injection molding offers advantages in material handling and ease of automation. However, compression molding retains a distinct advantage when processing reinforced polymers [14]. Moderate flow during compression molding helps to avoid high stresses and strains, therefore, reinforcing fibers are not damaged by the flow during mold filling. Thus, high concentrations of reinforcing fibers and long fibers can be incorporated into composite materials. A number of handbooks have been written describing the compression molding process [15,16].

Compression molding basically involves the pressing (squeezing) of a deformable material charge between two halves of a heated mold to fill and cure the material in the mold, and subsequent part removal (Fig. 1). In the manufacturing of thermoset products, transformation of flowable material into a solid product under the effect of the elevated mold temperature takes place. In this case, compression molding temperatures range from 140°C to 200°C. Mold pressures can vary from about 20 bar to 700 bar and curing times can vary from about 1 min for thin parts to over 1 hr for very thick rubber parts. Recently, the development of thermoplastic matrix composites to produce strong, lightweight structures has increased interest in compression molding. In thermoplastic matrix composite molding, temperatures as high as 350°C are utilized [17,18].

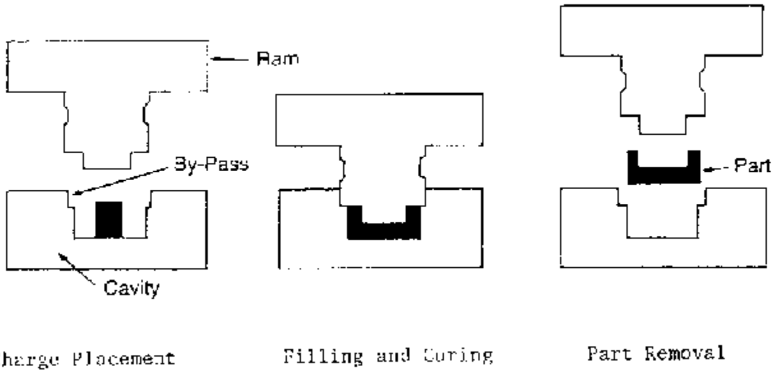


Figure 1 Schematic representation of compression molding process.

Depending on mold geometry, compression molding can be divided into flash, positive, and semipositive molding [9]. Compression molding is carried out using compression molding presses. Two types of presses are used—downstroking and upstroking. Molds usually operate using a clamping ram or cylinder with clamping capacities ranging from a few tons to several thousand tons. In addition to the clamping capacity, two other characteristics of the press are: the amount of daylight characterizing maximum platen separation, associated with stroke, and the platen size, ranging from a few centimeters to several meters. The temperature of the platens are controlled by built in heating or cooling elements or by separate heaters.

There are five stages of the compression molding process: (1) material preparation; (2) prefill heating; (3) mold filling; (4) in-mold curing; and (5) part removal. Material preparation includes compounding a resin with fillers, fibers, and other ingredients, or impregnating a reinforcing cloth or fibers with a resin. This stage controls the rheology of material and the bonding between fibers and resin. The prefill heating stage is carried out to speed up the molding process. This stage can occur outside or inside the mold before the mold is closed and flow begins. The mold filling starts with the material flow and ends when the mold is full. The effect of flow is critical to the quality and the performance of the molded product. It controls the orientation of fibers, which has a direct effect on the mechanical properties of the part [14]. In processes involving lamination of the long fiber-reinforced composites, there is little flow, since the initial charge almost completely conforms to the mold [17,19]. In the case of a thermoset matrix, some curing may occur during the mold filling stage. The in-mold curing stage follows the mold filling. In this stage, the part is cured in the mold, while the final stage of cure may be completed during a postcure heating after the part removal. The in-mold curing converts the polymer from a liquid into a solid with the rigidity of the product sufficient for removal from the mold. Part removal and cool-down is the final stage. This stage plays an important roll in warpage of the part and the residual stress development, which arise due to the difference in thermal expansion in different portions of the part. The temperature distribution and rate of cooling affect these residual stresses.

Figure 2 shows a typical curve for the variation of the plunger force required for the mold closing as a function of time at a constant closing rate during

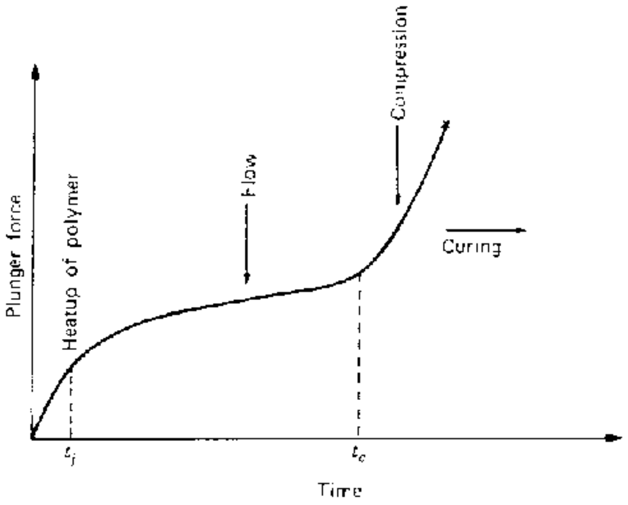


Figure 2 Schematic representation of the plunger force during compression molding at a constant mold closing speed.

molding of polymers not containing fibers [2]. In the first region at time $t < t_f$, corresponding to softening of the material, the force increases rapidly as the preform is squeezed and heated. At t_f the polymer is apparently in the molten state and is forced to flow into the cavity and fill it. Filling is completed at t_c, corresponding to the initiation of curing. At this stage, compression of the polymer melt occurs to compensate for the volume contraction due to curing. In the case of the sheet molding compounds (SMCs), a typical dependence of the force on time at various squeezing speeds shows the behavior depicted in Fig. 3 [20]. From comparing this figure with Fig. 2, it is seen that the squeezing force behavior in the presence of fibers is more complicated than that without fibers. The latter is due to void squeeze, breakage of molecular bonds, and fiberglass compaction.

Compression molds are generally made of tool steel. Cavities are often chrome plated. The mold should withstand high forces and pressures. The mold is heated by electric heaters, steam, or oil heating. Three mold configurations are utilized, typically referred to as flash molds, positive molds, and intermediate configurations [9]. Compression molds are often equipped with ejection systems. Mold dimensions should take into account cure and cooling shrinkage, typically ranging from 0.1% to 0.8%. In molding small parts, multiple cavity molds are usually utilized loaded through loading boards. In molding large parts, single cavity molds are typically used.

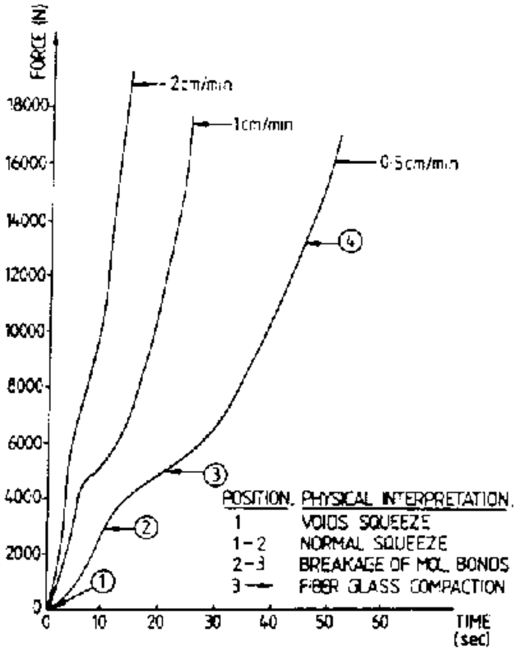


Figure 3 Squeezing force as a function of time for SMC squeezed between parallel disks. (From Ref. 20, courtesy of the Society of Plastics Engineers.)

For convenience of sizing and handling, preforms are typically used. The fibrous preforms are initially shaped to approximate the shape of the parts. The most widely used preforms are based on SMC, and bulk molding compound (BMC) [21]. The SMC contains resin, fibers, and other ingredients are prepared into a sheet form for easy loading into the mold. The BMC includes resin, fibers, catalyst, and other ingredients mixed into a puttylike mass which can be extruded into a ropelike form for easy handling and placement into the mold. Most recently, the high-performance thermoplastic and thermosetting polymers containing 60–70% by volume of continuous carbon fiber reinforcement are compression molded into structural composite parts for high-performance aerospace and industrial applications [17–19,22–24]. Matrix polymers for these composites are polyetheretherketone (PEEK), polyetherimide (PEI), polyarylene sulfide (PAS), polyphenylene sulfide (PPS), polyamideimide (PAI), polyethersulfone (PES), and thermoplastic or thermosetting polyimides (TPI). Preimpregnated prepregs, where the reinforcing carbon fibers are already embedded in the resin matrix, and postimpregnated prepregs, where resin and reinforcing carbon fibers are hybridized or cowoven, are utilized. The prepregs are laid up and compression molded into the laminates. In addition, a new technology for making self-reinforced or in-situ prepregs based on thermotropic liquid crystalline polymers (LCP)/thermoplastic (TP) has been recently proposed [25]. The LCP/TP prepregs are first made by extrusion followed by extension to achieve large aspect-ratio LCP fibrils in TP. Then, the prepregs are compression molded into laminates with the required packing sequence. Unidirectional or quasi-isotropic laminates can be obtained in a way very similar to conventional fiber-reinforced laminates.

8.2.2 Modeling

The main goal of modeling in compression molding is to predict filling patterns, pressure, stress and velocity distributions, orientation of fibers in the case of short-fiber reinforced composites, solidification or curing and residual stress development during process with relevance to mechanical properties of the molded products [26]. The majority of the simulations of compression molding is usually based on the lay-flat approximation due to the fact that the compression molded parts are typically long in planar dimensions in comparison with their lateral dimensions. Recently, the three-dimensional simulations have been attempted. In the case of the two-dimensional simulations, two approaches are used [14]. In the first case, a lubrication type model is utilized in which the shear stress terms in the planes through the thickness of the part dominates [14,21]. In the second case, a thin lubricating layer near the mold wall is assumed, leading to the absence of shear stresses in the thickness plane, with the shear and normal stresses in the plane of the part being dominant [21,27,28]. In the first case, the generalized Hele–Shaw flow model, originally utilized for the thermoplastic injection molding [29], is used, which allows one to combine the continuity and momentum equations into the following equation:

$$\frac{\partial}{\partial x}\left(S\,\frac{\partial P}{\partial x}\right)+\frac{\partial}{\partial y}\left(S\,\frac{\partial P}{\partial y}\right)=-\frac{dh}{dt} \tag{1}$$

where x and y are the planar co-ordinates, P is the pressure, dh/dt is the platen closing speed, and S is the fluidity which is determined by

$$S=\int_0^h \frac{(z-\lambda)^2}{\eta(z)}\,dz \tag{2}$$

where $\eta(z)$ is the viscosity variation in the thickness direction, z, h is the thickness of the part and λ is the value of z at which the shear stresses vanish. In the second case, the lubricating squeezing flow approximation where the charge slips along the mold surfaces is assumed. The following equations of motion are used:

$$\frac{\partial P}{\partial x}=\frac{\partial \tau_{xy}}{\partial x}+\frac{\partial \tau_{yx}}{\partial y} \tag{3}$$

$$\frac{\partial P}{\partial y}=\frac{\partial \tau_{xy}}{\partial x}+\frac{\partial \tau_{yy}}{\partial y} \tag{4}$$

where τ_{ij} are the shear and normal stress components. These equations are coupled with the following energy equation:

$$\rho C_p\left(\frac{\partial T}{\partial t}+u\frac{\partial T}{\partial x}+v\frac{\partial T}{\partial y}\right)=k\,\frac{\partial^2 T}{\partial T^2}+\eta\dot{\gamma}^2+\dot{Q} \tag{5}$$

where ρ is the density, C_p is the heat capacity, k is the thermal conductivity, u and v are the planar velocity in the x and y direction, respectively, $\dot{\gamma}$ is the strain rate, T is temperature, and t is time. The energy equation includes heat conduction and convection terms along with viscous dissipation, $\eta\dot{\gamma}^2$, and heat released due to chemical reaction, $\dot{Q}$. The last two terms in the energy equation require knowledge of the viscosity and reaction kinetics. Various reaction kinetic equations are utilized [30], In solving Eq. (1), information on the shear viscosity as a function of shear rate, temperature and state of cure is needed. In solving Eqs. (3), (4), and (5), information concerning the biaxial viscosity as a function of strain rate, temperature, and state of cure is needed. However, such information for both types of viscosity is not easily accessible for the fiber-reinforced composites. Therefore, some approximations are used to simplify this situation. In particular, the lubricated [31] and unlubricated [20] squeezing flow apparatuses have been utilized to measure the variation of the force as a function of time at various closing speeds during closing of two disks with a sample of composite material placed between them. Curves similar to those depicted in Fig. 3 can be obtained. By the best fit of the measured force-time data to a certain viscosity function, the material parameters can be obtained. However, it should be noted that the material properties, such as the viscosity, elastic modulus, strength, and elongation, are functions of the fiber orientation in the composite. The calculation of shrinkage and warpage is also dependent on the fiber orientation. This orientation is coupled with the flow field. Some approaches are proposed to calculate the fiber orientation. Most approaches in use today are based on Jeffery's equation [32] with an additional term included to represent the effect of fiber–fiber interactions on the orientational motion [33–37]. This term is in the form of an isotropic rotary diffusion, allowing interactions between fibers to cause randomization.

In the filling simulation of compression molding, the finite-element, finite-difference, and boundary-integral methods are used [14,21,26,38,39]. Compression molds are usually heated by means of channels drilled in the mold platens. The finite-element and boundary-integral methods are also used for thermal design of molds [40].

Figure 4 shows a typical example of the simulated mold filling pattern during compression molding of a car fender with a three-dimensional complex geometry [38]. The charge is initially placed in the central region as indicated in this figure by the contour number 1. The preheating time of the charge is 20 sec and the mold closing speed is 0.5 cm/sec. The calculated filling time is 5.05 sec. The initial charge area and location are found to be the most significant factors affecting the flow pattern. Thus, these factors must to be considered as important design variables for compression molding.

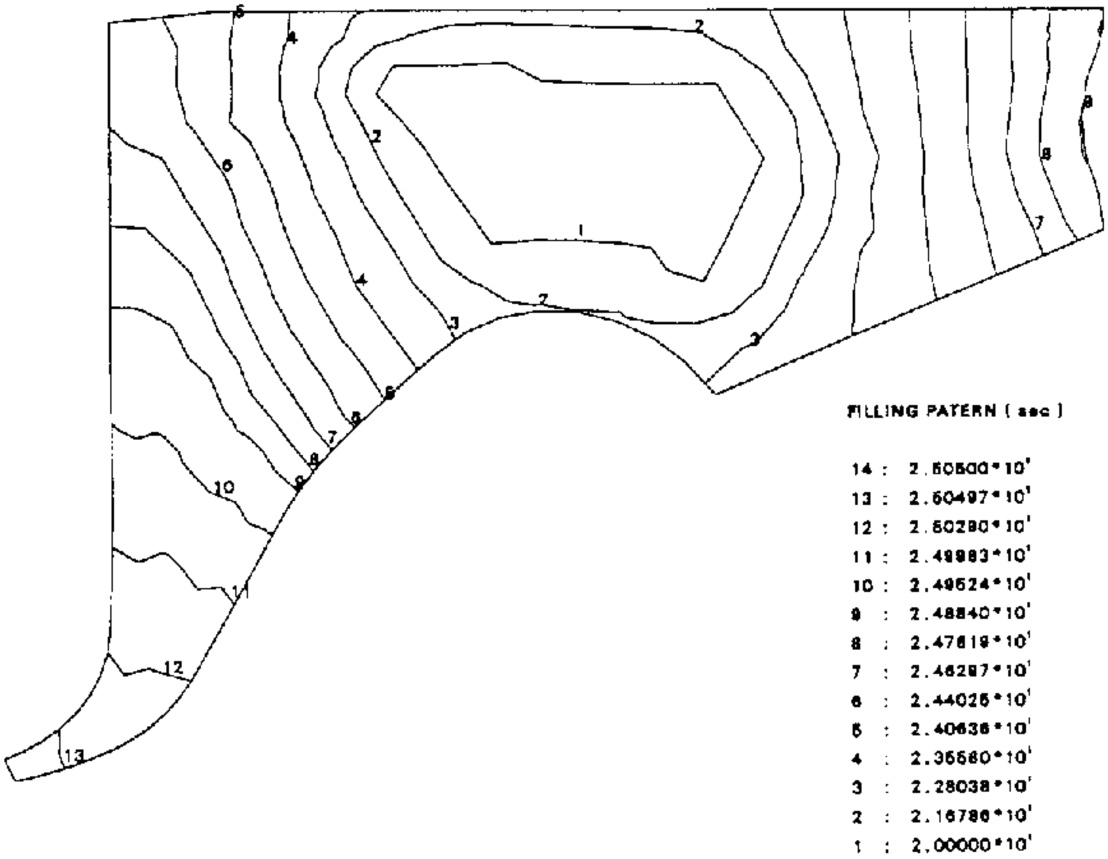


Figure 4 Melt front propagation with time during compression molding of a car fender (From Ref. 40, courtesy of the ASME.)

8.3 INJECTION MOLDING

8.3.1 Technology

Injection molding is one of the most widely employed molding processes. Injection molding is used for processing of thermoplastics, elastomers, thermosets, ceramics, and metals in order to make articles of varying complexity. Several books are available giving a brief or detailed description of the process [1–11,13,41–48]. In addition, various practical aspects of injection molding are described in a handbook [16]. The advantages of injection molding are high production rate, large-volume manufacturing with little or no finishing operations, minimal scrap, and good dimensional tolerances.

Injection molding of thermoplastics is defined as the automatic feeding of pellets into the hopper, melting, plasticating melt, and feeding melt into injection barrel at a temperature above the glass transition temperature, T_g, for amorphous polymers or melting point, T_m, for semicrystalline polymers. The melt is then injected through a delivery system consisting of a nozzle, sprue, runner system, and gate or gates into a mold having a temperature below T_g or T_m. The melt solidifies in the mold. Then, the mold is opened and the molded product is ejected.

Injection molding of elastomers is defined as the automatic feeding of a preheated or plasticated rubber stock into an injection barrel at a temperature below the vulcanization temperature [13,49,50]. Then, the rubber is injected through a delivery system into a mold. The mold temperature is kept high enough to initiate the vulcanization and subsequently vulcanize the rubber inside the mold. After the rubber has vulcanized, the mold is opened and the molded part is ejected.

Injection molding of thermosets and reactive fluids, which are capable of forming infusible crosslinked or network structures by irreversible chemical reaction, is also carried out using a hot mold. Reaction injection molding is characterized by in-mold polymerization from monomeric or oligomeric liquid components by a fast polymerization reaction [44,45]. Thermosets are solid or highly viscous materials at ambient temperature which are frequently highly filled.

An injection molding machine consists of a clamping unit containing the mold and the injection unit for feeding, melting, and metering the thermoplastic (Fig. 5). The most widely used injection units utilize a rotating screw to plasticize the material. Rotation of the screw causes the plasticized material to accumulate in front of the screw, which is pushed back. The material is injected by forward motion of the screw acting as a plunger which pushes the melt into the mold. The mold serves two functions, namely, it imparts shape to the melt and cools the injection molded part. The mold consists of the cavities and cores and the base in which they are located (Fig. 6). The mold contains one or more cavities with stationary and moving mold halves. Both the mold halves have cooling or heating channels or heaters. There are several types of injection molds [16,51]. These include the cold-runner two- and three-plate molds, the hot-runner molds, and the stacked molds. In the cold-runner molds the melt in delivery system solidifies together with that in the cavity. In the hot-runner mold, the melt in the delivery system is kept hot. The latter allows separation of the runner system from the part with melt in the runner used for the next shot. This is the reason that such molding process is usually called runnerless injection molding. In the stacked mold, a multiple two-plate mold with molds located one on top of the other allows one to double the output from the single machine having the same clamping force. In many cases, molds may have multiple cavities. The latter is dictated by the process economics. The multicavity mold may be arranged with balanced or unbalanced cavity layouts (Fig. 7). In the balanced cavity layout, the filling process in each cavity can be completed almost at the same time, which leads to uniformity of the molded products from each cavity.

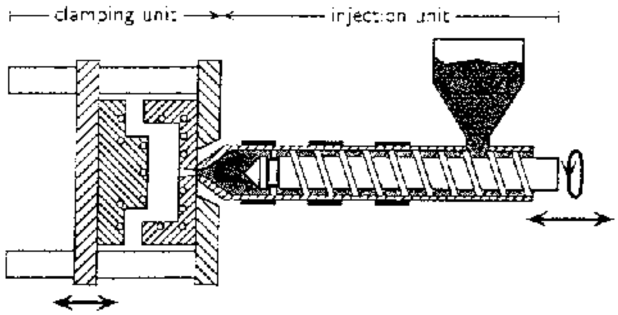


Figure 5 Schematic representation of an injection molding machine.

The connection between the runner and cavity is called the gate. In the mold making, the gate design is quite important. In addition, the size and the location of the gate is critical. The gate should allow the melt to fill the cavity and to deliver additional melt to prevent shrinkage due to cooling and should freeze at an appropriate time during molding cycle. The premature freezing will cause an undesirable phenomenon called underpacking, leading to the excessive shrinkage and sink marks. In addition, a mold requires a cooling or heating system and venting to remove air during the cavity filling and rapid and uniform cooling. Venting is usually achieved by arranging small gaps in the parting line, which allow air to escape quickly. In some cases, a forced removal of air is carried out by using vacuum venting. Mold cooling or heating is achieved by placing a number of channels in both halves of the mold through which the cooling or heating liquid flows to remove the heat from the melt or to add the heat to the melt. Mold heating is also done by placing electric cartridge heaters in the mold halves.

The injection molding cycle can be divided into three stages. These include cavity filling, packing (holding), and cooling. The three stages of molding cycle can be easily seen from Fig. 8, indicating schematically the pressure variation with time. In the filling stage, the pressure rises as the melt propagates into the cavity. This stage is followed by the packing stage where a rapid increase (typically within 0.1 s) in the pressure to its maximum is observed. Then, the cooling stage takes place at which the pressure slowly decays.

Molding variables such as injection speed, melt and mold temperatures, packing or holding pressure, and length of packing stage have a strong influence on the pressure development and properties of moldings. Frozen-in molecular orientation, residual stresses, polymer degradation, shrinkage, warpage, and weld line strength are influenced by the process variables [50,52–57]. In addition, in the case of injection molding of semicrystalline polymers, the molding variables strongly affect the crystallinity and microstructure development in moldings which in turn influence their performance characteristics [57–63].

Recently, a novel push–pull technology is proposed to increase the weld line strength in molded parts by imposing the oscillatory pressure on melt after the cavity filling which introduces additional movement of the melt at the weld line area [64,65]. The rheomolding technology based on melt vibration is also

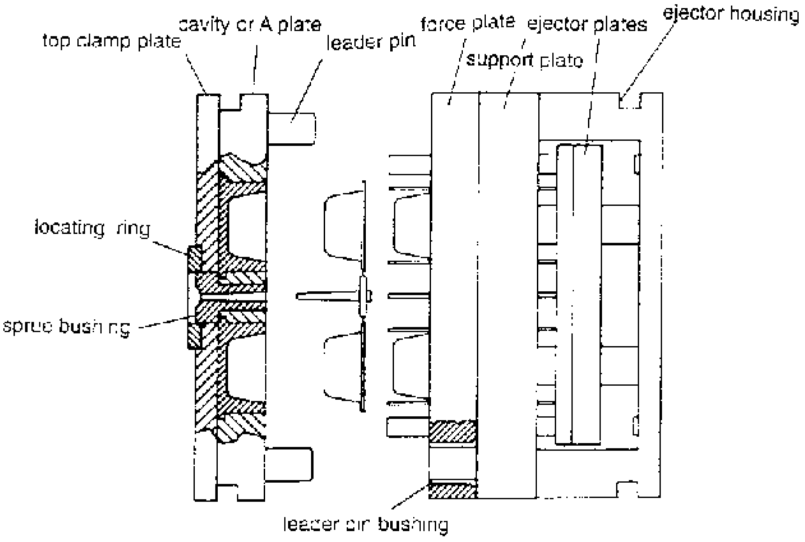


Figure 6 Schematic representation of a cold-runner, two-plate injection mold according to the DME.

proposed to control the rheology of a melt during its flow into the mold [66].

8.3.2 Modeling

In injection molding, very complicated three-dimensional parts are made. However, the planar dimensions of the molded parts are typically much larger than the cavity thickness. Thus, many available approaches for computer-aided design of molds are based on the lay-flat approximation by which the part can be represented as a two-dimensional object. In addition, recent attempts have been made to consider a three-dimensional cavity filling [67]. For two-dimensional flow in the mold the Hele–Shaw flow approximation is applied, leading to a single governing equation which is the combination of the equation of motion and continuity as [29,50,68–70]

$$\frac{\partial}{\partial x}\left(S\,\frac{\partial P}{\partial x}\right)+\frac{\partial}{\partial y}\left(S\,\frac{\partial P}{\partial y}\right)=0 \tag{6}$$

with S being the fluidity, which for strip flow is defined by

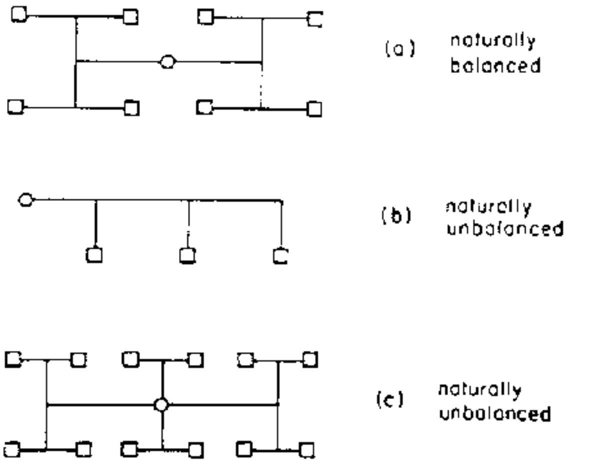


Figure 7 Naturally balanced and unbalanced runner systems.

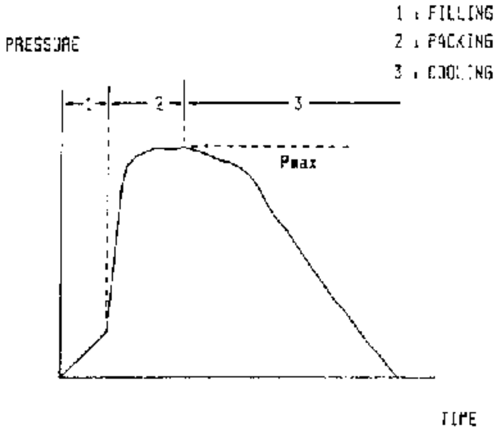


Figure 8 Schematic representation of pressure–time curve during filing, packing, and cooling stages of injection molding.

$$S = \int_0^b \frac{z^2}{\eta}\, dz \tag{7}$$

and η is the apparent viscosity, z is the gapwise coordinate, and b is the half gap thickness of the strip. This equation is coupled with the energy equation (5) which contains conduction, convection, viscous dissipation terms, and heat release due to crystallization in the case of molding of semicrystalline polymers and vulcanization or crosslinking in the case of molding of elastomers, thermosets, and reactive fluids. Additional information is required for simulation including the rheological equation and the equation of state. The majority of simulation techniques presently in use utilize the viscosity function according to the generalized Newtonian fluid equation based on the Cross model [50,71]:

$$\eta = \frac{\eta_0(T)}{1 + (\eta_0 \dot{\gamma}/\tau^*)^{1-n}} \qquad \eta_0(T) = B \exp(T_b/T) \tag{8}$$

where η_0 denotes the zero-shear-rate viscosity, $\dot{\gamma}$ is shear rate, T is temperature, and τ^*, B, T_b, and n are material constants. This equation describes the shear-rate and temperature dependence of the viscosity of thermoplastic melts and rubbers very well [50,68–70]. If required, the pressure dependence of the viscosity is also incorporated [50]. For the simulation of the packing stage, in addition to the rheological equation, the equation of state is required. The most widely used equations of state are the Spencer–Gilmore [72,73] and Tait [74] equations [75–80]. The Spencer–Gilmore equation has the form

$$(p + \bar{p})\left(\frac{1}{\rho} + \frac{1}{\bar{\rho}}\right) = \bar{R}T \tag{9}$$

where $\bar{p}$, $\bar{\rho}$, $\bar{R}$ are material constants. These constants can be determined from experimental p-v-T data obtained from dilatometric experiments. A recent book gives extensive p-v-T data for various thermoplastics [81].

Results of the flow simulation will give the pressure, temperature, and velocity fields and propagation of the melt front during the cavity filling, and the weld line or knit line formation and the shrinkage of molded products for specified processing variables. This information allows the determination of the clamp force requirement, positioning of venting and optimal molding variables, and molding cycle. In the majority of cases, the control volume finite-element method in the planar directions and the finite-difference method in the gapwise direction are used [70,82,83]. A vast amount of information is available related to simulation and comparison with experiments [50,56,60, 68,70]. Figure 9 gives a typical example of simulation of the pressure traces in comparison with experimental measurements based on data obtained in our laboratory for the injection molding of polypropylene. In rubber injection molding, efforts are described in recent publications [13,84,85]. There have also been significant advances in viscoelastic modeling of injection molding. In addition to process characteristics predicted in inelastic simulations, such simulations predict the frozen-in orientation and residual stresses in molding. These efforts are summarized in Refs. 50, 53 and 86.

8.4 INJECTION–COMPRESSION MOLDING

The injection–compression molding technique has been developed to utilize the advantages of both molding techniques [87–94]. This technique utilizes the conventional injection molding machine and a compression attachment (Fig. 10). At first, a polymer melt is injected in order to partially fill the mold which is partially open. Then, the compression stage is intro-

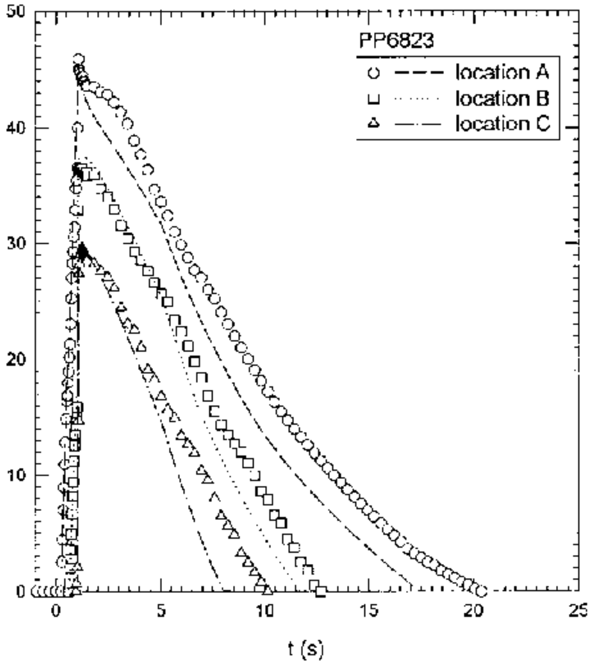


Figure 9 Comparison of predicted (lines) and measured (symbols) pressure–time curves in runner (A) and dumbbell cavity at different locations (B and C) from the gate during injection molding cycle for polypropylene.

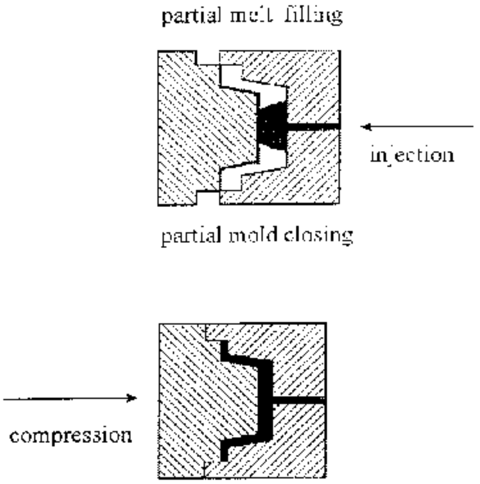


Figure 10 Schematic representation of injection–compression molding process.

duced, which leads to the final closing of the mold by squeezing flow of the melt. This compression stage is introduced to replace the packing stage of conventional injection molding. Since the pressure developed during the compression stage is significantly lower than that in the packing stage of conventional injection molding, the injection–compression molding introduces lower residual stresses, lower molecular orientation and birefringence, less, and more even, shrinkage, and better dimensional tolerances. At the same time, this process maintains high output, good process control, and automation inherent to conventional injection molding. The process is especially useful for molding thin parts that require high quality and accuracy. However, the process requires careful timing of injection clamp position and force. Injection-compression molding is presently employed in making the optical disks where the requirements for the dimensional tolerances and the optical retardation is very stringent. In production of the optical disks, this process is called coining. In comparison with injection molding, there are very few experimental studies in the literature on injection–compression molding. Concerning the simulation of the process, so far only one paper has reported such studies [95].

8.5 COINJECTION MOLDING

8.5.1 Technology

A major innovation in injection molding technology in recent years is the multicomponent molding process, sandwich injection molding or coinjection molding (Fig. 11). It was first introduced by the ICI in the early 1970s as an ingenious variation of the structural foam process leading to high surface quality of the product [96–101]. Schloemann-Siemag and Battenfeld have improved the process to fabricate moldings with tailor-made product characteristics, such as electromagnetic interference shielding and moldings with barrier properties obtained by a combination of two different materials. Recently, the process has been considered as an attractive method to recycle plastics. In coinjection molding, two different polymers are injected into a mold cavity sequentially or simultaneously in such a way that one polymer melt forms the skin and the other forms the core. The skin material completely encapsulates the core material, resulting in a sandwich structure.

In sandwich injection molding, two screw-injection machines are arranged to sequentially inject polymer melt through a valve within the nozzle and a single sprue into the mold cavity. Control means are provided by a control valve, so that melt from one screw-injection machine does not pass while the melt flows from the other injection machine. This procedure was classified as the single-channel technique. The major disadvantage of this technique is pressure drop and stagnation when switching from one injection machine to the other, resulting in a switching mark in the form of a dull ring on the surface.

A simultaneous injection process was developed in 1973 [102], permitting a continuous transition from one injection unit injecting the skin material and the other injecting the core material. Two polymers are injected simultaneously within 1–100% from an annular delivery system with the skin material through a ring nozzle and the core material through a central nozzle. Encapsulation is achieved by virtue of the design of the delivery system. In addition to eliminating the switching marks, this adjustable overlap allows the processor to control the thickness of the skin material in a given proximity. This process is classified as a two-channel technique and Schloemann-Siemag designed the first machine. Subsequently, Battenfeld, which was taken over by Schloemann-Siemag in 1977, has developed machines using two- and three-channel techniques [103,104]. Because of the simpler setup, the two-channel technique is preferred over the three-channel technique and used almost exclusively today.

Only limited studies on sandwich injection molding have been reported [105–109]. Experiments were performed in order to elucidate the effect of the material properties and the processing parameters on interface

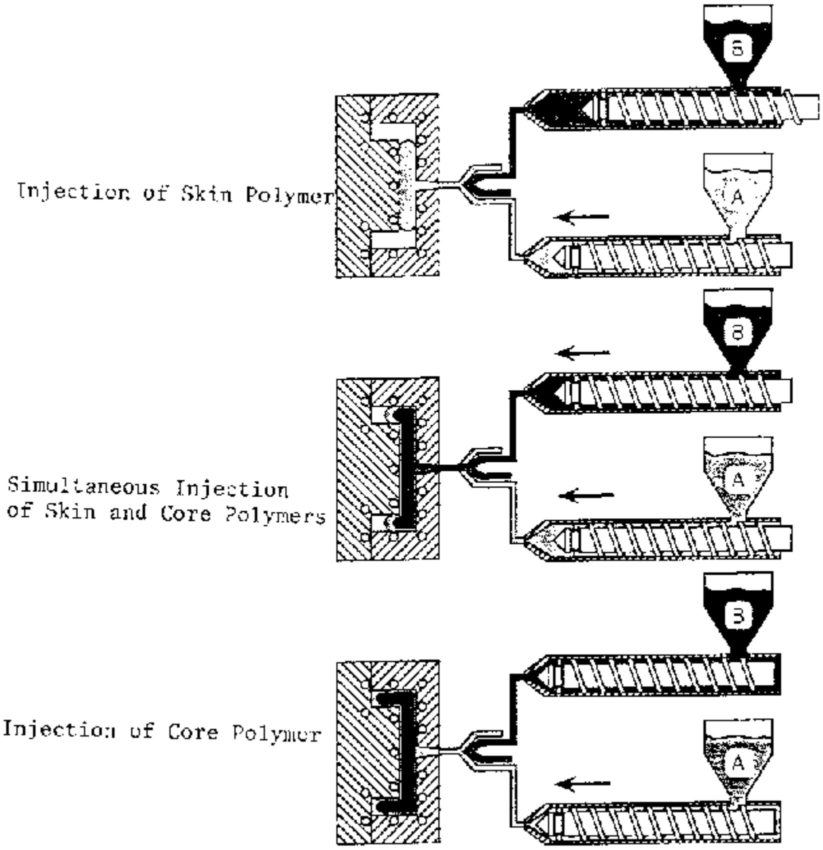

B
Injection of Skin Polymer
A
B
Simultaneous Injection
of Skin and Core Polymers
A
B
Injection of Core Polymer
A

Figure 11 Schematic representation of coinjection molding process according to Battenfeld.

distribution in molded parts. The viscosity ratio represents the primary influence of rheological properties on interfacial shape in sandwich injection-molded parts. Other rheological properties, such as the elasticity and normal stresses, may also have an effect. In addition, the processing parameters such as melt and mold temperatures, injection rates of each component and the length of simultaneous injection affect the interface evolution. To obtain evenly encapsulated skin/core structure in the molded parts, a proper choice of the viscosity ratio of the polymer melts and the control of the injection rate of polymer melts is required. The viscosity ratios should lie in the range of 0.82 to 1.83 [108]. Instabilities of the interface were also reported [109].

Recently, a novel fabrication process called lamellar injection molding (LIM) was developed [110–112]. A schematic of this process is given in Fig. 12. This process utilizes the manifold die earlier developed for multilayer extrusion combined with the injection molding process.

8.5.2 Modeling

Modeling of the coinjection molding process is a relatively recent undertaking [113–122]. The goal of the modeling is mainly to predict the interface evolution during cavity filling. Similar to single-phase injection molding, these efforts are based on the use of the Hele-Shaw approximation for sequential injection. Simulation of mold filling for the simultaneous sandwich process has been performed only recently. For the case of a one-dimensional filling of a rectangular cavity of width W and half-gap thickness h, the equations of continuity and momentum for each phase can be written as

$$\frac{\partial}{\partial x}\left[(h-\delta)\bar{v}_A\right]=0 \qquad \frac{\partial}{\partial x}\left(h\bar{v}_B\right)=0 \tag{10}$$

$$-\frac{\partial P}{\partial x}+\frac{\partial}{\partial z}\left(\eta_i\,\frac{\partial v_i}{\partial z}\right)=0 \tag{11}$$

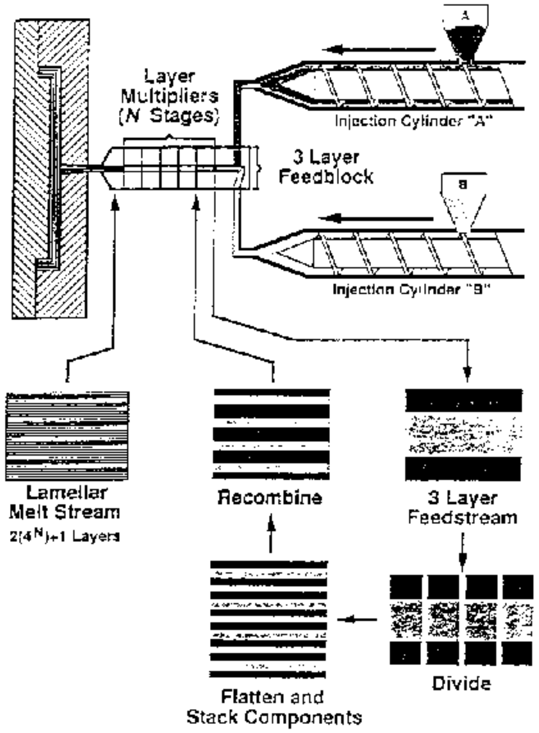


Figure 12 Schematic representation of lamellar injection molding reciprocating screw-injection molding machines. Simultaneous injection through a feedblock and layer multiplier is used to create a product with micron scale lamellar morphology. (From Ref. 111, courtesy of the Society of Plastics Engineers.)

where v_i is the average velocity across the half thickness of each phase, A and B, x and z are the streamwise and gapwise directions, respectively, and η_i is the apparent shear viscosity of each phase. The melt viscosities of both phase are shear-rate and temperature dependent [see Eq. (8)]. Similar to single-component injection molding, Eqs. (10) and (11) are coupled with the energy equation (5) and must be solved simultaneously. In addition, it is usually assumed that the shear stresses and heat fluxes are continuous at the interface, which is written as

$$\eta_A \frac{\partial v_A}{\partial z} = \eta_B \frac{\partial v_B}{\partial z} \qquad \text{at } z = \delta \tag{12}$$

$$k_A \frac{\partial T_A}{\partial z} = k_B \frac{\partial T_B}{\partial z} \qquad \text{at } z = \delta \tag{13}$$

where k_A and k_B are the thermal conductivity of each phase and T_A and T_B are the temperature of each phase.

The governing equations are usually solved by combination of the finite-element and finite-difference methods [113–117], finite-element analysis and particle tracking technique [120,121] or by the flow-analysis network (FAN) and finite-difference method [122–124].

Figure 13 shows the effect of the initial Newtonian viscosity ratio of the skin to core polymer, R, on the simulated and measured interface position, h/H, for injection of 40% of the HDPE or LDPE skin polymer into a strip cavity followed by simultaneous injection of the skin polymer HDPE or LDPE and core PS [122]. When the viscosity ratio increases, there is corresponding increase of the thickness of core phase.

8.6 GAS-ASSISTED INJECTION MOLDING

8.6.1 Technology

Gas-assisted injection molding is a relatively novel molding process invented over 20 years ago [125–127]. The process is deceptively simple to carry out but difficult to control due to the dynamical interaction between gas and polymer melt. A schematic of the process is given in Fig. 14. The process comprises three stages: cavity filling, gas injection, and packing. In the cavity filling stage, a predetermined amount of the polymer melt is injected into the cavity to partially fill it. Then, in the gas injection stage, nitrogen gas under high pressure is injected through the nozzle or mold wall into plastic. The nitrogen is typically

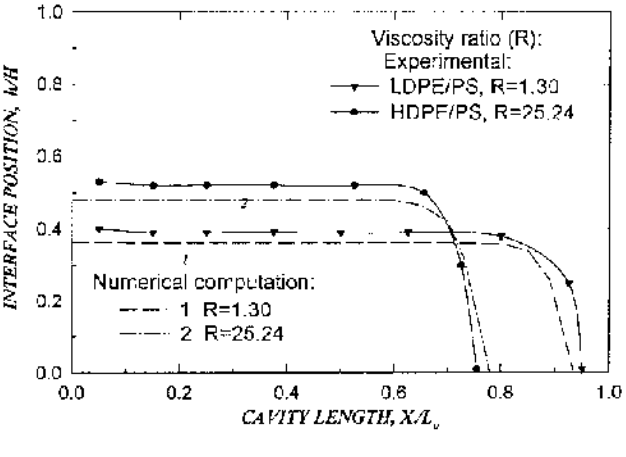


Figure 13 Effect of viscosity ratio of the skin to core polymer on the simulated and measured interface position, h/H, for injection of 40% of the HDPE or LDPE skin polymer into a strip cavity followed by simultaneous injection of the skin polymer HDPE or LDPE and core PS. (From Ref. 122.)

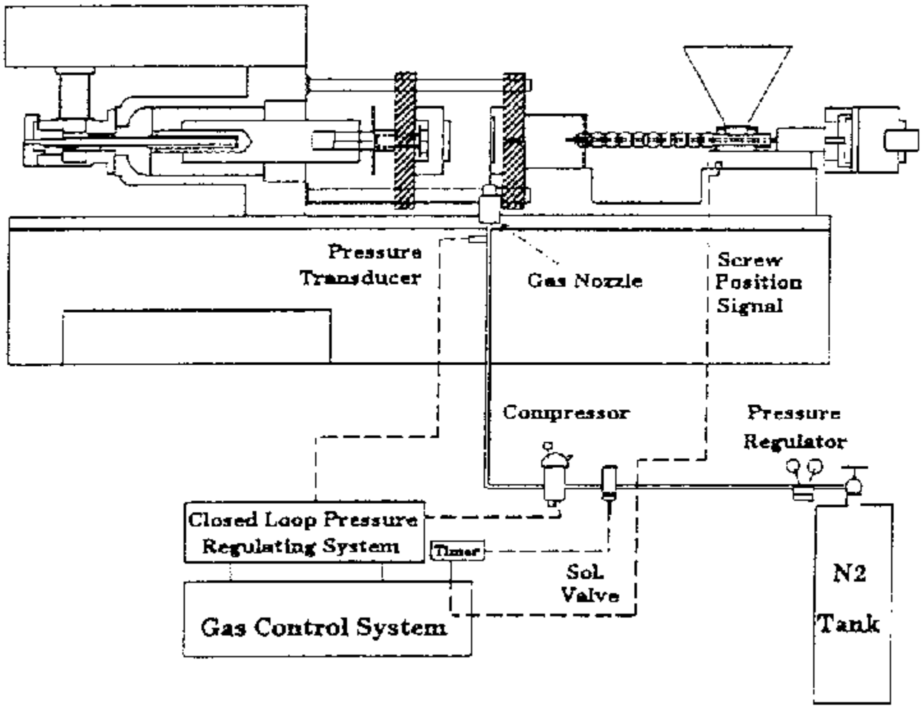


Figure 14 Schematic representation of gas-assisted injection molding. (From Ref. 128, courtesy of the Society of Plastics Engineers.)

injected at pressures ranging from 0.5 to 30 MPa. Gas can be injected sequentially or simultaneously during the cavity filling stage. The gas penetrates through the thick sections where the melt is hot and pushes the plastic melt to fill the mold. The polymer in the skin layer is stagnant due solidification upon contact with the cold mold surface. Due to the geometrical complexity of parts, multiple disconnected gas channels with a multiple gas injection system are frequently used. This allows one to transmit the pressure uniformly to various areas of the part. In the packing stage, after the cavity is filled, the gas pressure is maintained to compensate for shrinkage and just prior to mold opening, gas is vented. Advantages of the gas-assisted injection molding process in comparison with the conventional injection molding are the reduction of cycle time, part weight, shrinkage, warpage, injection, pressure, and clamping force. In addition, the process allows for the improvement of the surface finish and the elimination of sink marks.

The process parameters governing the gas-assisted molding are: melt and mold temperature, shot size, injection speed, gas pressure, and gas delay time [128–132]. The design parameters affecting the process are the diameter of gas channel and thickness of the cavity [130,133,134]. The main concern in gas-assisted injection molding is gas penetration length and skin melt thickness. The melt temperature has a variable effect on gas penetration length. In particular, an increase in the melt temperature is shown to decrease or increase the gas penetration length. Concerning the effect of the mold temperature on gas penetration length, studies are also contradictory. An increase of the mold temperature leads to an increase, decrease, or no effect on the penetration length. An increase in the shot size is found to decrease the penetration length. For the same shot size, an increase in the injection speed, which corresponds to the lower filling time, increases the penetration depth due to less cooling time available during injection of the melt. An increase in gas pressure level and time increases gas penetration length. Increasing delay time before the start of gas injection is found to increase or decrease the wall thickness and gas penetration length. Skin melt thickness typically increases with decreasing melt and mold temperatures [135]. Results concerning the effect of design parameters on the process are as follows. An increase in the diameter of the gas channel leads to shortening of the flow length. A decrease in the cavity thickness causes narrowing of the processing window, The fingering effect (multiple branches of gas penetration) are typical defects observed in the gas-assisted injection moldings. The fingering effect is typically observed during injection of small amount of melt. By increasing the amount of plastic injected, a well-balanced gas penetration is observed. Hesitation marks can be avoided by imposing a high injection speed or/and high mold temperature with a short or absent delay time. In many cases the multiple channel ribs are employed to improve the distribution of the gas through the molding. The study of the melt-filling phenomena of rectangular cavities with various arrangements of gas channels enabled the development of guidelines for the layout of gas channels ribs [136–139]. Self-calibrating needle valves are proposed to eliminate significant back pressure on the gas supply unit [140]. This valve allows the process to avoid the detrimental condition of choked gas flow.

In order to understand the gas-assisted molding process better, more work is required concerning the interaction between the rheological properties of melt and processing parameters and their effect on wall thickness and gas penetration length. In particular, the shape of the shear-rate-dependent viscosity curve was found to affect the sensitivity of change of wall thickness by the gas bubble velocity. In turn, the gas velocity is influenced by gas piston speed and gas pressure [141]. Polymer melts showing the shear thinning effect at low shear rates are more sensitive to changes in gas pressure and gas piston speed. Polymer melts maintaining the initial viscosity in wide range of shear rates are relatively insensitive to these changes. The elasticity of melt is also found to affect the process. Isothermal experiments performed in tubes indicated that, at Deborah numbers (a product of the shear rate of the process and the relaxation time of the melt) higher than unity, the fluid elasticity increases the hydrodynamical fractional coverage coated by a long penetrating bubble [142]. On the other hand, the coating of the shear-thinning fluid during the gas propagation is much thinner than in the case of a Newtonian fluid [139,143]. Gas-assisted injection molding is also utilized in molding of fiber-reinforced thermoplastics. Some relevant experimental studies are reported in Refs. 144 and 145.

8.6.2 Modeling

Gas-assisted injection molding is a complicated process to simulate due to the interaction between moving melt and gas boundaries which takes place during flow in a complex cavity. The goal of the model is to predict the gas penetration depth, the distribution of wall thickness in molded parts, and the location of undesirable air trap. The majority of available theoretical models are based on the two-dimensional Hele–Shaw flow approximation [145–164]. Some approaches are also proposed to handle three-dimensional simulations using the boundary-element method [165,166] and finite-element method combined with the pseudoconcentration method [167]. In the case of two-dimensional simulations, Eq. (6) is utilized in both melt filling and gas penetration phases. However, the use of this equation in the gas penetration phase is justified only with appropriate boundary conditions at the gas–polymer interface. The treatment of the melt filling phase is similar to the case of conventional injection molding. In the gas penetration phase, it is usually assumed that the gas transmits the pressure uniformly everywhere in the gas domain including the interface with the polymer. The geometry of the gas channel of a noncircular cross-section with the connection portion of thin part is usually approximated by an equivalent hydraulic diameter [168,169]. The control volume formulation is used in simulations with heat transfer solution based on the finite-difference method. The initial gas core temperature is assumed to be equal to the temperature of the polymer at the polymer–gas interface.

Figure 15 shows an example of a comparison between the simulated and measured gas penetration region during molding of a HDPE strip [155]. The numerical results are obtained by using the finite-element method for solving the flow equation and the finite-difference method for solving the energy equation. Good agreement between the predicted and experimental data for the length of the hollow channel is observed. However, the thickness ratio between gas and polymer layers is not well predicted.

8.7 TRANSFER MOLDING

8.7.1 Technology

Transfer molding is mainly utilized to mold products from thermosets and rubbers. It is related to compression and injection molding. In this process the polymer is placed in the transfer chamber and heated to achieve the flow state as indicated in Fig. 16. Then, the polymer is forced through the delivery system

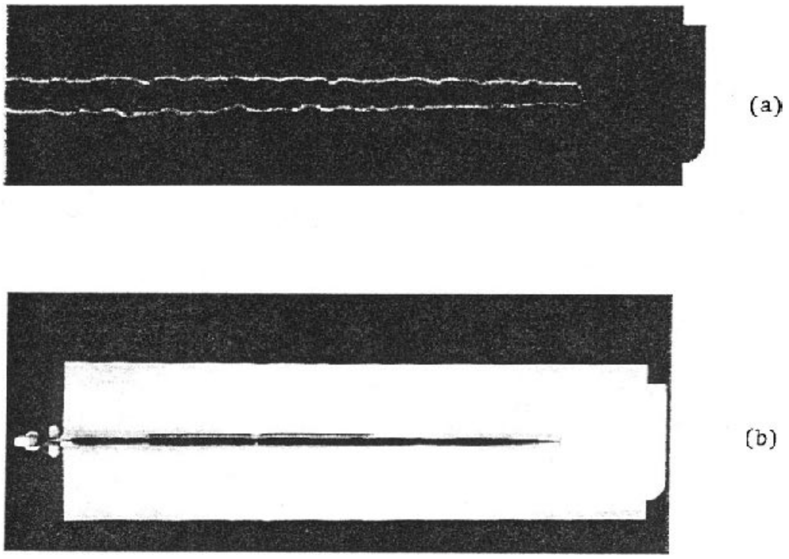


Figure 15 Experimental (a) and predicted (b) gas penetration regions for gas-assisted injection molding in a strip cavity. (From Ref. 155, courtesy of Hanser Publishers.)

(sprues, runners, and gates) into a closed mold by a hydraulically operated plunger. The chamber is usually loaded manually using a slug slightly exceeding the total volume of the molding and the delivery system. There are two basic methods for transfer molding called plunger transfer molding and pot transfer molding [9]. In comparison with compression molding, the cure time in the transfer molding is reduced due to heat dissipation taking place during the flow through the delivery system. Transfer molding is carried out by hydraulic presses similar to compression molding presses. Single and multicavity molds are used. In the case of the multicavity moldings, one transfer chamber with a manifold runner system connected to each cavity is used. Intricate moldings with metal inserts along with good dimensional accuracy are obtained with transfer molding. Transfer molding is widely used for manufacturing rubber seals, antivibration mountings, and other products. The antivibration molded parts often contain several elastomeric layers separated by metal plates [170]. In the case of rubber molding some comparisons of rubber properties between compression and transfer molding were made, indicating that the two molding methods produce distinctly different rubber properties for the same elastomer [171]. Evidently, this is due to the difference in the viscoelastic behavior of elastomers during mold filling. Transfer molding is also widely used for packaging of microelectronic devices [172–175].

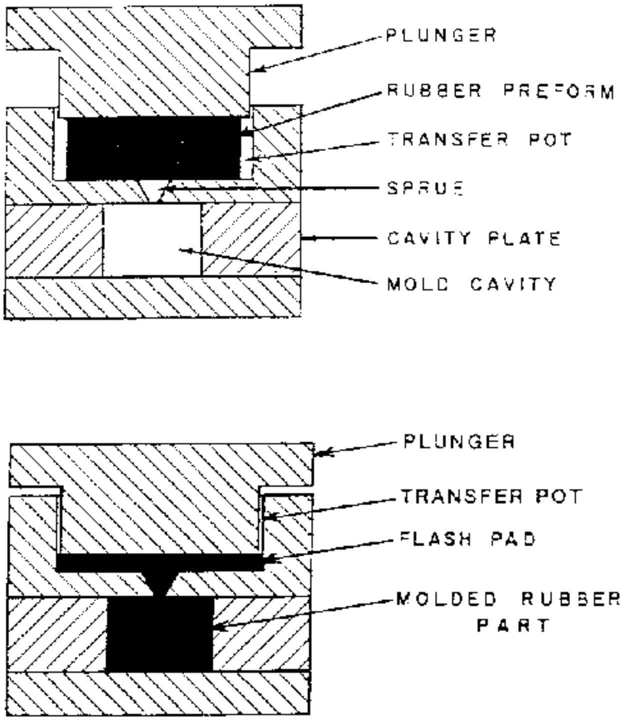


Figure 16 Schematic representation of transfer molding process.

8.7.2 Modeling

Modeling of flow in transfer molding is typically similar to those in the injection molding of thermosets and rubbers [50,68–70,84]. However, in the case of microchip encapsulation, the situation is different. In particular, in the latter case, a model is required to predict resin bleed and flash, wire sweep (deformation of the connecting wire), and paddle shift (gapwise movement of the microchip/leadframe assembly). These effects cannot be described by using the Hele–Shaw formulation as applied to injection molding. This is due to the fact that the Hele–Shaw approximation ignores the gapwise velocity component during the cavity filling. Therefore, recent efforts in modeling of encapsulation during transfer molding are made toward removing this restriction [176–183]. In particular, Fig. 17 presents the side and isometric views of the simulated fluid front advancement in the mold mimicking the microchip encapsulation and having the upper and lower cavities of different thicknesses [177]. In this case a three-dimensional flow simulation program was used. It is seen that the fluid front advances from the gate to the plate in the middle of the mold cavity and then splits into the upper and lower cavities. Since the upper cavity is thicker, the fluid front advances at a faster speed in the upper cavity. After the upper cavity is filled, the two fronts advance towards each other and form a weld line in the lower cavity.

8.8 RESIN TRANSFER MOLDING

8.8.1 Technology

Resin transfer molding (RTM) is a newly emerging technology which is used to manufacture continuous fiber composite products of complex geometry [26,44,184,185]. Sometimes, similar processes are also called structural reaction injection molding and liquid composite molding. A recent review of the RTM is given in Ref. 186. As indicated in Fig. 18, the RTM process is deceptively simple. A preform fabric or fiber mat is placed in the mold. The mold is closed. The resin supply vessel is attached to the mold by a pipe having a valve. Opening the valve allows the resin to be pumped from the vessel to the mold through one or several gates and to impregnate the preform. After the mold is filled and sealed, curing takes place. When the curing is completed, the part is taken out.

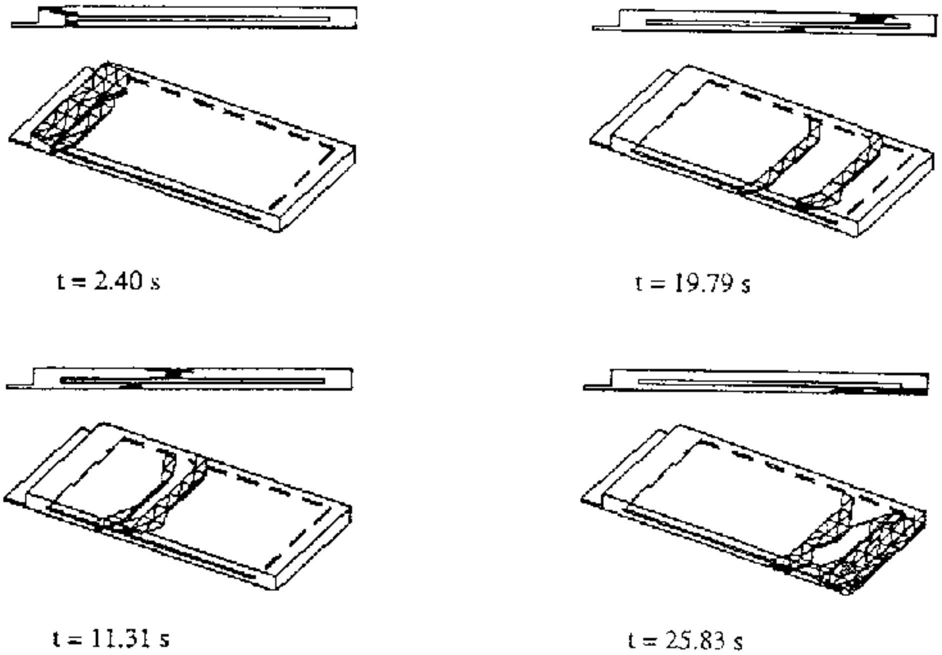


Figure 17 Three-dimensional simulation of the fluid front advancement in a flow domain similar to the mold cavity used for microchip encapsulation. (From Ref. 177, courtesy of the Society of Plastics Engineers.)

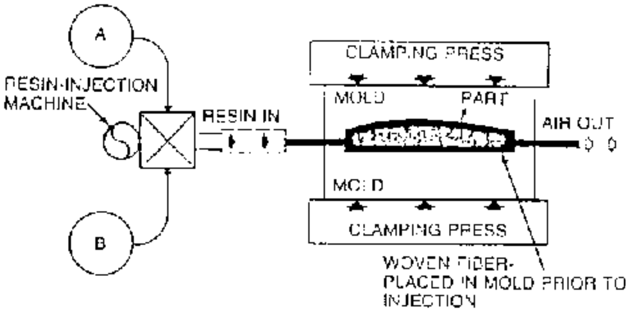


Figure 18 Schematic representation of resin transfer molding.

One major concern in the RTM is bleeding of air from the mold in order to avoid large air pockets trapped during flow. In particular, the air entrapment influences the porosity of the part and therefore its performance characteristics. Thus, mold design is an important consideration. Design of the injection port and vent is the most critical issue of the process. Sealing the mold to achieve a certain level of pressure is important in obtaining a porosity in the part below 1%.

Physical properties of the resin, such as viscosity, pot life, and T_g, determine the duration of the process cycle. The optimum viscosity of the resin for RTM lies below 0.5 Pa.sec. The low viscosity allows the process to be carried out without imposition of excessively high pressure. The pot life determines the time necessary for the viscosity during curing to stay within the optimum viscosity. In the case of thermosets, the curing temperature affects the T_g. Thus, it is desirable that the curing temperature be maintained above the steady-state value of T_g. The T_g of the molded product should be higher than the temperature of the intended use of the product.

The RTM is used primarily to manufacture large parts. It is friendly toward automation. Due to the low viscosity of impregnating reactive fluids, the process is carried out at low molding pressures. Therefore, mold cost is lower in comparison with injection or compression molding. The usefulness of the RTM has been proven in automotive applications. However, the main problem is in making low-cost adequate preforms. The preforms are made of randomly oriented fibers or woven fabrics. Fibers used are primarily glass fibers, but carbon fibers and synthetic fibers are also used. The reactive mixture must wet the fibers in order to achieve the required performance characteristics of the products. Therefore, bonding of matrix to fibers is an important issue. The bonding is typically achieved by using coupling (sizing) agents.

There have been substantial experimental efforts concerning the data generation related to resin impregnation useful for implementation of the RTM technology. These studies were concerned with rheology and reaction kinetics measurements [187,188], effect of structure fiber placement in preforms on the permeability [189–191], wetting of fiber mats [192], understanding of the effect of the race tracking and the fiber impinging, creation of fiber-rich or resin-rich areas [193,194], and the porosity [195,196] in moldings.

8.8.2 Modeling

The analysis of resin flow in the mold is very important from the view point of the mold design. In addition, process modeling is particularly useful in understanding, designing, and optimizing process conditions. An overview of the RTM process, its manufacturing problem, and modeling aspects is given in Refs. [197–199]. A number of two-dimensional flow simulation attempts are based on the shell model under the assumption that there is no flow in the thickness direction. Some three-dimensional attempts have also been made [200,201]. The analysis of flow is typically based upon finite-difference [202–205], boundary-element [206–209], control-volume [210–221] and finite-element methods [222–231]. In addition, numerical schemes using the body-fitted finite-element method [232–237] and combination of flow analysis network for movement of the free surface and finite-element method to solve governing equations for each successive flow front location [238] are used.

The impregnation of the preform by a resin is usually based on a flow of resin through anisotropic homogeneous porous media and governed by Darcy's law:

$$\bar{v} = -\frac{[k]}{\mu}\nabla P \tag{14}$$

where $\bar{v}$ is the velocity vector, p is the resin pressure, μ is the viscosity, and k is the permeability tensor given as

$$[k] = \begin{bmatrix} k_{11} & k_{12} \\ k_{12} & k_{22} \end{bmatrix} \tag{15}$$

Combining Eq. (14) with the continuity equation gives the equation governing the process of mold filling. The numerical solution of these equations along with the energy equation and curing kinetics equation allows one to determine the pressure, state of cure, and the location of the fluid front during the flow through porous media. This representation was further generalized to handling three-dimensional preforms [201]. In particular, a recent success with handling a complex three-dimensional situation has been proven in resin transfer molding of suspension links which control the location and orientation of a vehicle's wheels relative to the ground under various driving conditions. The preform for the trailing link was made from fiber mats different for I-beams and eyelet tubes. Figure 19 shows the simulated results for the fluid flow patterns during molding of composite trailing links for the following four cases: the fiber mats are (a) isotropic without race-tracking and fiber-impinging effects; (b) anisotropic without racetracking and fiber-impinging effects; (c) anisotropic with racetracking effects but without fiber-impinging effects; (d) anisotropic with race-tracking and fiber-impinging effects. Also, Fig 19e shows the flow front progression from short-shot experiments. Due to the symmetrical part geometry, a quarter of the part is taken as the simulation domain. In case (a), the flow front progression is like a plug flow. In case (b), corresponding to anisotropic and unequal permeability of each preform component, the flow front becomes irregular. However, the simulated and experimental flow fronts are very different. In case (c), the flow front lead–lag is reversed in each component, resembling the short-shot experiments. In case (d), the simulated melt front profile is like in case (c), except that the flow front lead–lag in the belt region is larger and a dry spot is formed near the outer side of the belt. This is in a good agreement with the short-shot experiments.

8.9 BLOW MOLDING

8.9.1 Technology

Blow molding is a process used to manufacture hollow parts, such as bottles and containers, by inflating a soften thermoplastic hollow tube or preform, usually called a parison, inside of the mold having a wall temperature below the T_m or T_g of the polymer [9–11,239–242]. An excellent survey of the development of blow molding is given in Ref. 243. Blow molding originates from the art of glass blowing and used by Egyptians as early as 1800 years BC. A patent describing blow molding of thermoplastic gutta-percha and its compounds as issued in the middle of the 19th century [244,245]. Later, a patent was granted for blow molding of celluloid from two sheets or a tube with steam used to soften and expand the preform [246]. The process is called extrusion blow molding if the preform is made by extrusion. The process is called injection blow molding if the preform is made by injection molding. The process is called stretch blow molding when the preform is stretched in the axial direction in addition to blowing. In extrusion, as well as in injection blow molding, the material deformation mainly occurs along the circumference which leads to anisotropic orientation in products. In stretch blow molding, a biaxial orientation is achieved in the axial direction and along the circumference, leading to controlled anisotropy of products. In recent years, parisons for blow molding are prepared by coextrusion or coinjection molding to attain the required permeability characteristics in multilayer blow-molded containers [247]. The process allows the reduction of the expensive plastics in the chemically resistant barrier layer which is combined with the structural layer and an intermediate tie layer. This is especially important for containers used in the food industry where an oxygen barrier or moisture barrier is required and in the chemical industry where a chemical barrier for packaging of chemicals in plastic containers is required.

An extrusion blow molding machine consists of a parison former comprising a screw extruder and an annular die which extrudes a tube vertically downward, and a mold handling device (Fig. 20). Machines of lower capacities extrude the parison continuously. Machines of larger capacities are typically equipped with an accumulator intermittently fed by an extruder. In this case, the accumulator is equipped with a plunger which extrudes the parison in a discontinuous manner. A mold closes around the parison. Air and, in some cases, nitrogen is injected into the melt parison until the melt fills the mold. The mold is cooled, the melt solidifies, the mold opens and the part is ejected. Molds for blow molding are constructed more lightly than injection molds due to the presence of moderate pressures and clamping forces during the process in comparison with injection molding. In extrusion blow molding, the swelling and sagging of parison are two important characteristics of the melt. The multiple pinch-off mold contributed to the basic understanding of parison sag and swell [248]. Motion pictures taken with a video camera were used to analyze the swelling of parison versus distance from the annular die [249–257]. Time dependence of the annular extrudate swell and sag of HDPE, one of the widely used plastics in blow molding, was also studied [258–261]. The inflation behavior was also studied using a transparent mold [262]. Several studies were made to measure orientation development in extrusion

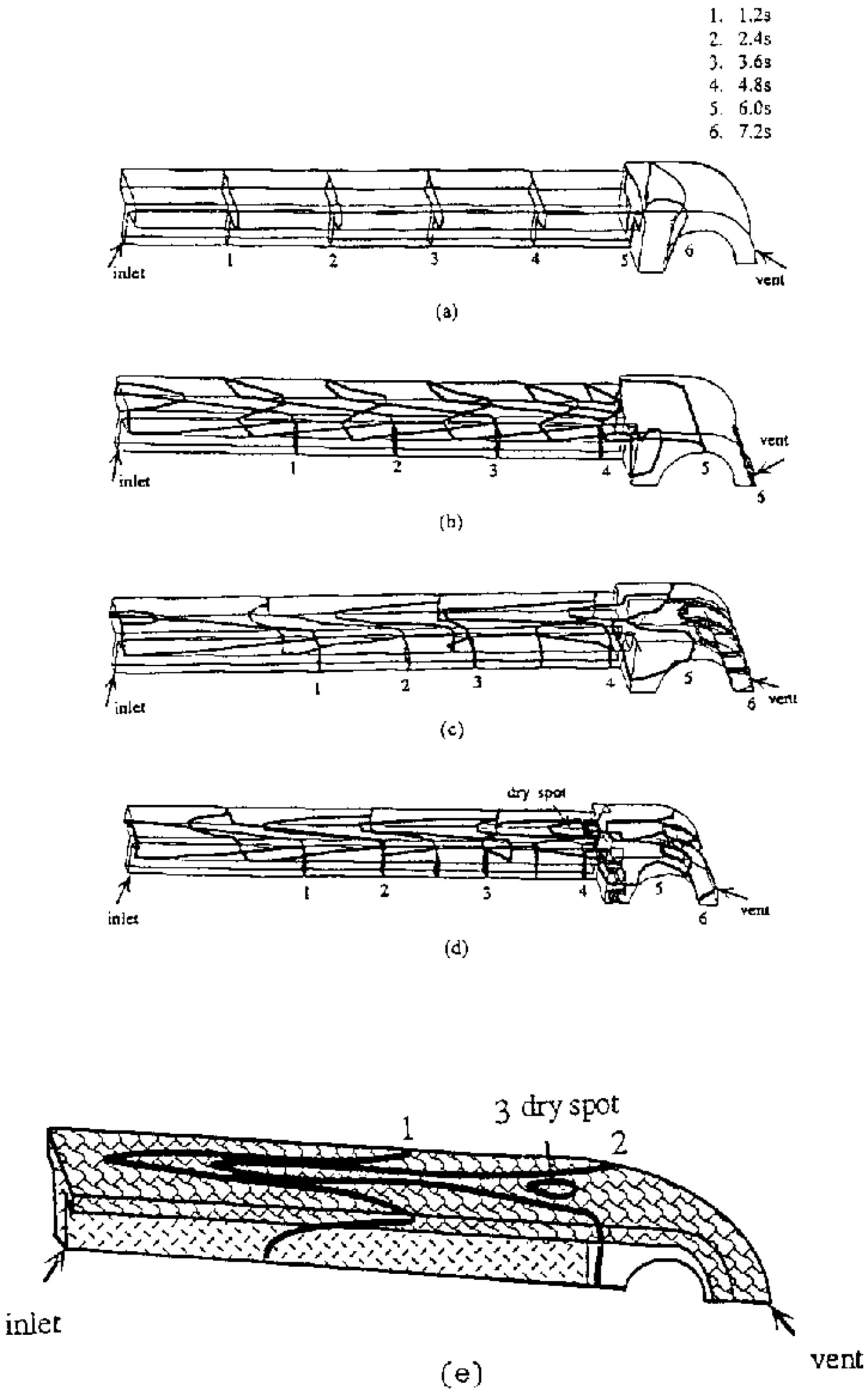


Figure 19 Simulated (a,b,c,d) and experimental (e) fluid flow patterns during resin transfer molding of a composite trailing link. The simulations are conducted for four cases of the fiber preform placement as indicated in the text. (From Ref. 201, courtesy of the Society of Plastics Engineers.)

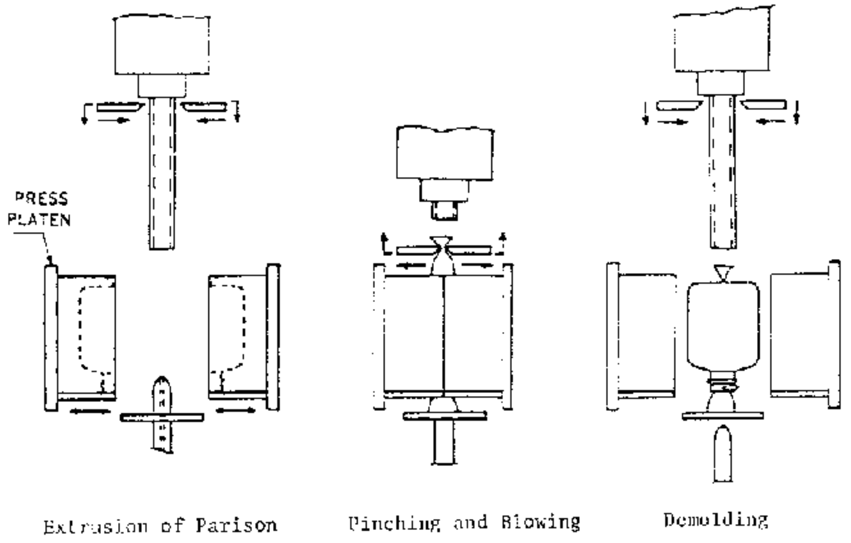


Figure 20 Schematic representation of extrusion blow molding process.

blow-molded bottles [263-266] as a function of process parameters such as the melt temperature and the inflation pressure. These also include the effect of the viscosity of the melt.

For injection blow molding, a parison is prepared and stored before use. In this case, the parison is preheated and blowing stage is carried out similarly as in extrusion blow molding.

In stretch blow molding, the axial stretching is introduced with the aid of a stretching rod and plug before inflation takes place. Orientation development in stretch blow-molded PET bottles has been investigated [267,268].

8.9.2 Modeling

The main objective of modeling blow molding is to predict and optimize the thickness distribution in the molded part. This will give a rational means for mold design and the setup of processing conditions. The latter can be done by comparing the simulated results for different polymers, molds, and parison geometry under different processing conditions. The analysis can be repeated until an acceptable thickness distribution is obtained. The predicted thickness can further be used for a structural analysis to indicate if the product is capable of withstanding the required load.

Depending on the material description, there are two main approaches to the simulation of blow molding. These approaches are based on nonlinear elastic (hyperelastic) behavior [269–275] and fluidlike behavior [276–281]. The parison inflation as a deformation of a fluid membrane can be considered as a simulation model, since the typical process temperature is above T_g and T_m. On the other hand, polymers deformed at high strain rates accumulate significant amounts of elastic strain, leading to rubberlike elastic behavior. Therefore, the formulation can be simplified using nonlinear elastic behavior. It is interesting to note that the calculated thickness distributions in the blow-molded parts according to the hyperelastic and fluid models are similar. Obviously, these two approaches introduce oversimplification, since the real polymer behavior during blow molding is viscoelastic. Thus, several attempts have been made to incorporate the viscoelastic fluid models based on integral [282–284] and differential [285–289] equations. For the case of the viscous fluid, the flow is governed by the mass conservation and momentum equation. Since the typical thickness of hollow products is significantly lower than the other dimensions, the membrane approach is typically used. Therefore, the local form of the mass conservation is given by

$$\frac{D\delta}{Dt} + \delta\nabla\bar{v} = 0 \tag{16}$$

where δ is the thickness, t is time, and $\bar{v}$ is the mean velocity vector. The momentum equation, weighted by the thickness, can be written as

$$\nabla F + f = \rho\delta\,\frac{D\bar{v}}{Dt} \tag{17}$$

where ρ is the fluid density, f is a surface force which stands for the inflation pressure and F is the tensor of contact forces per unit length. This force is obtained by weighting in δ the extra stress tensor $\underline{\underline{\sigma}}$, based on the assumption of stress free surfaces of the membrane. For a Newtonian and upper-convected Maxwell fluid, the extra stress tensor is respectively given by [290]

$$\underline{\underline{\sigma}} = 2\eta \underline{\underline{D}} \tag{18}$$

$$\underline{\underline{\sigma}} + \lambda \frac{\delta \underline{\underline{\sigma}}}{\delta t} = 2\eta \underline{\underline{D}} \tag{19}$$

where η is the viscosity and $\underline{\underline{D}}$ is the strain rate tensor, λ is the relaxation time, and $\delta \underline{\underline{\sigma}} / \delta t$ is the upper-convected Oldroyd's derivative.

For nonlinear elastic behavior, the equilibrium condition for the membrane is used by comparing the rate of work performed by the pressure on the membrane with the rate of increase in internal energy in the membrane:

$$\int_A P\bar{n} \cdot \bar{v}\, dA = \frac{\partial}{\partial t} \int_V \rho E\, dV \tag{20}$$

where p is the pressure, $\bar{n}$ is the normal to the membrane, $\bar{v}$ is its velocity, ρ is the material density, E is the internal energy density, A is the current surface area of the membrane, and V its volume. The internal energy density can be expressed through the second Piola–Kirchhoff stress tensor and the Green deformation tensor [291,292]. For the nonlinear elastic material, the second Piola–Kirchhoff stress tensor can be expressed through the derivative of the strain energy function, W. In the case of blow molding, two types of the strain energy function are most frequently utilized under the assumption that the material is incompressible. These are the Mooney–Rivlin and Ogden strain energy functions. The Mooney-Rivlin strain energy function is given as

$$W = C_1(I_1 - 3) + C_2(I_2 - 3) \tag{21}$$

where I_1 and I_2 are the first and second invariants of the deformation gradient tensor, and C_1 and C_2 are material constants.

The Ogden strain energy function is given as

$$W = \sum \frac{\mu_i}{\alpha_i} (\lambda_1^{\alpha_i} + \lambda_2^{\alpha_i} + \lambda_3^{\alpha_i} - 3) \tag{22}$$

where $\lambda_1, \lambda_2, \lambda_3$ are the principal elongations, μ_i and α_i are the material constants. The second Piola–Kirchhoff stress tensor can be transformed to obtain the Cauchy stresses through the principal elongations and the strain energy function. The material constants in the Mooney–Rivlin and Ogden strain energy functions are determined by fitting the Cauchy stress–strain relations to the results of the biaxial and uniaxial elongation experiments. However, it should be noted that to carry out such experiments for softened plastics at temperatures above the T_g or T_m is no simple task.

Finite-element analysis is typically used to perform numerical analysis of the blow molding process. General descriptions of the finite-element method can be found in Refs. 293 and 294. Detailed descriptions of the implementation of this technique to blow molding are given in Refs. 269 and 295.

A recent blow molding analysis focuses on the large parison deformations during blow molding using three-dimensional elements for predicting the clamping and inflation stages of the process [286]. In particular, Fig. 21 shows the parison meshes with two mold halves (a) and the progression of parison deformation during the clamping stage (b) and (c) followed by the inflation stages (d) and (e) of extrusion blow molding. The simulation is carried out using the two-term Mooney–Rivlin model. This calculation indicates that the three-dimensional simulation is able to handle large deformations, particularly in the pinch-off areas.

8.10 ROTATIONAL MOLDING

8.10.1 Technology

Rotational molding is a process for manufacturing hollow articles. The origin of the process dates back to inventions [296,297] for rotational casting of hollow articles from molten metals. Later, patented processes with biaxially rotating molds were proposed for making hollow articles from paraffin wax [298] and chocolate [299–302]. In the 1920s and 1930s other nonpolymeric materials were rotomolded [303–305]. The technology became available to the polymer industry in 1930s for making hollow rubber [302,306–309] and cellulose acetate [310] products. However, a major expansion of rotational molding took place with the emergence of vinyl plastisols [311,312] and powdered polyethylenes [313–315] and other thermoplastics [316]. Since then a number of books have been devoted to the description of the rotational molding process, materials and machines [7,317–321].

The process of rotational molding consists of charging a hollow closed mold with a plastic material in liquid or powder form (Fig. 22). Then, the mold is subjected to continuous biaxial rotation around both

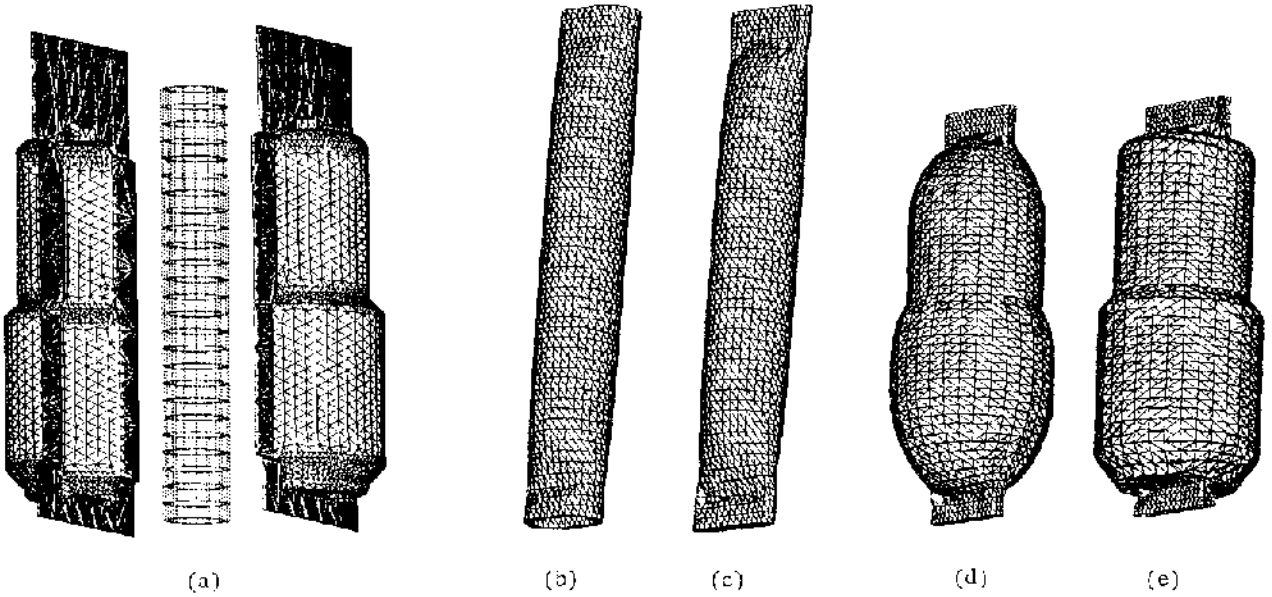


Figure 21 Parison mesh with mold halves (a) along with various steps of parison deformation during the clamping stage (b, c) and the inflation stage (d, e). (From Ref. 286, courtesy of Balkema Publishers.)

vertical and horizontal axes. The rotating mold is placed into an oven where the plastic material in the mold is heated up to melting under continuous rotation. Once a homogeneous layer of the polymer melt is formed, the mold is then moved to the cooling chamber or shower. While in the chamber, the mold is sprayed with air, water, or a mist of the two to cool it down. After cooling, the part is solidified to maintain the shape of the mold. The mold is then moved to the open station where it is opened and the part is removed. Then, the cycle is repeated.

The main process variables are temperature and duration of heat cycle, cooling cycle, and mold rotation. One major concern in rotational molding is degradation caused by a long heating cycle time and the presence of oxygen at the inner free surface of the moldings. Also, the process is characterized by the absence of shear during shaping and low cooling rates. These lead to poor mixing of additives and large spherulites in moldings. Overall cycle times in rotational molding are too long in comparison with those of injection and blow molding. The factors affecting wall thickness uniformity are rotational speed ratio of the major to minor axis and the magnitude of rotational speeds.

Various articles are presently made by rotational molding including large containers, fuel tanks, toys, furniture, road barrels, and many other hollow products. Different materials are used for rotational molding. Among them about 90% of the total plastics volume used in rotational molding consists of polyethylenes (PE). Other polymers include polypropylenes (PP), PP/PE copolymers, PVC plastisols, polycarbonates, nylons, ABS, acetal copolymers, etc. [316]. Liquid polymers, such as nylon reactive liquid (NYRIM) and polyurethanes, are also used [322–327]. For special applications, foams, composites, and multilayered articles can also be molded. The main advantage of rotational molding is the possibility to manufacture large articles at low pressures by means of low-cost molds and equipment. In addition, it is profitable to make products in small quantities.

The rotational molding is carried out by four types of machines: continuous, batch, shuttle, and jacketed mold [317,319]. The continuous machines with three or more arms are the most popular for making small-to-medium sized parts. Shuttle-type machines are popular for making extremely large parts.

The molds are made of cast aluminum, beryllium alloys, sheet mild steel, stainless steel, and sheet aluminum with a suitable mold surface treatment if necessary. Venting is provided by installing the vent tube in the mold. Venting eliminates pressure buildup during heating and permits air to enter the mold during cooling.

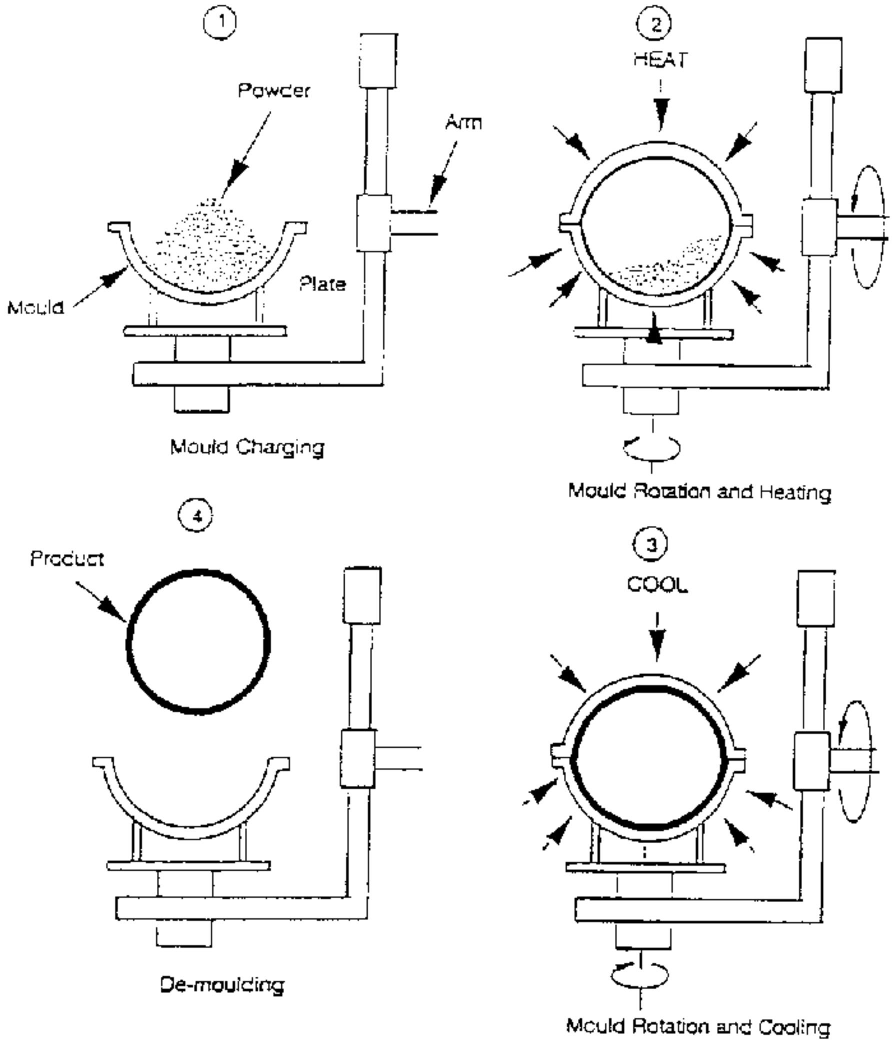


Figure 22 Schematic representation of rotational molding process. (From Ref. 319, courtesy of John Wiley and Sons, Inc.)

In spite of significant research efforts by various research groups, rotational molding is still considered an art. The molders experience is a driving force behind successful molding Few basic studies of the process are available. Some of these studies are summarized in a recent book [319]. The various steps of rotational molding have been studied, including heat transfer [328–333], sintering, densification, and leveling [334], and bubble formation and removal [335–337]. In addition, the effect of powder characteristics on the molding has been studied [7,338,339]. Recent studies also include extensive characterization of the performance of rotationally molded parts and their correlation with process conditions [325,331,340–351]. In addition, the relationship between structure and properties was discussed [352].

8.10.2 Modeling

The first attempts to model heat transfer in rotational molding were carried out [353–355] for uniaxially rotated cylindrical molds. These were based on one-dimensional transient heat conduction under the assumption that powder is in a simple circulation pattern or in static contact with the mold. Times for complete melting of the powder and cooling were predicted. A similar model was later extended for the rotomolding of liquids [356]. Furthermore, modeling of heat transfer in molds from different materials was carried out, indicating that the heating and cooling rates were dependent on the material and mold thickness [357]. The first heat transfer model for biaxially rotating molds was developed [358] by using a finite-difference method. The contact time was based on the period when the finite-difference mesh remained in contact with the moving powder pool in the rotating mold. Temperature profiles and melting times were predicted. Since conversion of powder into melt involves sintering, densification, and leveling, a model to describe each of these processes is required. An attempt to explain the sintering and densification of polymer powders on the basis of theories of metal sintering and glass densification has been made [353,359,360]. An excellent overview of activities in this area has been given [361,362]. In these papers, a sintering model due to Frenkel [363] based on Newtonian viscous flow has been modified to include the variation of the particle radius with time in the coalescence process. The proposed model is relatively simple and compares very well with experimental data [361]. Observations of coalescence in rotational molding suggest that melt elasticity slows down the process. Based on these findings, a mathematical model which is able to describe the complete polymer sintering process for viscoelastic fluids was recently developed [364]. Concerning bubble formation and removal in rotational molding, reviews of existing theories are provided in Ref. 365. However, these above-mentioned papers consider only each individual step of rotomolding instead of the whole process. In fact, the overall process is difficult to model due to the complexities caused by the biaxial rotation of the mold, the physical and phase changes which takes place in the plastic and the dynamic nature of the boundaries within the mold. Recent papers [366,367] consider an unsteady-state, nonlinear rotational molding heat transfer model. The model allows one to compute the internal air temperature profile, enabling an estimate of the oven time, cooling time, and the overall cycle time for various resins and for various thicknesses of molding, including very thick-wall products. In theory, during molding the heat is transferred from an oven, through a metal mold, then to a plastic bed and into the internal air space. Therefore, the heat transfer is broken down into a series of modules. The heat transfer in the metal mold and in the plastic bed is described by the following heat conduction equations, respectively [367]:

$$k_m \frac{\partial^2 T}{\partial r^2} + k_m \frac{\lambda}{r}\left(\frac{\partial T}{\partial r}\right) = \rho_m C_{pm}\left(\frac{\partial T}{\partial t}\right) \tag{23}$$

$$\frac{\partial}{\partial r}\left(k_p \frac{\partial T}{\partial r}\right) + k_p \frac{\lambda}{r}\left(\frac{\partial T}{\partial r}\right) = \rho_p C_{pp}\left(\frac{\partial T}{\partial t}\right) \tag{24}$$

where the subscripts m and p refer to the metal mold and plastics, respectively, k_m and k_p are thermal conductivities, C_{pm} and C_{pp} are heat capacities, ρ_m and ρ_p are densities, λ is the parameter which takes the value 0, 1, 2 for planar, cylindrical, and spherical shapes, T is temperature, t is time, and r is the radial co-ordinate in cylindrical or polar co-ordinates. To solve these equations, the initial and boundary conditions are required. The initial conditions for the initial temperatures are typically specified such that the temperature at all points in the metal mold wall is uniform. The boundary conditions include heat transfer between oven air and the metal mold surface, heat transfer at the interface of the metal mold and plastic bed, heat transfer between plastic bed and internal air and heat transfer of the internal air mass. The equations are solved by a finite-difference method [367]. Recently, the finite-element method was used in the simulation of rotational molding [368,369]. During rotational molding, the physical and thermal properties of the resin change significantly due to phase transformation. Therefore, the values of the density, heat capacity, and thermal conductivity as a function of temperature are required. In addition, the values of the heat transfer coefficients in the heating and cooling stages are required. However, applying the existing heat transfer correlations [370–373] to rotational molding is difficult due to their dependence on the geometry of the surface, the flow characteristics and physical properties of the fluid. Thus, the convection heat transfer coefficients are determined by fitting the simulated results with the experimental data [367]. The simulation enables the prediction of the oven time, cooling time, and the overall cycle time, thus, allowing evaluation and optimization of all machine variables. In particular, Fig. 23 shows the predicted and measured cycle and oven times for making cylindrical molding of various thicknesses [367]. The figure indicates a good agreement between the data, at least for moldings of lower thicknesses. This example shows that the computer simulation of rotational molding can be beneficial in evaluation and optimization of all machine variables.

8.11 THERMOFORMING

8.11.1 Technology

The origin of the thermoforming process dates back to ancient Egypt where food containers were made by heating tortoise shell in hot oil followed by shaping and cooling. Modern thermoforming began about 60 years ago with major developments in thermoplastic

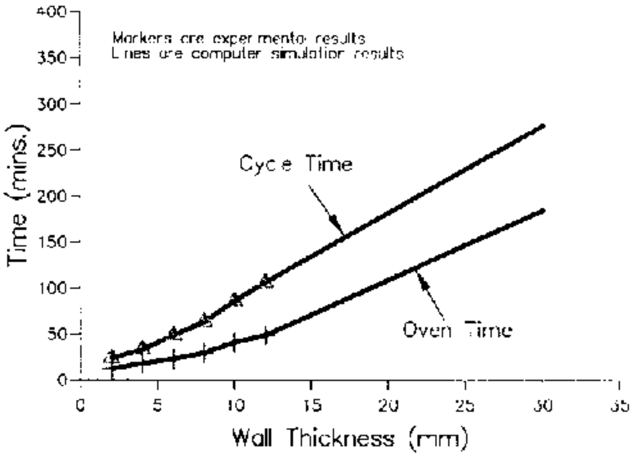


Figure 23 Comparison of computer simulation results with the experimental data for various thicknesses of cylindrical moldings made by rotational molding. (From Ref. 367, courtesy of the Institute of Materials.)

resins, screw extruders and roll-fed thermoforming machines. A number of monographs give an extensive description of thermoforming technology [374–377]. Thermoforming is defined as the process of heating a flat thermoplastic sheet, called a blank, to a softened state (above or close to T_g or T_m) and subsequent biaxial stretching of the sheet into a desired contoured shape of variable thickness by pneumatic and/or mechanical means and finally solidification into this shape by cooling. Thermoforming is not a single process, but rather a large number of related processes consisting of a few basic operations. One embodiment of this process is schematically shown in Fig. 24. Basic sequences of processes include heating a flat sheet, bending and/or stretching it into shape, cooling it to a solid state, and subsequent trimming of the part. Heating of the sheet above or close to T_g or T_m into a rubbery state is carried out using infrared, convection, and contact methods. Heaters with nickel–chrome wires, metal resistance rods, metal plates, ceramic bricks, quartz, gas-fired burners, heat lamps, and halogen bulbs are used. Bending and/or stretching is carried out by pneumatic and/or mechanical means. Pneumatic deformation is most commonly done by 710–735 mm Hg vacuum suction and/or positive pressure using dry air of carefully controlled pressure and flow rate. Mechanical deformation is used primarily for preforming or prestretching before final vacuum or air pressure forming.

Stretching is a primary mode of deformation in thermoforming. The main requirement for a polymer sheet subjected to thermoforming is to withstand sufficient elongation at the drawing temperature and applied pressure for making successful products. The stretching process (formability) is characterized by the draw ratio, which is defined by the areal and linear draw ratios and depths of draw. The areal draw ratio is the ratio of the area of the formed sheet to that of the unformed sheet. The linear draw ratio is the length of an imaginary line drawn on the formed sheet to its original length. The depth of draw is defined as the depth a sheet could be drawn into a female mold to the minimum dimension at the rim.

The cooling is the final stage of thermoforming. In the case of thin-gage products, cooling is done through contact with a cold mold provided with water flowing through channels. In the case of free surfaces of medium- and heavy-gage products, cooling is carried out by forced air, water mist, or water spray. After forming, products are separated from the surrounding web by a trimming operation by means of shearing cutting, saw, wheel, wire, router, or laser cutting.

Plastics used for thermoforming, in descending order of usage include polystyrenes, ABS, polypropy-

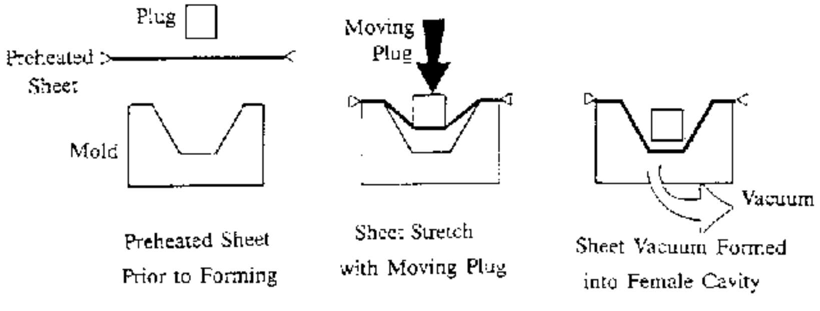

Plug
Preheated Sheet
Mold
Preheated Sheet Prior to Forming
Moving Plug
Sheet Stretch with Moving Plug
Vacuum
Sheet Vacuum Formed into Female Cavity

Figure 24 Schematic representation of plug-assisted thermoforming process.

lenes, PET, HDPE, PVC PMMA, cellulosics, and LDPE. Numerous plastic articles such as major appliance components (e.g., refrigerator cabinets and door liners), recreation products (e.g., swimming and wading pools) home products (e.g., tub and shower stalls), display items (e.g., signs), packaging, and disposable items used in the food industry and medical applications (e.g. skin and blister packs, containers, picnic plates) are made by thermoforming.

There are several methods of thermoforming, including one- and multistep process. A one-step process is defined as drape forming, vacuum forming, pressure forming, free blowing and matched die molding. The process of vacuum forming dominates the industry. Pressure forming allows very accurately reproducible depressions and protrusions of the mold at high forming speed and production rate. The match die molding is carried out between two mold halves (male and female) of somewhat similar contours. It is an ideal method for manufacturing containers with thin walls. Multistep thermoforming includes billowing, plug-assisted vacuum forming, plug-assisted pressure forming and plug-assisted drape forming. In the plug-assisted process a mechanically driven plug is used to stretch a sheet. Molds used in thermoforming are typically subjected to low stresses. Wood, plaster, and reinforced thermosets are used as mold materials in the case of short production runs and prototypes. Durable molds made of aluminum and steel are used in the case of mass production.

Recently, a monograph [374] has given an extensive overview of research and development activities in thermoforming until 1995. Since then a number of experimental investigations have been reported, related to various aspects of thermoforming. In particular, studies of thermoforming of a blend of styrene–butadiene block copolymer with polystyrene [378] and liquid crystalline polymers [379] and stamp forming of glass-fiber-reinforced polypropylenes [380,381] have been carried out. In an attempt to establish realistic processing targets for commercial processing operations, a statistical analysis of weights of thermoformed parts and their wall thickness has been performed [382]. These include thermoformed products made of a blend of a styrene–butadiene block copolymer with polystyrene, HDPE, and a blend of an acrylic and PVC resin. The potential of utilizing of white titanium dioxide pigment particles as a viable nucleation agent for polypropylenes in thermoforming has been considered [383]. The effect of processing parameters such as plug velocity, plug temperature, and film temperature on the compression resistance and wall thickness distribution of plug-assisted vacuum-formed containers of high-impact polystyrene has been investigated [384,385]. In particular, decreasing the stretching time and the temperature difference between plug and film is found to be important for good material distribution and compression resistance. Studies on microstructure, mechanical properties, and residual stresses in glass-fiber-reinforced polypropylenes as a function of thermoforming parameters and the fiber–matrix interface quality have been carried out [386]. The cooling rate was found to be a critical parameter of the molding process affecting these characteristics. Rubber-modified polypropylene cups have been prepared by vacuum thermoforming [387]. Fracture toughness in the longitudinal and transverse directions to the direction of drawing was measured. The presence of rubber particles results in reduced anisotropy, leading to an increase in the fracture toughness for cracks running transversally to the drawing direction with simultaneous reduction of the fracture resistance in the longitudinal direction. Cones have been made of neat and talc-filled HDPE, PP, and PPS [388]. The levels of orientation of the talc particles and polymer chains in the products were measured by the wide-angle X-ray diffraction pole figure technique [388]. The surface of the talc particles was found to be parallel to the surface of the thermoformed parts with the polymer chain orientation indicating crystal growth on the talc particle surfaces.

8.11.2 Modeling

The objective of modeling and computer simulation of the thermoforming process is to provide a rational means for mold and part design and setup of optimal processing parameters. Successful simulation methodology allows avoidance of numerous trial-and-error experimental tests required for designing product and development of manufacturing process. An extensive review of earlier simulation efforts in thermoforming is given in Ref. 269. As indicated in the review, initial efforts in simulating the thickness distribution in thermoformed articles were based on models without dependence on material behavior and were applicable only to simple geometry. Therefore, the thickness distribution is predicted from a simple mass balance and the imposed boundary constraints. The predictions based on these models were independent of material used and were unable to describe the thickness distribution in the real process. In plug-assisted thermoforming, in addition to accurate prediction of wall thickness distribution, the relation between the applied force on a plug and the sheet deflection and the locations of maximum stress and strain are important factors in process optimization.

Advances in simulation of thermoforming have been based on the finite-element method and incorporation of realistic material behavior. Various constitutive equations are employed to describe the deformation behavior of polymer in thermoforming. Similar to blow molding, the available approaches include the hyperelastic [389–399], viscoelastic [395,396,400,401], and viscoplastic [402,403] models. Since the stretching process in thermoforming is a primary mode of deformation, a suitable constitutive equation should describe behavior of polymers in various tensile tests at wide range of strains, strain rates, and temperatures. However, carrying out such tensile tests on polymers in temperature range of interest is a difficult task. A brief overview of the available techniques and their limitations are given in Ref. 269. Presently, it seems that the techniques based on a bubble inflation rheometer [404, 405], first developed in Refs. 406 and 407, and stretching apparatus with the rotary clamps [408] are most suitable methods. The stress–strain data obtained using these instruments are fitted to a constitutive model to determine model parameters necessary to carry out numerical simulation. The available simulation attempts of thermoforming are exclusively based on finite-element analysis. There are two main advantages of using a finite-element formulation for thermoforming. The formulation is applicable to any arbitrary geometry of the part. Also, the nonlinear behavior of polymer at large strains can be readily accommodated. Total lagrangian formulation with second Piola–Kirchhoff stress tensor and Green–Lagrange strain tensor are used in simulation. Earlier attempts of thermoforming are based on membrane approximation. This approximation is reasonable in the case of very thin sheet compared with its length without imposition of clamping boundary and mold wall boundary conditions. Thus, to describe the real situation full three-dimensional analysis is required [409,410]. More recent efforts in simulation of thermoforming devoted to prediction of process of manufacturing products from fiber-reinforced composites [411–415]. A finite-element model to describe radiant heating of polymer sheets is described in Ref. 316. As an example of a thermoforming simulation, the thermoforming of a rectangular three-dimensional box made of polyphenylene oxide is shown in Fig. 25 [269]. Due to symmetry, only the front half of the inflated sheet is shown. The mold is not shown for clarity. The figure gives a sequence of deformation experienced by the polymer sheet during thermoforming. The simulation is carried out using a finite-element analysis with an Ogden model for the description of the material behavior. The three-dimensional simulation is especially important in predicting the thickness variation that occurs in the corner due to severe thinning experienced by the polymer in this area.

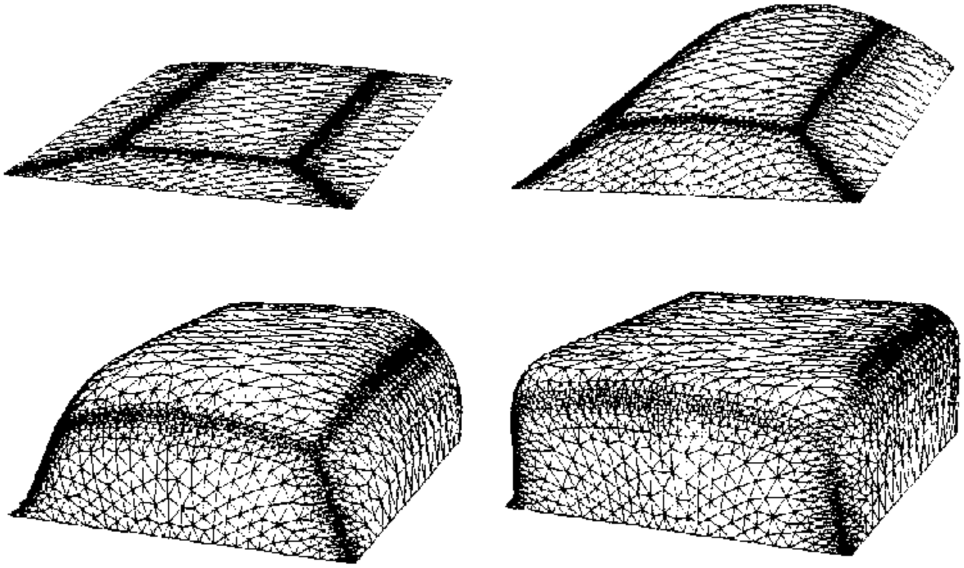


Figure 25 Simulation of thermoforming of a rectangular box. (From Ref. 269, courtesy of Hanser Publishers.)

8.12 CONCLUDING REMARKS

The present chapter on molding processes indicates that during the last two decades the development of science-based molding technologies has experienced major advances. These include development of new molding techniques and models for various existing molding processes along with their application to design processes in industrial environment. The major goal of this new development is to provide an optimal route for the quick delivery of new technologies and products to the marketplace. This is done through use of the computer-aided engineering which allows replacement of the old trial-and-error approaches with new scientific methodologies, leading to an increase in productivity and short stand-up times. However, a major challenge in this direction is to develop models with a more realistic description of polymer behavior and to include interactions between various stages of molding processes between their various stages and their influence on the microstructure and performance of molded products. Further progress in this area of the computer-aided engineering will have a major impact on the future expansion and productivity of molding processes.

REFERENCES

1. PG Progelhof, JL Throne. Polymer Engineering Principles: Properties, Processes, Tests for Design. Munich: Hanser, p. 1, 1993.
2. Z Tadmor, C Gogos. Principles of Polymer Processing. New York: Wiley, 1979.
3. DG Baird, DI Collias. Polymer Processing: Principles and Design. Boston: Butterworth-Heinemann, 1995.
4. S Middleman. Fundamentals of Polymer Processing. New York: McGraw-Hill, 1977.
5. CD Han. Rheology in Polymer Processing. New York: Academic Press, 1976.
6. JL. White. Principles of Polymer Engineering Rheology. New York: Wiley, 1990.
7. JL Throne. Plastics Process Engineering. New York: Marcel Dekker, 1979.
8. A Ram. Fundamentals of Polymer Engineering. New York: Plenum, 1997.
9. JM Charrier. Polymeric Materials and Processing. Munich: Hanser, 1990.
10. JRA Pearson. Mechanics of Polymer Processing. London: Elsevier, 1985.
11. J-F Agassant, P Avenas, J-Ph Sergent, PJ Carreau. Polymer Processing: Principles and Modeling. Munich: Hanser, 1991.
12. MM Denn. Process Fluid Mechanics. Englewood Cliffs, NJ: Prentice-Hall, 1980.
13. JL White. Rubber Processing: Technology, Materials, Principles. Munich: Hanser, 1995.
14. CL Tucker III. In: AI Isayev, ed. Injection and Compression Molding Fundamentals. New York: Marcel Dekker, 1987, chap 7.
15. T Reinhart, ed. Engineered Materials Handbook, vol 1, Composites, Metals Park: ASM International, 1987.
16. DV Rosato, DV Rosato. Injection Molding Handbook. New York: Van Nostrand, 1995.
17. H-J Kang, E Buchman, AI Isayev. SAMPE J 27: 21, 1991.
18. FN Cogswell. Intern Polym Process 1: 57, 1987.
19. A Buchman, AI Isayev. SAMPE J: 27, 19, 1991.
20. RJ Silva-Nieto, BC Fisher, AW Birley. Polym Eng Sci 21, 499, 1981.
21. EG Melby, JM Castro. In: SL Aggarwal, ed. Comprehensive Polymer Science. Oxford: Pergamon Press, vol 7, 1989, chap 3.
22. PV Malick, ed. Composite Engineering Handbook. New York: Marcel Dekker, 1997.
23. A Farouk, TH Kwon. Polym Compos 11: 379, 1990.
24. RH Pater. SAMPE J 30: 29, 1994.
25. AI Isayev. In: AI Isayev, T Kyu, SZD Cheng, eds. Liquid-Crystalline Polymer Systems: Technological Advances. Washington, DC: ACS Publishers, 1996.
26. SG Advani, ed. Flow and Rheology in Polymer Composites Manufacturing. Amsterdam: Elsevier, 1994.
27. MR Barone, DA Caulk. Polym Compos: 6, 105, 1985.
28. MR Barone, DA Caulk. J Appl Mech 53: 361 1986.
29. C. A. Hieber, SF Shen. J. Non-Newt Fluid Mech 7: 1, 1980.
30. MR Kamal, ME Ryan. In CL Tucker III, ed. Computer Modeling for Polymer Processing. Munich: Hanser, 1989, chap 2.
31. LJ Lee, LF Marker, RM Griffith. Polym Compos 2: 209, 1981.
32. GB Jeffery. Proc Roy Soc A102: 161, 1923.
33. F Folgar, CL Tucker III. J Reinf Plast Compos 3: 98, 1984.
34. M Gupta, KK Wang. Polym Compos 14: 367, 1993.
35. BE Ver Weyst, CL Tucker III, PH Foss. Int Polym Process 12: 238, 1997.
36. SG Advani, CL Tucker III. J Rheol 34: 367, 1990.
37. JS Cintra, Jr, CL Tucker III. J Rheol 39: 1095, 1995.
38. TH Kwon, CS Kim. Trans. ASME, J Eng Mater Technol 117: 239, 1995.
39. TA Osswald, CL Tucker. Polym Eng Sci 28: 413, 1988.
40. SJ Park, TH Kwon. J Manuf Sci Eng Trans ASME 120: 287, 1998.
41. II Rubin. Injection Molding: Theory and Practice, New York: Wiley, 1972.
42. EC Bernhardt, ed. Processing of Thermoplastic Materials. New York: Van Nostrand Reinhold, 1959.

43. DM Bryce. Plastic Injection Molding: Manufacturing Process Fundamentals. Dearborn: SME, 1996.
44. CW Macosko. Reaction Injection Molding. Munich: Hanser Publishers, 1989.
45. FM Sweeney. Reaction Injection Molding, Machinery and Processes. New York: Marcel Dekker, 1987.
46. BC Mutsuddy, RC Ford. Ceramic Injection Molding. London: Chapman and Hall, 1995.
47. RM German, A Bose. Injection Molding of Metals and Ceramics. Princeton, NJ: Metal Powder Industries Federation, 1997.
48. G Patsch, W Michaeli. Injection Molding: An Introduction. Munich: Hanser Publishers, 1995.
49. MA Wheelans. Injection Molding of Rubber. London: Butterworth, 1974.
50. AI Isayev, ed. Injection and Compression Molding Fundamentals. New York: Marcel Dekker, 1987.
51. G Menges, P Mohren. How to Make Injection Molds. Munich: Hanser Publishers, 1993.
52. R Wimberger-Friedl. Prog Polym Sci 20: 369, 1995.
53. GD Shyu, AI Isayev. SPE ANTEC 41: 2911, 1995.
54. H Janeschitz-Kriegl. Polymer Melt Rheology and Flow Birefringence. Berlin: Springer-Verlag, 1983.
55. TA Osswald, G Menges. Material Science of Polymers for Engineers. Munich: Hanser Publishers, 1995.
56. JF Stevenson, ed. Innovation in Polymer Processing: Molding. Munich: Hanser Publishers, 1997.
57. HEH Meijer. Processing of Polymers. Weinheim: VCH, 1997.
58. AI Isayev, X Guo, L Guo, M Demiray. SPE ANTEC 43: 1517, 1997.
59. CS Kwok, L Tong, JR White. Polym Eng Sci 36: 651, 1996.
60. CS Hindle, JR White, D Dawson, K Thomas. Polym Eng Sci 32: 157, 1992.
61. A Siegmann, A Buchman, S Kenig. Polym Eng Sci 22: 560, 1982.
62. M Fujiyama. Int Polym Process 10: 251, 1994.
63. Y Ulcer, M Cakmak. Polymer 35: 5651, 1994.
64. G Kalay, PS Allah, MJ Bevis. Kunstst Plast Eur 87: 768, 1997.
65. EM Grossman. SPE ANTEC 41: 461, 1995.
66. JP Ibar. Polym Eng Sci 38: 1, 1998.
67. KK Kabanemi, JF Hetu, A Garcia-Rejon. Int Polym Process 12: 182, 1997.
68. CL Tucker, ed. Computer Modeling for Polymer Processing. Munich: Hanser Publishers, 1989.
69. P Kennedy. Flow Analysis of Injection Molds. Munich: Hanser Publishers, 1995.
70. M Sobhanie, AI Isayev. In: AI Isayev, ed. Modeling of Polymer Processing. Munich: Hanser Publishers, 1991.
71. MM Cross. Rheol Acta 34: 329, 1979.
72. RS Spencer, GD Gilmore. J Appl Phys 20: 502, 1949.
73. RS Spencer, GD Gilmore,. J Colloid Sci 6: 118, 1951.
74. DV Van Krevelen. Properties of Polymers. Amsterdam: Elsevier, 1976.
75. MR Kamal, ME Ryan. In: CL Tucker, ed. Computer Simulation for Polymer Processing Munich: Hanser Publishers, 1989, chap 2.
76. D Huilier, WI Patterson. In: AI Isayev, ed. Modeling of Polymer Processing. Munich: Hanser Publishers, 1991, chap 6.
77. HH Chiang, CA Hiever, KK Wang. Polym Eng Sci 31: 116, 1991.
78. HH Chiang, CA Hiever, KK Wang. Polym Eng Sci 31: 1372, 1991.
79. AI Isayev, TW Chen, K Shimojo, M Gmerek. J Appl Polym Sci 55: 807, 1995.
80. AI Isayev, TW Chen, M Gmerek, K Shimojo. J Appl Polym Sci 55: 821, 1995.
81. P Zoller, DJ Walsh. Standard Pressure-Volume-Temperature Data for Polymers. Lancaster: Technomics, 1995.
82. VW Wang, CA Hieber, KK Wang. J Polym Eng 7: 21, 1986.
83. HP Wang, HS Lee. In: CL Tucker, ed. Computer Modeling for Polymer Processing. Munich: Hanser Publishers, 1989, chap 8.
84. AI Isayev, M Wan. Rubber Chem Technol 69: 277, 1996.
85. M Wan, AI Isayev. Rubber Chem Technol 69: 294, 1996.
86. RY Cheng, SY Chiou. Polym Eng Sci 35: 1733, 1995.
87. W Knappe, A Lampl. Kunststoffe 74: 7, 1984.
88. G Klepek. Kunststoffe 77: 13, 1987.
89. B Friedrichs, W Friesenbichler, K Gissing. Kunststoffe 80: 583, 1990.
90. SY Yang, MZ Ke. Adv Polym Technol 14: 15, 1995.
91. S Moore. Modern Plast 72: 23, 1995
92. SY Yang, L Nien. Adv Polym Technol 15: 205, 1996.
93. SY Yang, L Nien. Int Polym Process 11: 188, 1996.
94. B Friedrichs, M Horie, Y Yamaguchi. J Mater Process Manuf 5: 95, 1996.
95. Y Shiraishi, N Narazaki. Kobunshi Ronbunshi 53: 391, 1996.
96. E Escales. Kunststoffe 60: 847, 1970.
97. DF Oxley, DJH Standiford. Plast Polym 39: 288, 1971.
98. PJ Garner DF Oxley. British Patent (filed April 19, 1968) 1,156,217, 1971.
99. PJ Garner, DF Oxley. Offenlegungsschrift 1,778,451/7, 1971.
100. PJ Garner. U.S. Patent (filed December 6, 1968) 3,599,290, 1971.
101. PJ Garner. U.S. Patent 3,690797, 1972.
102. R Hanning. Offenlegungsschrift (filed August 21, 1972) 2,241,002, 1973.
103. E Langecker. Offenlegungschrift (filed September 27, 1972) 2,247,995, 1974.
104. W Filiman, H Eckardt. Offenlegungschrift 2,342,957, 1975.
105. RC Donovan, KS Rabe, WK Mammel, HA Lord. Polym Eng Sci 51: 774, 1975.

106. JL White, BL Lee. Polym Eng Sci 15: 481, 1975.
107. SS Yang, JL White, ES Clark, Y Oyanagi. Polym Eng Sci 20: 798, 1980.
108. P Somnuk, GF Smith. SPE ANTEC 41: 760, 1995.
109. GR Langecker. SPE ANTEC 43: 456, 1997.
110. WJ Schrenk, MA Barger, RE Ayers, RK Shastri. SPE ANTEC 39: 544, 1993.
111. MA Barger, WJ Schrenk, DF Pawlowski, CN Brown. SPE ANTEC 39: 550, 1993.
112. K Nichols, J Sole, M Barger, D Pawlowski, J Bicerano. SPE ANTEC 40: 1433, 1994.
113. LS Turng, VW Wang. SPE ANTEC 37: 297, 1991.
114. LS Turng, VW Wang, KK Wang. ASME HTD 175: 654, 1991.
115. LS Turng, VW Wang, KK Wang. J Eng Mater Technol 115: 48, 1993.
116. SC Chen, KF Hsu, KS Hsu. SPE ANTEC, 39: 82, 1993.
117. SC Chen, KF Hsu, KS Hsu. SPE ANTEC, 40: 182, 1994.
118. G Schlatter, A. Davidoff, J. F. Aggassant, M Vincent. SPE ANTEC, 41: 456, 1995.
119. E Vos, HEH Meijer, GWM Peters. Int Polym Process 6: 42, 1991.
120. GWM Peters, PJ van der Velden, HEH Meijer, P Schoone. Int Polym Process 9: 258, 1994.
121. WF Zoetelief, GW Peters, HEH Meijer. Int Polym Process 12: 216, 1997.
122. DJ Lee, AI Isayev, JL White. SPE ANTEC 44: 346, 1998.
123. E Broyer, Z Tadmor, C Gutfinger, Israel J Technol 11: 189, 1973.
124. E Broyer, Z Tadmor, C Gutfinger. Trans Soc Rheol 19: 423, 1975.
125. R Hanning. U.S. Patent (filed July 25, 1975) 4,033,710, 1977.
126. E Friederich. U.S. Patent (filed January 9, 1976) 4,101,617, 1978.
127. H Eckardt, J Ehritt. Plastverarbeiter 40: 14, 1989.
128. V Kapila, NR Schott, S Shah. SPE ANTEC 42: 649, 1996.
129. JF Stevenson. SPE ANTEC 42: 655, 1996.
130. WR Jong, JS Huang, YS Chang. SPE ANTEC 42: 668, 1996.
131. K Cutright, O Becker, KW Koelling. SPE ANTEC 44: 442, 1998.
132. J Zhao, X Lu, L Fong, HH Chiang. SPE ANTEC 44: 454, 1998.
133. SC Chen, JG Dong, WR Jong, JS Huang, MC Jeng. SPE ANTEC 42: 663, 1996.
134. O Becker, K Koelling, T Altan. SPE ANTEC 43: 610, 1997.
135. SC Chen, KS Hsu, JS Huang. Ind Eng Chem Res 34: 416, 1995.
136. SY Yang, SJ Liou. Adv Polym Technol 14: 197, 1995.
137. SY Yang, FZ Huang. Intern Polym Process 10: 186, 1995.
138. SY Yang, SJ Liou, WN Liou. Adv Polym Technol 16: 175, 1997.
139. SY Yang, FZ Huang, WN Liau. Polym Eng Sci 23: 2824, 1996.
140. CL Clark, R Williams, JS Dixon, S. Bi. Plast Eng 52: 35, 1996.
141. KW Koelling, RC. Kaminski. SPE ANTEC 42: 644, 1996.
142. PC Huzyak, KW Koelling. J Non-Newt Fluid Mech 71: 73, 1997.
143. AJ Poslinski, PR Oehler, VK Stokes. Polym Eng Sci 35: 877, 1995.
144. O Becker, D Karsono, K Koelling, T Altan. SPE ANTEC 43: 615, 1997.
145. DM Gao, A Garcia-Rejon, G Salloum, D Baylis. SPE ANTEC 43: 625, 1997.
146. SC Chen, KF Hsu, KS Hsu. Int J Heat Mass Transfer 39: 2957, 1996.
147. SC Chen, NT Cheng, MJ Chang. Mech Res Commun 24: 49, 1997.
148. YK Shen. Int Commun Heat Mass 24: 295, 1997.
149. LS Turng. Adv Polym Technol 14: 1, 1995.
150. LS Turng. Eng Plast 8: 171, 1995.
151. SC Chen, NT Cheng, KS Hus. Int Commun Heat Mass 22: 319, 1995.
152. SC Chen, NT Cheng, KS Hsu. Int J Mech Sci. 38: 335, 1996.
153. SC Chen, NT Cheng. Int Commun Heat Mass 23: 215, 1996.
154. SC Chen, NT Cheng, MJ Chang. Mech Res Commun 24: 49, 1997.
155. DM Gao, KT Nguyen, A Garcia-Rejon, G Salloum. Int Polym Process 12: 267, 1997.
156. SC Chen, SY Hu, WR Jong, MC Jeng, SPE ANTEC. 43: 620, 1997.
157. RY Chang, MH Tsai, CH Hsu. SPE ANTEC, 43: 598, 1997.
158. YS Soh and CH Chung. SPE ANTEC 43: 603, 1997.
159. RY Chang, SC Lin, KC Chen, CH Hsu, MH Tsai. SPE ANTEC 44: 438, 1998.
160. SC Chen, NT Cheng, SY Hu. J Appl Polym Sci 67: 1553, 1998.
161. DM Gao, KT Nguyen, JF Hetu, D Laroche, A Garcia-Rejon. Adv Perform Mater 5: 43, 1998.
162. SC Chen, SY Hsu, RD Chien, JS Huang. J Appl Polym Sci 68: 417, 1998.
163. TJ Wang, HH Chiang, X Lu, L Fong. SPE ANTEC 44: 447, 1998.
164. H Potente, M Hansen. Int Polym Process 8: 345, 1993.
165. FS Costa, W Thompson, C Friedl. In: SF Shen, P Dawson, eds Simulation of Materials Processing: Theory, Methods and Applications. Rotterdam: Balkema, 1995, p 1113.
166. R Khayat, A Dredouri, L Herbert. J Non-Newt Fluid Mech 57: 253, 1995.

167. GAAV Haagh, H Zuidema, FN van de Vosse, GWM Peters, HEH Meijer. Int Polym Process 12: 207, 1997.
168. AI Isayev, KD Vachagin, AM Naberezhnov. J Eng Phys 27, 998, 1974.
169. CA Hieber, RK Upadhyay, AI Isayev. SPE ANTEC 29: 698, 1981.
170. HSY Hsich, RJ Ambrose. In: AI Isayev, ed. Injection and Compression Molding Fundamentals. New York: Marcel Dekker, chap 5, 1987.
171. BN Dinzburg, R Bond. Int Polym Process 6: 3, 1991.
172. DP Seraphim, R Lasky, CY Li. Principles of Electronic Packaging. New York: McGraw-Hill, 1989.
173. LT Manzione. Plastic Packaging of Microelectronic Devices. New York: Van Nostrand Reinhold, 1990.
174. M Pecht. Handbook of Electronic Package Design. New York: Marcel Dekker, 1991.
175. LT Nguyen, M Pecht, EB Hakim. Plastic-Encapsulated Microelectronics: Materials, Processes, Quality, Reliability and Applications. New York: Wiley, 1995.
176. WR Jong, YL Chen. SPE ANTEC 44: 1206, 1998.
177. R Han, M Gupta. SPE ANTEC 44: 1211, 1998.
178. E Suhir. J Reinf Plast Compos 12: 951, 1993.
179. LT Nguyen A Danker, N Santhiran, CR Shervin. ASME Winter Annual Meeting, Nov 1992, p 27.
180. WR Jong, SC Chen, WF Hsu, TY Chen. SPE ANTEC 43: 1443, 1997.
181. JH Wu, AAO Tao, KS Yeo, TB Lim. IEEE Trans Comp Pack Manuf Technol Pt B Adv Pack 21: 65, 1998.
182. SJ Han, KK Wang. J Electr Pack 117: 178, 1995.
183. AMK Reddy, M Gupta. SPE ANTEC 42: 1453, 1996.
184. AB Strong. Manufacturing. In: SM Lee, ed. International Encyclopedia of Polymer Composites, vol 3. New York: CCH, 1989, p 102.
185. D White. Resin Transfer Molding. In: SM Lee, ed. International Encyclopedia of Polymer Composites, vol 3. New York: VCH, 1989, p 506.
186. LJ Lee. In: T Gutowski, ed. Advanced Composite Manufacturing. New York: Wiley, Ch. 10, p. 393, 1997.
187. JM Castro, CW Macosko. AIChE J., 28, 250 (1982).
188. LJ Lee. In: SL Aggrawal, ed. Reaction Injection Molding in Comprehensive Polymer Science, vol 7, London: Pergamon, 1989, p 379.
189. TL Luce, SG Advani, JG Howard, RS Parnas. Polym Compos. 16: 446, 1995.
190. DL Woerdeman, FR Phelan, RS Parnas. Polym Compos 16: 470, 1995.
191. WB Young, SF Wu. J Reinf Plast Compos 14: 1108, 1995.
192. YT Chen, HT Davis, CW Macosko. AIChE J 41: 2261, 1995.
193. K Han, LJ Lee. J Compos Mater 30: 1458, 1996.
194. A Hammami, F Trochu, R Gauvin, S Wirth. J Reinf Plast Compos 15: 552, 1996.
195. V Rohatgi, N Patel, LJ Lee. Polym Compos 17: 161, 1996.
196. N Patel, LJ Lee. Polym Compos 16: 386, 1995.
197. W Michaeli, V Hammes, L Kirberg, R Kotte, TA Osswald, O Specker. Process Simulation in the RTM Technique. Munich: Hanser Publishers, 1989.
198. WK Chui, J Glimm, FM Tangerman, JS Jardine, JS Madsen, TM Donnellain, R Leek. SIAM Rev 39: 714, 1997.
199. CL Tucker. Polym Compos 17: 60, 1996.
200. RR Varma, SG Advani. In: Advances in Finite Element Analysis in Fluid Dynamics. New York ASME, FED 200, 1994, p 21.
201. JM Castro, S Conover, C Wilkes, K Han, HT Chiu, LJ Lee. Polym Compos 18: 585, 1997.
202. S Li, R Gauvin. J Reinf Plast Compos 10: 314, 1991.
203. WB Coulter, SI Guceri. Manufacturing Science of Composites. ASME Proc 4: 79, 1988.
204. B Friedrichs, SI Guceri. Polym Eng Sci 35: 1834, 1995.
205. H Aoyagi, M Uenoyama, SI Guceri. Int Polym Process 7: 71, 1992.
206. MK Um, WT Lee. Polym Eng Sci 31: 765, 1991.
207. H Hadavinia, SG Advani, RT Fenner. Eng Analys Bound Elem. 16: 183, 1995.
208. FM Schmidt, F Berthet, P Devos. In: J Huetink, FPT Baaijens, eds. Simulation Materials Processing: Theory, Methods and Applications. Rotterdam: Balkema, 1998, p 453.
209. YE Yoo, WI Lee. Polym Compos 17: 368, 1996.
210. LJ Lee, WB Young, RJ Lim. Compos Struct 27: 109, 1994.
211. TJ Wang, LJ Lee, WB Young. Int Polym Process. 10: 82, 1995.
212. TJ Wang, RJ Lin, LJ Lee. Int Polym Process 10: 364, 1995.
213. WB Young, K Rupel, K Han, LJ Lee, MJ Liou. Polym Compos. 12: 20, 1991.
214. WB Young, K Han, LH Fong, LJ Lee, MJ Liou. Polym Compos 12: 29, 1991.
215. KM Pillai, SG Advani. Polym Compos 19: 71, 1998.
216. SW Kim, KJ Lee, JC Seferis, JD Nam. Adv Polym Technol. 16: 185, 1997.
217. K Han, LJ Lee, A Shafi, D White. J Compos Mater 30: 1475, 1996.
218. VMA Calado, SG Advani. Compos Sci Technol 56: 519, 1996.
219. VR Voller, S Peng. Polym Eng. Sci. 22: 1758, 1995.
220. MV Bruschke, SG Advani. Polym Compos 11: 398, 1990.
221. R Lin, LJ Lee, M Liou. Int Polym Process 6: 356, 1991.
222. F Trochu, R Gauvin, DM Gao, JF Bourealut. J Reinf Plast Compos 13: 262, 1994.
223. M Lin, HT Hahn, H Huh. Compos. Pt A Appl Sci Manuf 29: 541, 1998.

224. A Hammami, R Gauvin, F Trochu. Compos Pt A Appl Sci Manuf 29: 603, 1998.
225. A Hammami, R Gauvin, F Trochu, O Touret, P Ferland. Appl Compos Manuf 5: 1161, 1998.
226. F Trochu, P Ferland, R Gauvin. Sci Eng Compos Mater 6: 209, 1997.
227. YM Ismail, GS Springer. J Compos Mater 31: 954, 1997.
228. W Chang, N Kikuchi. Int J Comput fluid 7: 49, 1996.
229. S Ranganathan, FR Phelan, SG Advani. Polym Compos 17: 222, 1996.
230. WB Young, MT Chuang. J Compos Mater 29: 2192, 1995.
231. HGH van Melick, GAAV Haagh, FN van der Vosse, T Peijs. Adv Compos Lett 7: 17, 1998.
232. LW Hourng, CY Chang. J Reinf Plast Compos 12: 1081, 1993.
233. CJ Wu, LW Hourng, JC Liao. J Reinf Plast Compos 14: 694, 1995.
234. CY Chang, LW Hourng, CJ Wu. J Reinf Plast Compos 16: 566, 1997.
235. CY Chang, LW Hourng. J Reinf. Plast Compos 17: 165, 1998.
236. LW Hourng, CY Chang. J Reinf Plast Compos 17: 2, 1998.
237. CY Chang, LW Hourng. Polym Eng Sci 38: 809, 1998.
238. FR Phelan. Polym Compos 18: 460, 1997.
239. DV Rosato, DV Rosato, eds. Blow Molding Handbook. Munich: Hanser, 1989.
240. SI Belcher. In: SL Aggarwal, ed. Comprehensive Polymer Science, vol 7. Oxford: Pergamon, 1989, p 489.
241. SN Lee. Blow Molding Design Guide. Munich: Hanser, 1998.
242. H Belofsky. Plastics: Product Design and Process Engineering. Munich: Hanser, 1995.
243. GPM Schenkel. Int Polym Process 3: 3, 1988.
244. C Hancock. British Patent 11,208, 1846.
245. ST Armstrong. U.S. Patent 8180, 1851.
246. WB Carpenter. U.S. Patent 237,168, 1881.
247. DE Mouzakis, J Karger-Kocsis. J Appl Polym Sci. 68: 561, 1998.
248. N Sheptak, CE Beyer. SPE J 21: 190, 1965.
249. NR Wilson, ME Bentley, BT Morgen. SPE J 26: 34, 1970.
250. KC Chao, WCL Wu. SPE J 27: 37, 1971.
251. ED Henze, WCL Wu. Polym Eng Sci 13: 153, 1973.
252. G Ajroldi. Polym Eng Sci 18: 742, 1978.
253. A Dutta, ME Ryan. J Non-Newt Fluid Mech 10: 235, 1982.
254. D Kalyon, V Tan, MR Kamal. Polym Eng Sci. 20: 773, 1980.
255. MR Kamal, V Tan, D Kalyon. Polym Eng Sci 21: 331, 1981.
256. DM Kalyon, MR Kamal. Polym Eng Sci 26: 508, 1986.
257. RW DiRaddo, WI Patterson, MR Kamal. Adv Polym Technol 8: 265, 1988.
258. AH Wagner, DM Kalyon. Polym Eng Sci 36: 1897, 1996.
259. A Garcia-Rejon, JM Dealy. Polym Eng Sci 22: 158, 1982.
260. N Orbey, JM Dealy. Polym Eng Sci 24: 511, 1982.
261. JM Dealy, N Orbey. AIChE J 31: 807, 1985.
262. PL Swan, JM Dealy, A Garcia-Rejon, A Derdouri. Polym Eng Sci 31:L 705, 1991.
263. JL White, A Agrawal. Polym Eng Rev 1: 267, 1981.
264. MR Kamal, D Kalyon, V Tan. Polym Eng Sci 22: 287, 1982.
265. MR Kamal, D Kalyon. Polym Eng Sci 23: 503, 1983.
266. KJ Choi, JE Spruiell, JL White. Polym Eng Sci 29: 463, 1989.
267. C Bonnebat, G Roullet, AJ de Vries. Polym Eng Sci 21: 189, 1981.
268. M Cakmak, JE Spruiell, JL White. Polym Eng Sci 24: 1390, 1984.
269. HG de Lorenzi, HF Nied, In AI Isayev, ed. Modeling of Polymer Processing. Munich: Hanser, 1991, p 117.
270. HG deLorenzi, CA Taylor, Int Polym Process 8: 365, 1993.
271. WR Haessly, ME Ryan. Polym Eng Sci 33: 1279, 1993.
272. RW Diraddo, Garcia-Rejon. Polym Eng Sci 34: 1080, 1994.
273. K Hartwig, W Michaeli. In: SF Shen, P Dawson, eds. Simulation of Materials Processing: Theory, Methods and Applications. Rotterdam: Balkema, 1995, p 1029.
274. P Wriggers, RL Taylor. Eng Comput 7: 303, 1990.
275. N Santhanam, H Himasekhar, KK Wang. In: SF Shen, P Dawson, eds. Simulation of Materials Processing: Theory, Methods and Applications. Rotterdam: Balkema, 1995, p 1059.
276. ME Ryan, A Dutta. Polym Eng Sci 22: 569, 1982.
277. ME Ryan, A Dutta. Polym Eng Sci 22: 1075, 1982.
278. A Dutta, ME Ryan. Polym Eng Sci 24: 1232, 1984.
279. AJ Poslinski, JA Tsamopoulos. AIChE J 36: 1837, 1990.
280. B Debbaut, B. Hocq, JM Marchal. SPE ANTEC 41: 908, 1995.
281. B Debbaut, T Marchal. SPE ANTEC 43: 802, 1997.
282. A Wineman. J Non-Newt Fluid Mech 6: 111, 1979.
283. Y Otsuki, T Kajiwara, K Funatsu. Polym Eng Sci 37: 1171, 1997.
284. B Debbaut, O Homerin, A Goublomme, N Jivraj. SPE ANTEC 44: 784, 1998.
285. FM Schmidt, JF Aggasant, M Bellet, L Desoutter. J Non-Newt Fluid Mech 64: 19, 1996.
286. D Laroche, F Erchiqui. In: J Huetink, FPT Baaijens, eds. Simulation of Materials Processing: Theory, Methods and Applications. Rotterdam: Balkema, 1998, p 483.
287. S Tanoue, T Kajiwara, K Funatsu, K. Terada, Y Yamabe. Polym Eng Sci 36: 2008, 1996.

288. A Cohen, JT Seitz. Int Polym Process 6: 51, 1991.
289. M Bellet, JF Aggasant, A Rodriguez-Villa. In: J Huetink, FPT Baaijens, eds. Simulation of Materials Processing: Theory, Methods and Applications. Rotterdam: Balkema, 1998, p 489.
290. RB Bird, RC Armstrong, O Hassager. Dynamics of Polymeric Liquids. New York: Wiley, 1987.
291. AC Eringen. Nonlinear Theory of Continuous Media. New York: McGraw-Hill, 1962.
292. AD Green, JE Adkins. Large Elastic Deformation. Oxford: Oxford University Press, 1960.
293. OC Zienkiewicz. The Finite Element Method. London: McGraw-Hill, 1977.
294. JT Oden. Finite Elements of Nonlinear Continua. New York: McGraw-Hill, 1972.
295. HG deLorenzi, HF Nied. Comput Struct 26: 197, 1987.
296. LR Whittington, Dictionary of Plastics. Lancaster: Technomics, 1978.
297. TJ Lovegrove, U.S. Patent 48,022, 1865.
298. FAVoelke. U.S. Patent 803,799, 1905.
299. GS Baker, GW Perks, U.S. Patent 947,405, 1910.
300. JJ Jensen, U.S. Patent 1,812,242, 1931.
301. BL Zousman, EJ Finnegan. Chocolate and Cocoa. In: RE Kirk, DE Othmer, eds. Encyclopedia of Chemical Technology, vol 6. 1979, p 17.
302. PT Dodge. Rotational Molding. In: II Rubin, ed. Handbook of Plastic Materials and Technology. New York: Wiley, 1990.
303. RL Powell. U.S. Patent 1,341,670, 1920.
304. JH Emery. U.S. Patent 1,373,211, 1921.
305. D Fauerbach. U.S. Patent 1,784,686, 1930.
306. WH Campbell. U.S. Patent 1,792,813, 1931.
307. W Kay. U.S. Patent 1,998,897, 1935.
308. GW Trowbridge, U.S. Patent 2,035,774, 1936.
309. GR Kelm. U.S. Patent 2,262,431, 1941.
310. GE West, U.S. Patent 2,042,975, 1936.
311. RP Molitor. U.S. Patent 2,629,134, 1953.
312. S Zweig. Rotational Molding of Plastisols. Mod Plast 33: 123, 1953.
313. RE Gulick. The Economics of Rotomolding Powdered Polyethylene. Mod Plast 39: 102, 1962.
314. RV Jones, RL Rees. Applications of Polyolefin Powder. SPE J June: 80, 1967.
315. RO Ebert. Progress in Powder Molding, Plast Technol 12: 58, 1966.
316. PT Dodge. Rotational Molding. In: Encyclopedia of Polymer Science and Engineering, New York: Wiley, vol 14, 1988, p 659.
317. PF Bruins, ed. Basic Principles of Rotational Molding. New York: Gordon and Breach, 1971.
318. D Ramazatti. Rotational Molding. In: E Miller, ed. Plastic Product Design Handbook. New York: Marcel Dekker, 1983, chap 4.
319. RJ Crawford, ed. Rotational Moulding of Plastics. New York: Wiley, 1992.
320. PT Dodge. Materials for Rotational Molding. Denver, CO: Plastics Design Forum, 1994.
321. G Bell. The Engineers Guide to Designing Rotationally Molded Plastic Parts. Chicago, IL: Association of Rotational Molders, 1982.
322. JL Throne. Polym Eng Sci 20: 899, 1980.
323. E Harkin-Jones, RJ Crawford. Adv Polym Technol 15: 71, 1996.
324. E Harkin-Jones, RJ Crawford. SPE ANTEC 44: 1148, 1998.
325. E Harkin-Jones, RJ Crawford. Polym Eng Sci 36: 615, 1996.
326. E Harkin-Jones, RJ Crawford. Plast Rubber Compos Process Appl 24: 1, 1995.
327. E Harkin-Jones, RJ Crawford. Plast Rubber Compos Process Appl 23: 211, 1995.
328. JL Throne. Polym Eng Sci 12: 335, 1972.
329. JL Throne. Polym Eng Sci 16: 192, 1976.
330. RJ Crawford, JA Scott. Plast Rubber Process Appl 5: 239, 1985.
331. K Iwakura, Y Ohta, CH Chen, JL White. Int Polym Process 4: 163, 1989.
332. S Bawiskar, JL White. Int Polym Process 10: 62, 1995.
333. L Xu, RJ Crawford. Plast Rubber Compos Process Appl 21: 257, 1994.
334. RC Progelhof G Cellier, JL Throne, SPE ANTEC 28: 627, 1982.
335. RJ Crawford, JA Scott. Plast Rubber Process Appl 7: 85, 1987.
336. AG Spence, RJ Crawford. Polym Eng Sci 36: 993, 1996.
337. L Xu, RJ Crawford. J Mater Sci 28: 2067, 1993.
338. MA Rao, JL Throne. Polym Eng Sci 12: 237, 1972.
339. MS Sohn, JL Throne. Adv Polym Technol. 9: 181, 1989.
340. JL Throne, MS Sohn. Adv Polym Technol. 9: 193, 1989.
341. K Iwakura, Y Ohta, CH Chen, JL White. Int Polym Process 4: 76, 1989.
342. Y Ohta, CH Chen, JL White. Kunststoffe 79: 1349, 1989.
343. CH Chen, JL White, Y Ohta. Polym Eng Sci 30: 1523, 1990.
344. CH Chen, JL White, Y Ohta. Int Polym Process 6: 212, 1991.
345. S Bawiskar, JL White. Polym Eng Sci 34: 815, 1994.
346. Sj Liu, CY Ho. SPE ANTEC 44: 1156, 1998.
347. PJ Nugent, RJ Crawford, L Xu. Adv Polym Technol 11: 181, 1992.
348. M Kontopoulou, M Bisaria, J Vlachopoulos. Int Polym Process 12: 165, 1997.
349. RJ Crawford, PJ Nugent. Plast Rubber Compos Process Appl 17: 33, 1992.
350. RJ Crawford. J Mater Process Technol 56: 263, 1996.

351. RJ Crawford, P Nugent, W Xin. Int Polym Process 6: 56, 1991.
352. MJ Oliveira, MC Cramez, RJ Crawford. J Mater Sci 31: 2227, 1996.
353. MA Rao, JL Throne. Polym Eng Sci 12: 237, 1972.
354. JL Throne. Polym Eng Sci 12: 335, 1972.
355. JL Throne. Polym Eng Sci 16: 257, 1976.
356. RC Progelhof, JL Throne. Polym Eng Sci 16: 680, 1976.
357. S Bawiskar, JL White. Int Polym Process 10: 62, 1995.
358. RJ Crawford, P Nugent. Plast Rubber Process Appl 11: 107, 1989.
359. FS Ribe, RC Progelhof, JL Throne. SPE ANTEC 32: 20, 1986.
360. FS Ribe, RC Progelhol, JL Throne. SPE ANTEC 32: 25, 1986.
361. CT Bellehumeur, MK Bisaria, J Vlachopoulos. Polym. Eng. Sci. 36: 2198, 1996.
362. O Pokluda, CT Bellehumeur, J Vlachopoulos. AIChE J. 43: 3253, 1997.
363. J Frenkel. J Phys 9: 385, 1945.
364. CT Bellehumeur, M Kontopoulou, J Vlachopoulos. Rheol Acta 37: 270, 1998.
365. RJ Crawford, JA Scott. Plast Rubber Process Appl 7: 85, 1987.
366. DW Sun, RJ Crawford. Plast Rubber Compos Process Appl 19: 47, 1993.
367. L Xu, RJ Crawford. Plast Rubber Compos Process Appl 21: 257, 1994.
368. LG Olson, G Gogos, V Pasham, X Liu. SPE ANTEC 44: 1116, 1998.
369. G Gogos, X Liu, LG Olson. SPE ANTEC 44: 1133, 1998.
370. HY Wong. Heat Transfer for Engineers. New York: Longman, 1977.
371. F Kreith, M Bohn. Principles of Heat Transfer. New York: Happer International, 1986.
372. MN Ozisik. Heat Transfer—A Basic Approach. New York: McGraw-Hill, 1987.
373. JP Holman. Heat Transfer. New York: McGraw-Hill, 1990.
374. JL Throne. Technology of Thermoforming. Munich: Hanser Publishers, 1996.
375. JL Throne. Thermoforming. Munich: Hanser Publishers, 1987.
376. J Florian. Practical Thermoforming: Principle and Applications. New York: Marcel Dekker, 1987.
377. G Gruenwald. Thermoforming: A Plastics Processing Guide. Lancaster: Technomic, 1987.
378. MJ Stephenson, ME Ryan. Polym Eng Sci 37: 450, 1997.
379. M Mogilevsky, A Siegmann, S Kenig. Polym Eng Sci 38L: 322, 1998.
380. M Hou. Compos Pt A Appl Sci Manuf 28: 695, 1997.
381. K Friedrich, M Hou. Compos Pt A Appl Sci Manuf 29: 217, 1998.
382. ME Ryan, MJ Stevenson, K Grosser, LJ Karadin, P Kaknes. Polym Eng Sci 36: 2432, 1996.
383. NJ MaCauley, EMA Harkin-Jones, WR Murphy. Polym Eng Sci 38: 516, 1998.
384. A Aroujalian, MO Ngadi, JP Emod. Adv Polym Technol 16: 129, 1997.
385. A Aroujalian, MO Ngado, JP Emond. Polym Eng Sci 37: 178, 1997.
386. Y Youssef, J Denault. Polym Compos 19: 301, 1998.
387. A Pegoretti, A Marchi, T Ricco. Polym Eng Sci 37: 1045, 1997.
388. CH Suh, JL White. Polym Eng Sci 36: 2188, 1996.
389. HF Nied, CA Taylor, HG deLorenzi. Polym Eng Sci 30: 1314, 1990.
390. HG deLorenzi, HF Nied, CA Taylor. J Press Vessel Technol Trans. ASME 113: 102, 1991.
391. CA Taylor, HG deLorenzi, DO Kazmer. Polym Eng Sci 32: 1163, 1992.
392. WN Song, FA Mirza, J Vlachopoulos. J Rheol. 35: 92, 1991.
393. WN Song, FA Mirza, J Vlachopoulos. Int Polym Process 7: 248, 1992.
394. K Kouba, O Bartos, J Vlachopoulos. Polym Eng Sci 32: 699, 1992.
395. F Doria, P Bourgin, L Coincenot. Adv Polym Technol 14: 291, 1995.
396. P Bourgin, I Cormeau, T SaintMatin. J Mater Process Technol 54: 1, 1995.
397. GJ Nam, HW Ree, JW Lee, SPE ANTEC 44: 690, 1998.
398. D Laroche, F Erchiqui. SPE ANTEC 44: 676, 1998.
399. M Rachik, JM Roelandt. In: J Huetink, FPT Baaijens, eds. Simulation of Materials Processing: Theory, Methods and Applications. Amsterdam: Balkema, 1998, p 447.
400. A Rodriguez-Villa, JF Agassant, M Bellet. In: SF Shen, P Dawson, eds. Simulation of Materials Processing: theory, Methods and Applications. Amsterdam: Balkema, 1995, p 1053.
401. FM Schmidt, JF Agassant, M Bellet, L Desoutter. J Non-Newt Fluid Mech. 64: 19, 1996.
402. MH Vantal, B Monasse, M Bellet. In: SF Shen, P Dawson eds. Simulation of Materials Processing: Theory, Methods and Applications. Amsterdam: Balkema, 1995, p 1089.
403. S Wang, A Makinouchi, M Okamoto, T Kotaka, T Tosa, K Kidokoro, T Nakagawa. In: J Huetink, FPT Baaijens, eds. Simulation of Materials Processing: Theory, Methods and Applications. Amsterdam: Balkema, 1998, p 441.
404. W Michaeli, K Hartwig. Kunststoffe-Plast. Europe 86: 85, 1996.
405. A Derdouri, R Connoly, R Khayat. SPE ANTEC 44: 672, 1998.
406. DD Joye, GW Pohlein, CD Denton. Trans Soc Rheol 16: 421, 1973.

407. DD Joye, GW Pohlein, CD Denton. Trans Soc Rheol 17: 287, 1973.

408. J Meissner, J Hosettler. Rheol Acta 33: 1, 1994.

409. B Debbaut, O Homerin. SPE ANTEC 43: 720, 1997.

410. TM Marchal, NP Clemeur, AK Agarwal, SPE ANTEC 44: 701, 1998.

411. AK Pickett, T Queckborner, P Deluca, E Haug. Compos Manuf 6: 237, 1995.

412. AA Polynkine, F vanKeulen, H DeBoer, OK Bergsma, A Beukers. Struct Optim 11: 228, 1996.

413. SJ Liu. Polym Compos 18: 673, 1997.

414. SW Hsiao, N Kikuchi. J Eng Mater Technol Trans ASME 119: 314, 1997.

415. J Schuster, K Friedrich. Compos Sci Technol. 57: 405, 1997.

416. MA Thrasher. J Reinf Plast Compos 16: 1363, 1997.

Chapter 7.1

Material Handling and Storage Systems

William Wrennall
The Leawood Group Ltd., Leawood, Kansas

Herbert R. Tuttle
University of Kansas, Lawrence, Kansas

1.1 INTRODUCTION

Material handling and storage systems planning and design are subsets of facilities planning and design. Material flow has both internal and external effects on a site. There are influences for the site plan and the operations space plan. Conversely, the material handling system impacts the facility plans, as illustrated in Fig. 1.

In the facilities design process the material movement determines the flow paths. The material movement origins and destinations are layout locations. The storage locations and steps are effects of the operations strategy and thus the organization structure. A lean manufacturing system may have material delivery direct to point of use replenished daily and a pull system cellular manufacturing process that produces to order with a TAKT* time of 5 min. Such a system could have inventory turns of 300 per year. A more traditional system would have a receiving inspection hold area, a raw material/purchased parts warehouse, a single shift functional layout batch manufacturing system, inspection and test with 90% yield, a separate packing department, and a policy of one month's finished goods inventory. The space plans for the traditional system *should be* very different from the *lean* approach and *so should* the material handling and storage plans and systems. A "*pull*" system also indicates unnecessary material in the system. *If it does not pull it should not be there*.

Material handling determines the capacity of a manufacturing plant. From the receiving dock to the shipping platform the material flow routes are the circulation system. Flow restrictions can act as capacity limiters. The material handling and storage plan determines handling and storage methods, unit loads and containerization to support the operations and business strategy.

The product volume plot—the plot of volume/quantities of materials by product typically shows a negative exponential distribution, the important few and the trivial many Pareto distribution. The plot can be overlaid with the most suitable production mode as illustrated in the *product volume (PV)/mode curve*, Fig. 2.

We have suggested the following modes:

1. Project
2. Cellular
3. Linked cellular
4. Line
5. Continuous.

*TAKT time is the rate at which your customer requires product.

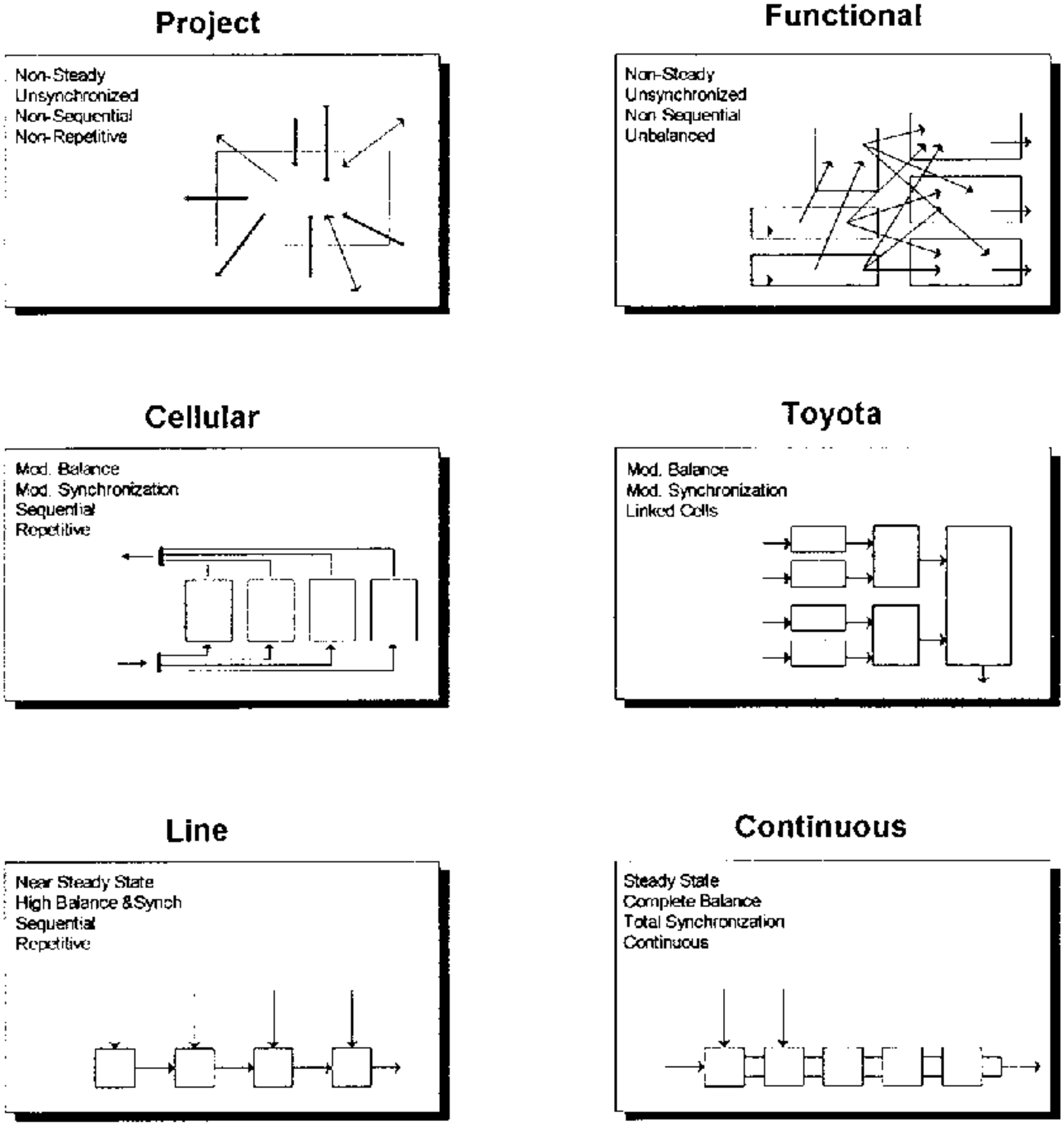


Figure 1 Material flow patterns.

We find these classifications more meaningful than fixed location, job shop, functional, mass production line, and process flow.

In the manufacture of discrete solid parts their transportability is improved by placing them in containers. This contained regular shape becomes the unit load and the material handling method is matched to the load. As material flow volumes increase, the characteristics trend to those of continuous flow, i.e., from solid discrete parts to bulk (flowable) solids, liquids, and gases. Solids are handled as pieces or contained in baskets, cartons etc. Ware [1] describes how solids can also be conveyed by vibratory machines. Liquids and gases are always contained and they conform to the shape of the container. The container may also be the channel of material transfer, such as a pipeline or duct, particularly for flowable bulk solids. Bulk solids can also be transported along pipes or ducts with the aid of a suspension agent, such as a gas or liquid.

In just-in-time (JIT)/lean manufacturing the aim of batch sizes of one is to emulate continuous liquid or gaseous molecular flow characteristics to achieve simplification.

Designs for material handling in liquids, gases, and bulk solids are an integral part of the process. Examples are chemical and food processes. In an oil

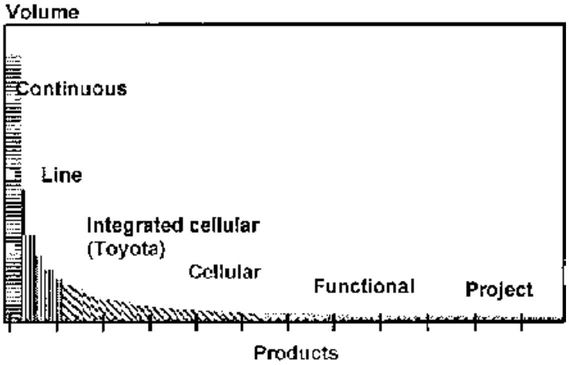


Figure 2 Product volume/mode curve.

refinery the input of raw materials to shipment of products is often totally integrated.

In process design it is always desirable to move up the PV/mode curve. Process efficiency increases from project mode to continuous flow. This *economy* was achieved originally by increasing volumes and thus increasing inventories. More recently, economies of operations with low volumes have been achieved by a change of focus/paradigm through rapid added value to materials yielding increased return on assets employed. The assumption had previously been made that the economies came from volume only. However, material handling unit volumes and storage requirements have shrunk with the use of:

Family manufacturing
Cellular operations
Product focus
Rapid setups
Pull systems
Batch sizes of one
Make to customer order
Mixed product less than full load shipments
Half machines
Lean operations.

Material movement dominates the design of many facilities. Conversely, the layout design sets the locations and origins of each material move. Movement adds cost, time, and complexity to manufacturing and distribution. It adds no ultimate or realized value until the final move—delivery to the customer. The priority therefore is to reduce material movement.

The minimizing of material movement requires an effective layout based on a sound manufacturing strategy. The affinities and focus approaches which can reduce both the amount and complexity of handling are powerful tools. They are described in Wrennall [2].

Layout affects material flow in three ways. First, the space planning units (SPUs) definitions set a pattern for overall material flows. Figure 1, given earlier, illustrates how production mode impacts both the intensity and complexity of flows. Second, the arrangement of SPUs in the layout can increase or decrease particular route distances. Third, the arrangement of SPUs sets a large-scale flow pattern (or nonpattern in some cases). Since layout design and material handling are interdependent, so is a discussion on the design of either or both of them.

1.2 MATERIAL FLOW ANALYSIS

Material flow analysis (MFA) examines the movement of materials over time. It helps develop affinities for the layout and evaluation of layout options, and is basic to the design of material handling systems. Unlike other reasons for affinities, material flow is tangible and measurable. The use of quantitative methods adds rigor to the facility layout planning process. After rigorous analysis and simplification, the remaining and necessary material moves are economical and timely.

Ultimately, all assembly processes are material handling. This discussion limits the handling of materials to and from a site and from place to place within the site. Material handling at the workplace and associated automation is discussed elsewhere.

The objectives of material flow analysis (MFA) are to compute affinities based on material flow, evaluate layout options, and assist handling system design. A macrolayout of 30 SPUs has 435 possible material flow routes. In addition, most facilities have many materials, processes, handling methods, and multiple movement paths with flows in both directions. Figure 3 illustrates some of the possible material variety.

Handling equipment, containers, route structures, and control methods all present additional variety. Therefore, analysis and the subsequent development of handling systems can be complex and difficult. This chapter explains how to bring order and structure to the process.

The MFA steps, shown in Fig. 4 are:

1. Classify materials
2. Define flow unit
3. Determine data source
4. Extract data
5. Format and analyze
6. Calibrate flows
7. Represent graphically.

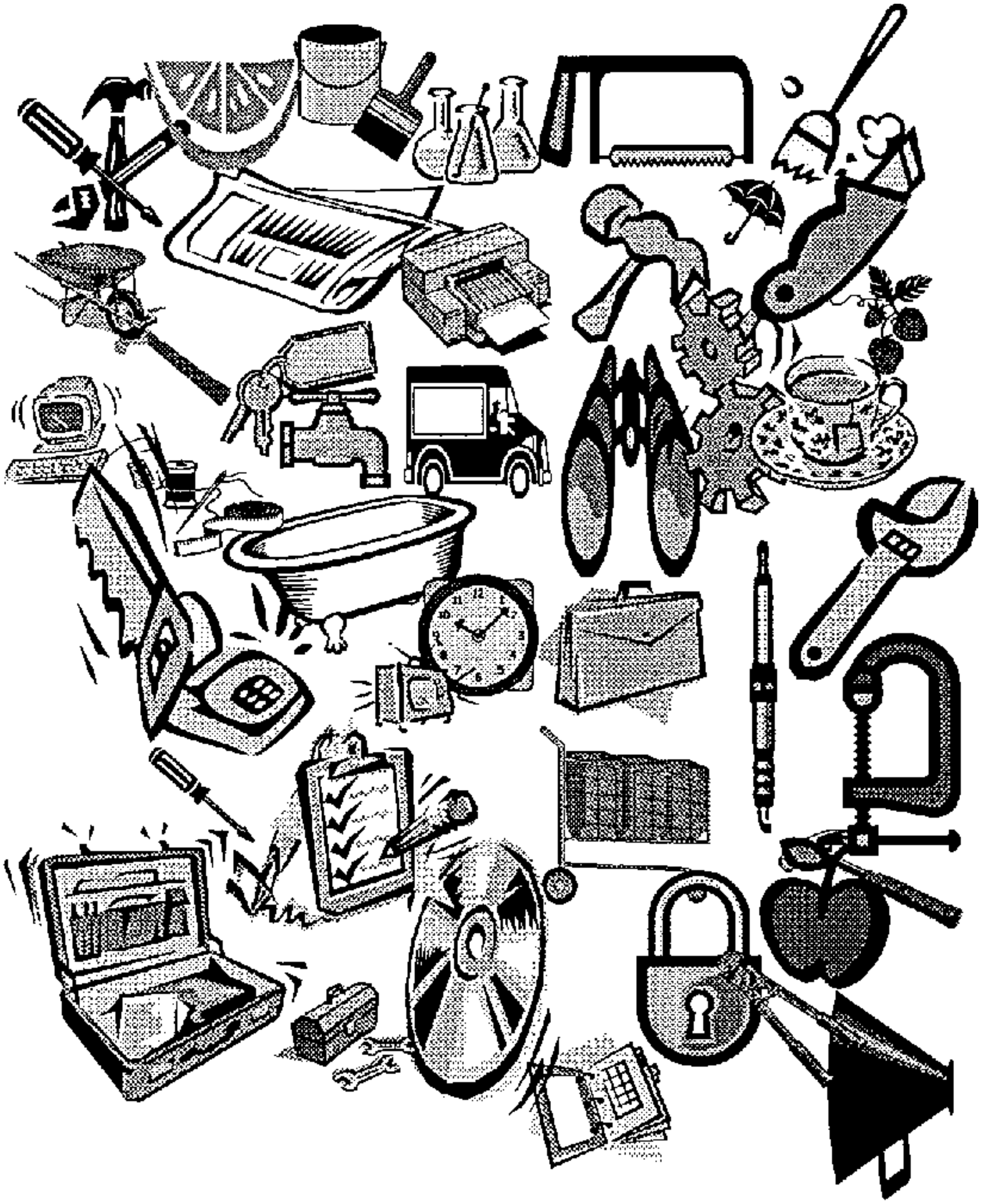


Figure 3 Material varieties.

These seven steps provide an understanding of the material flows in the facility. The calibrated flows are used to develop affinity ratings. These initial steps are also the basis for subsequent evaluation of layout options and material handling system design.

Step 1. Classify Materials. Most manufacturing and warehouse operations have a large variety of products and materials. Situations with 20,000 or more distinct items are not unusual. To analyze flow or design a material handling system around so many individual items is not practical. Classification reduces materials to a manageable number of items so that the classes then become the basis for determining flow rates, containers, and handling equipment.

The initial classifications stratify materials for common handling methods and container design. Weight, size, shape, ‘‘stackability,’’ and special features are

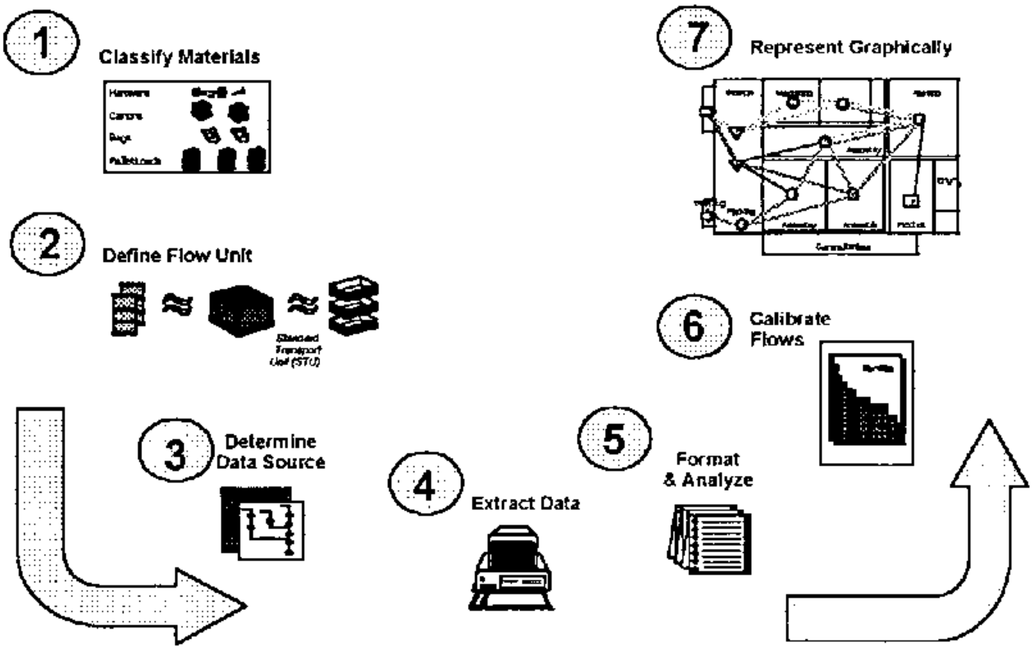


Figure 4 Material flow analysis.

defining criteria. Figure 5 shows a classification based on handling characteristics for a four-drawer cabinet.

In addition, similarities in product, process sequence, and raw material are bases for grouping items that move over the same routes.

Step 2. Identify Flow Units. Material flow is measured in units of material over a unit of time and the analyst chooses appropriate units for both parameters. The time unit is usually a matter of convenience and depends largely on data availability. Typical examples are cases per hour, tons per day, pallets per week.

Selection of the material flow unit is more problematic. Where only one type of material moves, the selection is straightforward, for example, the bushel for a grain mill. But few facilities have only a single material or material type. A wide variety of size, shape, weight, and other handling characteristics must be considered, as illustrated earlier in Fig. 3. For example, integrated circuits are tiny, delicate, expensive, and highly sensitive to electrostatic discharge (ESD), but the operations that use integrated circuits also use large metal cabinets. Between these extremes is a wide range of diverse items to move.

Various items of the same size may have different handling requirements and costs. A resistor and an integrated circuit (IC) are very close in size. But resistors are moved in bulk, in ordinary containers, and without special precautions. The individual IC is sensitive to ESD. It requires an enclosed, conductive and expensive container. It may have a special tube or bag to further protect it. Humans may touch it only if they wear a grounded wrist strap and a conductive smock.

Individual items or materials are seldom handled separately. Most items are in boxes, tote boxes, cartons, bundles, bales or other containers. These containers then are what need to be handled. But layout design requires a standard unit of flow. This is the equivalent flow unit (EFU) which should have the following characteristics:

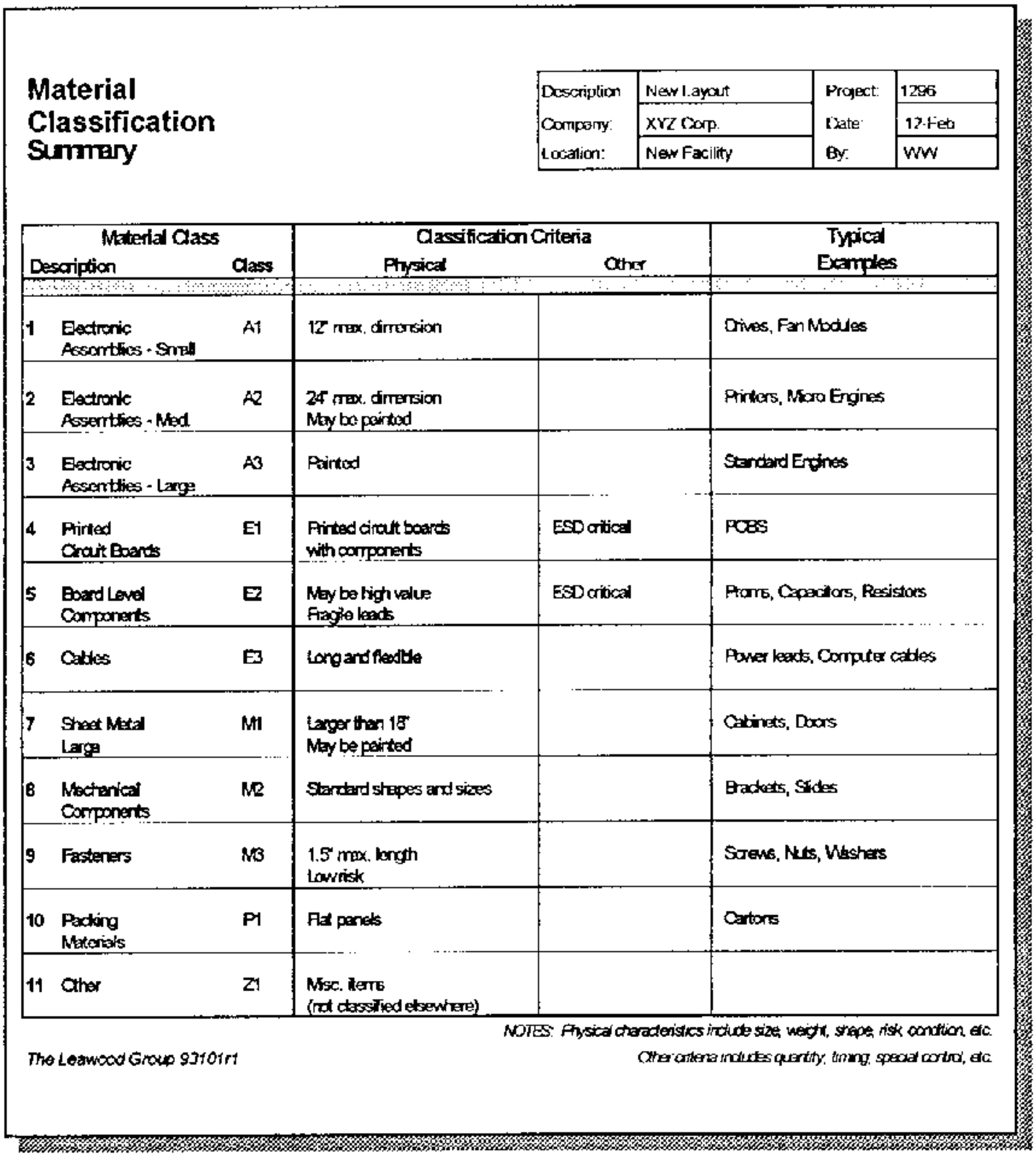
Applicable to all materials and routes
Easily visualized by the users
Independent of the handling method.

The equivalent flow unit should account for weight, bulk, shape, fragility, value, special conditions and other factors:

Weight is a common unit for most materials and is usually available in a central database.

Bulk, or density, relates weight and size. Overall dimensions determine bulk density.

Shape impacts handling difficulty. Compact regular shapes such as boxes stack and handle most easily. Round and irregular shapes stack with



Material Classification Summary

Description	New Layout	Project:	1296
Company:	XYZ Corp.	Date:	17-Feb
Location:	New Facility	By:	WW

Material Class			Classification Criteria		Typical Examples
	Description	Class	Physical	Other	
1	Electronic Assemblies - Small	A1	12" max. dimension		Drives, Fan Modules
2	Electronic Assemblies - Med.	A2	24" max. dimension May be painted		Printers, Micro Engines
3	Electronic Assemblies - Large	A3	Painted		Standard Engines
4	Printed Circuit Boards	E1	Printed circuit boards with components	ESD critical	PCBS
5	Board Level Components	E2	May be high value Fragile leads	ESD critical	Proms, Capacitors, Resistors
6	Cables	E3	Long and flexible		Power leads, Computer cables
7	Sheet Metal Large	M1	Larger than 18" May be painted		Cabinets, Doors
8	Mechanical Components	M2	Standard shapes and sizes		Brackets, Slides
9	Fasteners	M3	1.5" max. length Low risk		Screws, Nuts, Washers
10	Packing Materials	P1	Flat panels		Cartons
11	Other	Z1	Misc. items (not classified elsewhere)		

NOTES: Physical characteristics include size, weight, shape, risk, condition, etc.
Other criteria includes quantity, timing, special control, etc.

The Leawood Group 93101r1

Figure 5 Material classification summary.

difficulty. Long thin shapes (high aspect ratio) are awkward to handle.

Fragility refers to the sensitivity of objects to damage. Fragility influences handling difficulty and expense; 100 lbs of sand and 100 lbs of glassware are very different handling tasks.

Value for a wide range of objects and materials has little influence. But high value or special security items such as gemstones require protection from theft, damage or loss.

Special conditions that affect material handling difficulty and expense are stickiness, temperature, slipperiness, hazard, and ESD sensitivity.

As material moves through the production system, its aspects change and the handling effort, as measured by equivalent flow units, can change drastically. For example:

Bulk cellulose acetate flake may be received and plastic film rolls or sheets may be shipped.

Tree trunks may be received and newsprint shipped.
Bulk liquids and gases may be received but pharmaceutical intravenous packs or bottles of tablets are shipped.
Bauxite ore is received and aluminum ingots are shipped.
Plywood is received, entertainment centers are shipped.
Wood pulp and naphtha are received, chemicals, textiles, and plastics are shipped.

What seems a minor change in the item sometimes brings a dramatic change in the equivalent flow units.

Figure 6 is a schematic flow diagram that illustrates changes in flow intensity as the material is processed for a four-drawer cabinet.

Figure 7 is a river diagram illustrating material flow for all products in an entire plant. The diagram shows how flow intensity increases after the material is painted and decreases after the parts are assembled. Painted sheet metal parts are easily damaged and difficult to handle. Once assembled and packaged, the units become protected, compact, and stackable and their flow in equivalent flow units decreases dramatically for the same quantity and weight.

When a decision is made on an equivalent flow unit, convenience and familiarity often take precedence over accuracy. The primary purpose of this analysis is to rate flow intensities into one of four categories. We use the vowel letter rating system A, E, I, and O. Accuracy of the order of ±20% is therefore sufficient. For this level of accuracy, the following procedure is used:

Review potential data sources.
Interview production and support personnel.

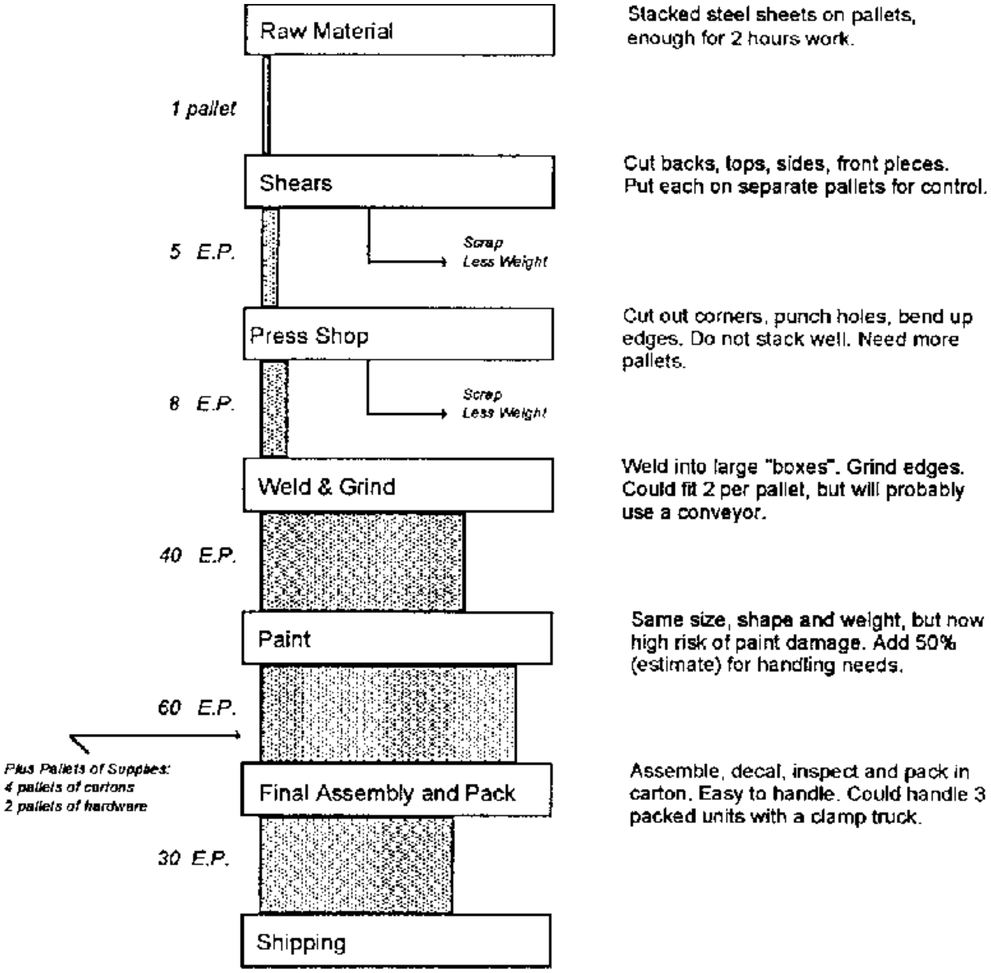

Raw Material
Stacked steel sheets on pallets, enough for 2 hours work.
1 pallet
Shears
Cut backs, tops, sides, front pieces. Put each on separate pallets for control.
5 E.P.
Scrap Less Weight
Press Shop
Cut out corners, punch holes, bend up edges. Do not stack well. Need more pallets.
8 E.P.
Scrap Less Weight
Weld & Grind
Weld into large "boxes". Grind edges. Could fit 2 per pallet, but will probably use a conveyor.
40 E.P.
Paint
Same size, shape and weight, but now high risk of paint damage. Add 50% (estimate) for handling needs.
60 E.P.
Plus Pallets of Supplies:
4 pallets of cartons
2 pallets of hardware
Final Assembly and Pack
Assemble, decal, inspect and pack in carton. Easy to handle. Could handle 3 packed units with a clamp truck.
30 E.P.
Shipping

Figure 6 Equivalent unit flow analysis.

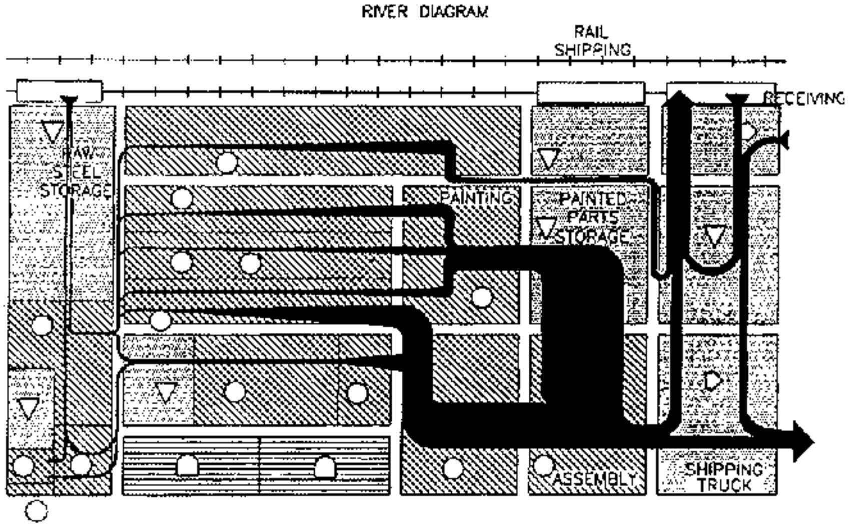


Figure 7 River diagram.

Observe operations.
Define the equivalent flow unit.

Some examples of equivalent flow units are pallets, bales, paper rolls, tote-boxes, letters, tons of steel, and computer cabinets.

The analysis now requires factors to convert all materials into the equivalent flow unit. Conversion may come from experience, work measurement, or benchmarking. An example from a jet engine overhaul facility is given in Fig. 8.

The graph converts item size to equivalent flow units. Pallets and pallet-size containers were the most commonly moved items and the basis for most records. The equivalent pallet was, therefore, the most sensible equivalent flow unit. The pallet container labeled "PT" is 1.0 equivalent flow unit on the vertical scale. Its volume is 60 ft^3 on the horizontal scale.

In this system, volume is the basis for equivalent flow unit conversion. Several tote pans of different sizes are labeled "2T," "4T," "8T," and "8S." An assembled jet engine has a volume of about 290 ft^3 and an equivalent flow unit value of 1.6 equivalent pallets. The relationship between volume and equivalent flow unit is logarithmic rather than linear, which is not unusual. Jet engines are 4.8 times the bulk of a pallet load. On dollies they require only a small tow tractor or forklift to move. The cost and effort is about 1.6 times that for moving a pallet load.

Additional factors can affect the logarithmic volume relationship. This accounts for differences in density, shape or other handling modifiers.

Work standards can be used as conversion factors. The time and cost of moving representative items are calculated and compared, and become benchmarks for all other items in the facility, or the data might be the basis for a graphical relationship similar to the one illustrated previously in Fig. 8.

Step 3. Determine Data Source. Almost every facility is unique with respect to the material data source. Products, volumes, and mix vary; practices are diverse, as are recording methods. Accuracy may be good, suspect, or demonstrably poor, and individuals who control data sources may be co-operative or protective.

This diversity necessitates extensive interviews with people who collect and compile the data. A good selection of data source often makes the difference between a difficult or an easy analysis. Here are some possible data sources:

Process charts
Routing sheets
Material requirements database
Routing database
Direct observation
Handling records
Work sampling
Schedule estimates
Informed opinions.

When selecting the data source, the analyst must also decide on the range of included items. All items should be used if their number is small or when computerized records make it feasible to do so. When a few products

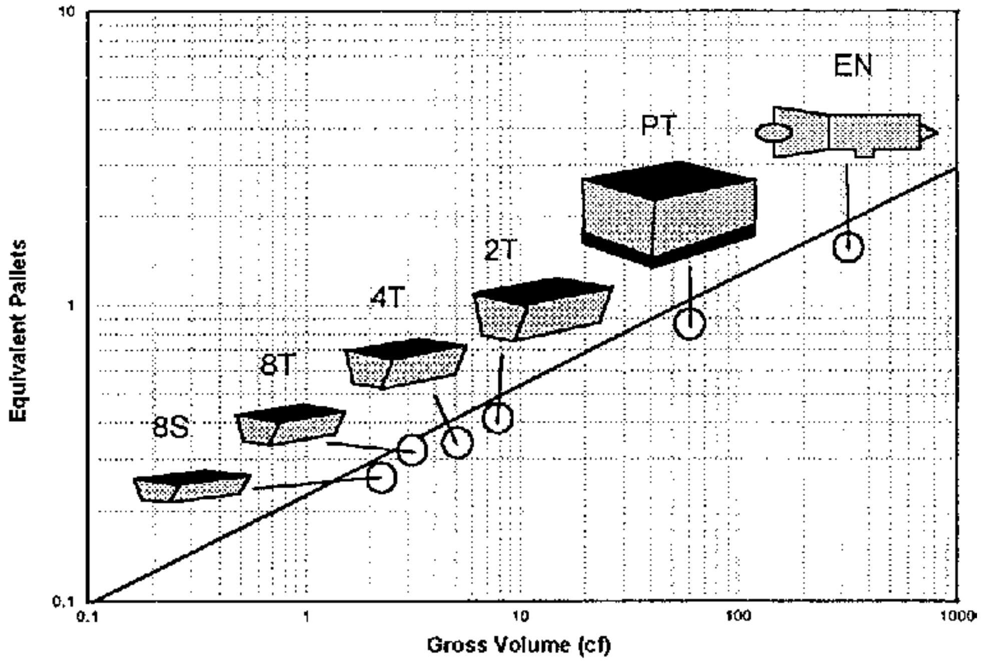


Figure 8 Equivalent flow units.

represent the largest volumes and are representative of others, data from the top 20–30% should be used.

Where groups of products have similar processes and flows, a representative item might portray an entire group. When the product mix is very large and diverse, random sampling may be appropriate. Figure 9 illustrates data selection guidelines

Process charts map the sequence of processes graphically; routing sheets often have much the same information in text form. With either source, each operation must be examined to determine in which SPU that operation will occur. This determines the route. From the product volume analysis or other information, the raw flow is determined which is then converted to equivalent flow units, as illustrated in Fig. 10.

This procedure is used directly if there are only a few products and where processes and flows are similar and a single item represents a larger product group. For large numbers of items, process charts with a random sample are used.

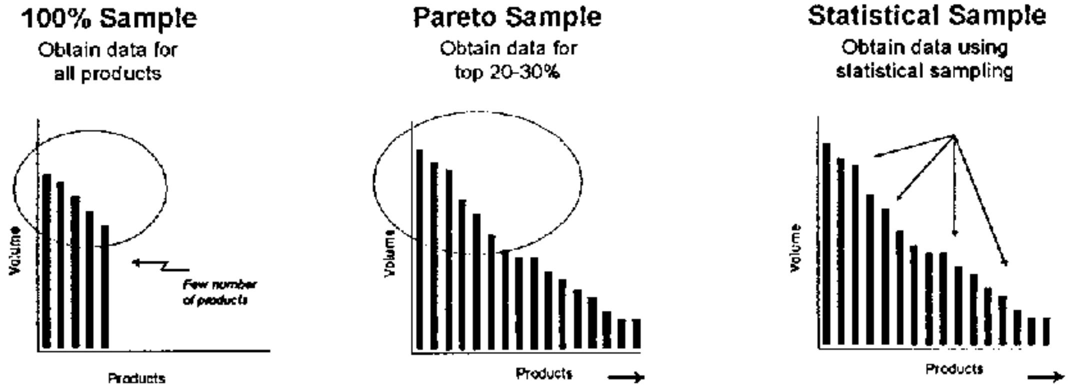


Figure 9 Data selection guidelines.

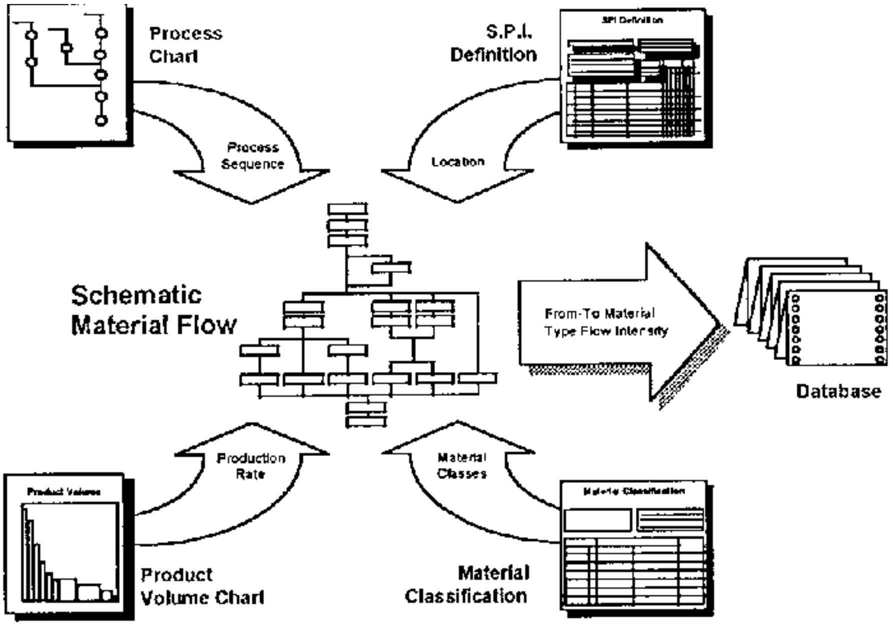


Figure 10 Equivalent flow units development process.

Most or all of the necessary information may exist in the databases of manufacturing requirements planning and other production and scheduling information systems. It may be necessary to change the data to a format suitable for flow analysis.

Material handling records or direct observation are good sources for data. If material is moved by fork truck, for example, and a central fork truck pool keeps good records of pickups and deliveries, these records contain the necessary information. In direct observation, the observer follows products through various moves and operations. In this way both process and material flow information are gathered simultaneously. The from–to chart of Fig. 11 documents flows obtained by direct observation.

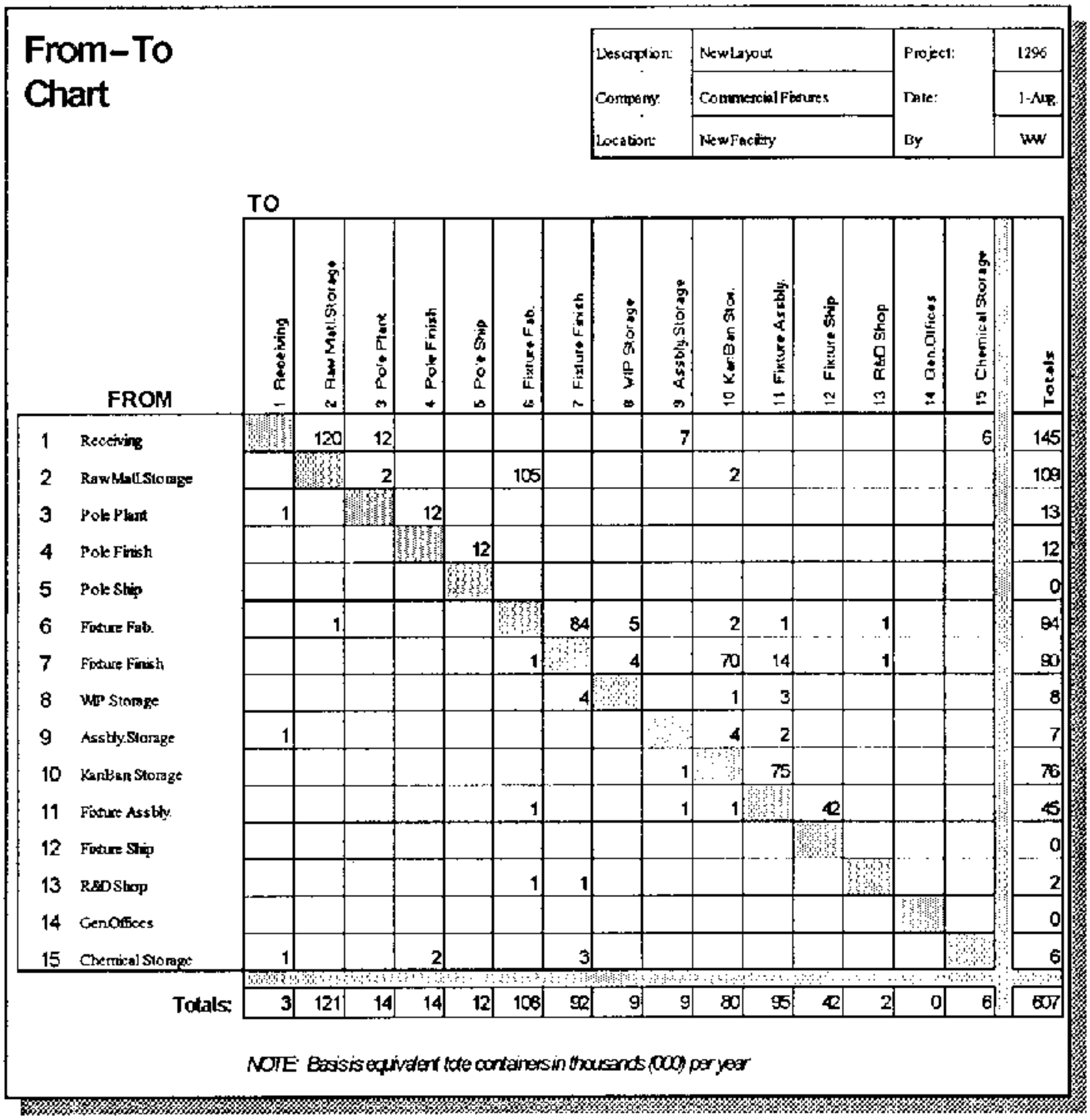
Several sources may be necessary to capture all flows. For example, an MRP database may contain flows for production items but not for scrap, maintenance, trash, or empty containers. These ancillary items are often significant and sometimes dominant, particularly in high-tech industries.

Step 4. Acquire the Data. After a data source is determined the data must be acquired. Computer-based data are accessed by information services. Other data sources may require considerable clerical effort. Direct observations or work-sampling derived data may require weeks to collect and process.

Step 5. Format and Analyze the Data. Manual methods can suffice for the entire material flow analysis. However, computer-aided analysis is necessary for facilities with a wide product mix, process focus and a complex process sequence. Spreadsheet programs are suitable for most analyses. Database programs are sometimes better than spreadsheets because of their reporting and subtotaling capabilities.

With computerized analysis, data can be entered as the project progresses. Initial data may come from downloaded information or manual collection and consist of product information such as names and part numbers and perhaps annual volume, weights, and routing. The analyst should consider ancillary uses for the database as well. The database may assist later in developing handling systems or determining storage areas. It might also be part of a group technology (GT) study for cell design.

Figure 12 is an example of a material flow report used for the layout of a mail processing facility. Data came from a series of schematic material flow charts, in turn derived from process charts, SPU definitions and a product volume analysis, as shown earlier in Fig. 2. Fields 1 and 2 of Fig. 12 are SPU numbers which define the flow path for that entry. Fields 3 and 4 are descriptors corresponding to the SPU numbers. Field 5 is a type code; field 6 is the equivalent flow unit; field 7 is the daily volume in pieces per day. All mail with the same



From–To Chart

Description:	New Layout	Project:	1296
Company:	Commercial Fixtures	Date:	1-Aug.
Location:	New Facility	By	WW

FROM \ TO		1 Receiving	2 Raw Matl.Storage	3 Pole Plant	4 Pole Finish	5 Pole Ship	6 Fixture Fab.	7 Fixture Finish	8 WIP Storage	9 Assbly.Storage	10 KanBan Stor.	11 Fixture Assbly.	12 Fixture Ship	13 R&D Shop	14 Gen.Offices	15 Chemical Storage	Totals
1	Receiving		120	12						7						6	145
2	RawMatl.Storage			2			105				2						109
3	Pole Plant	1			12												13
4	Pole Finish					12											12
5	Pole Ship																0
6	Fixture Fab.		1					84	5		2	1		1			94
7	Fixture Finish						1		4		70	14		1			90
8	WIP Storage							4			1	3					8
9	Assbly.Storage	1									4	2					7
10	KanBan Storage									1		75					76
11	Fixture Assbly.						1			1	1		42				45
12	Fixture Ship																0
13	R&D Shop						1	1									2
14	Gen.Offices																0
15	Chemical Storage	1			2			3									6
	Totals:	3	121	14	14	12	106	92	9	9	80	95	42	2	0	6	607

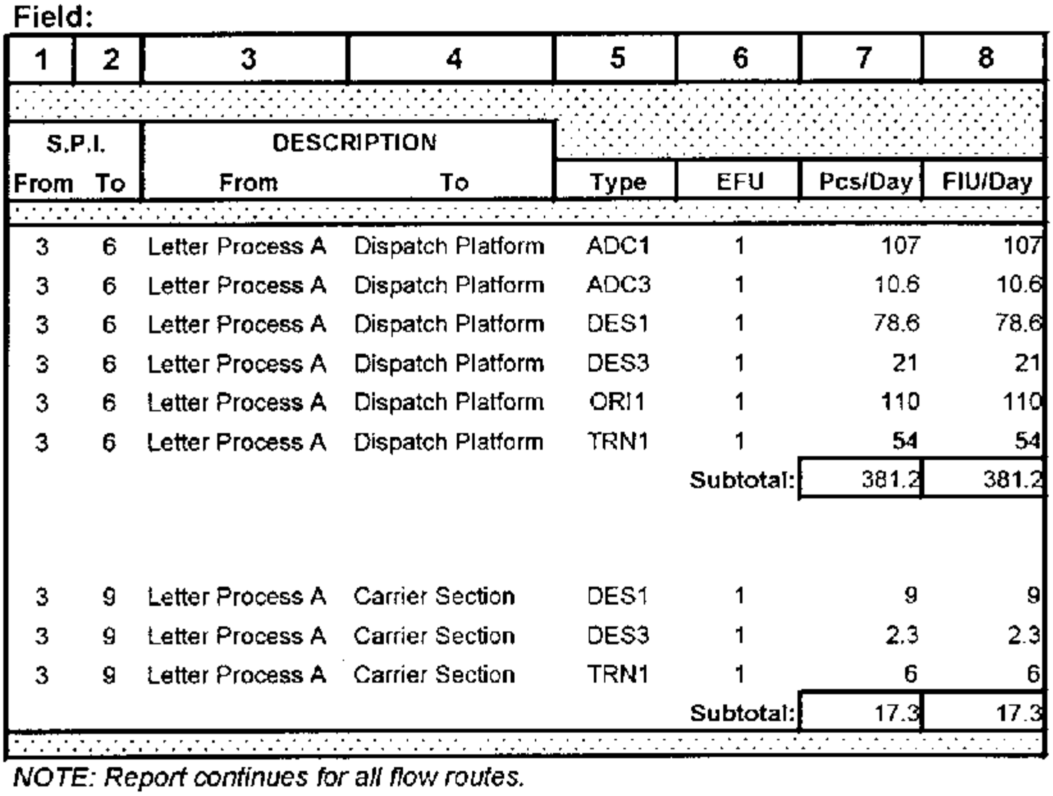
NOTE: Basis is equivalent tote containers in thousands (000) per year

Figure 11 Material flows from–to chart.

size, type, and class codes uses the same process and follows the same route. Field 8 is the number of equivalent flow units per day for each route and size. These subtotals are the basis for subsequent affinity ratings, staffing, and for material-handling-system design.

Other possible fields might contain information on the time required per trip, distance for each route and speed of the equipment. From this the database manager can derive the numbers and types of equipment and containers required.

Step 6. Calibrate Flows. This step includes the calculation of material flow from each route origin to each destination. It also includes conversion of calculated flows to a step-function calibration for use in layout planning. The calibration scale can be alphabetical or numerical. The vowel rating convention AEIO is used here. The intensities of flow distribution may indicate the important few and trivial many. The calibrations can be used for relative capacity of material-handling-system selection.



Field:

1	2	3	4	5	6	7	8
S.P.I.		DESCRIPTION					
From	To	From	To	Type	EFU	Pcs/Day	FIU/Day
3	6	Letter Process A	Dispatch Platform	ADC1	1	107	107
3	6	Letter Process A	Dispatch Platform	ADC3	1	10.6	10.6
3	6	Letter Process A	Dispatch Platform	DES1	1	78.6	78.6
3	6	Letter Process A	Dispatch Platform	DES3	1	21	21
3	6	Letter Process A	Dispatch Platform	ORI1	1	110	110
3	6	Letter Process A	Dispatch Platform	TRN1	1	54	54
					Subtotal:	381.2	381.2
3	9	Letter Process A	Carrier Section	DES1	1	9	9
3	9	Letter Process A	Carrier Section	DES3	1	2.3	2.3
3	9	Letter Process A	Carrier Section	TRN1	1	6	6
					Subtotal:	17.3	17.3

NOTE: Report continues for all flow routes.

Figure 12 Material flow report.

For the calibration the flow rates are ranked on a bar chart, as shown in Fig. 13.

The breakpoints are a matter of judgment and should be made near natural breaks. Experience from a range of projects suggests that the following proportions are a useful guideline:

A	5–10%
E	10–20%
I	20–40%
O	40–80%

Transport work. Total material handling cost is roughly proportional to the product of flow intensity and distance. In physics force multiplied by distance defines work. For layout planning, material flow intensity I multiplied by distance D equals "transport work" TW:

$$\mathrm{TW} = DI$$

In an ideal layout all SPUs with affinities would be adjacent. Since an SPU occupies finite space, proximity but not necessarily adjacency is possible. Placing two particular SPUs together forces other SPUs farther away. The theoretical optimum relative locations occur with equal transport work on all routes where total transport work is at the theoretical minimum. Transport work, then, is a metric for evaluating the layout. For evaluation, transport work is calculated along every path on a layout and the summation made. Layout options may be evaluated by comparing their total transport work.

Transport work is useful in another way. In Fig. 14 distance is plotted on the horizontal axis and flow intensity on the vertical axis. Each route on the layout plots as a point. As mentioned above, the ideal layout would have constant (or iso-) transport work, such a curve being a hyperbola. Routes with low intensity have long distances; those with high intensity, short distances. The product of distance and intensity for either is then equal.

A "good" layout, from strictly a material flow perspective, is one which has most or all points close to the same hyperbolic isotransport work curve. Routes which are significantly distant from the hyperbola indicate an anomaly in the layout.

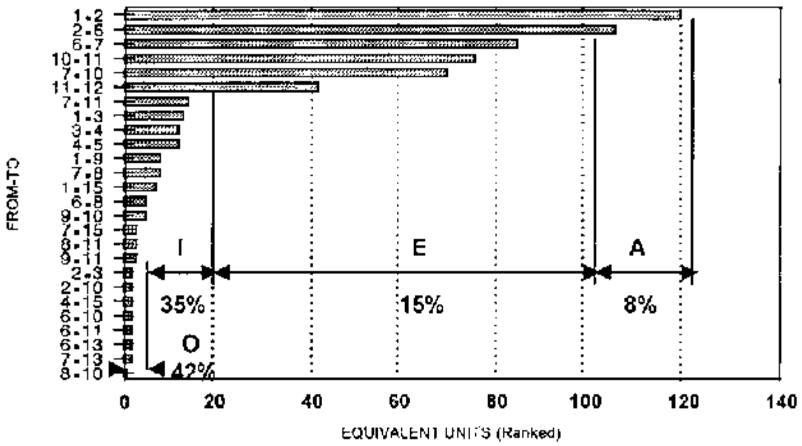


Figure 13 Material flow calibration.

Step 7. Graphical Representation. Several types of charts, plots, and diagrams present material flow information visually. The visual representation assists the layout designer during the creative portion of the layout project. It helps to evaluate layout options and design the material handling system.

Material handling and facility layout are inextricably intertwined. Layout determines the distances materials must be moved. Handling methods may affect the total handling and cost on any particular route.

Material flow diagrams and isotransport work diagrams are used to visualize material flow graphically. They show *sequence*, *distance*, *intensity*, or a combination thereof. They assist with evaluation and handling-system design. There are at least eight common types of diagrams:

Schematic
Quantified schematic
Locational
River

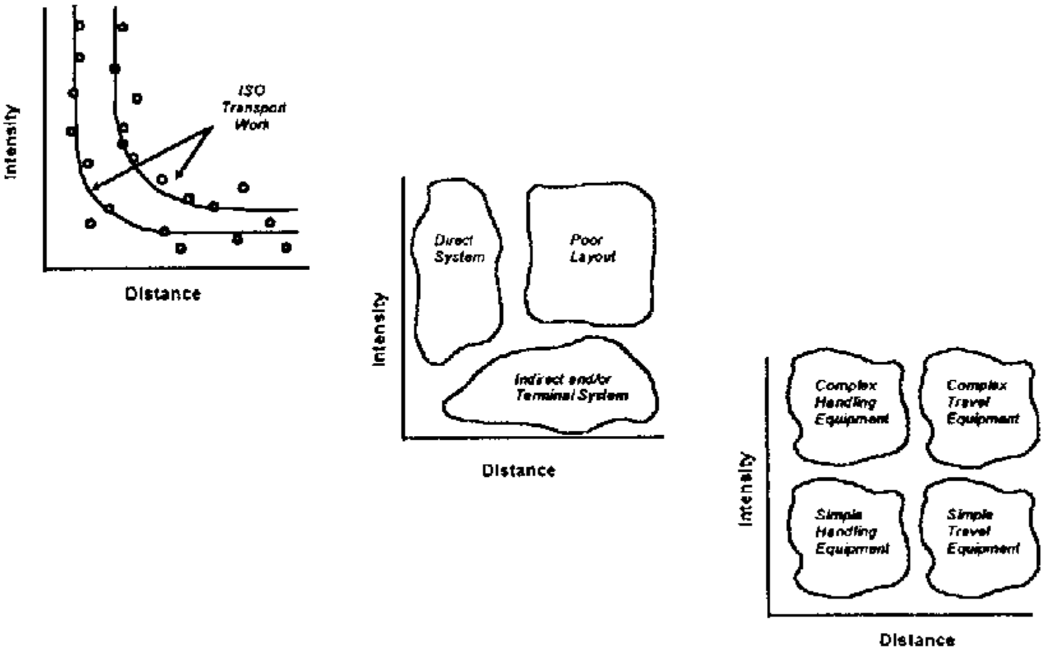


Figure 14 Distance–intensity plots.

String
Three-dimensional
Distance–intensity plot
Animated.

Figure 15 is a schematic diagram. The blocks represent locations on the layout and the arrows are material move routes. In this example a single arrow represents all materials. But different line styles or colors might show different materials, or separate diagrams might represent different material classes. Schematic diagrams are most useful in the early stages of a project when they help the analyst and others to document, visualize, and understand the material flows.

Figure 16 is a quantified schematic diagram. In addition to routes it illustrates flow intensity by the thickness of shaded paths. The thicker the path, the higher the flow intensity. The quantified schematic may derive from the schematic as the project progresses and data become known.

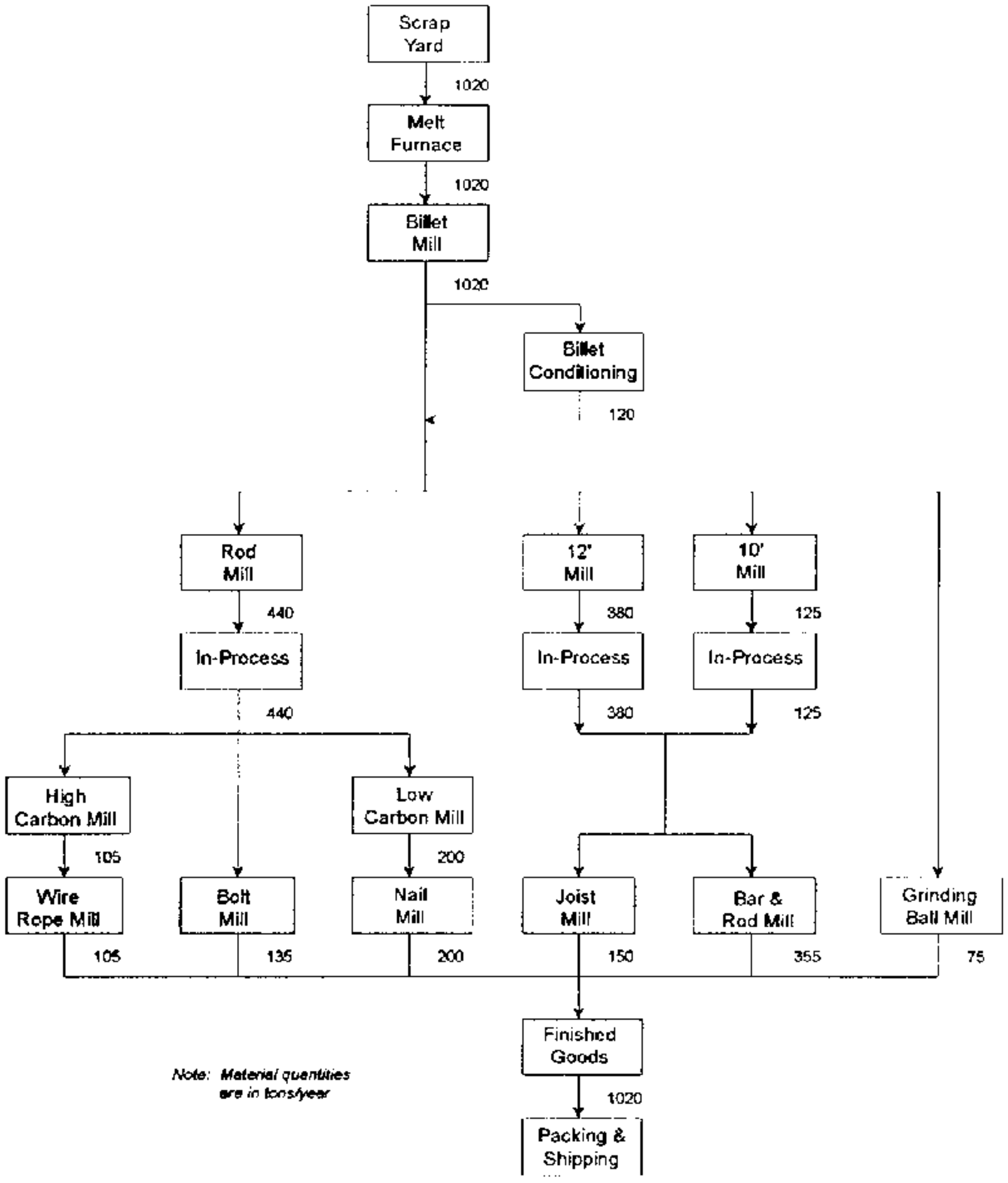


Figure 15 Schematic flow diagram.

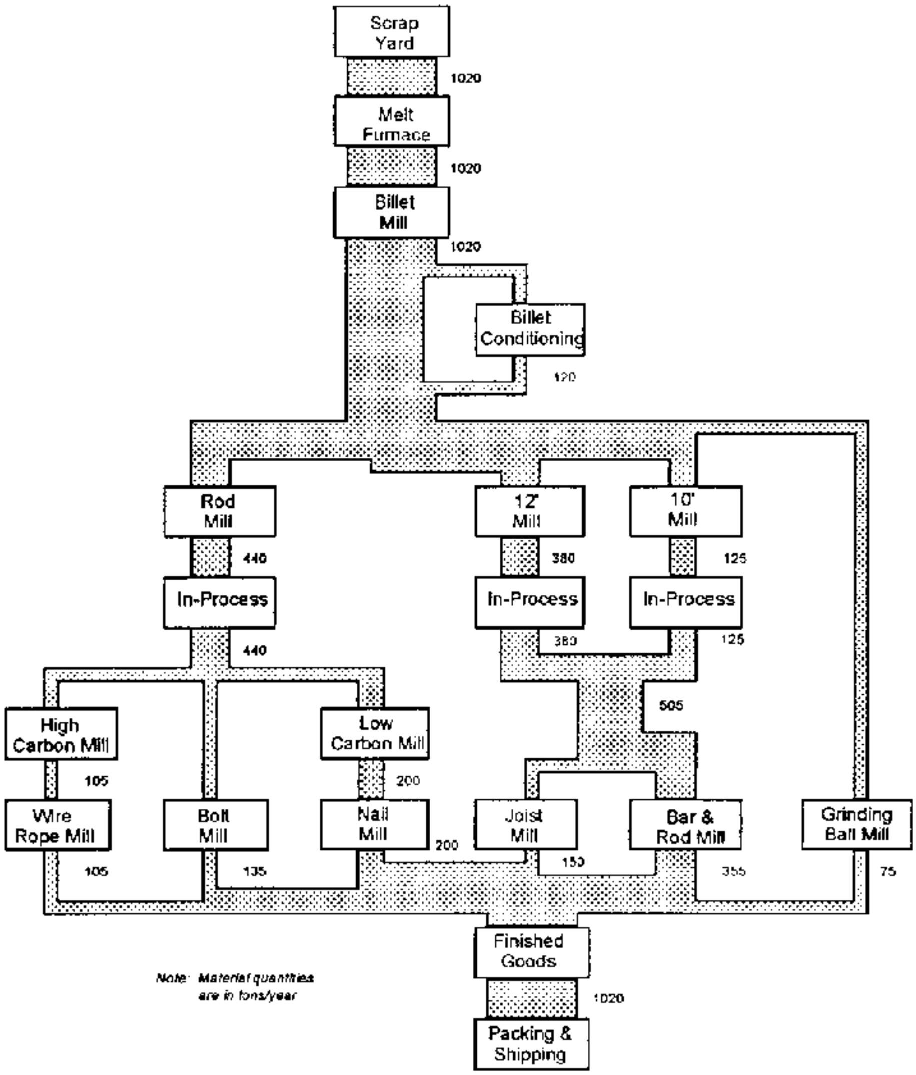


Figure 16 Quantified schematic flow diagram.

The locational diagrams of Figs. 17 and 18 superimpose material flows on a layout, showing the additional variable of distance. The layout may be existing or proposed. Locational diagrams illustrate the effect that layout has on flow complexity and transport work. The width of the lines is proportional to flow intensity. Colors or patterns can indicate either intensity or material classes. Alternatively, multiple lines may be used to represent various intensities, as in Fig. 18. These examples show no sequence information or the direction of flow. Location diagrams are appropriate for complex situations.

The river diagram of Fig. 19 presents sequence, intensity, and distance. It is suitable for simple flow situations with few routes and a fixed sequence.

The string diagram, Fig. 20, traces the path of individual products or items through the plant. A separate line for each item illustrates complexity and total distance. Where flow paths coincide, the lines are thicker. This diagram shows intensity and possibly sequence.

Figure 21 is a locational diagram in three dimensions. The river and string diagrams can also have three dimensions when vertical movement is important.

Distance–intensity plots are shown on Figs. 14 and 27 and explained in Step 6, Transport.

Computer simulation and animation software presents flow dynamically. Arena, Simple++, and Witness VR are popular packages. Simulation is an important tool in demonstrating and selling layouts

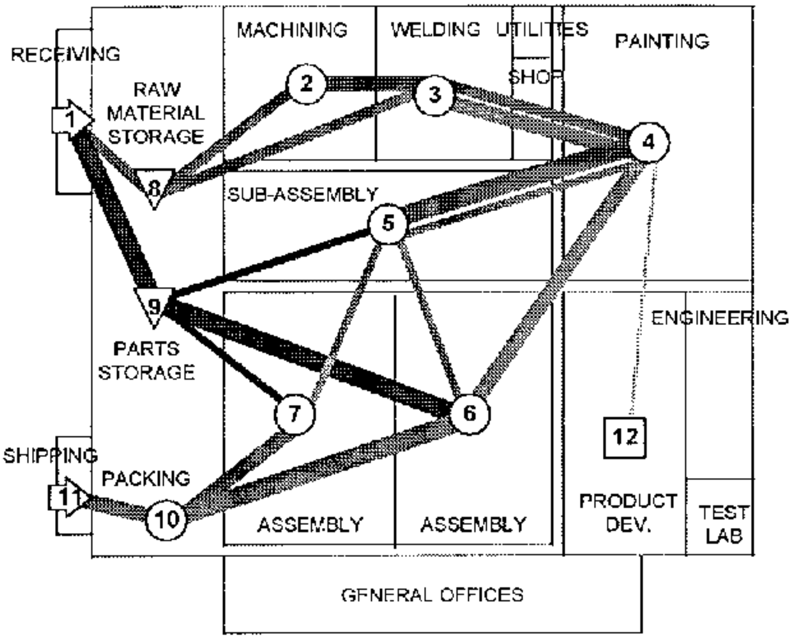


Figure 17 Locational flow diagram (shaded lines).

and flow patterns. It can also assist in designing certain types of handling systems, such as automatic guided vehicles (AGVs).

Macrolevel flow patterns. The facility layout affects sequence and characteristics of material flow. Indeed, the flow pattern dictates the shape or arrangement within a facility. Figure 22 shows the basic flow patterns: straight-through flow, L-shape, U-shape or circular, and hybrids.

With *straight-through* or *linear flow*, material enters and exits at opposite ends of the site or building. Flow deviates little from the shortest straight line path. Material movement is progressive. Receiving and shipping areas (entrances and exits) are physically separate.

Straight-through flow is simple and encourages high material velocity. Operations are typically sequential. This flow pattern has been a hallmark of

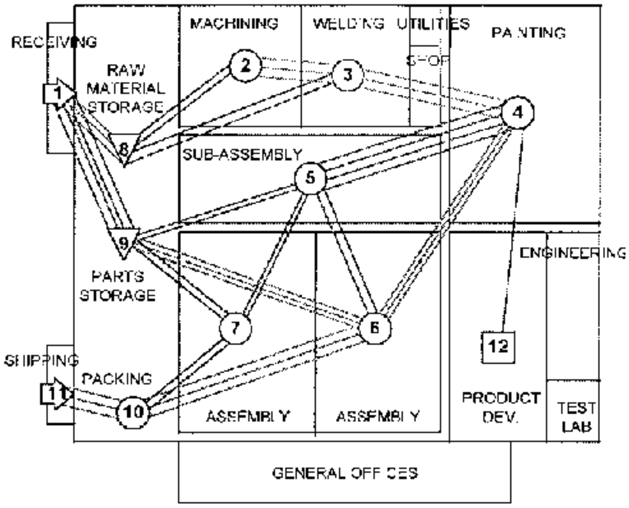


Figure 18 Locational flow diagram (multiple lines).

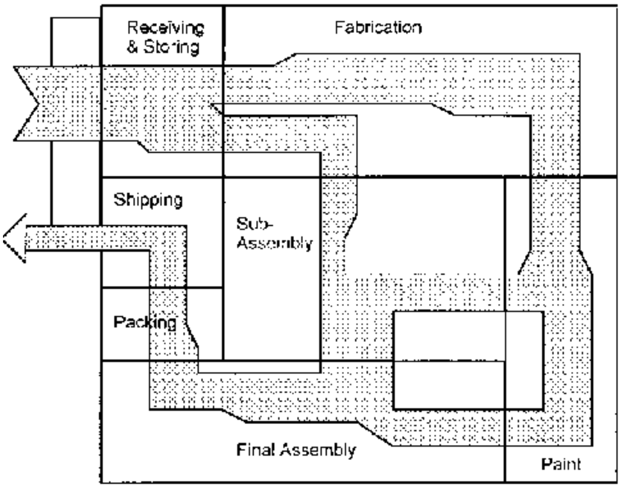


Figure 19 River diagram.

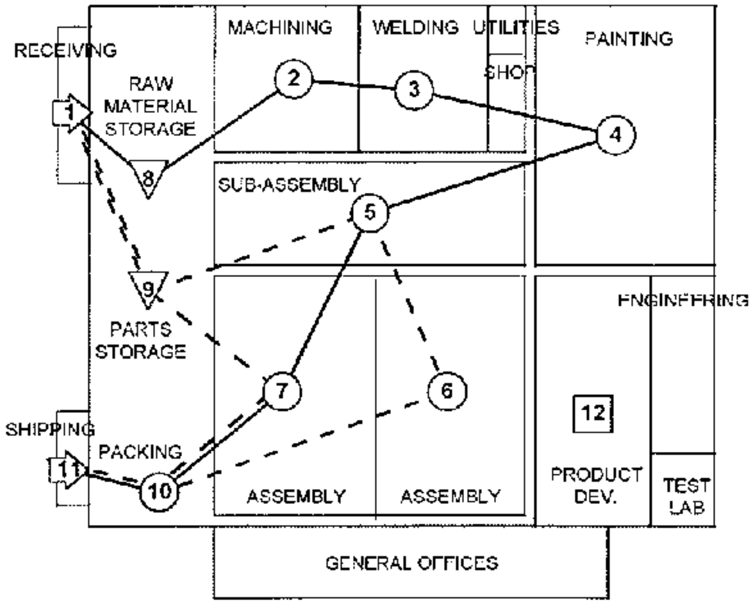


Figure 20 String diagram.

mass production. With this type of flow, material tracking and handling are relatively simple. In fact, Henry Ford and Charles Sorensen invented the assembly line to solve a material flow problem. Straight-through flow can also be vertical movement in a high or multistory building. This flow pattern was used in some of the first water-powered textile factories in England.

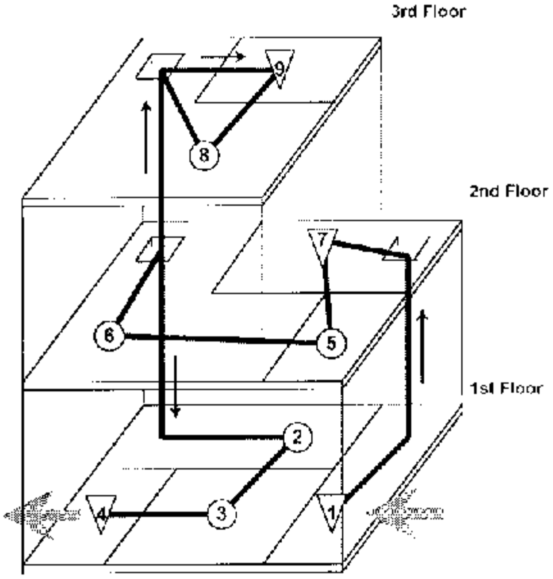


Figure 21 Three-dimensional material flow diagram.

L-shape flow has a 90° directional change. This pattern results from multiple material entry points along the flow path and a need for direct access. It is a flow pattern sometimes used in paint shops.

U-shape or circular flow is an extension of the L-shape flow. The loop may be open or closed. Materials return to their starting vicinity. These patterns combine receiving and shipping docks with shared personnel and handling equipment. Conversely, one set of truck docks in a building can create a U or circular flow, for example, morning receiving and afternoon shipping patterns.

The use of common receiving and shipping personnel is not conducive to good security. In pharmaceutical manufacturing regulations may require strict separation of receiving and shipping facilities. Incoming material handling, storage, and material physical characteristic differences may also require different personnel skills from those required at shipping.

Hybrids, such as X, Y, Z, or star, are combinations or variations of the basic flow patterns.

Flow complexity. Simple material flow patterns have fewer routes, fewer intersections and shorter distances. River and locational diagrams show flow complexity. These can be used to evaluate the relative complexity inherent in various layouts.

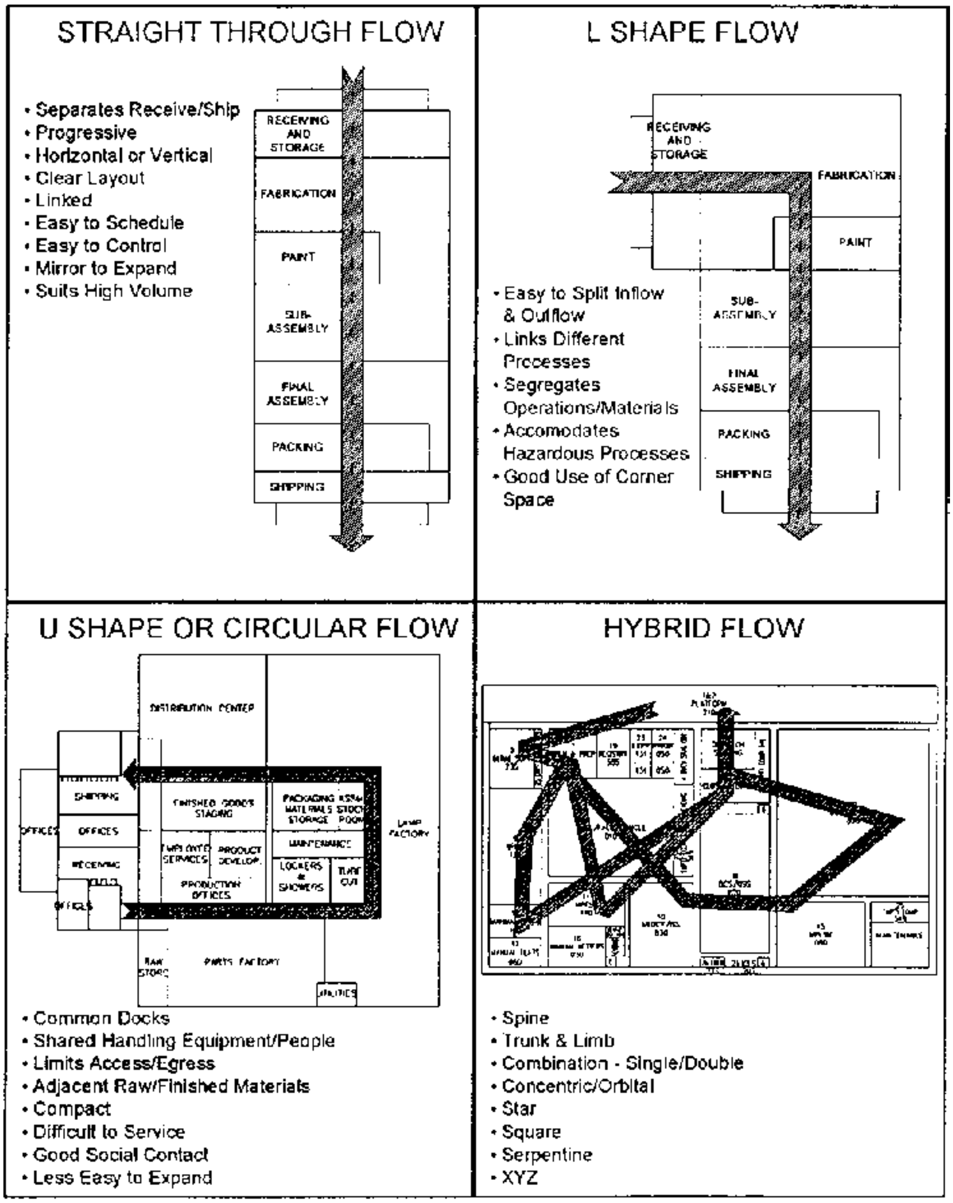


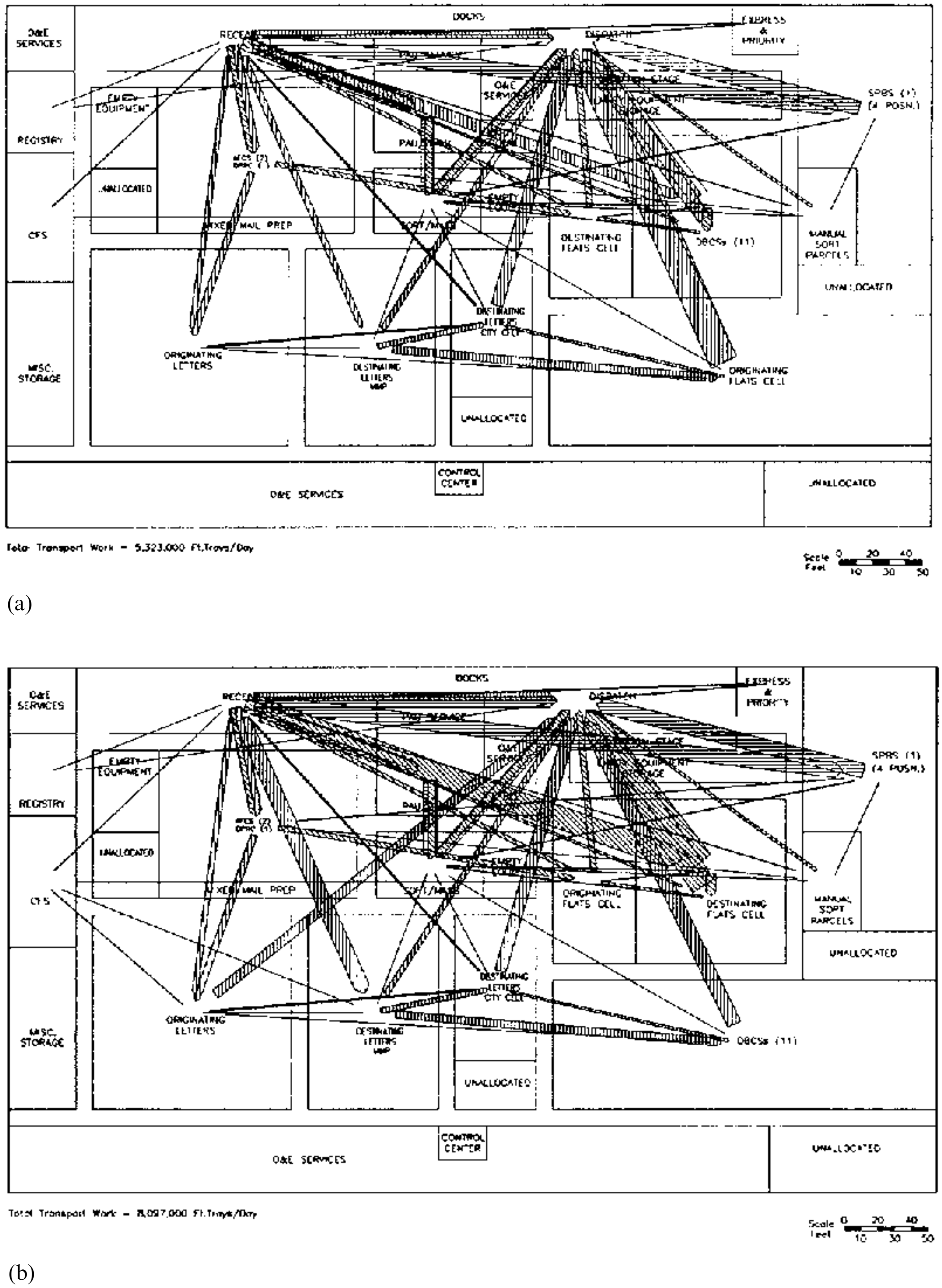
Figure 22 Macrolayout basic flow patterns.

Figure 23a and 23b shows locational flow diagrams for two postal facility layouts. A visual comparison indicates that Fig. 23b has a more complex flow pattern than Fig. 23a. The diagrams illustrate that the flow from SPU 27 to SPU 2 is the most intense and yet has the longest distance. This is verified by a comparison of total transport work: 5.323 million equivalent flow units/day shown in Fig. 23a versus 8.097 million equivalent flow units/day for Fig. 23b. The option shown in Fig. 23b is superior *from a material flow perspective*.

1.3 MATERIAL-HANDLING-SYSTEM DESIGN

Optimum material handling requires a macro or plant-wide system design. The system approach examines all materials and all routes. It fits and supports the firm's manufacturing strategy and considers many options.

In the absence of a comprehensive approach, factories usually default to a simple-direct system using forklift trucks. The system designs itself and is quite



(a)

(b)

Figure 23 Transport work material flow evaluation.

convenient in that respect. However, the convenience for the designer becomes a high-cost system that encourages large lots and high inventories. It seldom supports just-in-time and world-class manufacturing strategies.

The macrolevel handling plan specifies the route, container and equipment for each move. It then accumulates the total moves by equipment type and calculates equipment requirements. To prepare a handling plan:

1. Assemble flow analysis output:
   a. Routes and intensities
   b. Material flow diagrams
   c. Layout(s).
2. For each route and material class select:
   a. Container
   b. Route structure
   c. Equipment.
3. Calculate equipment requirements.
4. Evaluate and select equipment.

1.3.1 Containers

Materials in industrial and commercial facilities move in three basic forms: singles, bulk, and contained. Singles are individual items handled and tracked as such. Bulk materials, liquids, and gases assume the form or shape of their container. Fine solids such as flowable powders are also bulk. In containerized handling, one or more items are in or on a box, pallet, board, tank, bottle, or other contrivance. The container restrains the items within and, for handling purposes, the container then dominates.

Some materials take several forms. Nails and screws can move on belt conveyors almost like powders. Later in the process, handling may be individually or in containers. Containers offer several advantages:

Protecting the contents
Improving handling attributes
Improved use of cube
Standardizing unit loads
Assisting inventory control
Assisting security.

Pallet and palletlike containers have in the past been the most widely used. In many industries they still are.

"Tote pans" and "shop boxes" have evolved into sophisticated container families. They are versatile for internal and external distribution and are an important feature of kanban systems. Because of their wide use they should be standardized and selected with great care.

Just-in-time, cellular manufacturing and time-based competition strategies require moving small lot sizes to point of use, which calls for smaller containers.

For broad use in large plants, a family of intermodular units is important. The International Organization for Standardization (ISO) and the American National Standards Institute (ANSI) have set size standards. The most popular families use 48 in. × 40 in. and 48 in. × 32 in. pallet sizes. Figure 24 shows one system.

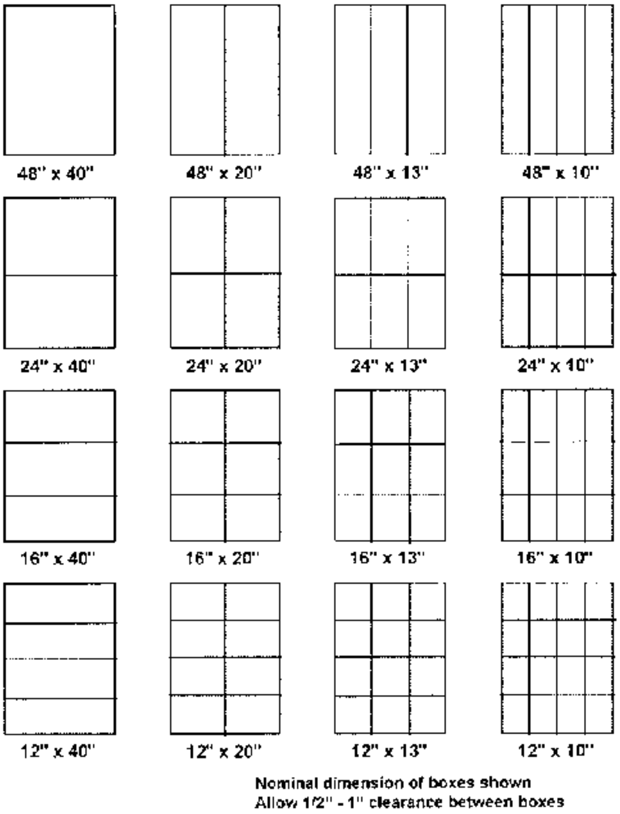


Figure 24 Box configurations for standard pallets.

Larger-than-pallet containers are primarily for international trade and ISO standardized unit have been designed. There is, of course, a large variety of nonstandard loads and containers.

The key to container selection is integration. Container, route structure, and equipment are intimately connected; they should fit with and complement each other. Other issues such as process equipment and lot size also influence container selection. Unfortunately, most container selections occur by default. Certain containers pre-exist and new products or items get thrown into them. Existing route structures may also dictate container selection.

Manufacturing strategy should influence container selection. Conventional cost-based strategies indicate large containers corresponding to large lot sizes; contemporary strategies emphasize variety and response time. The smallest feasible container corresponding to small process and movement lot sizes should be selected.

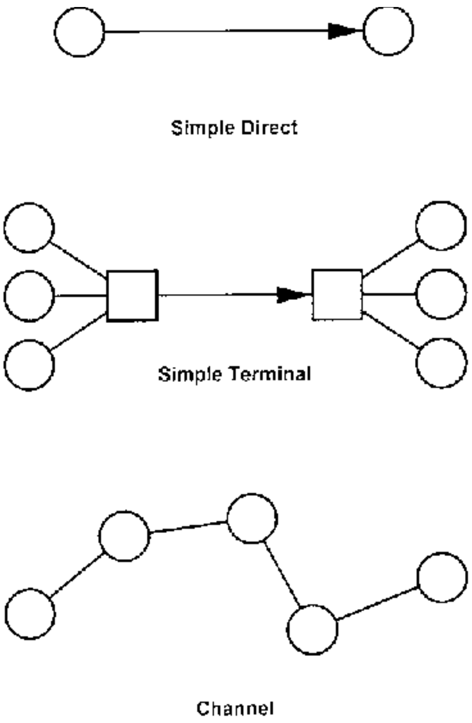


Figure 25 Basic route structures.

1.3.2 Route Structure

Route structure influences container and equipment selection. It impacts costs, timing and other design issues. Figure 25 shows the three basic route structures: direct, channel, and terminal. In a direct system, materials move separately and directly from origin to destination. In a channel system which has a pre-established route, loads move along it, often comingled with other loads. In a terminal system, endpoints have been established where the flow is broken. Materials may be sorted, consolidated, inspected, or transferred at these terminals. In practice, many hybrids and variations of these basic route structures occur, as Fig. 26 shows.

1.3.2.1 Direct Systems

Direct systems using fork trucks are common. In operation, a pallet of material needs moving to another department; the foreman hails a fork truck driver who moves it to the next department. An analogy for a direct system is how taxis operate, taking their customers directly from one location to another without fixed routes or schedules.

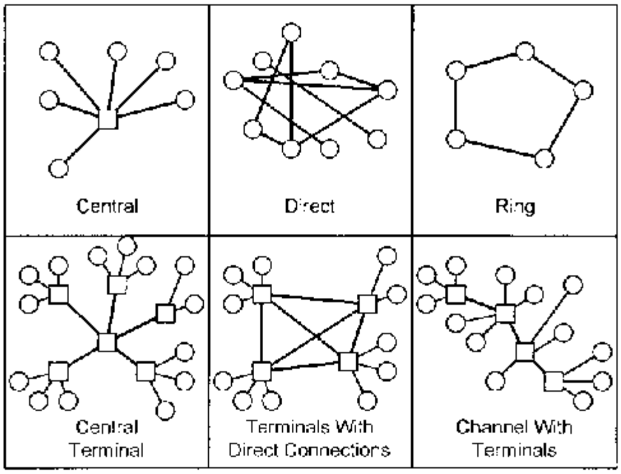


Figure 26 Hybrid route structures.

Direct systems are appropriate for high flow intensities and full loads. They also have the least transit time and are appropriate when time is a key factor, provided there is no queuing for transit requests.

1.3.2.2 Channel Systems

Channel systems use a predetermined path and schedule. In manufacturing some automatic guided vehicle systems work this way. Manned trailer trains and industrial trucks fit channel systems. They follow a fixed route, often circular. At designated points they stop to load and unload whatever is originating or reaching a destination at that point. City bus systems and subway systems use the channel structure.

Channel systems are compatible with just-in-time (JIT) and world class manufacturing strategies. Many JIT plants need to make frequent moves of small quantities in tote boxes. They may use a channel system with electric industrial trucks or golf carts. These carts operate on a fixed route, picking up material and dropping off loads as required. Externally, over the road trucks make several stops at different suppliers to accumulate a full load for delivery. Simultaneously, they return kanban signals and empty containers for refill.

Lower flow intensities, less-than-full loads and long distances with load consolidation benefit from channel systems. Standardized loads also indicate the use of a channel system.

1.3.2.3 Terminal Systems

In terminal systems loads move from origin to ultimate destination through one or more terminals. At the terminal material is transferred, consolidated, inspected, stored, or sorted. The United States Postal Service, Federal Express, and United Parcel Service all use terminal systems.

A single central terminal can control material well. Multiple terminals work well with long distances, low intensities and many less-than-full loads. Airlines use terminal systems for these reasons.

A warning about multistep distribution systems. The characteristics of ultimate demand assume seasonal characteristics and lead to misinterpretations in production planning.

1.4 EQUIPMENT

There are hundreds of equipment types each with varied capacities, features, options, and brands. The designer chooses a type which fits the route, route structure, containers, flow intensity, and distance. These design choices should be concurrent to assure a mutual fit.

Any material move has two associated costs—terminal and travel. Terminal cost is the cost of loading and unloading and does not vary with distance. Transport cost varies with distance, usually in a direct relationship as Fig. 27 illustrates. Equipment is suitable for either handling or transporting but seldom both.

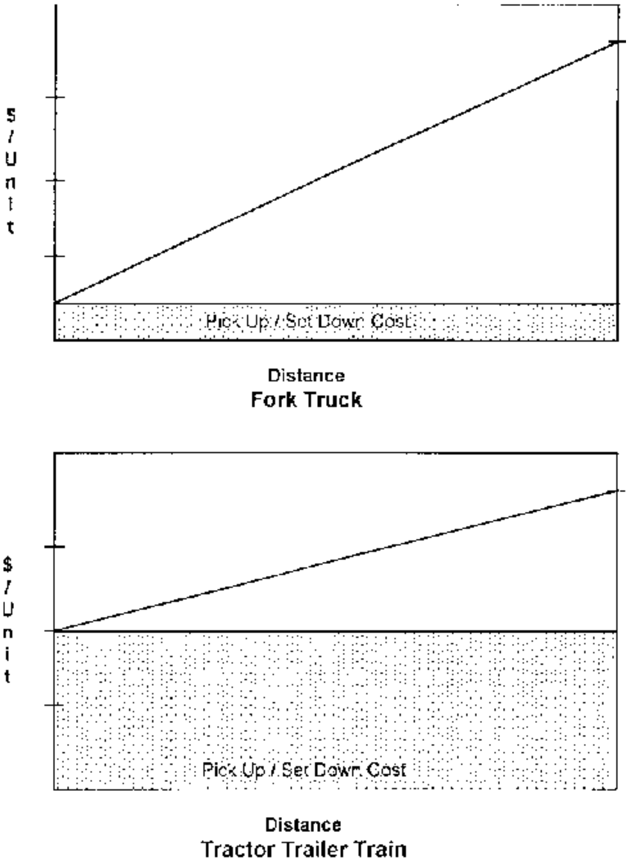


Figure 27 Terminal/travel cost comparisons.

1.4.1 Using the Distance–Intensity Plot for Selection

The distance–intensity (D-I) plot is useful for equipment selection. Figure 28 is a representative D-I plot with isotransport work curves. Each route on a layout plots as a point on the chart. Routes which fall in the lower-left area have low intensity and short distances. Typically these routes would use elementary, low-cost handling equipment such as hand dollies or manual methods. Routes in the upper left have short distances but high intensity. These require equipment with handling and manipulating capabilities, such as robots, or short-travel equipment, such as conveyors. Routes on the lower-right have long distances and low intensities. Equipment with transport capabilities like a tractor trailer train is needed here. Plots in the middle area indicate combination equipment such as the forklift truck. In the upper-right area, long distances and high intensities indicate the need for a layout revision. If substantiated, long routes with high intensities require expensive and sophisticated equipment.

Figure 29 is from a recent study on handling costs. In this study, representative costs for handling pallets with several common devices were calculated. These costs included direct, indirect, and capital.

Shaded areas on the diagram show regions where each type of equipment dominates as the lowest cost. This chart is generic and may not apply to a particular situation; nor does the chart account for intangibles such as flexibility, safety or strategic issues.

1.4.2 Using Material Flow Diagrams for Equipment Selection

Locational, river and string diagrams also help with equipment selection. Here, the flow lines indicate distance, flow intensity and fixed and variable paths. Figure 30 shows how to interpret this information.

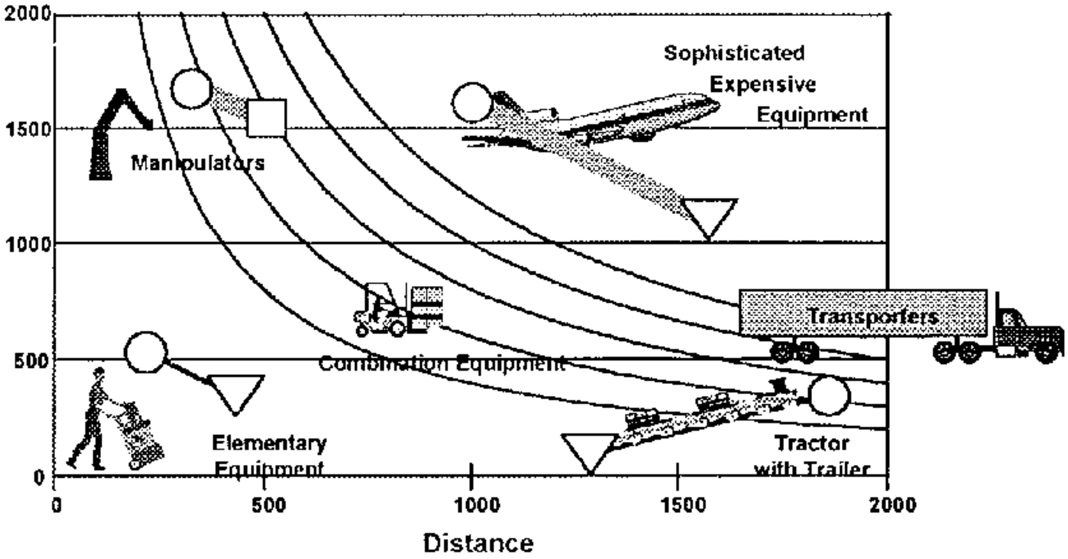


Figure 28 Distance–intensity plot equipment classifications.

1.4.3 Equipment Selection Guide

The range of equipment choice is broad. There is no substitute for equally broad experience when selections are being made. Nevertheless, Fig. 31 can assist the novice to some extent. The chart uses a modified Bolz [3] classification system for handling equipment with a three-digit hierarchical code. The first digit represents a primary classification based on design features:

100—Conveyors: fixed-path equipment which carries or tows loads in primarily horizontal directions.

200—Lifting Equipment: cranes, elevators, hoists, and similar equipment designed to move or position material in a primarily vertical direction.

300—Positioning/Weighing/Controlling: handling equipment used for local positioning, transferring, weighing, and controlling of material movement. Included are manipulators, robots, positioning platforms, and transfers. Also included are scales and weighing equipment, float controls, bin indicators, counters, and other control devices.

400—Industrial Vehicles: this class includes all types of vehicles commonly used in and around industrial and commercial facilities. Excluded are "Motor Vehicles" intended for over-the-road use. Examples are forklift trucks, tow tractors, trailers, and excavators.

500—Motor Vehicles: highway passenger and cargo vehicles customarily used on public roads.

600—Railroad Cars: all rolling stock suitable for use on national railroads. Excludes narrow-gage cars and locomotives for in-plant use.

700—Marine Carriers: all waterborne vessels used on canals, rivers, oceans and lakes.

800—Aircraft: all types of aircraft used to transport, lift, or position materials.

900—Containers/Supports: containers, platforms, pallets, coil supports, securement, bulkheads, dunnage, and related items.

The second and third digit positions represent finer classifications. For example, the 100 series indicates conveyors; the 110 series indicates belt conveyors; the 111 code indicates a bulk material belt conveyor. Similarly, 422 indicates a platform hand truck.

The list gives codes that would normally be used in a commercial or industrial facility. It updates the original Bolz system to include new types of equipment and exclude older, rarely used items.

To use Fig. 31, identify the load characteristics in each category across the top of the charts. These charactenstics are: Load Unit, Frequency, Distance, Path, Location. Moving down the columns note the equipment types which match the required characteristics. The columns for Equipment Characteristics show additional information.

While the load characteristics shown are important, other factors impact the decision. Moreover, many

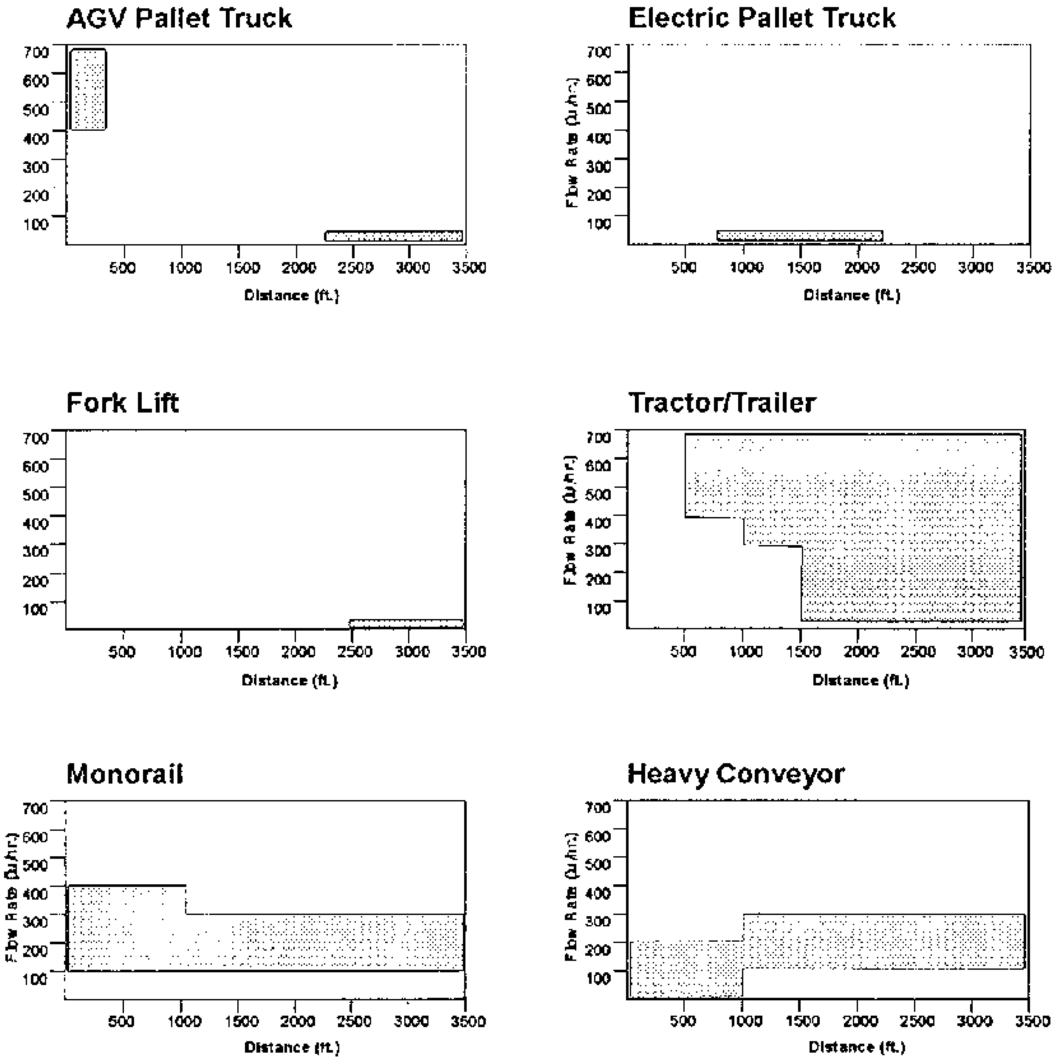


Figure 29 Equipment handling cost comparisons.

types may meet the load requirements. Here are some additional criteria to narrow the selection:

Equipment reliability
Capital investment
Operating expense
Safety
Flexibility
Training requirements
Maintainability
Standardization.

The final task in system design is to calculate equipment requirements. For fixed-path equipment this is straightforward. For variable path equipment such as fork trucks it requires estimating the time and distance on each route and then applying utilization factors. For sophisticated systems such as automatic guided vehicles and automatic storage and retrieval systems, a computer simulation should be used to test the feasibility.

1.4.4 Industrial Forklift Trucks

Forklift trucks (FLTs) belong to a category of equipment which is so common, versatile and useful that it warrants further discussion.

The *counterbalanced FLT* is the most universal. These trucks come with many options, some of

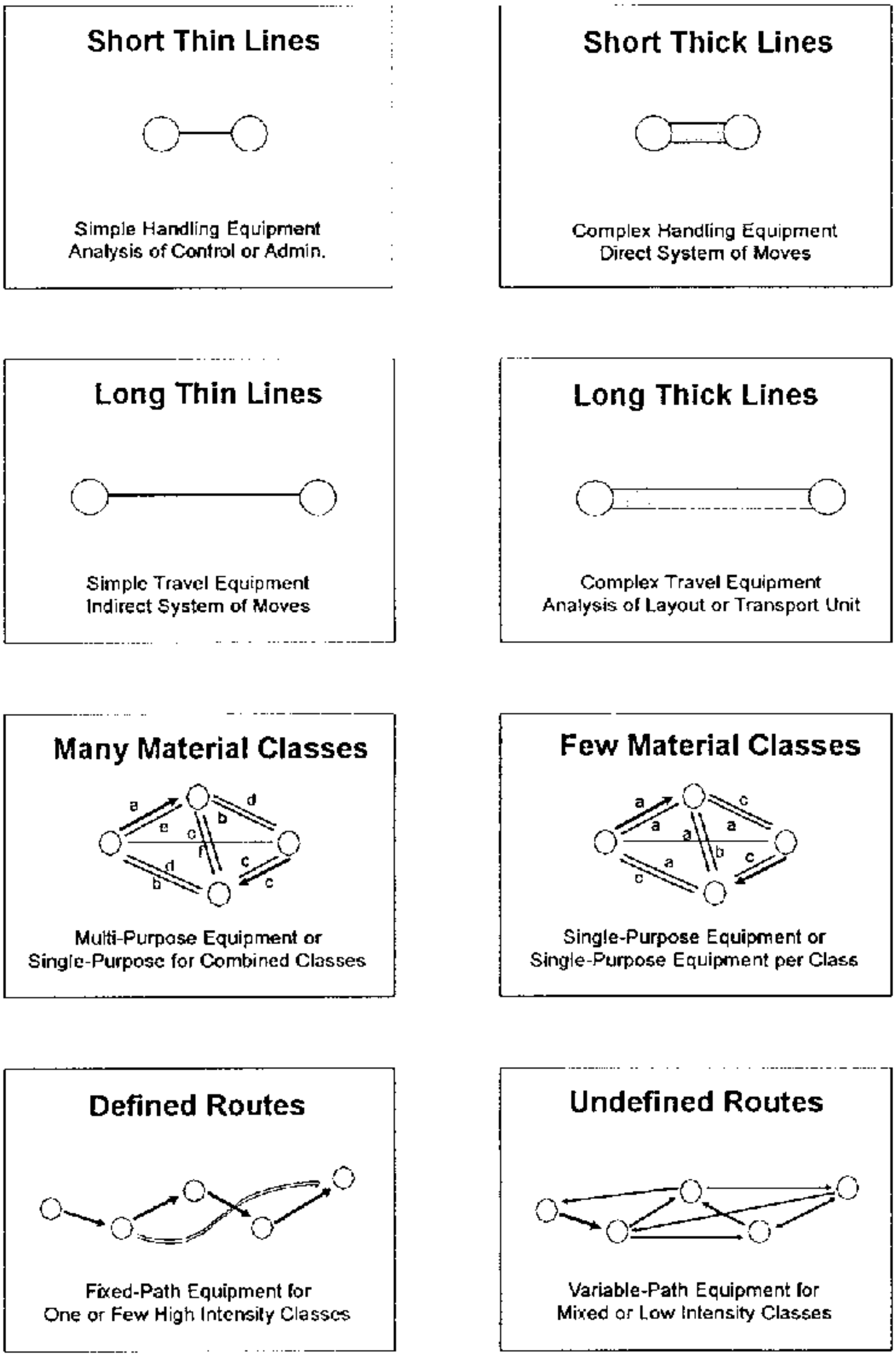

Short Thin Lines
Simple Handling Equipment
Analysis of Control or Admin.
Short Thick Lines
Complex Handling Equipment
Direct System of Moves
Long Thin Lines
Simple Travel Equipment
Indirect System of Moves
Long Thick Lines
Complex Travel Equipment
Analysis of Layout or Transport Unit
Many Material Classes
a
d
e
b
c
f
c
d
b
c
Multi-Purpose Equipment or
Single-Purpose for Combined Classes
Few Material Classes
a
c
a
a
a
a
b
c
a
c
Single-Purpose Equipment or
Single-Purpose Equipment per Class
Defined Routes
Fixed-Path Equipment for
One or Few High Intensity Classes
Undefined Routes
Variable-Path Equipment for
Mixed or Low Intensity Classes

Figure 30 Equipment selection guide.

			Material	Frequency	Distance	Path	Power	Operator	Mobility	Selfload	Tier/Stack
100	Conveyors										
	111	Bulk Material	B	CH	ELM	F	E	N	AN	YN	N
	112	Package	IPT	H	LMS	F	E	N	N	N	N
	112	Metallic Belt	IPT	H	LMS	F	E	N	N	N	N
	119	Other	-	-	-	-	-	-	-	-	-
	121	Bucket Elevators	B	C	MS	FV	E	N	AN	Y	N
	129	Other	-	-	-	-	-	-	-	-	-
	131	Apron	IPT	H	MS	F	E	N	N	N	N
	132	Slat	IPT	H	MS	F	E	N	N	N	N
	133	Crossbar	IPT	H	MS	F	E	N	N	N	N
	134	Carrier Chain	IP	H	LM	F	E	N	N	N	N
	135	Pallet	P	H	MS	F	E	N	N	N	N
	136	Car	P	HM	LM	F	E	N	N	N	N
	137	Trolley	P	HM	LM	F	E	N	N	N	N
	138	Aerial Tramways	P	HM	EL	F	E	N	N	N	N
	141	Drag	B	C	MS	FV	F	N	AN	N	N
	142	Flight	B	C	MS	F	E	N	N	N	N
	143	Tow	P	CH	LM	F	E	N	N	N	N
	144	Cable Tramways	P	HM	EL	F	E	YN	N	N	N
	145	Car Hauls	P	HM	LM	F	E	YN	N	N	N
	149	Other	-	-	-	-	-	-	-	-	-
	151	Straight Roll	IPT	H	MS	FV	ME	N	AN	N	N
	152	Concave Roll	IPT	H	MS	FV	ME	N	AN	N	N
	153	Herringbone Roll	IPT	H	MS	FV	ME	N	AN	N	N
	154	Skewed Roll	IPT	H	MS	FV	ME	N	AN	N	N
	155	Troughed Roll	IPT	H	MS	FV	ME	N	AN	N	N
	156	Wheel	IPT	H	MS	FV	ME	N	AN	N	N
	159	Other	-	-	-	-	-	-	-	-	-
	161	Horizontal	B	C	MS	FV	E	N	AN	YN	N
	162	Vertical/Inclined	B	C	MS	FV	E	N	AN	YN	N
	163	Feeders	B	CHM	MS	FV	E	N	AN	YN	N
	164	Mixers	B	CHM	MS	FV	E	N	AN	YN	N
	169	Other	-	-	-	-	-	-	-	-	-
	171	Hydraulic	B	C	ELM	F	O	N	N	NY	N
	172	Gas	B	C	ELM	F	O	N	N	Y	N
	173	Pneumatic Bulk	B	C	ELM	F	O	N	N	Y	N
	174	Pneumatic Package	I	HM	ELM	F	O	N	N	N	N
	175	Air-Jet Hydraulic	I	HM	-	-	O	N	N	-	-
	179	Other	-	-	-	-	-	-	-	-	-
	181	Electric	BIT	C	MS	FV	E	N	AN	YN	N
	182	Mechanical	BIT	C	MS	FV	E	N	AN	YN	N
	183	Oscillating	BIT	C	MS	FV	E	N	AN	YN	N
	184	Reciprocating	BIT	C	MS	FV	E	N	AN	YN	N
	189	Other	-	-	-	-	-	-	-	-	-
	191	Feeders	-	-	-	-	EM	N	-	YN	N
	192	Screens	-	-	-	-	EM	N	-	YN	N
	193	Hopper-Car Shakeouts	-	-	-	-	EM	N	-	YN	N
	194	Chutes	-	-	-	-	M	N	-	YN	N
	195	Hoppers/Troughs/Spouts	-	-	-	-	M	N	-	YN	N
	196	Flumes/Sluices	-	-	-	-	M	N	-	YN	N
	197	Air Cushion Transfers	IPT	HML	MS	V	EO	YN	UAN	YN	N
	199	Other	-	-	-	-	-	-	-	-	-
200	Cranes/Elevators/Hoists										
	211	Pedestal	IPT	ML	MS	-	EM	Y	N	N	-
	212	Pillar Jib	IPT	ML	MS	-	EM	Y	N	N	-
	213	Supported Jib	IPT	ML	MS	-	EM	Y	N	N	-
	214	Revolving Jib	IPT	ML	MS	-	EM	Y	N	N	-
	215	Derricks	IPT	ML	LMS	-	EM	Y	N	N	-
	219	Other	-	-	-	-	-	-	-	-	-
	221	Overhead Bridge	IP	ML	LMS	-	EI	Y	A	N	Y
	222	Gantry Portal	IP	ML	LMS	-	EI	Y	A	N	Y
	223	Gantry Storage	IP	ML	LMS	-	EI	Y	A	N	Y
	224	Overhead Monorail	IPT	HML	LMS	-	EI	Y	N	N	Y
	229	Other	-	-	-	-	-	-	-	-	-
	231	Railroad	IPT	ML	EL	F	I	Y	A	N	Y
	232	Crawler	IP	ML	LMS	V	I	Y	UA	N	Y
	233	Wheeled	IPT	ML	LMS	V	I	Y	UA	N	Y
	234	Floor	IPT	ML	LMS	V	E	Y	UA	N	Y
	235	Floating	IP	ML	LMS	V	I	Y	UA	N	Y
	239	Other	-	-	-	-	-	-	-	-	-
	241	Personnel	-	-	-	F	EO	YN	N	N	N
	242	Self-Supported	-	-	-	F	E	YN	N	N	N
	243	Dumbwaiters	IPT	HM	-	F	EM	YN	N	N	N
	244	Construction	IPT	ML	-	FV	E	YN	AN	N	N
	245	Freight	P	HML	-	F	EO	YN	N	N	N
	246	Skip Hoists	-	-	-	F	E	YN	AN	-	-
	249	Other	-	-	-	-	-	-	-	-	-
	251	Rope-Trolley Tautline	IPT	HM	LM	FV	EI	Y	UAN	N	N
	252	Man-Trolley Tautline	IPT	HM	LM	FV	EI	Y	UAN	N	N
	253	Slackline	IP	HM	LM	FV	EI	Y	UAN	N	N
	254	Drag Scrapers	B	C	MS	V	EI	Y	A	N	N
	259	Other	-	-	-	-	-	-	-	-	-
200	Continued										
	261	Cylinder	IT	HM	MS	F	E	Y	N	N	N
	262	Chain	IPT	HML	MS	FV	EM	Y	A	N	NY
	263	Cable	IPT	HML	MS	FV	EM	Y	A	N	NY
	264	Rope	IPT	HML	MS	FV	EM	Y	A	N	NY
	269	Other	-	-	-	-	-	-	-	-	-
	271	Windlasses	WIP	ML	MS	FV	EMI	Y	AN	N	N
	272	Capstans	WIP	ML	MS	FV	EI	Y	AN	N	N
	273	Single-Drum	WIP	ML	MS	FV	EM	Y	AN	N	N
	274	Multi-Drum	WIP	ML	LMS	FV	EI	Y	AN	N	N
	279	Other	-	-	-	-	-	-	-	-	-
	281	Sling	IPT	HML	MS	V	-	Y	-	-	-
	282	Grabs	BI	HML	MS	V	-	N	-	-	-
	283	Buckets	BI	HML	MS	V	-	N	-	-	-
	284	Magnets	BI	HML	MS	V	E	N	-	-	-
	285	Hook Auxiliary	-	ML	-	-	-	Y	-	-	-
	286	Boom Attachments	-	ML	-	-	-	N	-	-	-
	289	Other	-	-	-	-	-	-	-	-	-
300	Position/Weigh/Control Equipment										
	311	Manipulators	IT	HM	S	FV	EO	N	-	Y	N
	312	Electric Robots	IT	HM	S	FV	E	N	-	Y	NY
	313	Hydraulic Robots	IT	M	S	FV	O	N	-	Y	NY
	319	Other	-	-	-	-	-	-	-	-	-
	321	Upenders/Turnovers	IPT	HM	S	FV	EM	YN	AN	YN	N
	322	Car Dumpers	W	ML	S	FV	E	Y	AN	N	N
	323	Truck Dumpers	W	ML	S	FV	E	Y	AN	N	N
	324	Box/Barrel/Bag Dum	PT	HML	S	FV	EM	YN	AN	YN	N
	329	Other	-	-	-	-	-	-	-	-	-
	331	Tables	IT	-	-	-	EM	-	-	-	-
	332	Platforms	IPT	-	-	-	-	-	-	-	-
	339	Other	-	-	-	-	-	-	-	-	-
	341	Transfer Cars	IP	ML	LMS	F	EMI	Y	AN	N	N
	342	Turntables	IPT	HML	S	F	EM	YN	N	YN	N
	343	Ball/Caster	IP	ML	MS	FV	M	YN	AN	YN	N
	344	Walking Beam	IPT	CH	S	F	E	N	N	YN	N
	349	Other	-	-	-	-	-	-	-	-	-
	371	Yard	-	-	-	-	EM	Y	N	N	-
	372	Platform	-	-	-	-	EM	NY	N	N	-
	373	Portable	-	-	-	-	M	Y	A	N	-
	374	Counter	-	-	-	-	EM	NY	-	YN	-
	375	Batch	-	-	-	-	EM	NY	N	YN	-
	376	Conveyor	-	-	-	-	E	NY	N	Y	-
	377	Crane	-	-	-	-	M	Y	-	N	-
	378	Spring	-	-	-	-	M	Y	-	N	-
	379	Other	-	-	-	-	-	-	-	-	-
400	Industrial Vehicles										
	410	Powered Trucks	IPT	-	LMS	V	EI	Y	U	N	Y
	411	Platform	IPT	-	LMS	V	EI	Y	U	N	Y
	420	Hand Trucks	IPT	-	MS	V	M	Y	U	N	N
	421	Two-Wheel	IT	-	MS	V	M	Y	U	N	N
	422	Platform	-	-	MS	V	M	Y	U	N	-
	423	Wheelbarrows	-	-	MS	V	M	Y	U	N	N
	424	Carts	IT	-	MS	V	M	Y	U	N	N
	425	Dollies	IT	-	MS	V	M	Y	U	N	N
	426	Bicycles/Tricycles	IT	-	EL	V	M	Y	U	N	N
	429	Other	-	-	-	-	-	-	-	-	-
	430	Indust. Trailers	IPT		EIL	V	M	-	U	N	N
500	Motor Vehicles										
600	Railroad Cars/Locomotives										
700	Marine Carriers										
800	Aircraft										
900	Containers/Supports										

Material:

B	Bulk
I	Single Item
P	Pallet/Container
T	Tote/Box/Tray
L	Liquid
G	Gas
W	Wagon/Car/Truck

Frequency:

C	Continuous
H	>100 Loads/hour
M	>10 Loads/hours
L	>1 Load/hour

Path:

F	Fixed
V	Variable
A	Area

Distance:

S	Short < 10'
M	Med. < 100'
L	Long < 1,000'
E	Extra Long < 10,000'
O	Over the Road

Power:

I	Internal Combustion
S	Steam
E	Electrical
M	Manual or Unpowered
O	Other

Mobility:

N	None
A	Area
U	Unlimited

Others are Yes / No

Figure 31 Material handling selection guide.

which are: three or four wheel; battery driven or internal combustion engine; rider, stand-up or walkie; duplex, triplex or quad mast; and pneumatic, cushioned, or solid tires.

The counterbalanced design puts a large force on the front wheels and can cause floor loading problems. Lifting capacity diminishes with height and there is some danger of overbalancing. Carton clamps, side shifters and coil handlers are some of the available attachments.

Reach trucks have small wheels near the forward edge and on each side of the load, thus requiring less counterbalancing. In operation, the forks or the entire mast extends to pick up or set down a load. The truck does not travel in this extended position. Some characteristics of reach trucks as compared with counterbalanced trucks are:

5%–15% slower
Have nontilting forks
Require better floors
Use smaller batteries
Have poorer ergonomics
Are lighter weight
Work in narrower aisles.

Other forklift trucks include the following:

Pallet trucks are small, inexpensive machines which pick up pallets resting on the floor or on low stacks. Both manual and battery-powered models are available. Pallet trucks cannot handle double faced pallets and require almost perfect floor conditions.

Stackers are small manual or electric machines similar to pallet trucks. Unlike pallet trucks, they can elevate and thus stack their loads. Outriggers or legs support the weight of the load. Outriggers straddle the load; legs are underneath the forks. Stackers are often used in maintenance or tool changing.

Four-way trucks are useful for carrying long items lengthwise through relatively narrow aisles. They are variations of the reach truck with rotatable wheels that allow them to travel in four directions.

Turret trucks have a mast that rotates on a track without extending beyond the width of the machine. These trucks can operate in an aisle only 6 in. wider than the truck and access pallets on both sides. Turret trucks are used for high rise storage operations.

Side-loader trucks pick up and carry the load on the side. The forks are at right angles to the travel direction, which is useful for long, narrow loads such as pipe or lumber. The side loader carries such loads lengthwise down the aisle.

1.4.5 Conveyors

Conveyors are fixed devices which move material continuously on a pre-established route. These systems range from short, simple lengths of unpowered conveyor to vast networks of conveyor sections with sophisticated controls.

Belt conveyors have a flexible belt which rides on rollers or a flat bed. The belt may be cloth, rubber, plastic, wire mesh, or other material. Most articles can ride a belt conveyor up to 30° inclination.

With *roller and skate-wheel conveyors*, objects ride on rollers or wheels. Any objects on the conveyor should span at least three sets of rollers. Movement can come from powered rollers, gravity, or operators.

Chain conveyors carry or push objects with a chain. Many varieties are available.

Overhead conveyors use an I-beam or other shape as a monorail. Carriers roll along the monorail with loads suspended underneath. A chain connects the carriers and pulls them along. In a power-and-free system, the chain and carriers are independent. A disconnection mechanism stops the carrier. Power-and-free systems offer more flexibility than standard monorails but at a much higher cost. Recent designs of power and free conveyors are inverted and floor mounted.

1.4.6 Vibratory Machines

Ware [1] defines a vibratory machine as "any unit intentionally or purposely vibrated in order for it to perform useful work. Vibration induces a material to move instead of forcing it."

The two distinct categories of vibratory machines that are most often used in material handling systems are those for accomplishing induced vertical flow and induced conveying.

1.4.7 Automatic Guided Vehicle Systems

Automatic guided vehicle systems (AGVS) use driverless vehicles to transport materials within an operation.

AGV size can vary from small, light-duty vehicles that carry interoffice mail to heavy-duty systems that transport automobiles during assembly. Several types of guidance are available with a range of sophistication in logic and intelligence.

Most AGVs move along a predetermined track system not unlike a model railroad. Optical tracking systems use reflective tape or paint on the floor to define the track. A photosensitive device on the vehicle detects drift from the track and actuates the steering mechanism for correction. Optical systems are inexpensive and flexible. They are sensitive to dirt, however, and many users consider them unsatisfactory.

Electromagnetic guidance systems follow a magnetic field generated by conductors laid in the floor. The frequency of this field can vary in each track section and thus identify the vehicle's location. Sensors on the vehicle detect the field, its location and perhaps the frequency. The guidance system corrects the vehicles track accordingly. Electromagnetic guidance systems are somewhat expensive to install or relocate, but AGV owners generally prefer electromagnetic guidance systems for their reliability.

A newer type of guidance system optically reads "targets" placed high on walls and columns. The system then computes vehicle position with triangulation. In the future, guidance systems may use the satellite navigation systems.

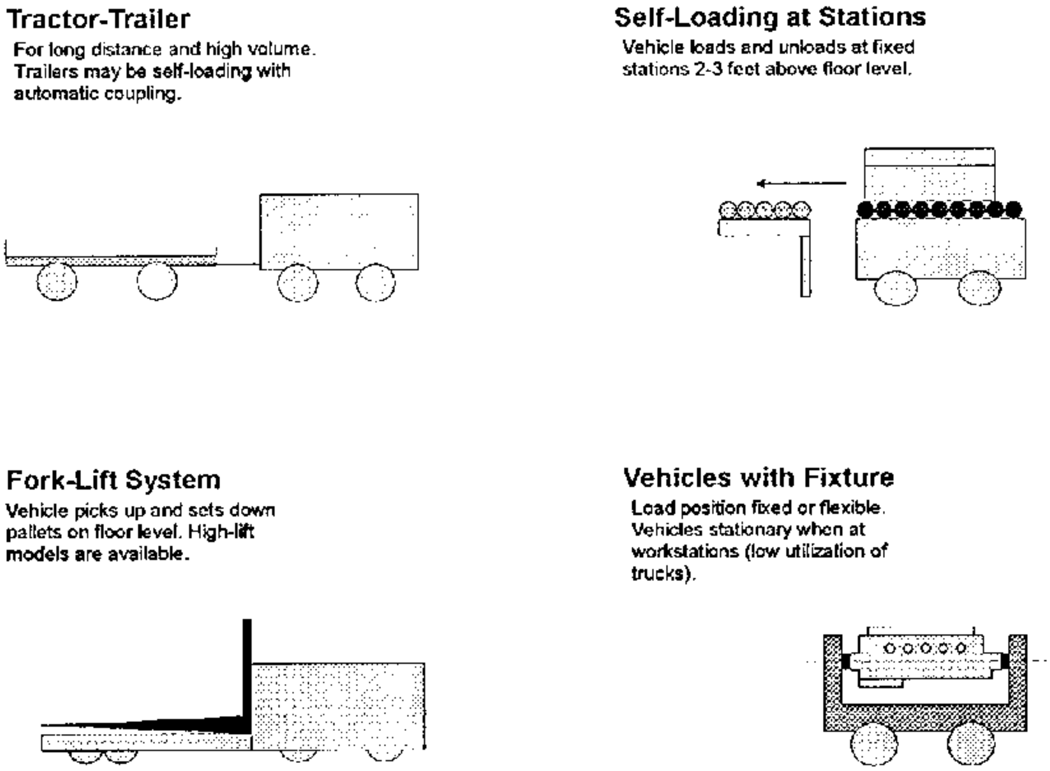
Figure 32 illustrates some of the vehicles available for AGV systems. Tractor–trailer systems use a driverless tractor to tow one or more trailers, using manual or automatic coupling. Such systems are best for large loads and long distances. Some vehicles serve as assembly stations in addition to moving loads.

Self-loading vehicles stop at fixed stations and load or unload containers. These are normally pallet-size loads.

AGV forklift systems use vehicles similar to pallet trucks. They can pick up a pallet, carry it to a new location and lower it automatically. All or part of the cycle may be automatic.

Special systems may use fixtures to carry engines, automobiles or other large products through a production process.

At the lowest level of control vehicles follow a single path in a single direction. They stop at predetermined stations, at obstructions or when encountering



Tractor-Trailer

For long distance and high volume. Trailers may be self-loading with automatic coupling.

Self-Loading at Stations

Vehicle loads and unloads at fixed stations 2-3 feet above floor level.

Fork-Lift System

Vehicle picks up and sets down pallets on floor level. High-lift models are available.

Vehicles with Fixture

Load position fixed or flexible. Vehicles stationary when at workstations (low utilization of trucks).

Figure 32 Automatic guided vehicles.

another. Intelligent trucks have preprogrammed destinations, locating their position by sensing the magnetic frequencies. These vehicles can use multiple paths to navigate to and stop at their assigned destination. Centralized control systems use a computer to track vehicles and control movement. Vehicles broadcast their current location and the computer sends control signals back to the vehicle controlling both movement and route.

Towveyors were the precursors to AGVs. They are powered by a cable or chain which moves continuously in a floor groove. A pin or grip mechanism connects and disconnects the vehicle.

1.4.8 System Design and Documentation

When the flow analysis is complete and a layout selected, it is time to prepare the macrolevel material handling plan.

Now that the handling for each route has been identified, equipment requirements are estimated. In the case of fixed-path equipment, such as roller conveyors, this estimation is simple and straightforward. Where variable-path equipment is used on multiple routes, estimate the total time required for each route and material class as well as the effective equipment utilization. In Fig. 33 an example estimate is shown for a bakery ingredients warehouse.

1.5 WAREHOUSING AND STORAGE

The most successful manufacturers and distributors now recognize that inventory often camouflages some form of waste. The causes of waste are in the structure of the inventory systems. The ultimate goal is to restructure and eliminate all storage of products.

Restructuring for minimum inventory is usually more fruitful than pursuing better storage methods, although compromises must be made and a requirement for some storage often exists.

1.5.1 Stores Activities

This section explains how to design optimum storage systems for the inventory which remains after a suitable restructuring effort.

Storage operations have two main functions: *holding* and *handling*. Holding refers to the stationing of materials in defined storage positions. Handling is the movement to and from the storage position. Ancillary activities such as inspection, order picking, or receiving are also part of handling.

Average turnover is the ratio of annual throughput to average inventory over the same period. Throughput and inventory may be in dollars, production units, or storage units ($, pieces, pallets, cartons).

$$\text{Turnover} = \frac{\text{Annual throughput}}{\text{Average inventory}}$$

The relative importance of holding and handling in a particular situation guides the analysis. With high turnover, *handling* dominates; with low turnover, *holding* dominates.

Handling-dominated warehouses call for detailed analysis of procedures and material handling. These warehouses use more sophisticated handling devices such as automatic storage and retrieval systems (ASRS) and automated conveyors.

Holding-dominated warehouses call for simple, inexpensive, and flexible handling equipment. These warehouses often require high-density storage methods, such as drive-through racking.

1.5.2 Storage Equipment

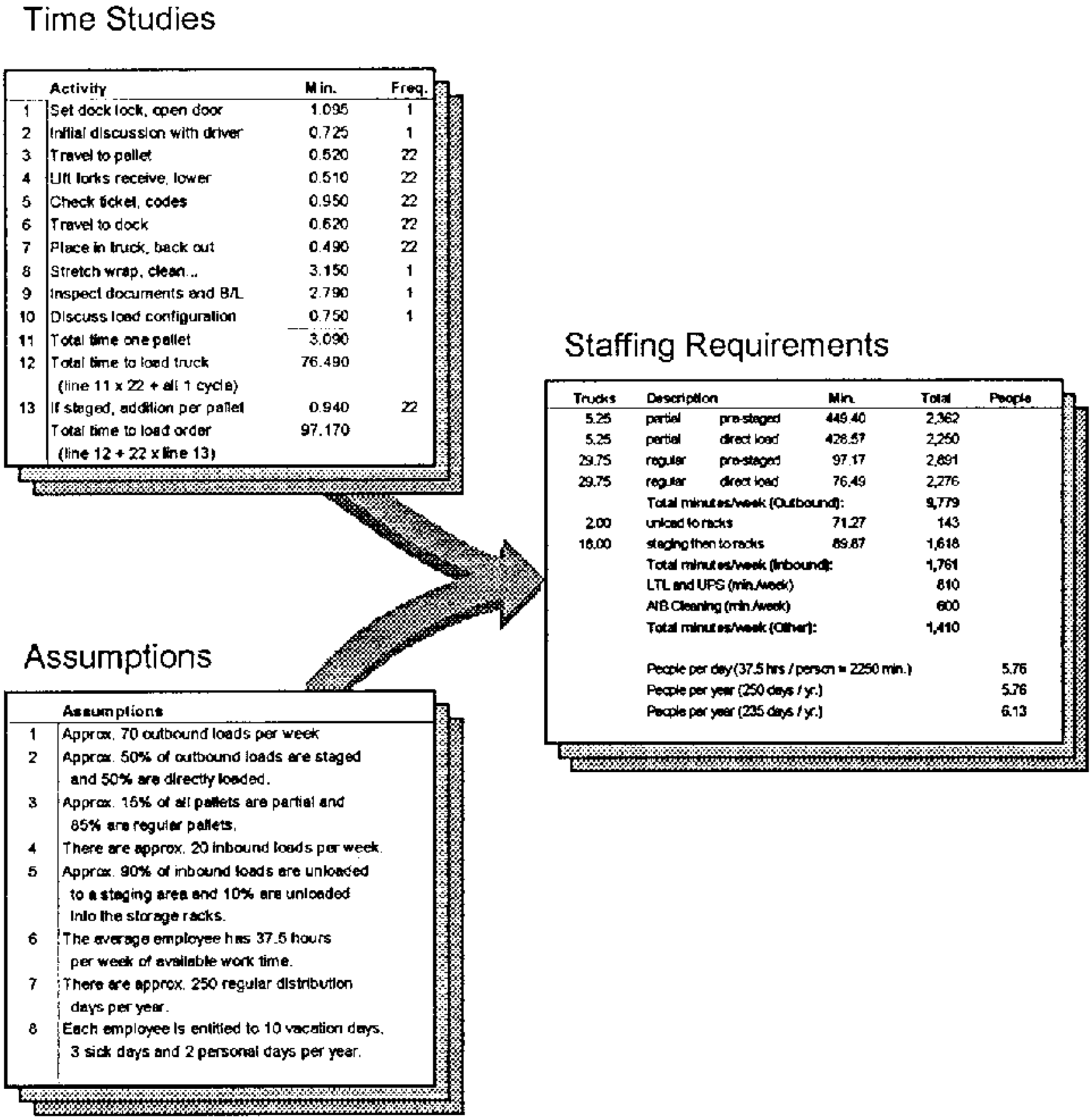
The types of storage equipment available are almost as diverse as the types of containers and handling equipment. The selection of both storage equipment and containers is interrelated.

1.5.3 Analysis and Design of Storage Systems

The design of storage systems should co-ordinate with the layout design of the total facility. Layout planning has four phases: orientation, macrolayout, populated layout, and implementation.

Orientation. Storage planning during this phase is at a high level. In this phase the planners are oriented to the entire scope of the project, for example, the building size estimates, planning assumptions, project staffing, and policies and strategies to be supported.

Macrolayout. This is the main planning phase where the major storage area SPUs are determined. In addition to determining storage space these SPUs can include pick and pack areas, docks, and receiving areas, although some of them may be in a separate location. The designer refines estimates of storage space and co-ordinates them with other design and strategic decisions.



Time Studies

	Activity	Min.	Freq.
1	Set dock lock, open door	1.095	1
2	Initial discussion with driver	0.725	1
3	Travel to pallet	0.520	22
4	Lift forks receive, lower	0.510	22
5	Check ticket, codes	0.950	22
6	Travel to dock	0.620	22
7	Place in truck, back out	0.490	22
8	Stretch wrap, clean...	3.150	1
9	Inspect documents and B/L	2.790	1
10	Discuss load configuration	0.750	1
11	Total time one pallet	3.090	
12	Total time to load truck (line 11 x 22 + all 1 cycle)	76.490	
13	If staged, addition per pallet	0.940	22
	Total time to load order (line 12 + 22 x line 13)	97.170	

Assumptions

	Assumptions
1	Approx. 70 outbound loads per week
2	Approx. 50% of outbound loads are staged and 50% are directly loaded.
3	Approx. 15% of all pallets are partial and 85% are regular pallets.
4	There are approx. 20 inbound loads per week.
5	Approx. 90% of inbound loads are unloaded to a staging area and 10% are unloaded into the storage racks.
6	The average employee has 37.5 hours per week of available work time.
7	There are approx. 250 regular distribution days per year.
8	Each employee is entitled to 10 vacation days, 3 sick days and 2 personal days per year.

Staffing Requirements

Trucks	Description		Min.	Total	People
5.25	partial	pre-staged	449.40	2,362	
5.25	partial	direct load	428.57	2,250	
29.75	regular	pre-staged	97.17	2,891	
29.75	regular	direct load	76.49	2,276	
	Total minutes/week (Outbound):			9,779	
2.00	unload to racks		71.27	143	
18.00	staging then to racks		89.87	1,618	
	Total minutes/week (Inbound):			1,761	
	LTL and UPS (min./week)			810	
	A/B Cleaning (min./week)			600	
	Total minutes/week (Other):			1,410	
	People per day (37.5 hrs / person = 2250 min.)				5.76
	People per year (250 days / yr.)				5.76
	People per year (235 days / yr.)				6.13

Figure 33 Calculating requirements.

Populated layouts. Populating the layout is where each piece of equipment is located within each space planning unit. Detail planning of designated storage occurs during this phase, which may or may not include the identification of storage locations for each item number. In situations where handling and picking dominate, the specific storage location may be very important for effective operation of the overall system. In other situations, assignment of storage locations is considered an operating decision.

Implementation. In the final phase, equipment is purchased, installed, and operations initiated.

The detailed storage plan should include the following:

Populated layout
Material handling plan
Equipment requirements summary

Information systems plan
Staffing plan.

Populated layouts show the location of all racks, aisles, doors, offices, and other features. The layout should have sufficient information to prepare architectural and installation drawings.

The material handling plan for the storage operations is similar to that made for the macrolayout of any facility. It shows all origin and destination points for materials. It shows flow rates, equipment, and containers used on each route. For many warehousing operations, the material handling plan is simple and can be overlaid on a layout plan.

The equipment requirements summary lists the types and numbers of storage and handling equipment. It should also include a summary specification for each type of equipment.

The information systems plan specifies the type of information which is to be available, equipment required and other data necessary to purchase equipment and set up systems. It should include manual as well as computer-supported systems.

Preparing a complete storage plan requires the following steps:

1. Acquire data/information.
2. Classify storage materials.
3. Calculate material and order flows.
4. Calculate storage requirements.
5. Select equipment.
6. Plan the layout.
7. Specify information procedures and systems.

Step 1. Acquire Data/Information. Information required for the storage analysis covers products, volumes, inventory, orders, and current and past operations.

Products and volumes. Information on products includes general orientation material on the types of products to be stored and any special characteristics. A detailed list or database should be included with every item number, and products should be included by size, brand, or other classification. Volume information should include historical sales (or throughput) volumes for each line item or product group as well as total sales. This is often the same product volume information used for facility planning and material handling analysis. A product profile showing items or groups and their volumes on a ranked bar chart is useful. Forecasts by product group should be obtained or prepared.

Inventory. Historical inventory information may be available when there is a similar existing operation. The information should include average and peak inventory for each item or product group over a meaningful period. When historical information does not apply, policies or judgnent must suface. A decision to keep ‘‘two months on-hand’’ or ‘‘maintain an average 10 turns’’ can help establish inventory requirements.

Orders. An order refers to any withdrawal request. It may be a sales order, production order or verbal request for incidental supplies. An order profile shows the average line items and line item quantity per order. The profile may also include weekly or seasonal order patterns and should include forecast trends and changes. Identifying urgency or delivery requirements may be necessary in some cases.

Current and past operations. This information includes staffing, space usage, procedures, operation sequence, equipment, policies, and any other pertinent information not included above.

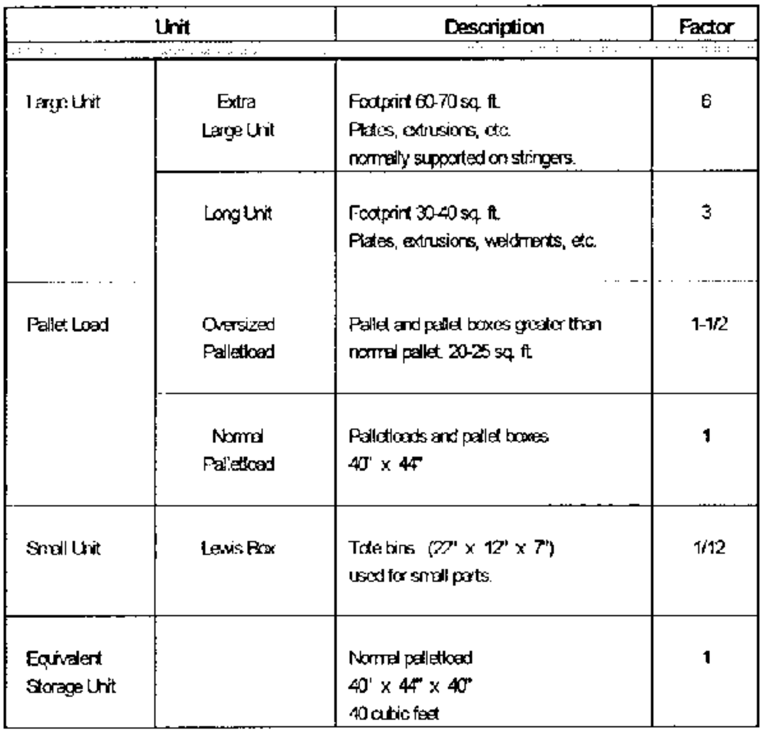
Step 2. Classify Materials. The classification of materials is similar to classification activities used for material flow analysis. There may be slight differences, however, since the primary concern here is storage characteristics. Figure 34 shows one classification scheme. Categories to select from are function, destination, work-in-process, finished goods, high turnover items and slow movers.

Step 3. Calculate Material and Order Flows. Material flows for a storage operation are calculated in the same way as for any other layout. Orders are an additional parameter. Order flows in a storage operation affect the timing of an order and picking patterns.

Step 4. Calculate Storage Requirements. For each storage class the storage space requirement must be calculated. This may be done by using turnover rates, existing data, or computer simulation. It is necessary in this step to confirm inventory policies and storage area utilization levels—random storage with high space utilization or dedicated locations with lower overall space utilization.

A ‘‘pull’’ replenishment system with certified vendors requires less space for operating, janitorial, maintenance, and office supplies.

Step 5. Select Equipment. In a warehouse operation handling and storage equipment are interrelated and should be selected together. Handling equipment types were discussed previously. Storage equipment types are discussed in Sec. 1.5.3.



Unit		Description	Factor
Large Unit	Extra Large Unit	Footprint 60-70 sq. ft. Plates, extrusions, etc. normally supported on stringers.	6
	Long Unit	Footprint 30-40 sq. ft. Plates, extrusions, weldments, etc.	3
Pallet Load	Oversized Palletload	Pallet and pallet boxes greater than normal pallet. 20-25 sq. ft.	1-1/2
	Normal Palletload	Palletloads and pallet boxes 40" x 44"	1
Small Unit	Lewis Box	Tote bins (22" x 12" x 7") used for small parts.	1/12
Equivalent Storage Unit		Normal palletload 40" x 44" x 40" 40 cubic feet	1

Figure 34 Material classification.

Step 6. Plan the Layout. Planning a storage or warehouse layout follows the same procedure as planning factory layouts. Figure 35 is helpful for estimating space for storage.

Step 7. Specify Management/Operating Systems. Figure 36 illustrates the external flows of material and information from and to the warehouse.

Figure 37 traces the high-level internal flows within a typical warehousing operation. When the storage system is designed these flows should be refined and specified. Operation process charts and information flow charts are useful for this documentation.

Information systems require specification. For simple manual systems a notation on the flow chart may suffice; for computer-based systems a more detailed specification is required. Figure 38 illustrates the overall operation of a computerized warehouse information system.

A staffing study should also be performed. This study may range from an informal estimate to detailed work-measurement study. Figure 39 is a time standard for unloading a typical truck. The MTM-based EASE software system was used to generate these data. From such time studies and throughput information an estimate of staffing is generated.

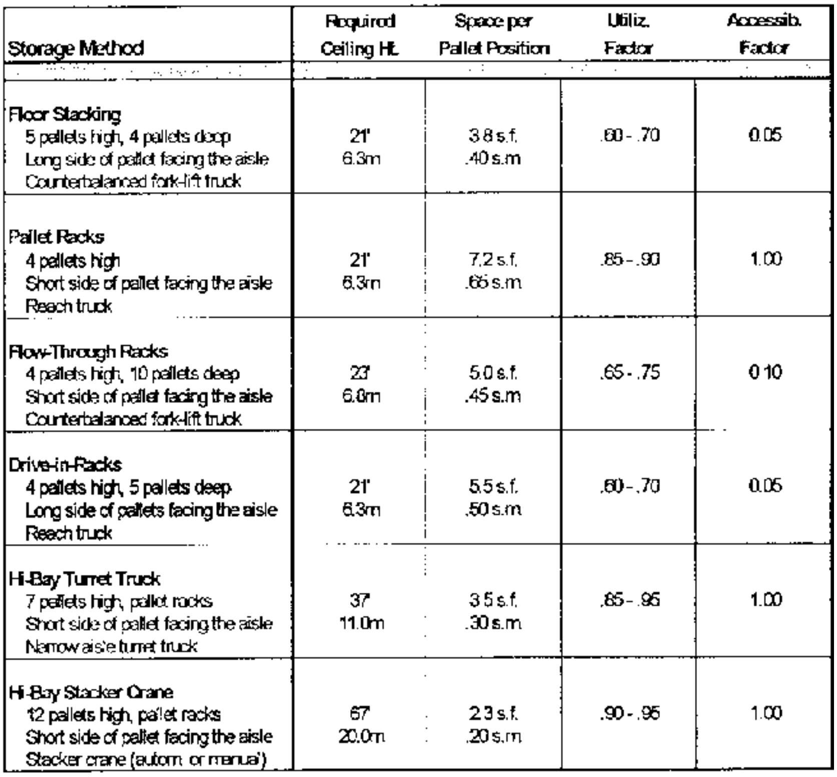
1.5.4 Pallet Storage

Floor stacking is the simplest way to store pallet loads. This method utilizes floor space and building volume effectively and can tolerate pallets with overhang. To achieve these advantages requires stacking three to five pallets deep which gives a high storage volume per item. In the right location, with access from both sides, floor stacking can operate on a first-in-first-out (FIFO) basis. Floor stacking also requires strong, stable, stackable, and unbroken loads, illustrated in Fig. 40.

Pallet racks should be used when loads are unstackable or storage volume is too small for deep floor stacking. *Double-deep racks* achieve higher density storage but require a reach truck causing problems of access to the rear pallet.

Flow-through racks are used when FIFO is important and turnover is high. In these racks the pallets or cartons ride on rollers or wheels and flow by gravity from the load to the unload end. They require high-quality pallets, and the racks are relatively expensive.

Drive-in racks should be used for large quantities of unstackable pallet loads. The rack openings must be wide enough for the truck and special narrow trucks



Storage Method	Required Ceiling Ht.	Space per Pallet Position	Utiliz. Factor	Accessib. Factor
Floor Stacking 5 pallets high, 4 pallets deep Long side of pallet facing the aisle Counterbalanced fork-lift truck	21' 6.3m	3.8 s.f. .40 s.m	.60 - .70	0.05
Pallet Racks 4 pallets high Short side of pallet facing the aisle Reach truck	21' 6.3m	7.2 s.f. .65 s.m	.85 - .90	1.00
Flow-Through Racks 4 pallets high, 10 pallets deep Short side of pallet facing the aisle Counterbalanced fork-lift truck	23' 6.8m	5.0 s.f. .45 s.m	.65 - .75	0.10
Drive-in-Racks 4 pallets high, 5 pallets deep Long side of pallets facing the aisle Reach truck	21' 6.3m	5.5 s.f. .50 s.m	.60 - .70	0.05
Hi-Bay Turret Truck 7 pallets high, pallet racks Short side of pallet facing the aisle Narrow aisle turret truck	37' 11.0m	3.5 s.f. .30 s.m	.85 - .95	1.00
Hi-Bay Stacker Crane 12 pallets high, pallet racks Short side of pallet facing the aisle Stacker crane (autom. or manual)	67' 20.0m	2.3 s.f. .20 s.m	.90 - .95	1.00

Figure 35 Storing space planning guide.

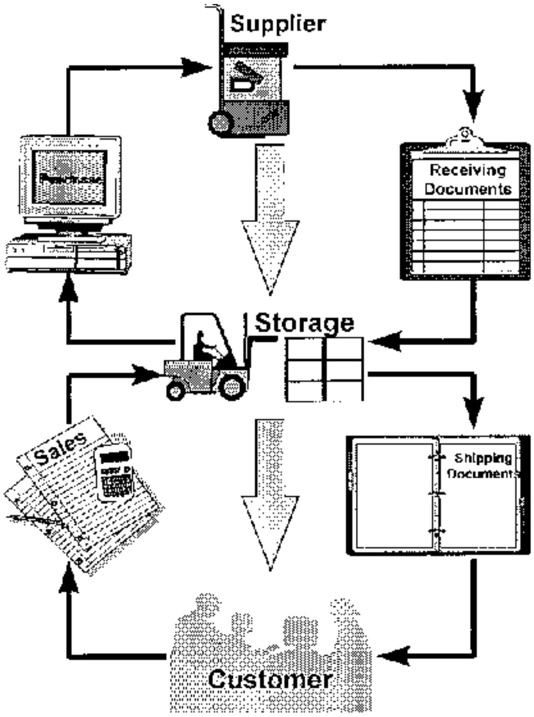


Figure 36 External material and information flows.

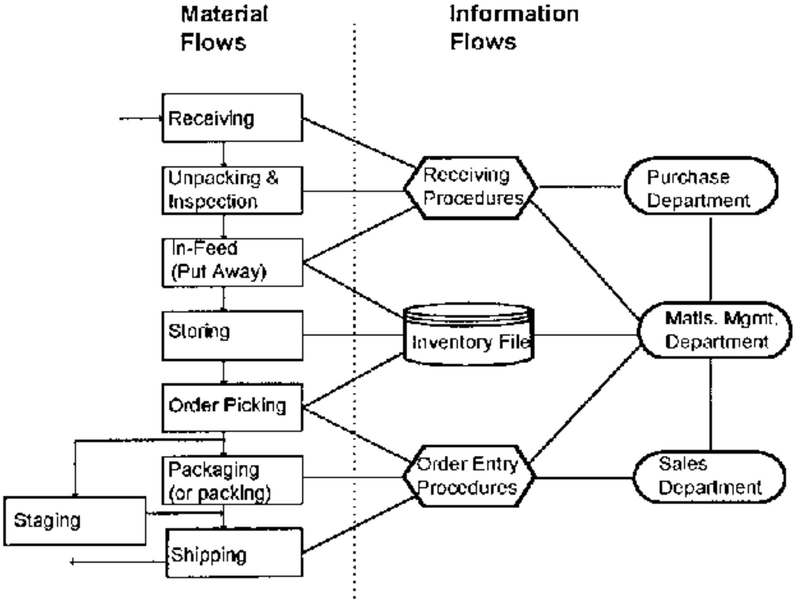


Figure 37 Internal material and information flows.

may be necessary. Since pallets are supported only on their edges, pallet quality must be high. Limited FIFO is possible if there is access to both sides.

1.5.5 Small Parts Storage

Small parts storage systems are either static or dynamic. Static systems include shelving and drawers in various configurations. Dynamic systems are vertical carousels, horizontal carousels, mini-trieves and movable-aisle systems.

Shelving is a basic inexpensive and flexible storage method. It often does not use space effectively and is costly for intensive picking operations.

Modular drawer systems offer denser storage than shelving. They are more expensive than shelves and more difficult for picking.

1.5.6 Automatic Storage and Retrieval Systems

Automatic storage and retrieval systems (ASRS) store materials in a high-density configuration. These systems use a stacker crane or other mechanical device to carry each load to its location and place it in storage. The same crane retrieves loads as required and delivers them to an output station. A computer system controls movements and tracks location. The ASRS computer often is in communication with a production control system such as MRP. Such systems usually work with pallet-size loads. Mini-trieve systems are similar in concept to automatic retrieval systems but use smaller loads such as tote boxes.

1.6 CONCLUSION

Materials movement is a key consideration in facility planning. The material flow analysis is necessary for proper facility design and is a prerequisite for the design of material handling systems and storage areas. It is also an important means of evaluating design options.

It is important to select the material handling equipment to fit the determined material flow system. Often the flow and handling are forced to fit the material handling methods you have been sold.

Even though we want to eliminate material handling and storage *waste* product storage may be required for aging, quarantine, or qualifying processes. In other cases storage serves as a buffer in an improperly designed and maintained system.

Following the step-by-step procedures outlined in this chapter will support the operating strategy by reducing costs, time and material damage. This is basic to achieving world class and being a time-based competitor.

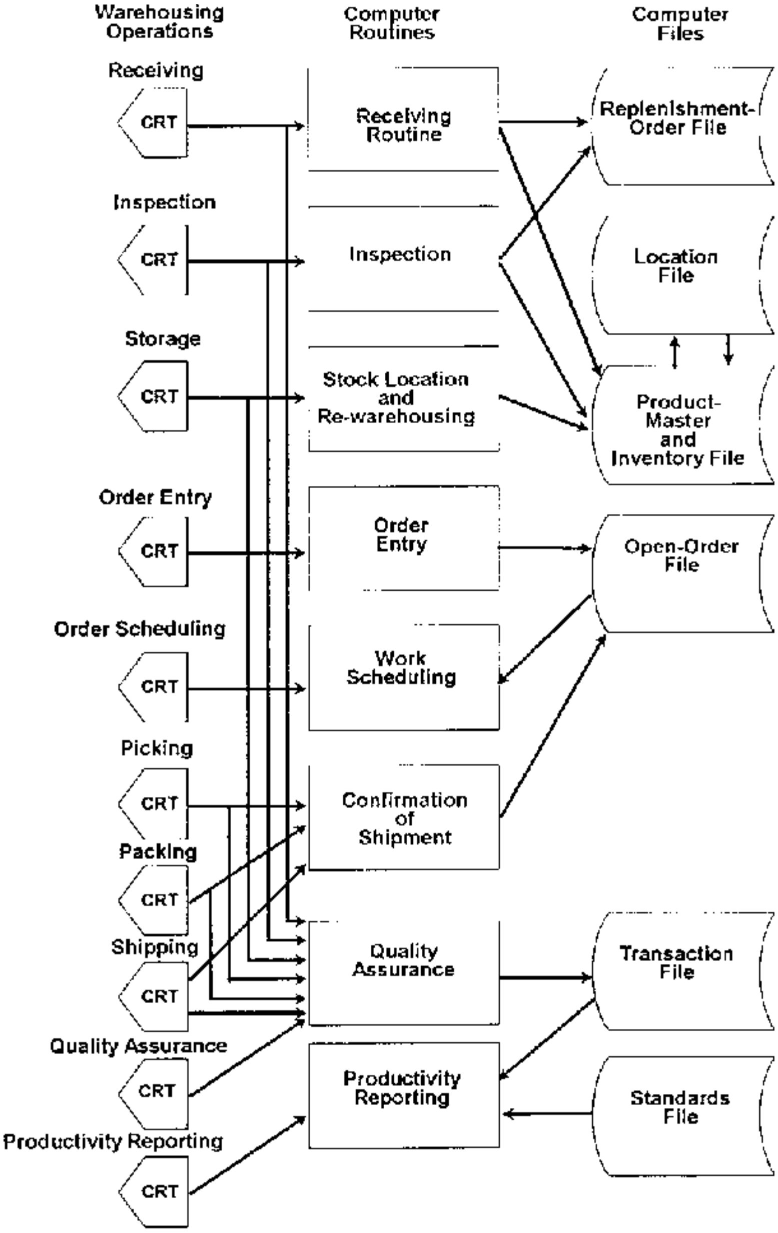
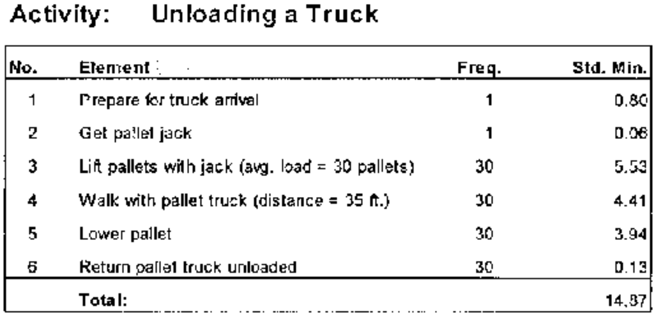


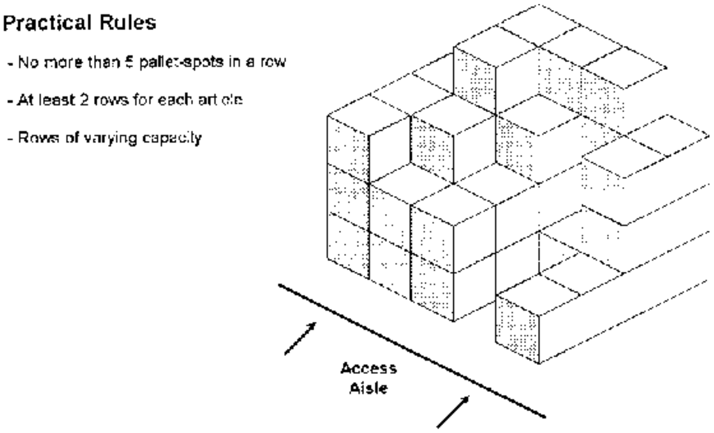
Figure 38 Computerized warehouse information system.



Activity: Unloading a Truck

No.	Element	Freq.	Std. Min.
1	Prepare for truck arrival	1	0.80
2	Get pallet jack	1	0.06
3	Lift pallets with jack (avg. load = 30 pallets)	30	5.53
4	Walk with pallet truck (distance = 35 ft.)	30	4.41
5	Lower pallet	30	3.94
6	Return pallet truck unloaded	30	0.13
	Total:		14.87

Figure 39 Ease^{TM} generated time standard.



Practical Rules

- No more than 5 pallet-spots in a row
- At least 2 rows for each article
- Rows of varying capacity

Figure 40 Floor stacking.

REFERENCES

1. B Ware. Using vibratory machines to convey bulk solids. Chemical Processing, Itasca, IL: Putman Publishing Company, 1998, pp 74–79.
2. W Wrennall, Q Lee, eds. Handbook of Commercial Facilities Management. New York: McGraw-Hill, 1994.
3. HA Bolz, GE Hagemann, eds. Materials Handling Handbook. New York: The Ronald Press, 1958, pp 1.5–1.16.

FURTHER READING

CR Asfahl. Robots And Manufacturing Automation, New York: John Wiley, 1985.

A Carre. Simulation of Manufacturing Systems, Chichester: John Wiley, 1988.

H Colijn. Mechanical Conveyors for Bulk Solids, Amsterdam: Elsevier, 1985.

G Hammon. AGVS at Work. Bedford, UK: IFS (Publications), 1986.

NL Hannon. Layout Needs: An Integrated Approach. Mod Mater Handling April, 1986.

WK Hodson, ed. Maynards's Industrial Engineering Handbook. 4th ed. New York: McGraw-Hill, 1992.

M Hulett. Unit Load Handling. London: Gower Press, 1970.

AL Kihan. Plant Services and Operations Handbook. New York: McGraw-Hill, 1995.

Modern Dock Design. Milwaukee: Kelly Company, 1997.

W. Müller. Integrated Materials Handling in Manufacturing. IFS (Publications), UK: Bedford, 1985.

U Rembold. Robert Technology and Applications. New York: Marcel-Dekker, 1990.

G Salvendy. Handbook of Industrial Engineering. 2nd ed. New York: Wiley-Interscience, 1992.

ER Sims. Planning and Managing Industrial Logistics Systems. Amsterdam: Elsevier, 1991.

JA Tompkins, JD Smith. The Warehouse Management Handbook. New York: McGraw-Hill, 1988.

W Wrennall. Requirements of a Warehouse Operating System. In: JA Tompkins, JD Smith, eds. The Warehouse Management Handbook. New York: McGraw-Hill, 1988, pp 531–559.

W Wrennall, Q. Lee. Achieving Requisite Manufacturing Simplicity. Manufacturing Technology International. London: Sterling Publications, 1989.

Chapter 7.2

Automated Storage and Retrieval Systems

Stephen L. Parsley
ESKAY Corporation, Salt Lake City, Utah

2.1 DEFINITION

A 20-year-old definition of automated storage and retrieval (AS/R) systems states that the technology is "... a combination of equipment and controls which handles, stores, and retrieves materials with precision, accuracy, and speed under a defined degree of automation" [1]. While basically sound, the definition somewhat limits the reader's imagination when it comes to the entire spectrum of functionality an AS/R system can offer to the planner designing a new logistics process.

Using today's "logistics-speak," the AS/R system is a device which automatically receives material arriving at an often anomalous rate, securely buffers the material in a controlled access structure, resequences and conditionally and responsively releases material out to points of consumption—all under a high degree of automation so as to eliminate the need for human resources in the process of performing these non-value-added functions.

2.2 A BRIEF HISTORY

Automated storage and retrieval systems were initially introduced in the late 1960s, and rose to popularity between the early 1970s and early 1980s. Their primary use and justification was in the control and automated handling of pallets or tote pans of material. The goal was to minimize product damage, free floor space, control and track the inventory—protecting it from pilferage or unauthorized disbursement, and minimize the cost of material handling labor.

During the first wave of popularity, other logistics practices were causing a significant demand for storage capacity. For one, MRP (material requirements planning) systems were being introduced that tended to cause large quantities of material to flow into the organization because of errant use of EOQ (economic order quantity) procedures and faulty forecasting practices. Additionally, problems with the supply chain and the ability to track on-hand inventories caused managers to adopt overly conservative safety stock policies which also inflated inventory levels. It was absolutely unacceptable to stop production for any reason, let alone for the shortage of material.

Systems were large, and new systems were often simple extrapolations of requirements based on current inventory levels without first determining if the process would actually require the inventory levels this method projected.

Probably the most infamous story of a distribution warehouse plan that went wrong is a repair parts distribution center planned for a large heavy equipment manufacturer. That system was to have over 90 aisles of very tall and very long AS/R systems. It was planned and a contract was awarded, but the system was canceled before it was completed in the mid-1980s. Located adjacent to a major interstate highway, the structural steel for that facility stood for years. Like a skeleton, it silently reminded those that saw it to think about the entire process before staking your career on a plan that assumes the future is a simple factor of growth from where one stands today.

In the mid 1980s, an economic recession caused manufacturers and distributors to pull back plans for expansion and automation due to a shortage of capital. At that same time, the production philosophies of just-in-time (JIT) were being introduced to this country. Together, these two events led planners to consider the AS/R system technology a weapon of destruction—especially if deployed in their own companies. After all, it was a *storage* technology, and storage of material was to be avoided at all costs.

More on JIT later. But to summarize, the technology of AS/R systems grew rapidly up until these events, and then almost disappeared in the United States until the early 1990s. At one time the industry included nearly 20 companies providing equipment and automation that meets the classic definition of AS/R systems. Today less than a third of them remain, but the number of systems being planned and installed is at an all-time high, both in the United States and worldwide.

2.3 A STATE OF MIND

Perhaps the biggest reason for the decline of the industry is the fact that material handling and, more specifically, storage, have always been regarded as cost adders to the overall distribution process. As a cost factor, the limit of our interest has been to minimize the cost.

Aside from the fact that proximity can add value to material, most would respond that the best way to address material handling is to eliminate it. Since it cannot be eliminated in all cases, however, the next best thing is to design the systems for handling such that they are not dependent on scarce resources in order to function properly, and that they operate with total control and predictability. In other words—automate.

But automation costs money, and we have been inclined (or instructed) to not spend money on non-value-adding functions. So another way had to be found. We had little success eliminating these functions. We have even tried to pass on (outsource) the requirements to our suppliers, in the hope that they would, at least, make the problem go away.

The pressure to implement just-in-time manufacturing methods spawned a panic in the 1980s to reduce inventory below historical levels. We forced our suppliers to deliver just in time in the belief that reducing inventory was the key to cost reductions and increased control of the supply chain. The price we paid, however, was reduced reliability of supply, higher costs, and reduced quality.

One of the interesting "truths" to grow out of this era was the platitude: "... there are three attributes to every opportunity: Good, Fast, and Cheap. You can have any two of the three ..." (see Fig. 1). While few people realized that understanding this relationship was the beginning of true system-based reasoning, there were underlying causes for the presence of inventory that few people could see or address. They were narrowly focused on only one element of a properly designed logistics pipeline. They tried to fix the problem by changing the rules under which only a portion of the system operated—without re-engineering the entire system to behave in a way consistent with the new goals.

It is quite natural for the human mind to decompose problems into components. We are taught as beginners that we "eat an elephant—one bite at a time." The problem with this approach is that if the first bite does not disable the elephant, it will probably react in a violently defensive way.

Systems are no different. The difficulty with designing systems "one bite at a time" is that we often fail to see the impact a decision may have on other aspects of the system. As soon as a portion of the system is changed, it may start reacting in unpredictable ways. It is usually at this point that all improvement efforts take a back seat to the efforts of just trying to keep the system running and shipping product.

When components of the system are dealt with independently, we have very little success reassembling the components and making them work in concert with the rest of the process. It is much like trying to reassemble a broken mirror in order to see a true reflection. The result just never resembles reality [2].

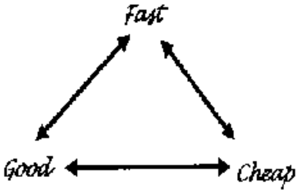


Figure 1 Conundrum of conflicting goals.

2.4 "INVENTORY HAPPENS"

Joking about inventory does not make its presence any less painful. The fact is, there are few warehouses in existence today that are not periodically bursting at the seams for lack of space to store more material. Even in today's environment where JIT rules, the stories of hidden inventory, and warehouses on wheels, abound. It is well known that left to the best systems available, inventory will expand to fill the space available.

In Japan, where we like to think the concept of JIT started, the first experiences were not the result of wanting to reduce the costs of holding an inventory. Just-in-time was developed out of the necessity to free up space for value-adding manufacturing. The result was near chaos, again, because of the lack of consideration for what the change to JIT did to the overall system.

While it used to be acceptable to ship 100 units of material on Monday for an entire week's supply, the new paradigm wants five shipments of 20 units delivered over the course of 5 days. This means that the ordering and delivery costs are factored up by 5, as are the number of trucks on the roads to complete these deliveries. In the beginning, Japan was plagued with a transportation infrastructure that could not handle the added traffic, and lateness and delivery failures abounded. The government even proclaimed that the reason no one is on time anymore is because of just-in-time.

In summary, most people simply tried to reduce inventory through edicts. The companies that have succeeded with JIT implementations, however, learned to use inventory as an asset, not as a waste element in their process. To achieve the goals of inventory reduction, however, they have turned to the root cause of inventory, and altered the process in ways that correspondingly reduce a smooth running process's need for inventory.

2.5 THE EQUATIONS OF INVENTORY

But using inventory to your advantage does not mean a wanton disregard for common sense or economics. Most manufacturing engineering curriculums courses taught in this country include a healthy dose of the operations research equations used to compute economic lot quantities for production and procurement. Known as ELQ or EOQ, these ancient equations are based on sound mathematics that are designed to maximize the probability of actually having material at the right place when it is needed, but absolutely minimizing the cost of material procurement, ownership, and control.

By and large, these equations have lost popularity because of misunderstanding. Many inventory planners view them as obsolete, or as inconsistent with modern logistics techniques. As we examine the math behind these equations, however, we find they are particularly useful in helping us define system-based plans.

To understand them, however, one must realize that the inventory a given system will require is totally a function of that system's theoretical utilization, the variability of the material supply, and the variability of the value-adding process itself.

As an extremely abstract way of illustrating this point, consider the simplest of queues, the *M/M/1*. This is a single-line queue ahead of a single server resource. The assumptions are that the arrival process is exponential, and that the traffic intensity (arrival rate/service rate) $\rho < 1$. In other words, the number of items arriving per period of time demanding service are always less than the capacity of the server to provide service.

If we only look at the work-in-process (WIP) buildup that can develop as a result of equipment utilization, the length of the line ahead of the server is estimated by the equation [3]

$$L_q = \rho^2/(1 - \rho)$$

The significance of this example is to show that as the process's utilization (i.e., the traffic intensity ρ) approaches 100%, the length of the queue waiting for service grows to an infinite length (see Fig. 2).

At first, it may not be clear how this can occur. It occurs because there is not enough capacity to accommodate surges. The actual utilization may be below 100%, but if the value-adding resource sits idle for

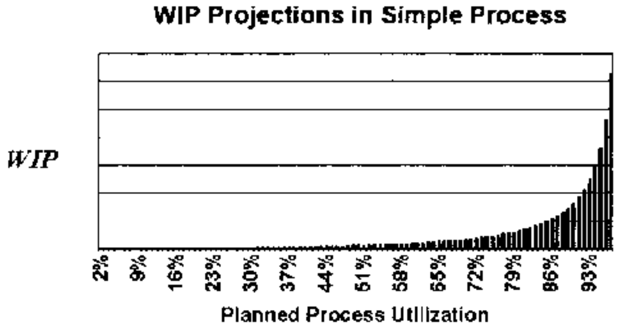


Figure 2 WIP as a function of system utilization.

lack of material, that idleness cannot be bought back. The capacity to add value during the idle period is lost forever.

Add to this the effects of variance associated with scheduling, material availability, and process downtime, and you begin to get the picture, WIP happens, even in the best of planned systems.

2.6 FUNDAMENTAL DIFFERENCE IN DESIGN PHILOSOPHIES

The primary focus of re-engineering the logistics supply chain has been too centered on cost reduction. In today's U.S. economy, many of the factors that led the Japanese manufacturer to embrace JIT and continuous flow technologies are affecting domestic manufacturers. In particular, labor shortages and space shortages are pushing logistics planners into a new philosophy of design that tends to favor a new look at automation.

In this excerpt from a JTEC report [4], the required change in design philosophy is summarized:

> In general, automating a task is a way to create labor by freeing people from non-value added work. While the U.S. views labor as a cost to be minimized, the Japanese seem to view labor as a resource to be optimized. The Unites States asks, "Given a specific task, what is the lowest annual cost for performing it?" The Japanese ask, "Given a fixed number of people, what is the most value I can add with the best assignment of skills?" The Japanese treat human capital the way we manage financial capital.

To this I would add an observation: we tend to design our systems to utilize our most valuable resource from the neck down. We rarely see systems that take advantage of the human ability to dynamically problem solve. If we disagree with this view, we should remember that it has only been a generation since we commonly referred to our employees as "hired hands."

In that same JTEC report, it explains that the Japanese are between 2 and 5 years ahead of the United States in deployment of automated technologies. This is not to say that the United States should run out and try to catch up. The pressures have been different. The Japanese have built and automated their systems based on known shortages of land, labor, and other resources.

With today's sub-4% U.S. unemployment levels, those "hands" are becoming scarce. Of the people available to work, many are better educated than in the past, and will not stay with a job that does not use their minds, as well as their backs. Additionally, the existing workforce is gettmg older, and the arriving replacements are better educated about ergonomic issues. Today's workforce is increasingly resistant to sacrificing their long-term well-being for the few premium dollars an ergonomically hazardous "hard-job" offers.

Finally, as a shortage, land may not seem to be a problem for most manufacturers. For planners that do not already have available space under roof, however, the capital to add brick and mortar is almost as unattainable in the United States as land is in Japan.

2.7 TECHNOLOGY SHIFT

In summary, AS/R systems technology is a tool that can automate some of the non-value-added tasks associated with material management, thus freeing scarce resources (humans) to be directed to value-adding tasks. In particular, it eliminates the linear handling of material from receiving to a storage location, the expediting function of finding the material and moving it to a point of use, and the process of accounting for, monitoring, and protecting material from unauthorized distribution.

As mentioned before, the technology took a severe hit in the 1980s during the time we were all trying to implement JIT, and survive an economic recession. But it was not a worldwide demise. While the United States saw application of the technology stagnate and over half its AS/R systems industry suppliers disappear, in the world market the technology found a new niche—speed.

The rest of the world was awakening to the need for continuous flow manufacturing, which yielded a demand for very responsive systems that could serve a variety of missions without significant reconfiguration. Part of the answer was inventory servers that could distribute much smaller quantities of material at very high transaction rates. This made possible the concept of flexible manufacturing where the order, and material requirements to satisfy the order, were conceivably known only a few minutes in advance.

Obviously, material stocks needed to be kept close at hand to supply the value-added process, but the supplies had to be efficiently handled, and quickly available to the point of use. To make this possible, extremely high transaction rates were demanded of any system so as to minimize the lead time between the customer's order and the final product delivery.

The initial problem that speed created was a decline in reliability. Just getting the vehicles up to speed created engineering challenges. Higher acceleration rates could not be achieved because of drive wheel slip, or load stability issues.

Since the technology was not expanding in this country, little was done to systematically address these problems. In the world market, however, the AS/R systems markets were booming, and money was poured into R&D to create new AS/R systems products made from exotic materials and utilizing unheard-of design concepts to minimize the mass of the vehicles used to move the loads.

Current technologies include crane-based AS/R systems that can reliably operate at speeds approaching 350 m/min with accelerations and decelerations exceeding 0.8 g. As recently as the early 1990s, sustained—reliable 350 m/min operation was only a dream, and the best acceleration sustainable without wheel-slip or load-skew was 0.1 gs.

Even more exotic are the ultrahigh-speed miniload and tote-load AS/R systems devices that can approach nearly 160 loads moved into and out of storage per aisle, per hour (see Fig. 3).

Perhaps the most dramatic difference is in overall system height and number of aisles used to create a single system. Through the end of the 1980s and into the early 1990s, systems built in this country tended to average more than six aisles of storage and were trending towards an average height of 65 ft. Some were in excess of 100 ft and supported their own building enclosure, making them a stand-alone facility for warehousing. In 1990, less than 300 AS/R systems machines were built and shipped in the United States. Around the world, however, over 4500 were shipped [5] (see Fig. 4).

Large and tall systems are still being installed today and still reach those heights, but a new trend has emerged that uses AS/R systems in a more distributed role within the flow process. The need to densify existing operations is creating demand for smaller, shorter systems that will fit in existing facilities.

Oftentimes, not having to penetrate a roof, or build a new building to house a tall system is advantageous because of local zoning laws. Major structural revisions to a part of a building often precipitate a requirement that the entire facility be brought up to current code and compliance standards. Constructing a machine in an existing building without modifying the building can often circumvent these regulations.

One such system in operation today is located between three different, but interdependent production groups. It receives, buffers, and distributes all material to and from these operations, and manages the inventory to maximize the nachine utilization between all three departments (see Fig. 5).

These differences indicate that the world is adopting a more dynamic approach to AS/R systems application—one of freeing the human worker to produce more value-added work, as opposed to a role of securely storing and retrieving material and reducing total manpower. Smaller, more efficient systems near the worker's point of use are being built. They are sized

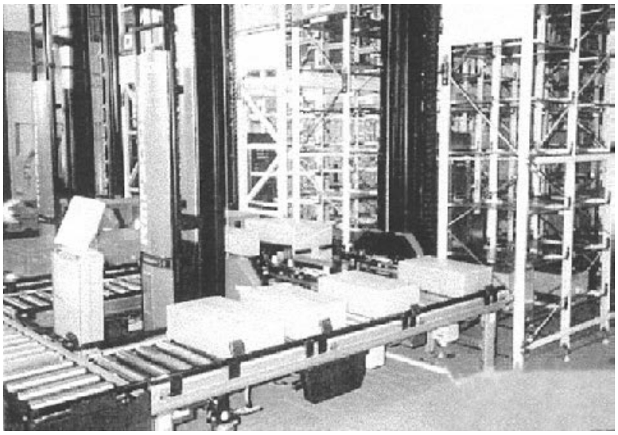


Figure 3 Mini-load AS/RS "front end."

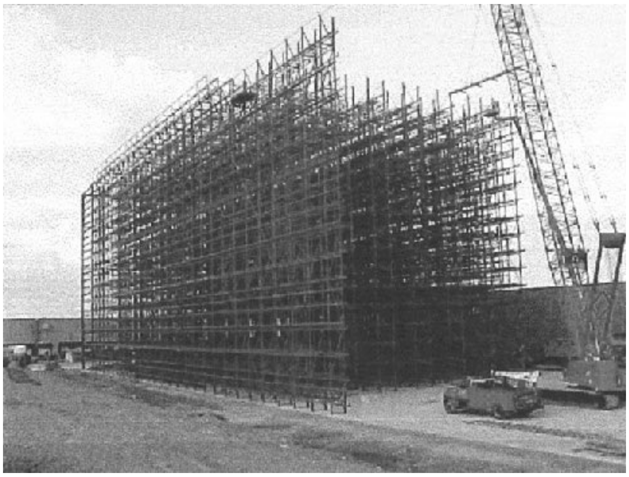


Figure 4 Rack supported building erection.

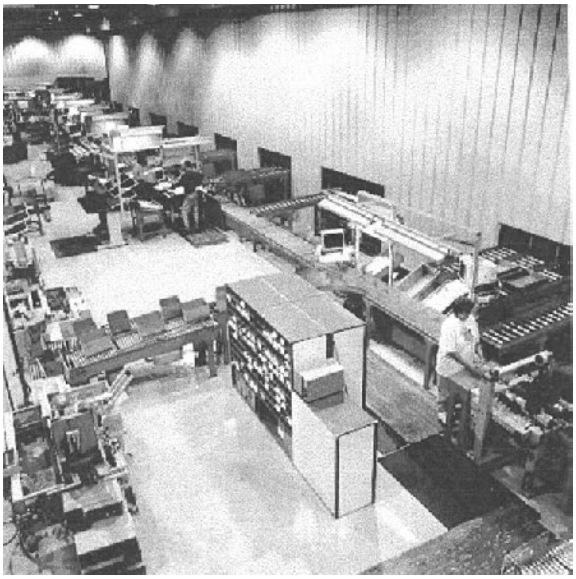


Figure 5 Work place process buffer.

as a function of the elements of the process at that point of use—rather than as a consolidation of requirements to be satisfied by some centralized warehouse remote from the user of the materials stored therein.

Much the same as distributed computing technologies have revolutionized the information systems industry, distributed warehousing is changing the way we manage our logistics and support our processes. As a final example, a fairly common application trick used outside of the United States is to locate the AS/R system along the column lines between production bays. This is traditionally difficult to use space because material tends to settle in, but not flow around the perimeters of our production bays. These systems are often applied with only one side to the rack, but deliver material along the production lines, direct to the points of use.

Looking at this type of application, it is apparent that value-added delivery (proximity value) to the points of use were the driver for the application, not efficient storage, centralized control, or labor reduction.

2.8 DESIGNING AN AS/R SYSTEM

To conceptually develop an AS/R systems solution today, one needs to start with the logistics process itself. The philosophy of continuous flow manufacturing needs to be extended to the warehouse and distribution center operations in order to avoid a mistake by oversizing the warehouse based on what is seen in the existing storage facility.

Once the total process is defined for all material flow, it can be examined for storage activities that might be eliminated through automation. Examples of these are the material co-ordination that goes on between two or more sets of internal customers and suppliers. If they each tend to have dynamic, but variable processes, WIP will develop between the operations. If you try to operate them in a pure "demand pull" constraint set idleness and lost capacity to add value can be experienced. These are the prime spots to evaluate the use of dynamic buffers, or AS/R systems.

Once these balances and penalties are documented, you then have the makings of a planned system process. Too often, automation of these "eddies" in the flow of material through the process have been ignored because of a perceived lack of justification for special non-value-adding equipment. In this type of example, however, eliminating the idleness will indirectly add capacity to the value-adding process.

2.8.1 Data Requirements

Assuming that you have an application for which an AS/R system might apply, there are several things to consider before beginning. Obviously, you will need to dimensionally define the load being handled, its weight, and the rate and activity patterns for the material arrival and distribution. This would include the normal quantity of receipts and disbursements per period of time.

Few planners have the luxury of all loads being the same exact size and weight, or arriving at the same rate. Therefore, the planner will need to prepare data tables showing the probable interarrival rate of loads to be stored:—by some classification system—either weight, size, supplier source, or some combination thereof.

Historically, one drawback to AS/R system application has been load size stability. If load sizes varied greatly, problems existed in economically handling the variety of loads in the system. Today, particularly with small carton and tote systems, technology has been developed that minimizes the impact of load size variety. Still, the planner needs to chart the size classes, and the percentage each class contributes to the planned inventory level.

Additionally, the designer must document the required output rate and responsiveness of the system. This is the maximum time allowed to retrieve any load in inventory and distribute it to a point of take-away or consumption. Exceeding this time will cause idleness in the value adding process. Again, this needs to be tabled in a classification that relates back to the inbound activities.

The planner needs to define the materials involved in this process—from a fire commodity standpoint. While some systems are built with the purchaser taking the responsibility for fire protection design, the racks will need to be designed to leave space for the level of protection the customer's insurance carrier will require. In most cases, the fire protection plumbing and equipment will be assembled with the rack before it is erected, making it a costly choice for the customer to add the system after the AS/R system is erected and commissioned.

Boundary conditions of the site are also necessary. While estimates can be used during initial planning, a formal civil engineering survey of the site and any encroaching features, complete with floor load capacity analysis will be required for the final construction engineering, and final firm price. At the very least, the footprint of the area in which the system could reside must be known from the beginning. Also, any vertical limits must be known, such as roof truss height, ductwork, piping, etc.

2.8.2 Flow Data

The warehouse must be part of an overall logistics plan. From this plan, flow rates to and from the warehouse should be available. Additionally, distances from the source and to the consuming points will help the planner optimize the horizontal transport systems used to get material to and from the warehouse (see Fig. 6).

When tabulating flow and responsiveness requirements, it is important to consider the effects of queuing. While planned queues seem to run contrary to the tenants of continuous flow, consumption point queues significantly impact overall warehouse design. They may increase the cost of the workstation by the addition of queuing stations or the requirement for some additional floorspace but the impact of the queue is to decrease the instantaneous response requirements of the AS/R systems. This can mean a significant cost tradeoff.

2.8.3 Activity Data

It is desirable to have the design team evaluate actual data tapes of transactions that occur in the process being considered for improvement with AS/R systems technology. In today's computer environment, activity records are often available for download to a PC data file, directly from an existing activity database. Oftentimes, analysis of this "real-world" data reveals trends and requirements that are overlooked by averages and derived statistical models.

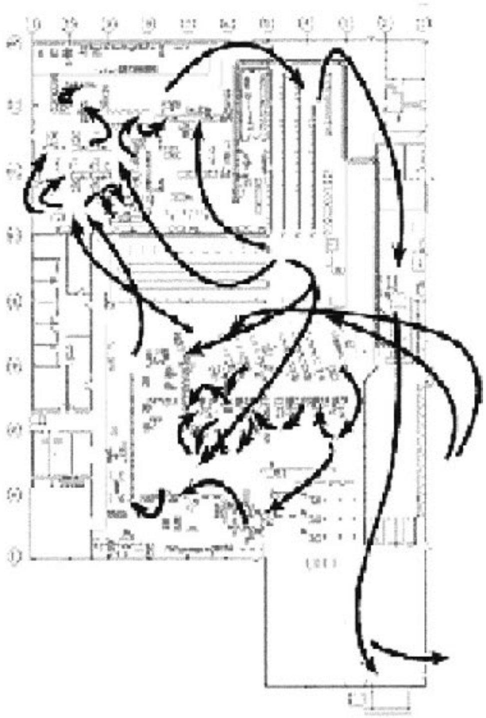


Figure 6 Process flow diagram.

2.9 TECHNOLOGY CLASSES

While certain types of carousel applications and vertical column extractor systems legitimately fit the definition, AS/R systems are normally thought of as having a static rack system on which the material or loads are stored, and a retrieval machine designed to pick and place the load to and from points of input and output. The terms unit load, miniload and micro/toteload are commonly used to verbally classify the systems (see Fig. 7).

2.9.1 Unit-Load Cranes

Typically available in world-class capacities of 1000 kg, 2000 kg, and 3000 kg, these devices have often been used to handle much heavier loads. The most common

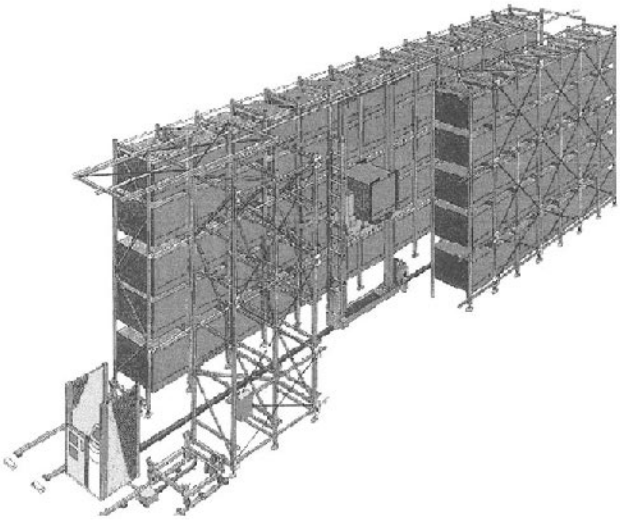


Figure 7 Typical unit load system.

size class is 1000 kg, even though most U.S. firms tend to believe their unit loads are approaching 2000 kg. The higher weight, in most cases, appears to be based on the heaviest possible load in the system, even though that load may only represent less than 5% of the total loads stored (see Fig. 8).

Obviously, the disparity between possible weight and actual weight needs to be reconciled before specifying the system's requirements. The cost of overspecifying the load will significantly influence the rack, crane, rail, and installation costs. A Pareto analysis should be conducted if only a few loads are causing the heightened weight specification. Many world class systems "cull out" those exception loads so as to reduce the cost of the overall system.

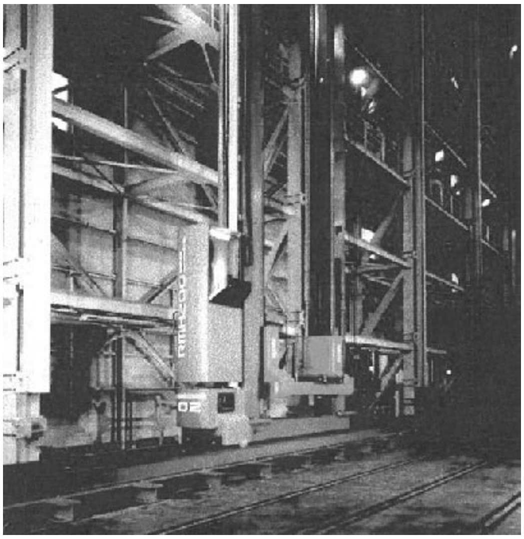


Figure 8 Unit load crane.

The most common unit load size being stored in today's system is the 42 in. × 48 in. pallet with a height of 48 in. including the pallet. But standardized technology exists to handle loads well beyond these sizes, as well as much smaller sizes.

The unit-load crane can also specified for heights up to and exceeding 100 ft. Except for those sites that are truly landlocked, the value of this height system is mostly exaggerated. When all costs are tallied, including the costs of high-altitude maintenance procedures, one must consider if a shorter or longer system would be more cost efficient.

Unit-load cranes typically use a shuttle fork to pick up and deposit the load from the input/output station to the rack storage location.

2.9.2 Miniload Cranes

These cranes are typically smaller versions of the unit-load cranes discussed above. They differ, however, in that they can use a variety of shuttle designs that effectively allow the more rapid transfer of loads from input/output station to the crane and from the crane to the rack.

Miniloads are generally considered to be for loads of 300 kg or less. In fact, the average weight specification for these systems is trending towards 100 kg. They usually store material in tote pans which are often divided, allowing the storage of multiple SKUs (stock keeping unit) per tote. Some manufacturers offer systems of higher weight capacity, but other factors seem to negate the efficiency of trying to handle a heavier load. Most "miniloads" are applied in order picking applications, and the ergonomics of dealing with heavy loads (which tend to be larger—precipitating longer reaches) tends to discourage this (see Fig. 9).

With all loads, especially those heavier than 300 kg, the structure of the tote or pan must be such that it does not sag or deform under maximum load pressures. When the pan is that sturdy, it is usually no longer easily handled manually (most miniload systems have operators moving totes in and out of circulation manually.) Additionally, the rack system must be much sturdier if the heavier loads are specified.

As mentioned, the miniload is typically applied to order picking situations. There is no "most common" size of pan. The pan is usually a direct function of an

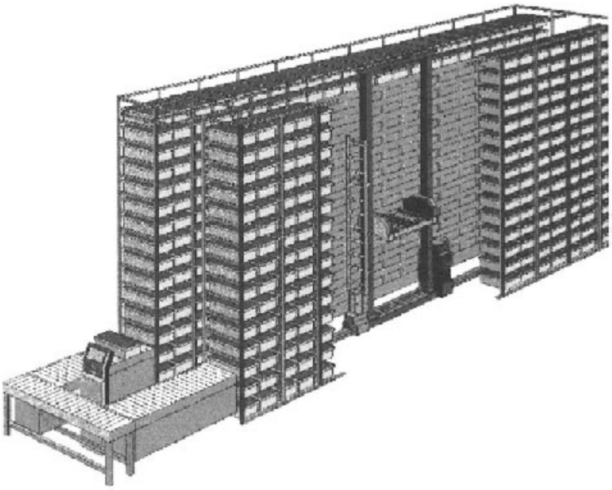


Figure 9 Typical mini-load system.

analysis of the material being stored, but the 18 in. × 26 in. tote pan seems to be used more often than not because of the ready availability of standard totes. Custom-sized totes are often desired, but the tooling and fabrication costs generally drive the design towards an already established standard size. Systems have been justified and built with totally custom totes, but remember that humans are usually picking out of the pan, as well as replenishing it. The presence of the human gives the system a great deal of flexibility to adapt that the mathematical evaluations often do not take into consideration when computing averages (see Fig. 10).

Tote pan height is also a consideration. Most totes for order picking miniloads are between 4 in. and 8 in. tall. Again, the primary consideration is the material being transported, but remember that a higher wall or a wider pan may create a poor ergonomic lifting and reaching ratio [6].

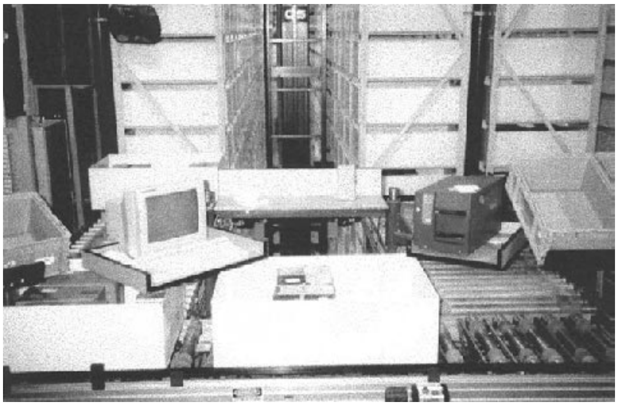


Figure 10 Typical mini-load picking station.

2.9.3 Rack-Supported Versus Free-Standing Buildings

Most miniload and smaller systems are free-standing structures built in existing facilities. Large unit load systems, however, are often used to support the building that encloses them. Since the rack is often heavier to support the loads being stored, it is usually a marginal additional cost to strengthen it enough to support the structural loads imposed by the building's roof and skin surfaces.

While tax incentives are not as great as in the past, a rack supported building may be partially treated as capital equipment—at least for that part of the building that is directly attached to the rack and part of the structural integrity of the system. This usually allows for accelerated depreciation methods that will help in system justification.

The popularity of rack-supported structures, however, is also related to the simplified construction and erection of the AS/R system itself, the speed with which an enclosed system can be constructed, and the fact that being more machine than building, building occupancy codes are often easier to comply with. An example of the latter is when a rack-supported structure is built in a zone where the area under the roof dictates rest room quantities, parking spaces, etc. Rack-supported warehouses are often excluded from plant space for the purposes of these calculations.

A hybrid of this concept is the partially rack-supported structure. This occurs when part or all of the system is to be installed in a building with insufficient ceiling clearance. The existing roof is often framed and opened and the new AS/R system in installed up through the hole. The rack protruding above the roof is then enclosed as a rack supported structure from the old roofline to the new top.

If the building already exists, then the freestanding AS/R system solution may be a natural choice. Ecient, cost-justifiable systems that are only 11.5 ft tall have been installed, and more are on the way. Again, this emphasizes the emerging practice of using the technology to free-up labor at the workplace for other value-adding functions, as opposed to centrally consolidating production stores.

2.9.4 Micro/Tote-Load Systems

As the name implies, these systems tend to handle much smaller loads than are handled in a miniload or unit-load system. Typically used with totes bearing one SKU or a single kit of material, the systems are frequently used in workplace management systems, box buffers, or "asynchronous process buffers" (see Fig. 11).

Utilizing extremely fast-response vehicles, these systems are often used to deliver materials through the racks to workstations located along the sides of systems. Additionally, some models of these devices, because of their exceptionally high throughput rates, are used as box or carton buffers between production and final distribution operations.

2.10 DYNAMIC BUFFERS

A buffer system is just as the name implies—a device with the ability to accommodate and level surges in activity. Like a balloon, the inventory level in a buffer will expand and contract to level the actual flow of material to the point of consumption. This is the primary reason AS/R systems are seeing so much popularity today.

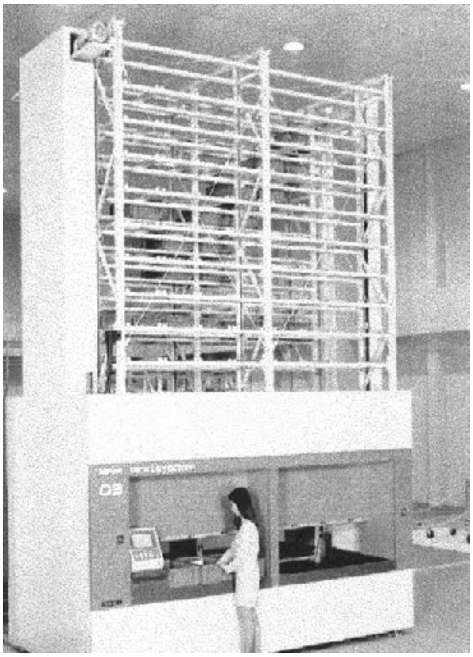


Figure 11 Micro/tote storage and picking buffer.

A lot of planners have attempted to apply double-ended AS/R systems as a means of not only buffering material, but as a way of transporting the loads across the plant, or from production to shipping.

Double-ended systems work, but are a very elegant (read: costly and difficult to optimize) solution. For one thing, unless the flow through the aisle is well balanced and bidirectional, the crane is limited to 50% productive utilization. For every load it delivers to the output station, it has to travel empty the length of the system to get a load to replace the load just distributed. Over a similarly planned single-ended system, the cranes will produce 30–40% more activity with no other changes than eliminating the opposite end transport stations.

2.10.1 Asynchronous Process Buffer Versus Storage System Paradigm

Asynchronous process buffers (APBs) are an increasingly popular way of simplifying the management of materials between interdependent operations. Usually consisting of a small, focused micro/tote-load AS/R system, the system usually uses side delivery stations to deliver materials to and receive product from production cells. The input and output of these cells are usually part of the normal finished product's production routing, and include extensive interdependence on the activity between cells to stay busy (see Fig. 12).

A significant distinguishing characteristic of these systems, however, is that the normal flow from cell 1 to cell 2 to cell 3, etc., is marked by the fact that the service rates in each cell are neither consistent between orders, or deterministically related to the preceding

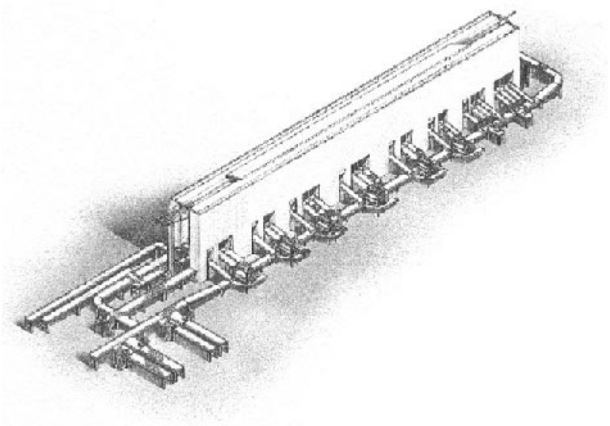


Figure 12 Asynchronous process buffer.

operation. The only consistency is that the material must follow the specific 1 > 2 > 3 routing. In these applications, the APB can not only handle the physical moves between cells, but can manage the storage of WIP that will develop between cells as a function of intercell variability.

In most APBs the use of closed system replenishment rules provides an automatic kanban that throttles the system from having a runaway cell. As a free side effect, however, these systems can be tuned by the addition of "free" totes (extra totes in the system for use between cells). These free totes provide some internal slack to the strict kanban control, allowing cells to operate more smoothly in the presence of brief interruptions in the planned continuous flow.

For example, one cell may produce a product that is placed in an empty tote and delivered to the next cell for the next process operation. To perform the first cell's function, it needs raw materials, and an empty tote in which to place the output to be transported to the next cell.

The second cell may remove the product from the tote, process it, and place it in a finished product tote for delivery to a packaging station for shipment. The empty tote created is then sent back to the first cell for replenishment.

Between each operation, the loads may need to be stored to prevent work buildup at the workstation that may make the station inefficient. Then, when it appears that the station will be able to accept the next load, the system needs to get it out to the cell before it is needed to prevent idleness.

The flow of product from cell 1 to cell 2 and so on, is balanced by the back flow of empties to the sending cells. If a backup stalls one of the cells, the backflow stops, which in turn, stops the forward flow of material. This provides for a self-metering system that needs little control logic to keep all cells operating in a balance with the total system's capacity. The ability of the system to keep running in lieu of single cell failures is then a function of the number of "free" totes held in the system between each cell.

2.10.2 Computing Cycle Times

The throughput, or cycle time of AS/R systems has been defined in numerous ways. There are techniques such as activity zoning to attempt to improve the overall efficiency of the device, but there are only a couple of industry benchmarks for computing cycle times.

The best way of analyzing the capacity of a proposed system is with a simulation of the system using actual data representing material arrivals and disbursements. In fact, the only way to analyze a side delivery system with multiple input and output stations is with a dynamic simulation.

An alternative manual method is to compute the probable time to complete each class of move that might be scheduled at each station, and then sum the probability weighted average time for each move based on expected activity. While this method does not always expose system interferences due to contention for resources caused by scheduling, it is a good first look at system capacity without the effort and expense of simulation.

For end-of-aisle systems (input and output occurs at one end of the AS/R system aisle) there are two methods that produce comparable results. The purpose of approximating cycle time is, of course, to provide a "first-pass" analysis of the adequacy of a design, and to allow a comparison of alternative solutions.

The first solution is based on recommended methods developed and published by the Material Handling Institute, Inc. (MHI) [7]. It refers to the calculation procedures to compute the single cycle and dual cycle moves typical of end of aisle systems (see Fig. 13).

The single cycle move is a complete cycle with the AS/R system machine in a home or P&D (pickup & deposit station) position, empty and idle. The single cycle time is measured by computing the time to move the crane to a rack location 75% of the length of the aisle away from the home position, and 75% of the height of the system above the first level of storage. In a 100-bay long, 12-tier-tall system, the crane would

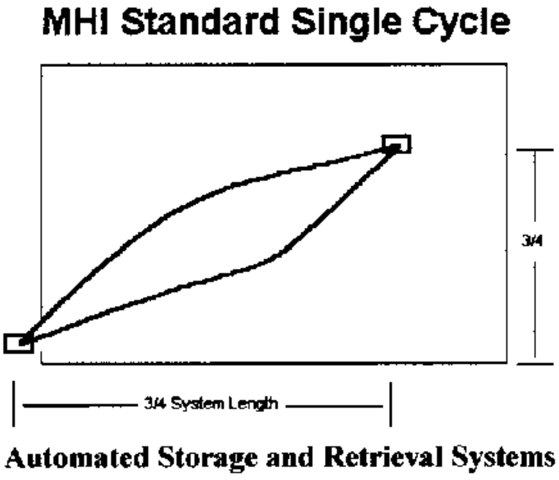

MHI Standard Single Cycle
3/4
3/4 System Length
Automated Storage and Retrieval Systems

Figure 13 Material handling institute AS/RS single cycle.

leave the home position, travel to the 75th bay and ninth tier. This is often referred to as the 75/75 position.

The total single cycle time is then computed as two times the time to make the 75/75 move, plus the time required to perform two complete shuttle moves. A shuttle move is the time required to extend the shuttle fork under the load, lift it off the rack, and then retract the shuttle with the load on board.

A caution in applying this algorithm: modern AS/R systems have the ability to control acceleration and vehicle speed as a function of whether the retriever is traveling empty or with load. Therefore, true cycle times for single or dual cycles must be computed based on the specific performance parameters of the product being analyzed.

The dual cycle, as defined by MHI is similar (see Fig. 14). The time is based on the crane starting empty at the home position. The cycle involves the crane picking up a load at the home $(0, 0)$ position, taking it and storing it in the 75/75 position. The crane then moves to the 50/50 position (50% of the length of the aisle, and 50% of the height of the aisle) to pick up a load. After picking it up, the crane then moves back to the home position and deposits the load picked up from the 50/50 position.

In summary, there are three crane moves and four shuttle moves making up the dual cycle.

There are no specified standards for the ratio of single to dual cycle commands performed by a given system. The use of input and output queuing conveyors can allow work to build up such that dual cycles are performed a majority of the time. Obviously, dual cycles are preferable to singles in that two loads are moved per three crane moves, but response requirements often result in a series of single cycle moves to process a sudden demand for output.

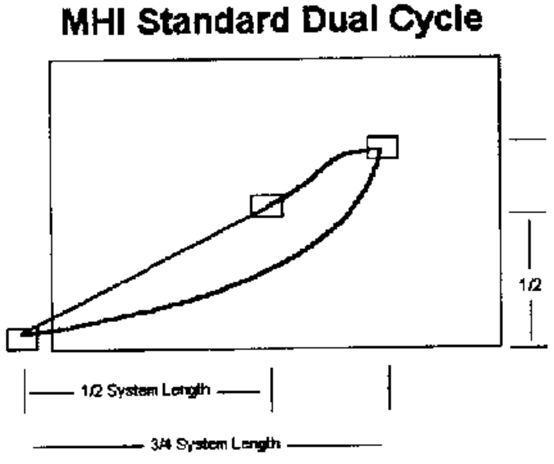


Figure 14 Material handling institute AS/RS dual cycle.

As a starting point, most planners will assume 30% of the moves will be single cycle moves, with the balance being duals.

Additionally, AS/R system performance is usually enhanced through the use of velocity zoning of the storage aisle. This is the practice of storing the fastest moving inventory nearest the input/output station at the end of the aisle. In practice, it is unusual for a Pareto effect to not be present in the inventory activity profile. This effect will significantly impact the overall requirements of the system design.

Using this rule of thumb to weight the single and dual cycle move times, the expected loads moved per hour (M) can be simply approximated as follows:

$$M = 3600/(0.30C_s + 0.70C_d)$$

where

$C_s =$ Seconds required to perform a single cycle move

$C_d =$ Seconds required to perform a dual cycle move

A second approach was more recently published that more directly approximates the cycle times for single and dual cycles of an end-of-aisle AS/R system. It takes into consideration the effects of randomized storage locations on cycle time and the probability of being commanded to store or retrieve to any location in the aisle [8]. It understates the overall capacity of a crane if the vehicle uses higher speeds and/or accelerations when moving in an unloaded condition. If used uniformly to analyze all options, however, it is useful for rough-cut analysis. These equations are

$$T_{SC} = T[1 + Q^2/3] + 2T_{p/d}$$

$$T_{DC} = [T/30][40 + 15Q^2 - Q^3] + 4T_{p/d}$$

where

$$T = \max(t_h, t_v)$$

$$Q = \min(t_h/t_v, t_v/t_h)$$

with

$T_{SC} =$ Single command cycle time

$T_{DC} =$ Dual command cycle time

$T_{p/d} =$ Time to perform a pick up or drop off shuttle move

$t_h =$ Time required to travel horizontally from the P/D station to the furthest location in the aisle

t_v = Time required to travel vertically from the P/D station to the furthest location in the aisle

Again, this provides a single cycle and dual cycle estimate, but makes no attempt to state how many loads will be moved by the system per hour. The planner must determine the mix of single to dual cycles. The starting point, in lieu of other factors is 30% single, 70% duals. A final rule of thumb for use in the feasibility stage of project design is to only apply equipment up to 80% of its theoretical capacity.

The important thing to remember is that all cycle time estimates are just that—estimates. The technique should be used to analyze the perceived efficiency of one concept or type of equipment over another. As long as the technique is used identically to compute throughput of all alternatives, it is an adequate tool to make a first comparison of alternatives. In all cases, however, mission-critical systems should be simulated and tested against real or expected transaction data to ascertain actual system capacity to handle activities in the real system.

2.10.3 System Justification Based on Flow Versus Static Costs

The rule of thumb is that if you put 15 engineers and accountants in a room, you will produce 347 different methods of computing the return on investment of a proposed project. The fact is: justification is simple. It is a function of the computed payback period, the capital available to fund the project, and the commitment of management that the process the system will support is a process that will support the vision of the company into the foreseeable future.

The only factor that the planner can deterministically project is the computed payback period. The balance of a payback analysis becomes subjective unless you realize that it is very difficult to justify any major material handling investment unless it is part of an overall process re-engineering effort.

There is a strong temptation to jump directly to an analysis of alternatives by reducing the cost of a warehouse system to the cost per storage location. Even if the expected costs of labor, utilities, and facility space are factored into the equation, this method will almost always push the planner to the sutoptimal solution that overly depends on manual (human) resources.

The inventory turns, and flexibility and responsiveness of the system, and the value adding capacity added by the system must be factored into the equation as well. Each of these factors must be approximated for each alternative at varying degrees of activity. And do not assume that each alternative has a linear response to increases in activity rates.

For example, it is common to see planners consider very narrow aisle (VNA) man-onboard order-picking systems technology over AS/R systems. At low rates, the cost per transaction is lower for VNA, primarily because the capacity of the AS/R system is available, but not being used.

As the activity rates approach the design capacity of the AS/R system, however, the cost per transaction of the VNA will actually increase and responsiveness decrease because of the activity induced congestion. (Remember the earlier reference to the attributes; good, fast, and cheap). Add to the reality of these systems the variability of nonautomated or semiautomated man-to-load systems, and it becomes clear why so many of these warehouses are not functioning today as they were envisioned when built only a few years ago.

The raw numbers (averages) may not clearly show the increased costs of VNA in this example. Only through complete system analysis can a correct decision be based, and this usually involves simulation. Simulation will not only help the planner understand the intrinsic behavior of the plan, but only through simulation will problems like gridlock be exposed that are not illustrated by the average throughput numbers often proposed in system concept summaries [9].

2.11 THE ROLE OF THE SUPPLIER IN PLANNING AN AS/R SYSTEM

As much as the role of AS/R system has changed in the way it is applied, the role of the AS/R system supplier has changed to that of a consultative partner in the project of determining the optimal system for the application. The reason for this is related to the earlier discussion about the ineffectiveness of trying to solve problems by breaking them apart into smaller subtasks and components. Asking a supplier to simply respond to concept specifications without having that supplier participate in the overall analysis of the logistics process will usually lead to a suboptimal concept proposal.

2.11.1 Objectivity of Solutions

There is still a belief that allowing the supplier in on the initial planning is a bit like letting the fox design the henhouse. In today's market, however, there is simply too much information being exchanged to seriously believe that a single supplier could substantially influence a project team to only consider one offering.

2.11.2 Real-Time Cost Analysis

There are multiple benefits from involving the supplier in the planning and analysis process. To begin, if the budget is known by everyone, the supplier, who works with the technology every day, is in a good position to keep the team on track by pointing out the cost impact of "features" that may not be economically feasible.

2.11.3 Use of Standardized Products

More specifically, the supplier will be in a role to help the team understand the application of the technology, including the use of standardized componentry designed to reduce the custom engineering costs of a new design.

Standardized products are often criticized as a supplier trying to hammer an old solution onto your problem. In fact, standardized products usually offer a wider set of standard functionality and variability than most custom engineered solutions. If the planner is able to use standardized solutions for the AS/R systems piece of the plan, substantial cost reductions can be realized in both engineering and total project cycle time.

Reduction in project cycle time is often an overlooked opportunity. If you consider that many projects are approved only if they pay for themselves in 30 months or less, a reduction in project implementation time of 3 months (over other alternatives) nets you a 10% savings by giving you the system sooner. The sooner you start using it, the sooner the returns from the investment start to come in.

2.11.4 Performance Analysis and Optimization

Another role of the supplier as a member of the team is the ability to use supplier-based simulation and analysis tools for rough-cut analysis and decision making. For example, a common assumption is that the fastest crane will make a system faster and more responsive. There is a tradeoff of cost for speed, but more specifically, there are system operational characteristics that will limit the ability to effectively use this speed. A person who does not work with the application of this technology on a regular basis will often miss the subtleties of these design limits.

In a recent analysis, one supplier offered an 800+ ft/min crane for use in an asynchronous process buffer. The crane could start from one end of the system, attain the top speed, slow down and accurately position itself at the end of the 130 ft long system. However, the average move under the actual design of the process was less than 18 ft, with an estimated standard deviation of less than 10 ft. This means that 97.7% of the moves will be less than 38 ft. The acceleration and deceleration rates were the same across all speed ranges, but the cost of the 800-fpm drive was wasted since the crane would only attain speeds of less than 350 ft/min on 98% of its moves. The cost difference between a 350 ft/min crane and an 800 ft/min crane will approach 21%.

2.12 CONCLUSION

The technology of AS/R systems has been reinvented in the last 10 years. As part of a strategically planned process, it can effectively serve to free up human resources to other value-adding operations.

The trend in application is towards smaller, more strategically focused systems that are located much closer to and integrated with the flow plan of specific processes. While large systems are still being designed and justified, these systems are less common that the small systems being installed within existing facilities without modification to the buildings (see Fig. 15).

The use of standardized system components has reduced the manufacturing and engineering costs of custom engineered, "one-off" designs, allowing planners a broader range of opportunity to use better, faster more reliable and productive equipment in the process of buffering the material flow.

To fully exploit the opportunity for improvement, the planner must evaluate the entire process before simply specifying a storage buffer. Use of the supplier

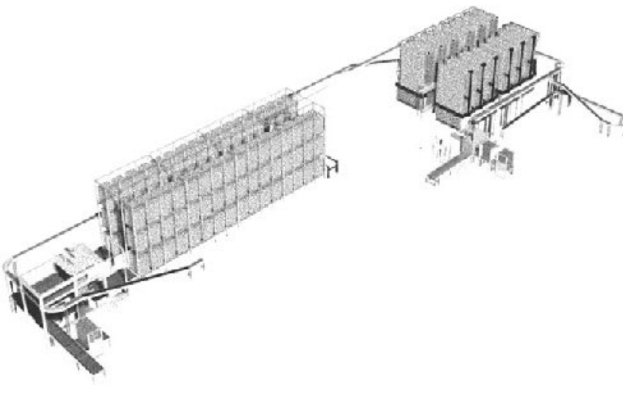


Figure 15

in the planning process will improve the quality of the recommendation for improvement, and will insure that solutions proposed are optimized, workable, and correct in terms of cost, schedule and overall system performance.

REFERENCES

1. Considerations for Planning and Installing an Automated Storage/Retrieval System. Pittsburgh, PA: Automated Storage/Retrieval Systems Product Section, Material Handling Institute, 1977.
2. PM Senge. The Fifth Discipline. New York: Currency Doubleday, 1990.
3. DT Phillips, A Ravindran, JJ Solberg. Operations Research Principles and Practice. New York: Wiley, 1976.
4. JM Apple Jr, EF Frazelle. JTEC Panel Report on Material Handling Technologies in Japan. Baltimore, MD: Loyola College in Maryland, 1993, p 29.
5. RE Ward, HA Zollinger. JTEC Panel Report on Material Handling Technologies in Japan. Baltimore, MD: Loyola College in Maryland, 1993, p 81.
6. Applications Manual for the Revised NIOSH Lifting Equation. Pub no 94-110, U.S. Department of Commerce—National Technical Information Service (NTIS), Springfield, VA, 1994.
7. JM Apple. Lesson Guide Outline on Material Handling Education. Pittsburgh, PA: Material Handling Institute, 1975.
8. JA Tompkins, JA White. Facilities Planning. New York: Wiley, 1984.
9. N Knill. Just-in-time replenishment. Mater Handling Eng. February: pp 42–45, 1994.

Chapter 7.3

Containerization

A. Kader Mazouz and C. P. Han
Florida Atlantic University, Boca Raton, Florida

This chapter reviews the design, transportation, and inventory of containers. Container design is a primary aspect of the handling and dispatching of containers. An efficient container design will keep adequately the quality of the product being carried. Two issues identified at the design stage are quality and economic issues. An offline quality control program will enhance the design and usage of the container. Section 3.1 of the chapter will focus on the design. In this situation we will provide guidelines to performing a design experiment on a dunnage, a plastic container mainly used in the automobile industry to transport parts. Similar approaches could be used design corrugated boxes or any other type of container. Section 3.2 focuses on statistical modeling of container inventory control in a distribution network. Example practical problems are included for an automobile maker and a fresh fruit company.

3.1 EXPERIMENTAL APPROACH TO CONTAINER DESIGN

First the issue of design of containers is addressed. The approach is developed to determine an optimal container design, to eventually realize a durable container. An analysis and development of a design experiment is performed to identify the major controllable variables to perform a statistical significance analysis on different containers. A container is modeled using finite-element techniques and tested to determine its durability under simulated conditions. A database is developed to help engineers to choose an optimal container design. The database includes the choice of structures, material process, wall thickness, shipping conditions, and any combinations of these. The method developed has been tested with different plastics using an illustrative example.

3.1.1 Introduction

With the increasing competition in industry more and more factories are taking a closer look at material handling for ways of cutting expenses. Container design, because it is only an auxiliary part of the product, has not received enough attention. Often containers are designed according to experience. As a result, the container is either too strong so that its life is much longer than the life of the product contained and therefore adding unnecessary cost, or too weak, causing product damage.

3.1.2 Procedure

Durability may be defined as a function of different variables. These variables may or may not have a great effect in the durability of the container. Once these variables are identified, a design of experiments is performed. A design experiment is a test or series of tests in which purposeful changes are made to the input for changes in the output response. To use the statistical approach in designing and analyzing experiments, an outline of a recommended procedure is described in the sections that follow.

3.1.3 Choice of Factors and Levels

Close attention must be paid in selecting the independent variables or factors to be varied in the experiment, the ranges over which these factors will be varied, and the specific levels at which runs will be made. Thought must also be given to how these factors are to be controlled at the desired values and how they are to be measured. Variables which have a major effect on the durability of the dunnage are the material, the process used to produce the dunnage, the nominal wall thickness, the load applied, and the ambient temperature. The first three are controllable variables and the other two are uncontrollable. The material may be limited to HDPE (high-density polyethylene), POM (acetal), or ABS (acrylonitrile butadiene styrene). Loads may be static to simulate the stacking of dunnages and impact loads or dynamic to simulate the transportation of parts via train, truck, or ship. Temperature conditions may be studied at −20°F, 68°F, and 100°F and the process reduced to four methods; vacuum, injection, rotational forming, and injection molding.

3.1.4 Choice of Experimental Design

The choice of design involves the consideration of sample size, the selection of a suitable run order for the experimental trials, and the determination of whether or not blocking or other randomization restrictions are involved. For this experiment it is known at the outset that some of the factors produce different responses. Consequently, it is of interest to identify which factors cause this difference and the magnitude of the response. For example, two production conditions A and B may be compared, A being the standard and B a more cost-effective alternative. The experimenter will be interested in demonstrating that there is no difference in strength between the two conditions. Factorial design can greatly reduce the number of experiments performed by looking at which combinations of factors have a greater effect in the durability of the dunnage.

3.1.5 Performing the Experiment

Using computer-aided design CAD and ANSYS (finite-element software) a model of the dunnage is constructed. The name finite element summarizes the basic concept of the method: the transformation of an engineering system with an infinite number of unknowns (the response at every location in a system) to one that has a finite number of unknowns related to each other by elements of finite size. The element is the critical part of the finite-element method. The element interconnects the degrees of freedom, establishing how they act together and how they respond to applied actions. A plastic quadrilateral shell may be used as an element. This element has six degrees of freedom at each node (translation and rotation), plasticity, creep, stress stiffening, and large defection capabilities.

Because of the incompleteness of current data in service life prediction, some tests are necessary to set up an engineering plastics durability database. A nondestructive experiment is performed on the dunnage. This experiment measured the deflection of the dunnage under different loading. The deflection is measured at several sections, in order to make sure that the model constructed on ANSYS correlates to the actual one. Theoretical results obtained from the computer model are used to verify the experimental results. Once the model in ANSYS is verified, the study under different loading conditions starts. Furthermore the ANSYS model can be brought to failure. Failure occurs when the stress level of the dunnage model is higher than the tensile yield stress. Stresses higher than this will cause permanent plastic deformation.

3.1.6 Data Analysis

Statistical methods provide guidelines as to the reliability and validity of results. Properly applied, statistical methods do not allow anything to be experimentally proven, but measure the likely error in a conclusion or attach a level of confidence to a statement. There are presently several excellent software packages with the capability to analyze data for the design of experiments. With the help of statistical data on the durability of a specific dunnage and the results of the ANSYS model, an optimal decision can be made regarding the durability of the dunnage.

3.1.7 Database

A database is used to generate the decision support system. A flowchart of the dunnage durability database is shown in Fig. 1. The user-friendly program guides the user where data needs to be input. Help menus are available at any instant of the program. The output comes in the form of a report that shows the durability of the dunnage under the specified con-

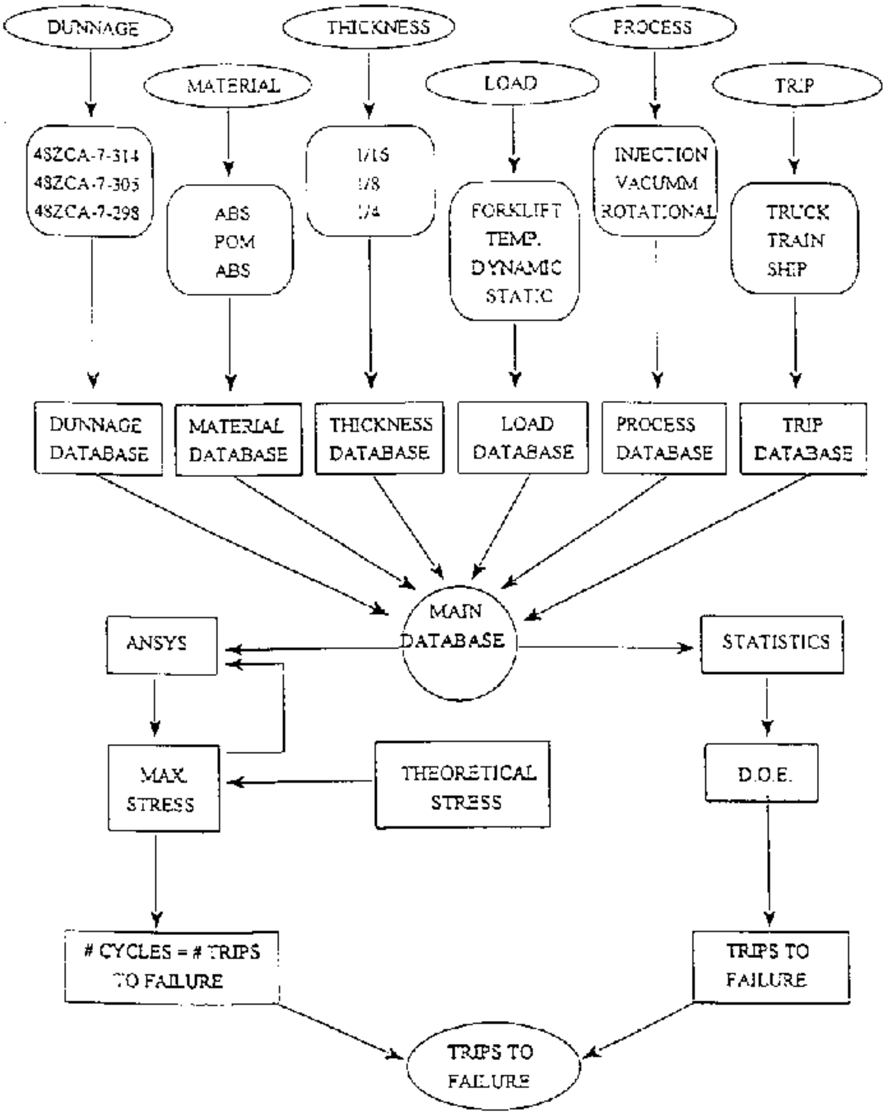


Figure 1 Dunnage durability solution flowchart.

ditions selected. The database includes the choice of structure, material, shipping condition, wall thickness, and the process used to manufacture the container or any combination of these choices. Using this database, the life of the dunnage can be predicted.

3.1.8 Dunnage Experiment

As a test base, dunnage was provided by The Ford Motor Co. Figure 2 shows a simplified CAD design of a typical dunnage. The dunnage geometry tends to be quite complicated, since the container has to hold automobile parts in place while withstanding two test, for impact and vibration, as shown in Fig. 3:

Vibration test: the container is placed in a motion table with a maximum displacement of 1 in. peak to peak, a frequency range of 1 to 4.5 Hz and an acceleration of 1.1 *g*. The motion of the table is circular for simulating train motion and angular for truck and transport.

Impact test: the container is placed on top of a sled with a 10° incline and it is let go to reach a velocity of 15 mph at the time it hits a wall at the bottom of the sled. Deceleration is almost instantaneous.

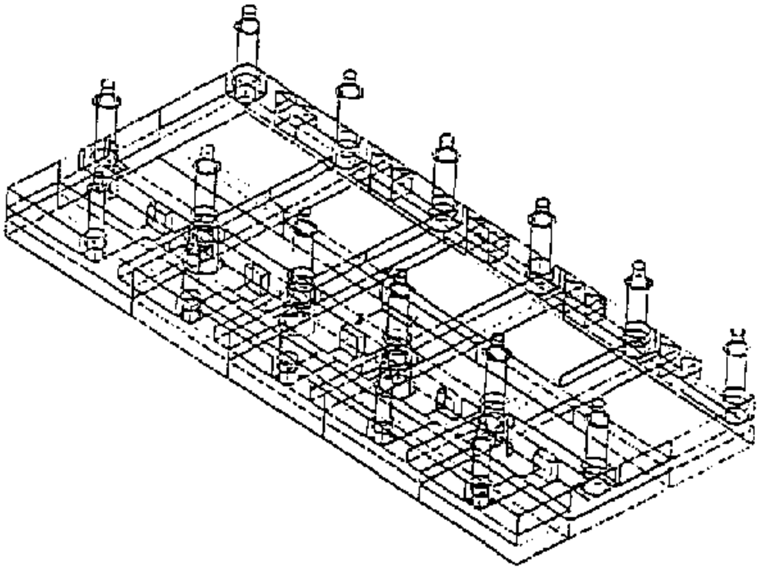


Figure 2 CAD drawing of a dunnage.

Factors and levels of study are shown in Table 1. Levels were set to cover a wide range of possible scenarios of what the dunnage may undergo. The result is a factorial system of 3^2 by 4^3. This means that two factors are at three levels and three factors area at four levels. A randomized factorial design was performed to obtain the set of experiments. Randomization is the corner stone underlying the use of statistical methods in experimental design. By randomization it is meant that both the allocation of the experimental material and the order in which the individual runs or trials of the experiment to the performed are randomly determined. By properly randomizing the experiment, the effects of extraneous factors that may be present are "averaged out." The randomized factorial design is shown in Table 2.

A small section of the dunnage meshed in ANSYS is shown in Fig. 4. The finite-element method solves for the degree-of freedom values only at the nodes so it will be convenient to increase the number of elements in the critical areas of the container. ANSYS will provide at each node information regarding deflection, stresses, and forces.

The ANSYS model was simplified to make it fail sooner than the actual container. After performing the nondestructive experiment, results were compared

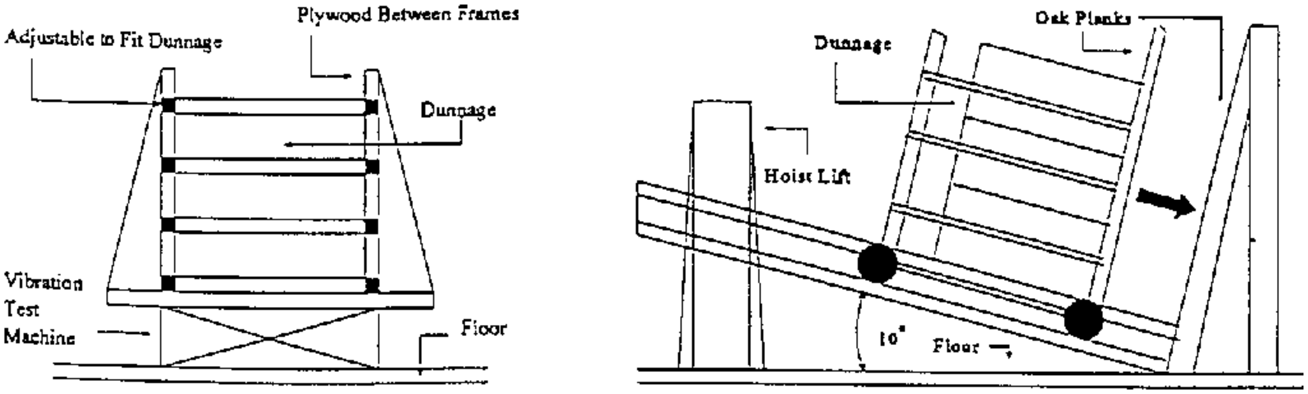


Figure 3 Vibration and impact test.

Table 1 $3^2 \times 4^3$ Design

Factors	Levels
Material	3
Wall thickness	4
Temperature	3
Static loading	4
Dynamic loading	4

Table 2 Randomized Factorial Design

Test	Mat.	Tem.	Thk.	Sta.	Dyn.
1st	1	1	1	1	1
2nd	3	2	1	2	2
3rd	2	2	1	3	3
4th	2	3	1	3	3
5th	2	2	2	1	2
6th	2	3	2	2	1
7th	3	1	2	3	4
8th	1	2	2	4	3
9th	3	3	3	1	3
10th	1	2	3	2	4
11th	2	2	3	3	1
12th	2	1	3	4	2
13th	2	2	4	1	4
14th	2	1	4	2	3
15th	1	3	4	3	2
16th	3	2	4	4	1

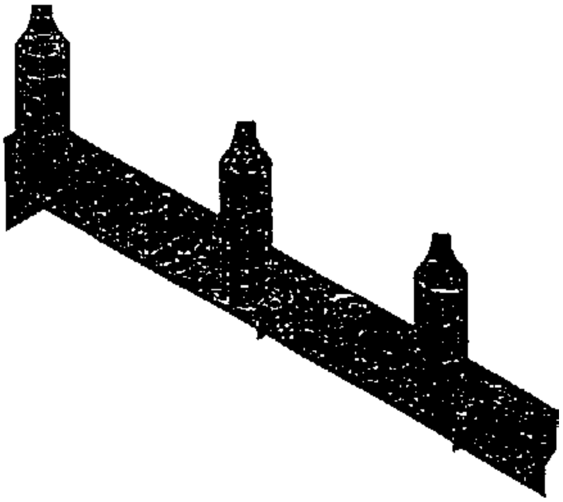


Figure 4 ANSYS finite-element model.

as shown in Fig. 5. The dotted line indicates the deflection obtained from the experiment, and the straight line was obtained from the ANSYS model.

3.1.9 Conclusion

A procedure to determine the durability of plastic containers has been shown. Using design experiments, an optimal container design can be obtained. The procedure is cost effective, and can be incorporated in the design of many different containers.

3.2 STATISTICAL MODELING OF CONTAINER INVENTORY CONTROL IN A DISTRIBUTION NETWORK

3.2.1 Introduction

In most material transportation processes, containers are used to facilitate transportation, and protect the material being carried from possible damage. To complete transportation of material or goods, one requires the availability of enough containers at the supply sites. Shortages of containers at a supply site will delay the transportation of the material.

Containers protect materials and facilitate handling. Containers are an integral part of the total transportation system. Less expensive materials such as paper, plastic bags, or carton boxes are used as containers to reduce cost. These containers are disposed of or recycled after transportation. With increasing freight volume and environmental concerns, the requirements for better quality and reusable containers are increasing rapidly. The cost of containers is becoming a significant part of the total transportation cost.

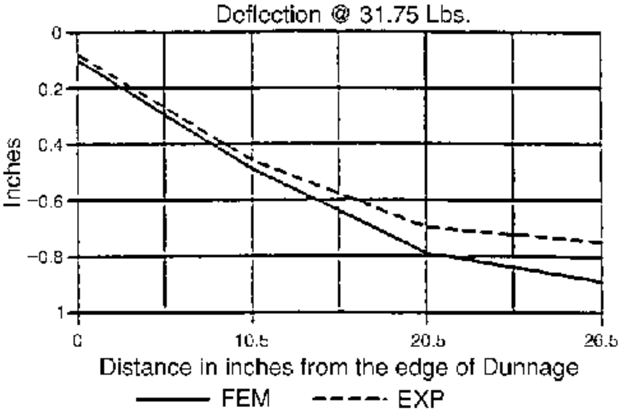


Figure 5 FEM versus experimental results.

A distribution network identifies a list of supply sites and destination sites connected by routes. When reusable containers are used in a distribution network, the containers are required to flow through road networks carrying the materials in demand. After transportation, the containers are not necessarily returned to the supply site. The containers can be sent directly to container inventories of the destination sites for future use.

A container inventory transportation network can be classified as either a *closed* system or an *open* system. The closed system is a network in which the total number of containers in the system does not change. The open system is a network in which the total number containers changes. A transportation network can also be classified as a balanced or unbalanced system. In a balanced system, the container inventory at each site is balanced, meaning that the number of containers shipped out by demand of a particular site is equal to the number of containers returned. The inventory level of containers remains unchanged at each site.

In an unbalanced system the inventory at some sites will keep increasing or decreasing. There are two reasons why a system can be unbalanced. One is the number of containers broken during usage. We have to add new containers into the system to compensate for broken containers. The other reason is that the demand shipment and the return of containers are not equal for some sites. After a period of time, these sites will have extra containers or will have a container shortage. If the system is a closed system, the total containers in the system will still be kept the same. Therefore, we can ship containers to the sites with container shortages from the sites with extra containers. The redistribution of the containers within such an unbalanced system to make the containers available at every site is essential to the performance of the whole system. Closed unbalanced transportation systems are the subject of this section.

When materials are transported between sites, the container inventory levels at each site will change. The container inventory control in a large transportation system is a type of network-location-allocation problem. The demand pattern of the containers is similar to the demand pattern of the materials. As with any of the other inventory items, container inventory also has its carrying cost, shortage cost, and replenishment cost. The container's carrying cost, shortage cost, and replenishment cost should be included into the total cost of the distribution network.

Obviously, if there are not enough containers in the network, it will cause transportation delays. However, using more containers than necessary results in higher initial investment and carrying costs. One of the fundamental problems of distribution network optimization is to know how many containers should be maintained in a particular system to make it efficient and economic. On the other hand, although there are sufficient containers in a system, if they are not located at proper sites, they are unavailable to the system at the moment when they are required. This will also cause transportation delays or give up optimal routes. An efficient way at reduce container inventory levels is to redistribute the empty containers to appropriate sites at appropriate times. The more frequently we redistribute empty containers, the lower the container inventory level that can be expected in the system. However, the cost for container transportation increases at the same time.

An additional focus is when and how to redistribute empty containers in the system to reach the lowest total cost. How to satisfy the requirement of transportation and maintain a minimum amount of container inventory are common issues in analyzing such a transportation system.

In this section we study the methods to minimize the total cost of a transportation distribution network. We use CIRBO as an acrony for Container Inventory contRol in a distriBution netwOrk.

3.2.2 Reusable Container Inventory Control in a Distribution Network

Reusable container inventory control in a distribution network presents the combination of the characteristics found in the transportation network system and the inventory control system. It deals with not only the inventory control but also the transportation systems management. In fact there are three major issues affecting the total cost considered here:

1. Optimal supply site selection for the commodity in demand
2. Control policy selection for the container inventory system
3. Optimal empty container redistribution method.

In most cases, the demand and transportation time are probabilistic. Issue 1 and issue 3 are transportation problems with probabilistic demands. Issue 2 is a special inventory control problem. If the system has infinite containers or if the containers are not used in the material transportation, this system becomes a pure transportation problem.

On the other hand, if the optimal routes have been selected for commodity shipment, the system degenerates into a problem of multiple inventory control and container redistribution in a distribution network. In this case the system performance is totally dependent on the inventory policy or the container management. Analyzing such a system will clearly demonstrate how container management affects the performance of a transportation system.

The framework of this section is to develop a simulation modeling procedure and address common problems of CIRBO systems. We first define the CIRBO problem and describe different inventory policies. Then, the simulation models for CIRBO are created using SIMAN© simulation language. A simulation code generator (SCG) system is then developed using SIMAN as a target program to systematically generate a CIRBO model based on a set of input conditions. The SCG itself is implemented by C + + language in an object-oriented window environment. The resultant framework is reusable, extendible and user friendly.

3.2.3 CIRBO Model Development

There are two steps in developing the CIRBO model. First, mathematical models are developed to describe the distribution network. Then a computer simulation code is generated. The mathematical models supply a theoretical foundation, while the simulation code creates a simulation model based on the user input specifications.

3.2.3.1 System Outline

Assume a typical transportation network with reusable containers which consists of *m* roads linking each site. Each site could be a commodity supply site and/or a commodity demand site. Each demand site can receive a commodity from multiple supply sites and each supply site can offer commodities to different demand sites. On each node, there can be a container inventory and commodity inventory, and it can also generate demand for commodities.

Each supply site contains both a commodity inventory and a reusable container inventory. The commodity is contained in reusable containers and then transported by some method (airplane, ship, truck, or train) among these sites.

When one site in the network requires materials, it looks for supply sites from all other sites in the transportation system. Some priorities for supply sites will be selected according to specific transportation rules. Here the rules should concern many features, such as transportation cost, material availability, container availability, material inventories, and container inventories for possible future demands, etc.

When the selected site has adequate commodity and containers available, the transportation takes place. However, if the commodity or container is not available at the selected site, the demand has to be sent to the secondary sites for supply. If, in some case, that demand cannot find adequate supply in the whole system, it causes an unsatisfied demand. A penalty will occur.

From the above statements, we can see that there are two main issues in the transportation network. They are commodity transportation and container management. In container management, the issues that need to be concerned are container inventory policies (when and how much of a replenishment should be made) and empty container redistribution (how a replenishment should be made). Actually, we can decompose the whole problem into three subissues:

1. Optimal schedule and route plan to minimize the total cost for commodity transportation
2. Optimal container inventory control policy to minimize the holding cost, shortage cost, and redistribution cost
3. Optimal redistribution route selection to minimize unit redistribution cost.

A network transportation problem can be studied in different ways. From the view of commodity demand and supply, it is basically a dynamic transportation problem. It mainly deals with the schedule and route problem of material transportation. The container availability and the container control policy can be handled as constraints for route and schedule optimization.

On the other hand, from the view of containers, the problem can be described as a multiple inventory control problem. The problem deals with the holding cost, the shortage cost, and the redistribution cost for the reusable container inventory in the system. The commodity transportation affects the container demand pattern, the lead time and the shortage cost of the container inventory. The redistribution of containers in a multiple inventory is another dynamic transportation problem. The cost of this transportation can be calculated and added to the total cost as replenishment cost. In this section, we discuss this problem from the view of containers.

3.2.3.2 Dynamic Transportation Models

If containers are not used, or there are infinite containers in each site, we never need to worry about container availability. Distribution networks with reusable containers become a pure dynamic transportation system. The issue becomes that for each moment, the flow of commodity from various sources to different destinations should be selected to minimize the total cost. The total cost consists of three parts: transportation cost, holding cost for commodity waiting in supply nodes, and penalty for unsatisfied demand.

3.2.3.3 Container Inventory System Analysis

There are two major issues in a transportation system with reusable containers. The first issue is to define how many containers should be invested in the system to make it economic and efficient. Another issue is to find the method to manage these containers to make them available when a supply site needs them. To highlight the effect of container and the effect of inventory policy, we assume that the optimal transportation route for commodity delivery has already been selected using some dynamic transportation solution method. If this optimal plan cannot be executed, the reason for that must be caused by the container shortages at some nodes. The difference between the optimal plan and suboptimal transportation plan is the effect of container availability.

3.2.3.4 Redistribution Modeling

In CIRBO the unit cost for replenishment depends on how the redistribution route is selected. Also a cost matrix form can be constructed. The issue is that we want to find the optimal transportation plan to satisfy the requirement of distribution and to minimize the redistribution cost.

3.2.3.5 Statistical Modeling and Optimization of the Container Inventory Control

Based on the mathematical models of the CIRBO system, the system performance measurement and various controllable variables can be identified. However, it is still very difficult to find the optimal solution using these models for such a complicated problem, especially when the system is a probabilistic system. A statistical systems modeling approach is therefore recommended as a tool to analyze such systems.

The first consideration in building a simulation model is to specify the goals or the purpose of the model. In the CIRBO system analysis, we can optimize the number of containers in the system by:

1. Minimizing the total cost, or
2. Reaching a specified service level, or
3. Reducing the time of redistribution of empty containers, etc.

Here, item 2 (service level) or item 3 (time of redistribution) can be the focus of study. However, they do not indicate the overall performance of the system. Take the service level as an example, in order to improve the service level, one of two methods can be used. The first one is to increase the number of containers in the system if the container carrying cost is small. The other method is to reduce the time period between the container redistribution if the redistribution cost is minimal. High service level is merely a measurement of the system performance. However, it makes no sense to seek high service levels without concerning the total cost of the system.

A statistical systems modeling method is used in this section. The key issue to make the simulation technology more acceptable is to make the simulation process significantly easier to learn and use. Here the simulation process includes not only the model building but also the experimental design and data analysis.

3.2.4 Case Studies

In this section, we present two case studies. One case study is performed for an automobile manufacturer and the another one is conducted for a fresh fruit company.

3.2.4.1 Modeling of a Transportation System for an Automobile Maker

Problem Description. The transmission and chassis division of an automobile manufacturer manages the transportation of a large number of automobile components and subassemblies. Reusable containers are employed in the component subassembly transportation system. One of these systems is the Mexico–Canada route. This route includes a main plant in the United States, denoted US, two plants in Mexico (MF1 and MF2) and another plant in Canada (CN). Car parts are shipped from US to MF1. After some part assembles are performed at MF1, containers are needed to ship these assembled parts to MF2. The extra empty containers will be shipped back to US.

More assembly work will take place at MF2. After that, they will be shipped to CN and then back to US using the amount of containers.

The demand from each plant and the average time the containers spend in each plant, and delays on the board of customs and on the road are listed in Table 3. The time spent for each period is a random variable, and these follow a normal distribution with the variance of 6 = 0.1 to 0.2 days. This system has a fixed schedule and transportation route. The plants usually work 5 days a week without holidays, and there are different holiday schedules in the United States, Canada and Mexico. During weekends and holidays, the plants only receive trucks but do not send any trucks out.

The automobile manufacturer is very interested in a decision support system that can study the effects of the number of containers in the transportation system. The ideal decision support system should represent the current transportation system and be able to stimulate several proposed changes. It should also be able to trace the availability of containers at a given moment in each plant. Different container management and optimization methods should be tested with various numbers of containers in the system.

This is a typical case of the CIRBO that has four sites with a fixed route and a fixed schedule. The demand size is also known. In this case, all the factors in the material transportation problem are fixed and given. We can concentrate on the container inventory control problem. The system's variables are the numbers of containers in the system and the period of redistribution.

Simulation Modeling and Optimization. Using the SCG for CIRBO, we can create a SIMAN model for the car manufacturer. In this case, the number of sites is four. Each site has a unique supply. If there are not enough containers available at the location when needed, the truck has to wait until containers become available. We give a very high penalty to the container shortage because the manufacturer does not want this to happen at any situation. The user can input initial amount of containers for each location, then run the simulation.

Using real demand data and assuring that there are 5000 containers in the system, the demand waiting time and container availability at each plant is collected. Figure 6 gives the average container availability for each plant over 5 years and Fig. 7 shows the average demand waiting time at each plant in the 5-year period. From Fig. 6 we see that most of the containers will be accumulated at MF1 while other plants have a container shortage. The demand waiting time in the United States and Canada will increase, while the time spent in the Mexico plant will decrease (see Fig. 7). There are two ways to avoid the accumulation of containers and elongated waiting time: one is to increase the container inventory and the other is to rearrange empty containers.

For the purpose of comparing, we assume that there is the same number of containers in the system, and we redistribute empty containers annually to make the container inventory level back to its optimum. Running simulation for the same period, we have the results shown that average container level keeping at

Table 3 Data Prepared for Automobile Maker Transportation Systems

	Time in Plant		Time on Road		Demand
	Mean	Deviation	Mean	Deviation	(Cont./day)
US	4.0	0.1			
US–MF1			4.5	0.2	101
MF1	3.0	0.1			
MF1–MF2			2.0	0.1	80
MF2	3.0	0.1			
MF2–CST			0.5	0.1	80
CST	2.0	0.1			
CST–CN			4.5	0.1	80
CN	3.0	0.1			
CN–US			2.0	0.1	80
MF1–CST	0.5	0.1	0.5	0.1	21
CST–US	6.5	0.2	4.5	0.2	21

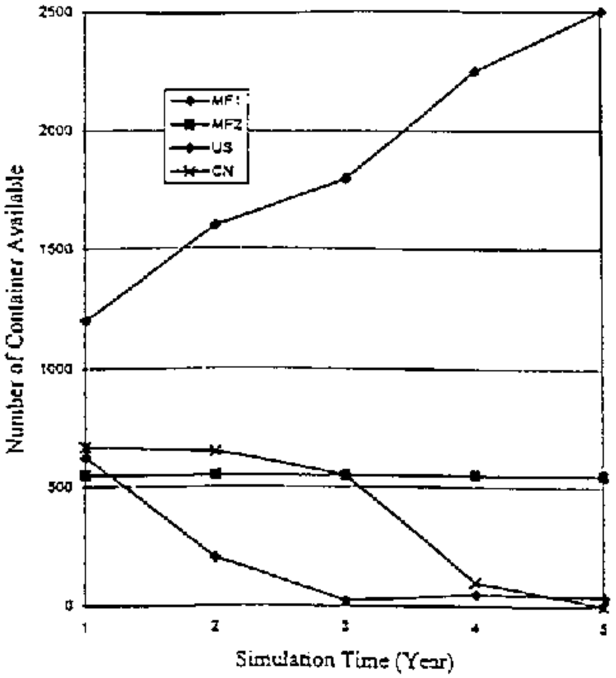


Figure 6 Average container availability for each plant during a 5-year period.

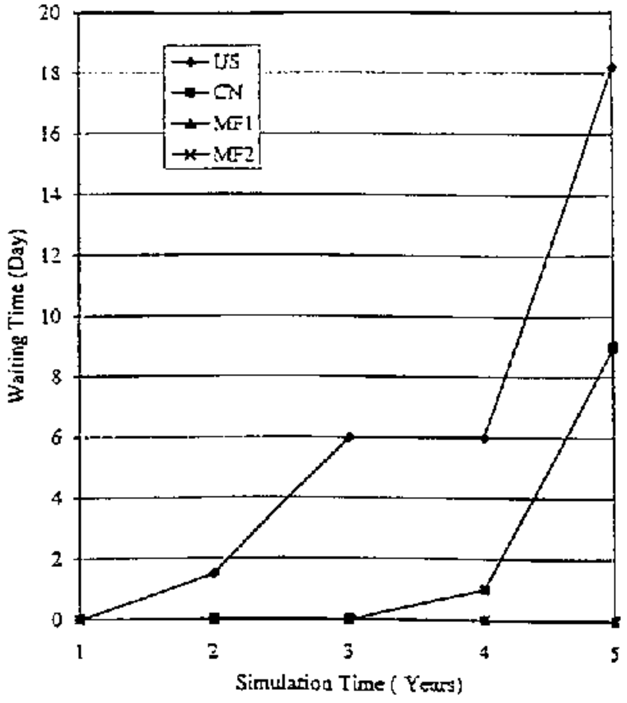


Figure 7 Average demand waiting time at each plant.

every plant is unchanged and the demand waiting time stays at zero.

Now, we go a step further to optimize the number of containers in the system. The possibility to reduce the total number of containers in the system is studied. By arranging simulation runs with different amounts of initial containers at each location, the demand waiting time can be compared. The results can be used to optimize the number of containers in the system as follows. We arrange 10 simulation runs, with each run having a different number of containers in the systems from 4000 to 3000. Assume that if there are not enough containers available when needed, the truck waits until containers become available. The container availability at each site and the truck waiting times are used as the service level measurements. To make the results reliable, simulation conditions are replicated in order to establish a required confidence level.

The results of the simulation output are presented in Fig. 8. This is the plot of average demand waiting time versus the number of containers in the system. From the curve we can see that when the containers in the system is reduced, the waiting time stays at zero until the containers in the system dropped to 3500. After that a very short waiting time, 0.05 day (1.2 hr) occurred. If the containers in the system are reduced to be less than 3000, the waiting time increases rapidly. It was obvious that the containers should not be less than 3000. A better service level could be achieved with at least 3500 containers in the system. However, keeping more than 3500 containers in the system does not further improve the service level but will only increase the inventory.

Summary. SCG for CIRBO has created a simulation model for an automobile maker successfully. This model can be used to analyze the performance of container inventory control in the transportation system.

1. An annual empty container redistribution is suggested to improve system performance.
2. Using the simulation model generated by SCG, we optimize the number of containers in the system. The model reduced the total number of containers by 30% and still maintained a high service level.

3.2.4.2 Modeling for a Fresh Fruit Company

Problem Description. Bananas and other fresh fruit are imported into the United States from a port in Latin America. The fruit is shipped on boats in

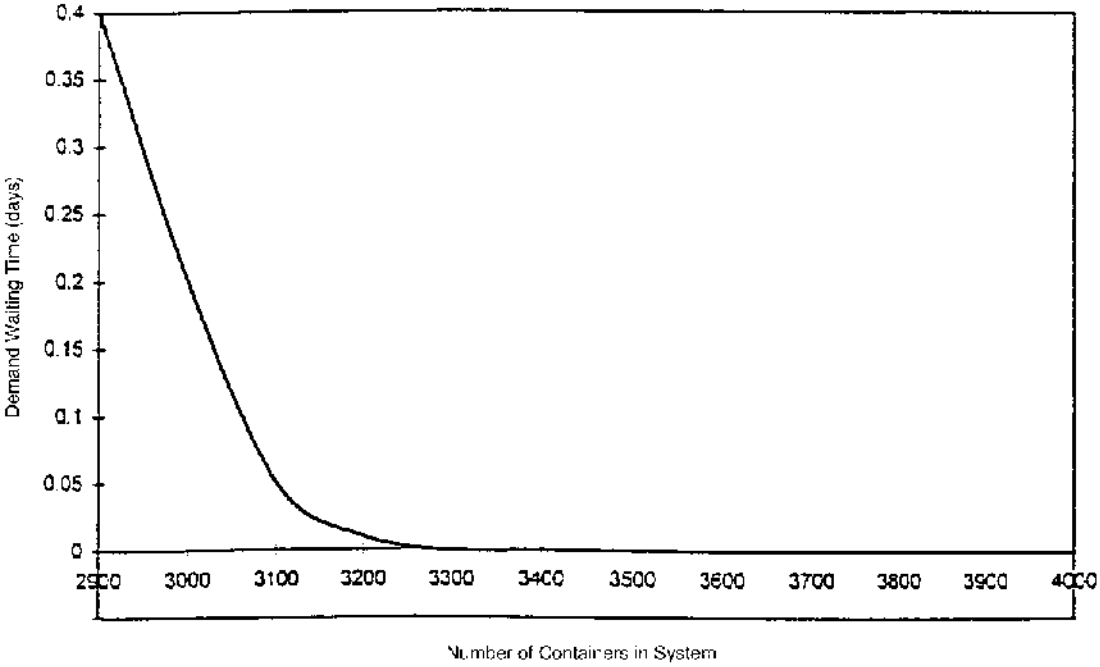


Figure 8 Optimize the number of containers in system.

marine-size shipping containers, and comes into a port in the Gulf of Mexico. Upon arrival the containers are distributed from the port to customer locations throughout the central part of the country.

There is an inherent problem in this fruit distribution system; the trade is unidirectional. The trade imbalance between the United States and those locations from which the bananas come makes shipping in both directions impracticable. Full containers are imported from the source and empty containers must be exported to replenish the container inventory. For the system to be operated efficiently, the boats returning to Latin America must return fully loaded with empty containers. An economical method is needed for keeping the number of containers in the Latin American port at a level high enough to ensure that the boats leaving for the United States will be fully loaded.

This dependence on return shipment of containers means that a stable inventory of empty containers has to be kept at the U.S. port when the ship arrives. Unfortunately the U.S. side of the distribution system has a large amount of variability associated with it. Many factors effect the amount of time when a container leaves and returns to port as outlined below:

1. The distance from the port to the customer's location
2. The amount of time that the customer keeps the container before returning it
3. The speed variability of the trucks and the ships that deliver the containers
4. The day of the week that the container leaves and returns to the port.

Currently, a high-level buffer inventory is required to overcome this variability so that any shortages of empty containers can be made up with empty containers from the buffer inventory. The size of buffer inventory is approximately one-half the capacity of a ship used in the system.

Objectives. The cost of owning and operating this fruit distribution system is tremendous. Each of the shipping containers costs approximately $20,000. Associated with each of the shipping containers is a refrigeration unit that costs approximately $7000–$10,000. In order for the refrigeration unit to operate there must be a generator to power it while it is in port. These cost approximately $5000 dollars per container. Lastly, for the containers to be moved there must be enough trailers. Trailers cost approximately $15,000 dollars each. The two container ships cost between

20 and 40 million dollars each. This brings the total equipment cost required to run the small system to the neighborhood of 70 to 80 million dollars.

The area targeted for cost reduction is the excess inventory of containers at the U.S. port. If the number of containers maintained in the buffer inventory could be safely lowered by 10 containers, the company would save approximately $350,000. It also saves the cost of maintaining those containers and the associate equipment over the life of the container.

On the other hand, with an investment of this size the system should look for maximum return on investment. To maximize the return in such a system, the system must be operated as efficiently as possible. Consider that a sufficient buffer inventory of empty containers in the U.S. port will be used to ensure against any possible loss of ship capacity. Current practice is to keep an excessively large buffer in container inventory at the U.S. port so the ships can be loaded efficiently.

This is a closed-loop system. If a company owns all the containers, there is no container replenishment in the system. The carrying cost and shortage cost are subject to control and are balanced. One of the policies is that container shortage is not allowed. The problem becomes that the company has to increase the number of containers and carrying cost.

Another method is to use a leasing program to reduce the number of containers the company owns, and leased containers are used to meet peak demands. This is another typical inventory control problem. The total cost consists of the following:

1. Carrying cost: the cost of investment in container inventories, of storage, of handling containers in storage, etc.
2. Shortage cost: the cost of lost ship capacity
3. Replenishment cost: the cost of leasing containers.

These three costs are subject to control. Thus the goal should be to optimize the total cost in such a way that the ships are filled to capacity. The shortage cost will always be less than the cost reduction of carrying cost and replenishment cost.

Simulation Modeling. To find the optimization solution, a simulation model has been constructed. The model uses two ships to simulate the transportation process and a network to simulate the distribution system in the United States. In order to approximate the actual system as closely as possible the original model had the following characteristics and capabilities:

1. Two ships, each with a capacity of 100 containers, were used to move containers between two ports. The ports were assumed to be 1500 miles apart and the ships operated at a variable speed. However, they work directly opposite each other so that the two ships never arrived at he same port at the same time.
2. The U.S. port was open for trucking 5 days a week, but the ships operate 7 days a week. Thus if a customer ordered a container of fruit and requested that it be delivered by a specific time, the delivery time was estimated. If the optimal departure time for the truck was to be a Saturday or a Sunday, the truck was forced to leave on Friday.
3. If a ship was to fully load on a weekend it would wait till the following Monday to allow trucks that had returned over the weekend to load their containers on the ship.
4. The speed of the trucks used to deliver the containers varied slightly with a normal distribution around 55 mph.
5. The amount of time that the trucker was allowed to hold on to the container before returning it was modeled with a normal distribution with mean based on the distance from the port.
6. The model can accept any kind of demand pattern. The information used for demand was a hypothetical demand as a function of distance from the port. This model can also use history data for the future forecast.

Control Policy 1: Company Owns All Containers. When the company owns all the containers, no leasing containers are added to the system. The reusable containers will remain unchanged in the system while the container inventory at the U.S. port will fluctuate (see Fig. 9).

In cargo shipping the shortage cost of not having enough containers is significant compared with the container carrying cost. This requires that a ship be fully loaded when it leaves the port. The only way to ensure that is to increase the containers in the system (in the U.S. port as buffer inventories).

Control Policy 2: Leasing Program to Reduce Buffer Inventory at the U.S. Port. When a leasing program is employed, the total containers in the system will change due to the leasing of containers. The inventory fluctuation is depicted in Fig. 10. Shortages are covered by leasing containers.

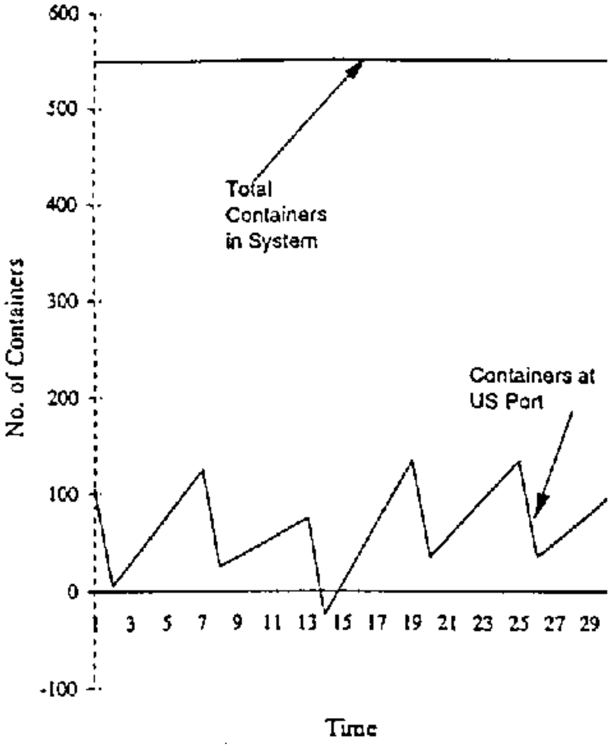


Figure 9 The inventory fluctuations with company's own program.

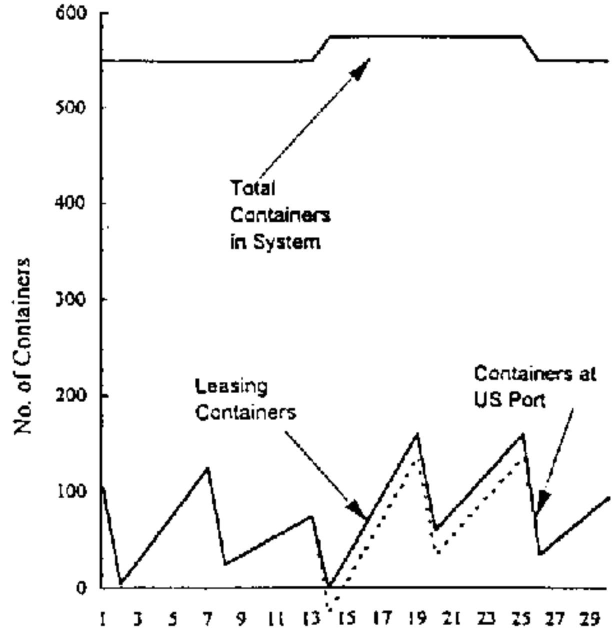


Figure 10 The inventory fluctuations with leasing program.

The container carrying and leasing cost is subject to control. Reducing the buffer inventory at the U.S. port means increasing the temporary leasing of containers. The amount of company owned containers and the leased containers are the decision variables. It depends on the ratio of carrying cost to leasing cost. Designing the simulation experiment in a way such that each simulation run will have different containers in the system and will lease containers when a shortage occurs, the final solution is attained by balancing the level of containers owned by the company and the level of leased containers based on the carrying cost and container leasing cost.

Summary. In this section the transportation problem of an inland–overseas fresh fruit operation is simulated. It is another example of the CIRBO. Simulation is successfully used to optimize the refrigerated container inventory at the U.S. port if the company owns all the containers. The leasing program is also introduced to make it possible to further reduce the buffer container level. It is suggested that the levels of container modeling can be reduced by 10–30% without impacting the delivery time (source level).

3.2.5 Conclusion

The objective of this section was to find methods to study reusable container inventory control in a distribution network. This system has many controllable variables and control policies. The container inventory control problem is a multivariable optimization problem. CIRBO is a special inventory control problem. To optimize a CIRBO system, mathematical models are presented to represent the CIRBO system in general. Various control policies are defined. Based on the mathematical description of the CIRBO system, simulation models can be constructed.

Theoretically, simulation is not an optimization tool. The simulation modeling of a real CIRBO system is not able to optimize CIRBO directly. However, to optimize a CIRBO, design of experiment and optimization theory is utilized. Simulation experiments are arranged using a design of experiment approach. Running simulation in this way, we can compare simulation results to find optimization solution.

Simulation code generation is introduced to make the simulation modeling process simple for nonexpert simulation code generators. It also saves time for simulation program development.

ACKNOWLEDGMENTS

The authors would like to acknowledge the Material Handling Research Center at Florida Atlantic University, The National Science Foundation, and the Ford Motor Company for supporting this study. And also acknowledge the work and assistance done by the following students: P. P. Aguilera, Weiming Feng and Pankaj Kanwar.

BIBLIOGRAPHY

KS Akbay. Using simulation optimization to find the best solution. IIE Solut May: 24–27, 1996.

ANSYS Manual Revision 4.3. Swanson Analysis Systems, Inc., Feb 15, 1994.

CB Basnet, SC Karacal. Experiences in developing an object-oriented modeling environment for manufacturing system. Proceedings of the 1990 Winter Simulation Conference, 1990, pp 477–481.

M Bogataj, L Bogataj. Inventory systems optimization for dynamic stochastic and periodical demand. Eng Costs Prod Econ 19(1–3): 295–299, 1990.

Bonelli P, Parodi A. An efficient classifier system and its experimental comparison with two representative learning methods on three medical domains. Proceedings of the Fourth International Conference on Genetic Algorithm. R Belew, LB Booker, eds. 1991, pp 288–296.

MD Byrne. Multi-item production lot sizing using a search simulation approach. Eng Costs Prod Econ 19(1–3): 307–311, 1990.

M Chen, WP Chen, DC Gong, M. Goetschalckx, L McGinnis. An AGV simulation code generator. Proceedings of Material Handling Research Center at Georgia Tech, Nov 1991.

C Das, SK Goyal. Economic ordering policy for deterministic two-echelon distribution systems. Eng Costs Prod Econ 21(3): 227–231, 1991.

N Erkip, WH Hausman, S Nahmias. Optimal centralized ordering policies in multiechelon inventory systems with correlated demands. Manag Sci 36(3): 381–392, 1990.

M Goetschalckx. Local User's Manual. Material Handling Research Center, GIT, Atlanta, GA, 1991.

JJ Gregenstette, C Ramsey, A Schultz. Learning sequential decision rules using simulation models and competition. Mach Learn J 5: 1990, 335–381.

Hutchison, et al. Scheduling approaches for random job shop flexible manufacturing systems. Int J Prod Res 29(5): 1053–1067, 1991.

RG Lavery. A simulation analysis of the effects of transportation system parameters on inventory levels. Proceedings of 90 Winter Simulation Conference, IEEE Service Center, Piscataway, NJ, 1990, pp 908–910.

CJ Liao, CH Shyu. Stochastic inventory model with controllable lead time. Int J Syst Sci 22(11): 2347–2354, 1991.

GE Liepins, AW Lori. Classifier system learning of Boolean concepts. Proceedings of the Fourth International Conference on Genetic Algorithms, R Belew, LB Booker, eds, 1991.

M Montazeri, LN Van Wassenhive. Analysis of scheduling rules for an FMS. Int J Prod Res 28(4): 785–802, 1990.

DC Montgomery. Design and Analysis of Experiments. 4th ed. New York: John Wiley, 1996.

CD Pegden, RE Shanon, RP Sadowski. Introduction to Simulation Using SIMAN. 2nd ed. McGraw-Hill, 1995.

D Porcaro. Simulation Modeling and DOE. IIE Solut September: 23–25, 1996.

R Riolo. Modeling simple human category learning with classifier system. Proceedings of the Fourth International Conference on Genetic Algorithms, R Belew, LB Booker, eds, 1991.

LW Robinson. Optimal and approximate policies in multiperiod, multiplication inventory models with transshipments. Operat Res 38(2): 278–295, 1990.

SM Semenov. Determination of prior probabilities in entropy models of a transportation system. Autom Remote Control 50(10): 1408–1413, 1990.

T Shimada, Yamasaki, Ichimori. Introduction of Packaging Design CAD System. Nippoh, 1990.

Chapter 7.4

Robotic Palletizing of Fixed- and Variable-Size/Content Parcels

Hyder Nihal Agha and William H. DeCamp
Motoman, Inc., West Carrollton, Ohio

Richard L. Shell and Ernest L. Hall
University of Cincinnati, Cincinnati, Ohio

4.1 INTRODUCTION

Warehousing is an expensive activity in the United States, where it accounts for nearly 5% of the Gross Domestic Product [1]. It can best be described as the material handling functions of receiving, storing, and issuing of finished goods. It is often viewed in industry as a necessary evil, since it does not add value to a product. However, the warehousing and distribution functions are critical to a successful manufacturing enterprise. Warehousing functions include information processing, receiving, storage, order picking, palletization, and shipping. The typical process for material handling in a warehouse is as follows:

1. Items are received at a warehouse in multiple pallet loads of identical items.
2. Loads are stored in the warehouse in some planned configuration.
3. When a customer's order arrives, an order picker goes through the warehouse to pick the desired items from separate pallets.
4. Items are routed to a load forming, palletizing, or palletization, station where items of various sizes and shapes are placed together on pallets for shipment to the customer. Although this palletizing operation has traditionally depended upon human labor, recent efforts at automating the palletization of parcels of mixed size and shape have proven very successful.

There are several disadvantages to human palletizing. One is related to cost. Even the most motivated and capable human can stack only about six parcels per minute, i.e., one parcel per 10 sec. Another disadvantage is related to safety and workers' compensation costs. A human who performs such a repetitive motion is at risk for cumulative trauma disorders, such as back and shoulder injuries. A typical human palletizer is shown in Fig. 1.

The advantages of robotic palletizing include: the maximization of the usage of the pallet cube; the retention of knowledge about each parcel throughout the distribution system; increased pallet load stability, insurance of forming pallets in accordance with regulations (i.e., not stacking poisons on top of food items, and control of parcel fragility, which reduces waste. Distribution centers are a necessary component in the logistics system of most manufacturing industries from food items, to dry goods, to computer or aircraft engine components or machine tool parts. All distributors, including the defense industries, parcel industries, and even medical industries, are potential users of a robotic palletizing system.

Palletizing may be defined as arranging products to form a unit load for convenient subsequent handling.

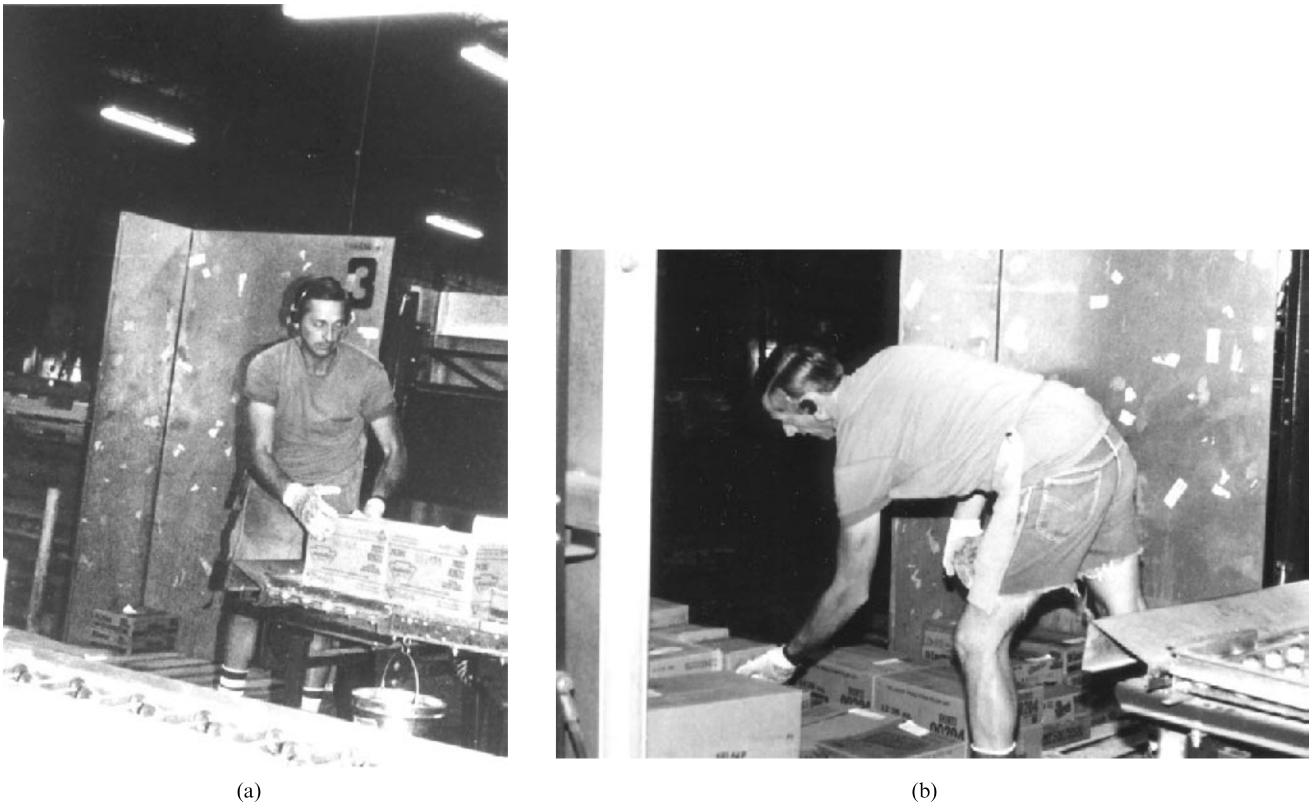


Figure 1 A typical human palletizing food products for distribution: (a) removal of box from conveyors; (b) stacking of box on pallet.

Research in palletizing focuses on determining the optimal pallet size and on packing methodology to maximize space utilization. This chapter concentrates on the palletizing of pallets using prismatic boxes. This restriction is not significant because the majority of stock-keeping units (SKUs) in distribution facilities are boxed. In addition, the methodologies developed for palletizing, with minimal changes, can apply to other palletizing approaches, such as container filling or pallet less stacking.

The University of Cincinnati was recently awarded U.S. Patent 5,175,692 for palletizing randomly arriving parcels of mixed size and content. This patent has been licensed to Motoman, who has developed the robotic solution into a product that is now available and promises to eliminate back injuries, reduce costs, and improve the quality of loads.

The purpose of this chapter is to describe robotic palletizing, including the algorithms for real-time pallet stacking of mixed-size/content parcels using an expert-system approach. The chapter is organized as follows. Previous research on robotic palletizing is reviewed in Sec. 4.2 The current palletizing methods are described in Sec 4.3. Future approaches to robotic palletizing are described in Sec. 4.4. Conclusions and recommendations are given in Sec. 4.5.

4.2 PALLETIZING

Palletization is a special case of a more general conformational problem called *space filling*. Space filling occurs in a variety of forms in industrial activities where the simplest one-dimensional case is called *stock cutting*.

4.2.1 Two-Dimensional Stock-Cutting Example

For this example, consider a continuous sheet of stock material of width, W, that is to be cut so as to satisfy the demands for lengths, D_i, of strips with a width, w_i, where $i = 1, \ldots, m$. In this case, the total demand or order is

$$D = D_1 + D_2 + \cdots + D_m$$

The demand D_i can be satisfied by supplying any number of pieces, n_i, of length, l_i, of the strips of width, w_i, so long as the total lengths, L_i sum to at least D_i:

$$D_i \leqslant L_i = n_i l_i \qquad \text{for } i = 1, 2, \ldots, m$$

The demands are met by deciding on various slitting patterns for the sheet of width W.

The jth slitting pattern is a way of dividing the width, W, into the smaller widths, w_i, for $i = 1, \ldots, m$. This pattern is applied to a length amount l_j of the sheet:

$$W \geqslant n_1 w_1 + n_2 w_2 + \cdots + n_m w_m$$

In the linear programming solution for this one-dimensional noninteger stock-cutting problem, the matrix A of the linear programming problem will have m rows and a large number of columns, k. One column will exist for each of the possible slitting patterns such that each vector. $N_i = [n_1, n_2, \ldots, n_m]$ of nonnegative integers satisfying the following conditions.

$$W \geqslant n_1 w_1 + n_2 w_2 + \cdots + n_m w_m$$

is a column of the matrix.

If X is a column vector of variables, each corresponding to a slitting pattern, one for each column of A, and if O is a row vector of all 1's, then the linear-programming problem may be stated:

$$\text{Minimize } O^T X = x_1 + x_2 + \cdots + x_k$$

subject to

$$A^T X = N$$

where N is the column vector $[n_1, n_2, \ldots, n_m]^T$.

Variations of this problem occur in both noninteger and integer forms. A linear-programming method may be used to solve the noninteger problem. However, a general difficulty is encountered due to the very large number of columns of possible solutions.

An integer problem is one in which the demands, D_i, are in integers and the variables, x_i are restricted to being integer. Rounded answers to the noninteger problem may be used to approximate the integer problem solution.

4.2.2 Three-Dimensional Space Filling

The general problem of filling a three-dimensional pallet with mixed-size parcels may be considered as a mathematical problem of finding the space that is filling the pallet's volume. That is, N parcels must be placed at positions (x_i, y_i, z_i) and the total volume filled as completely as possible. Other problems of this nature include the traveling salesman problem and the game of chess. In general, these problems are called NP-complete, that is, the computation time required for an exact solution increases exponentially with N. There is no method for finding an exact solution except exhaustive search of all possible solutions. Fortunately, modern artificial intelligent techniques provide a means to obtain good solutions. An expert system has been invented which provides solutions which satisfy a set of rules and consequently provide "good" solutions. Furthermore, the approach can be applied not only to single-product, mixed-layer, column or predefined order of arrival palletizing, but also to real-time, randomly arriving, and mixed-size and content palletizing.

4.2.3 Factors Affecting Palletizing

From the above discussion, it is apparent that different factors can affect the palletizing. The most important are:

Pallet size. Generally, the larger the pallet, the better are the chances of filling it efficiently.

Product proliferation. Contrary to initial intuition, a larger mix of sizes results in better load-forming efficiency, but at the expense of higher computer run time. Stated differently, if given an empty space, the chances of finding a box that closely fills that space are improved when a greater variety of box is available, but more time is needed to find that box. Note that boxes in an actual order typically present some correlation; for example, it is likely that there will be multiple boxes of a certain type. Putting this information to use will result in faster heuristics in generating load-forming layouts.

Standards. Establishing box/carton standards is essential because it greatly reduces the proliferation of boxes, thus allowing faster palletizing algorithms.

Algorithm. Exact algorithms are time consuming to the computer and difficult to implement. Heuristics often result in efficient solutions in relatively little time. Artificial intelligent methods could result in a better performance, especially if based on efficient heuristics.

Sequence of pick. Usually some pretreatment of the boxes can assist in the speed of reaching a solution. In many cases, the pretreatment may not even require additional work. For example, if boxes are stored and issued in a sequence that simplifies the allocation of space to the boxes (e.g., heavier boxes first, light ones later, boxes with identical sizes together, etc.), the solution could be reached more quickly and easily.

Look ahead. The ability to look ahead can also be used to speed up the search for space.

4.2.4 Palletizing of Identical-Size Parcels

Steudel [2] formulated the problem of loading uniform-sized boxes as a four-stage dynamic program that first maximizes the utilization on the perimeter of the pallet and then projects the arrangement inward. Correction steps were given for the cases where the projection resulted in overlapping boxes or in a large central hole. Smith and DeCani [3] proposed a four-corner approach to filling a pallet with identical boxes. The procedure determined the minimum and maximum number of boxes that could be placed starting from each corner of the pallet, and then iteratively evaluated the possible combinations that maximized the total number of boxes on the pallet. Although no claim of optimality is made in the paper, the results compare favorably with exact methods.

The results of these patterns are often summarized in a chart or table format. Apple [4] shows a set of patterns and a two-dimensional chart developed by the General Services Administration. The chart indicates which pattern is recommended for each box length–width combination. K. Dowsland [5] presented a three-dimensional pallet chart that works for different pallet sizes and indicates the sensitivity of the different patterns to variations in box sizes.

Researchers have tried to include some physical constraints to the pallet-loading problem. Puls and Tanchoco [6] considered the case where boxes are handled by opposite sides, and they modified the Smith and DeCani approach to start with three corners, resulting in layouts that are built with guillotine cuts. A guillotine cut is a straight line that cuts the pallet or rectangle across, resulting in two subrectangles. Carpenter and W. Dowsland [7] used a five-area approach that started from each of the corners and from the middle to generate alternative layout patterns. They evaluated the results based on criteria for load stability and clampability, i.e., the ability to handle the load with a clamp truck. It was deduced that layouts comprising two areas are the most suitable for clampability, but they also yield suboptimal utilization of the pallet volume. K. Dowsland [8] investigated the palletizing of boxes with a robot when it could handle one, two or four boxes at a time, and sought to determine the minimum number of transfers.

Gupta [9] investigated the problem of determining the pallet size when different box types are present, but each pallet was to hold only a single type of box. The problem was formulated as a two-stage mixed-integer programming model. The first stage seeks to optimize the placement of boxes along one side of the pallet and the second stage seeks to optimize the placement along the other.

4.2.5 Palletizing Boxes of Variable Sizes

In situations involving high volume and high complexity in terms of SKUs, the unit load to be formed is expected to contain items of different sizes. This problem has received much attention in operations research, especially under the closely related problems of bin packing, knapsack, stock cutting and plane tiling. The general form of the problem is far from being solved, and in fact can be shown to be *NP*-complete or "hard." As an outline proof, consider the simplified case where all the boxes have equal height and width, but differ in length. In this way, the problem is transformed into that of finding the combination of box lengths that best fill the pallet along its length. This problem is equivalent to the one-dimensional bin-packing problem, which was shown to be *NP*-complete [10]. *NP*-complete refers to the class of problems for which the only known solution involves enumerating all the possible combinations, which is time prohibitive because the number of alternatives grows combinatorially with increasing items. Consequently, these problems are solved using heuristics or expert system approaches, which yield nonoptimal solutions.

4.2.5.1 Heuristic Methods

Early efforts in the field include the work of Gilmore and Gomory [11, 12]. Their work investigated the two-dimensional stock cutting problem, which arises when a rectangular sheet of material is to be cut into smaller rectangles of different sizes. The problem is analogous to the palletizing of boxes of the same height. The authors formulated the problem as a linear program and suggested its solution by applying a knapsack function at every pivot step, recognizing that it would be computationally prohibitive.

Hertz [13] implemented a fast recursive tree search algorithm that optimized the solution obtained by using guillotine cuts. Note that this solution was not necessarily optimal for the general solution. Herz's algorithm assumed that the rectangles were positioned in one orientation only. When this assumption is applied to a box that can be rotated by 90°, a duplicate box with the length and width interchanged must be created. Christofides and Whitlock [14] also used a tree search routine to attempt to find the optimal layout that can be obtained using guillotine cuts. They narrowed the search space by eliminating redundant nodes that arise due to symmetry, the ordering of the cuts, and the location of the unused space. Applying this procedure to a problem with 20 boxes, the solution required 130 sec CPU time on a CDC 7600 computer.

Hodgson [15] combined heuristics and dynamic programming in the solution of a two-dimensional pallet layout. In this approach, the pallet is partitioned into a rectangular area, constituted by the boxes that were previously stacked starting from a corner, and into an L-shaped strip, the candidate to be filled. Dynamic programming was used to allocate boxes in the two rectangular sections forming the L. This approach restricted boxes to be placed in corridors around the starting corner, but because of the simple shape of the corridor, it resulted in significantly fewer partitions to be evaluated. Using the system, the operator interactively selects the first box (typically a large one) and the candidates for evaluation at each step. It was reported that the efficiency of packing increases with increasing number of box types, but at the expense of higher computer run time. In an adaptation of Hodgson's work, designed to run on a microcomputer, Carlo et al. [16] used a simpler heuristic of fitting boxes in order of decreasing size. The procedure was repeated by randomly varying the first box to be place and the orientation of the boxes, and the best result was saved. When allowed to run 1 min on a microcomputer, the procedure resulted in area utilization of about 95%.

Albano and Orsini [17] investigated the problem of cutting large sheets of material and proposed the approach of aggregating rectangles with an almost equal dimension into long strips. Then, a knapsack function was used to allocate strips across the width of the sheet. The procedure was fast and was found to result in very high area utilization (98%), especially when applied to larger problems.

The problem of packing three-dimensional pallets has been less thoroughly investigated. George and Robinson [18] studied the problem of loading boxes into a container. They developed a layer-by-layer approach. Following the selection of an initial box, all boxes with the same height become candidates, and are ranked first by decreasing width, second by quantity of boxes of the same type, and finally by decreasing length. The space in the layer is filled to preclude a face with pieces jutting by starting from one back corner and filling the area consistently to have a straight or steplike front. When evaluating their algorithm, George and Robinson found that it worked better with actual than with random or deterministic data, because actual shipments are likely to have correlated values.

4.2.5.2 Artificial Intelligence Approaches

Mazouz et al. [19–21] at the University of Cincinnati developed a rule-based expert system approach to palletize boxes arriving in a random sequence. The boxes are assigned locations on the pallet based on the criteria of size, toxicity and crushability. Toxicity is used to ensure that no toxic products are placed on top of edible goods, and crushability is used to ensure that no heavy loads are placed on top of soft or fragile boxes.

The system was developed using the OPS5 expert-system shell. The procedure first divided the available space into smaller discrete volume elements called voxels. Second, a relation table was generated for the box types in the bill of lading. The relations specify how many of one box type need to be stacked in order to obtain the same height as a stack formed with different box types. These relations become important in a layer approach to palletizing, in which a flat surface is required to form the next layer. Third, the boxes in the bill of lading were ranked according to the criteria of toxicity and crushability. Finally, at run time, for each box arriving on the conveyor, the procedure performed a search of the available space to determine where to stack the boxes. Boxes that could not satisfy the threshold requirement on toxicity and crushability were placed on a queue pallet. The expert system then downloaded the co-ordinates of the box to the interfaced Cincinnati Milacron robot that performed the palletizing. Test runs were made, and required 40 min on a VAX 11/750 to generate a pattern of 17 boxes arriving in a random sequence. Due to the layered approach, the loads formed with the system tended to be somewhat pyramid shaped, with larger layers at the bottom and smaller on top.

Another expert-system approach was developed at Georgia Tech University by Gilmore et al. [22] for use in palletizing boxes in a Kodak distribution center. The system was developed in Lisp-GEST and used a semantic frame representation. It considered the criteria of stability and crushability. The authors assumed that the order would be known in advance and that the boxes would arrive in a required sequence, and approached the building of pallets by columns rather than by layers. Using this approach, boxes of a similar type were stacked vertically in columns, which are then aggregated to form walls. A column approach is most applicable when there is some correlation between the boxes to be palletized. The column approach also requires simpler algorithms than a layer approach. The layer approach, on the other hand, provides stable pallets, even if they are moved before being wrapped. No report was provided on the speed or effectiveness of the Georgia Tech model. Other approaches, such as "simulated annealing" [23], could also be considered.

The goal of building an intelligent system for palletizing is fundamentally a problem of designing a decision maker with acceptable performance over a wide range of complexity in parcel sizes and uncertainty in parcel arrival sequences. Three approaches that have potential for this intelligent system are:

Expert system as a decision maker for palletizing.
Fuzzy logic as the decision-producing element.
Neural networks as decision-producing elements.

The expert system uses a rule-based paradigm built around "If-Then" rules. When the procedure works forward from a sequence of "If" conditions to a sequence of "Then" actions, it is called *forward chaining*. Forward chaining requires a database and a set of rules. This approach may be satisfactory for palletizing; however, it may be too slow for high-speed systems and has limited learning capability. *Backward chaining* starts with a desired sequence of "Then" actions and works backward to determine whether the "If" conditions are met.

The second approach deals with situations in which some of the defining relationships can be described by so-called *fuzzy sets* and *fuzzy relational equations*. In fuzzy set theory, the element membership decision function is continuous and lies between zero and unity. Fuzzy set theory is useful in situations in which data and relationships cannot be written in precise mathematical terms. For example, a "good stacking arrangement" may be difficult to quantify but provides significant fuzzy information that may be integrated into the decision-making process.

The third approach uses *neural networks* [24, 25]. With this approach, the input/output relationships can be modeled as a pattern recognition problem where the patterns to be recognized are "change" signals that map into "action" signals for specified system performances. This type of intelligent system can recognize and isolate patterns of change in real time and "learn" from experience to recognize change more quickly, even from incomplete data.

4.3 CURRENT WORK IN AUTOMATED PALLETIZING

An expert system is an excellent approach for palletizing, since it determines a solution that satisfies a set of rules. In the current system, both parcels and pallet space are represented by discrete volume elements, or voxels, that are equal to zero if the space is empty or unity if the space is full. The pallet is represented by a "blackboard" database that is changed as the pallet is filled. A bill of lading is used to represent the set of parcels which are to be stacked. A database of content information, size, fragility, etc. is also available for each parcel type. In addition, a relational database is formed, indicating size relationships between different parcel types. For example, one relationship between two small parcels placed together is that they could form a base for a large parcel.

The goal of the expert system is to determine where to place each randomly arriving parcel so that the overall center of mass coincides with the center of gravity or the pallet, and which satisfies all the other rules. Examples of rules include:

Toxic substances should not be placed on top of nontoxic products.
Boxes should not be crushed.
Glass containers should not be stacked on the bottom.
Fracture or fault lines should not be generated.
Interlocking of parcels should be done, if possible.

This expert system has been implemented in OPS5 and used to control a Cincinnati Milacron industrial robot, which was equipped with a vacuum gripper for palletizing food parcels. For all the tests conducted, a satisfactory stacking arrangement was obtained by the expert system. The major drawbacks at this time are computation time for the expert system. Speed of the robot was also a problem in the original implementation; however, a higher-speed Atlas robot was obtained. In the present research, we believe the computation time will be decreased by simplifying the algorithm, even though we expect to add additional rules throughout the study.

A conceptual diagram of a robotic palletizing workcell is shown in Fig. 2. The top-center block, the visual pallet, is the parent graphical user interface [26], the nerve center of the software system. From it, all data is relayed to and from the other software modules, such as the interface module, the barcode dynamic linking library (DLL), and the visual dynamic control interface (DCI) [27] (a robot control interface). In the case of a palletizing job of mixed size, or of content boxes arriving in random order, the interface module would come into play. As a job begins, the first box is scanned by the barcode reader. Then, the box SKU number is passed through a visual pallet to the interface, where its palletizing algorithm determines the box coordinates on the job pallet or a queue pallet. This data is passed through a visual pallet to a visual DCI which instructs the robot to palletize the box, return to the home position, and wait for the next instruction. After sending the co-ordinates to a visual DCI, the system determines if the palletizing algorithm has space on the job pallet for a box in the queue pallet. If it determines that it has adequate space, then it sends the data to a visual pallet, which relays the coordinates to the robot through a visual DCI. If there are not further instructions from the palletizing algorithm, a visual DCI instructs, through the barcode DLL, the barcode reader to scan the next box. The whole process starts over and continues until the last box is palletized.

In the past several years, a PC-based version of the expert system has been developed using the *Windows* development tool *Visual* C^{++} and integrated into the graphical interface described in this chapter [28,29]. The development of this PC-based palletizing algorithm was based on a revision of previously developed palletizing software, not a line-for-line conversion. Fortunately, all previously discovered rules can be included in this new software. Because of the recent improved processor capabilities in personal computers, the time required to process a solution for a pallet load has been greatly reduced. Processing time has been

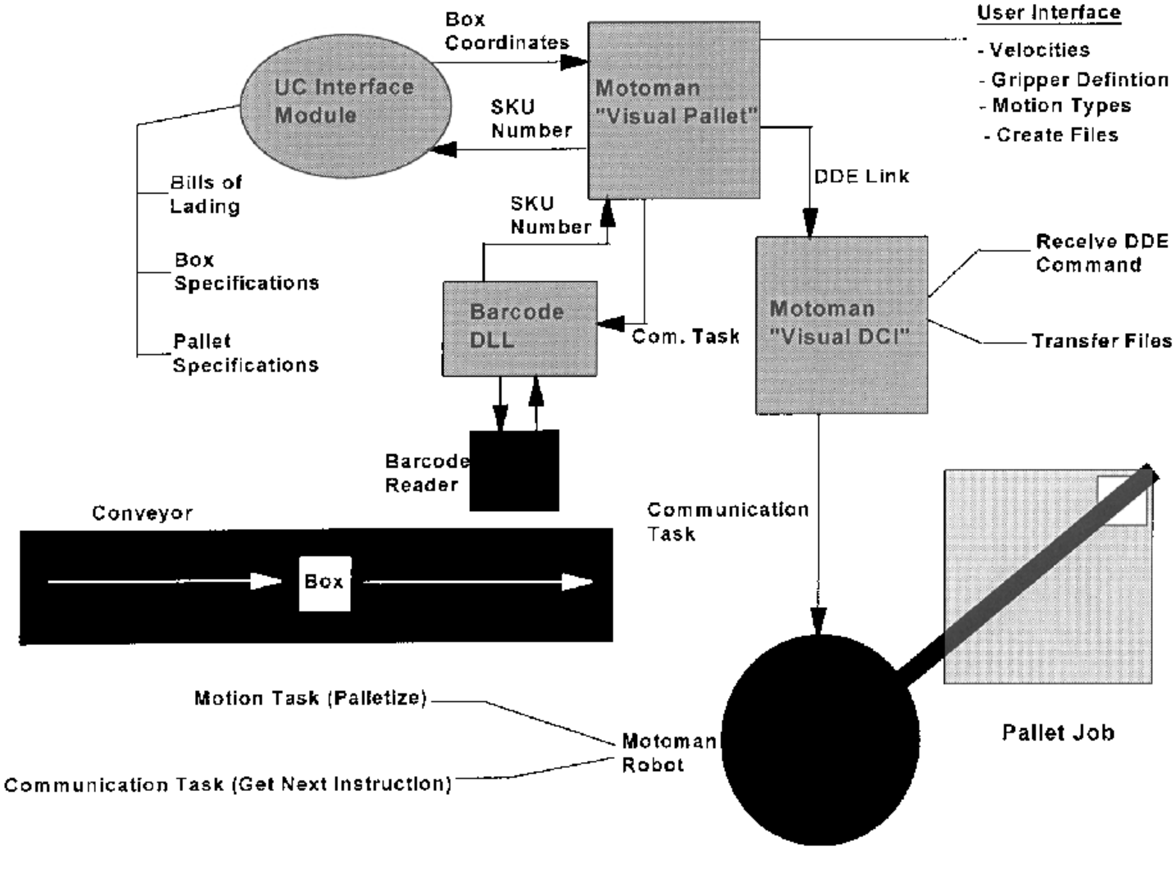


Figure 2 A conceptual diagram of a robotic palletizing workcell.

reduced from 2.35 min per box using the previous OPS5 expert system solution down to less than 5 sec per box using the presently developed PC-based palletizing solution. In light of these advancements, a robotic palletizing application becomes an even more attractive solution for every industry that utilizes this type of material handling.

An expert-system or rule-based approach was utilized in the development of the palletizing algorithm. These rules have been implemented directly in C language. This permits the system to run on a standard PC, and the code is transportable and expandable. A flowchart of the palletizing process is shown in Fig. 3. The overall logic of the expert system is shown in Fig. 4. The palletizing software system begins with system setup. This includes the first system setup in which pallet and box sizes are specified and the bill of lading specification and relationship determination. Then the real time loop is started in which a box is identified, and a search for an acceptable space is initiated. If an appropriate space is found, the co-ordinates are communicated to the robot and the space storage is updated. This loop continues until all the boxes from the bill of lading are placed. If space cannot be determined for any boxes, they are placed on a queue pallet. At the end of the loop, these boxes can be retrieved and placed on the pallet.

Two types of inputs are required for the algorithms. The first is a database of dimensional sizes and content information for the SKUs which are possible within the palletizing material handling stream. A separate effort is required to filter this data to ensure that all SKUs can be lifted by the particular robot gripper and placed by an industrial robot. Then, of the SKUs which can be handled, a relational database is prepared which examines spatial relationships, such as the number of boxes of one type that would form a stable base for a given number of boxes of another type. In addition, content-specific rules may be determined, such as those related to fragility, crushability, or contamination.

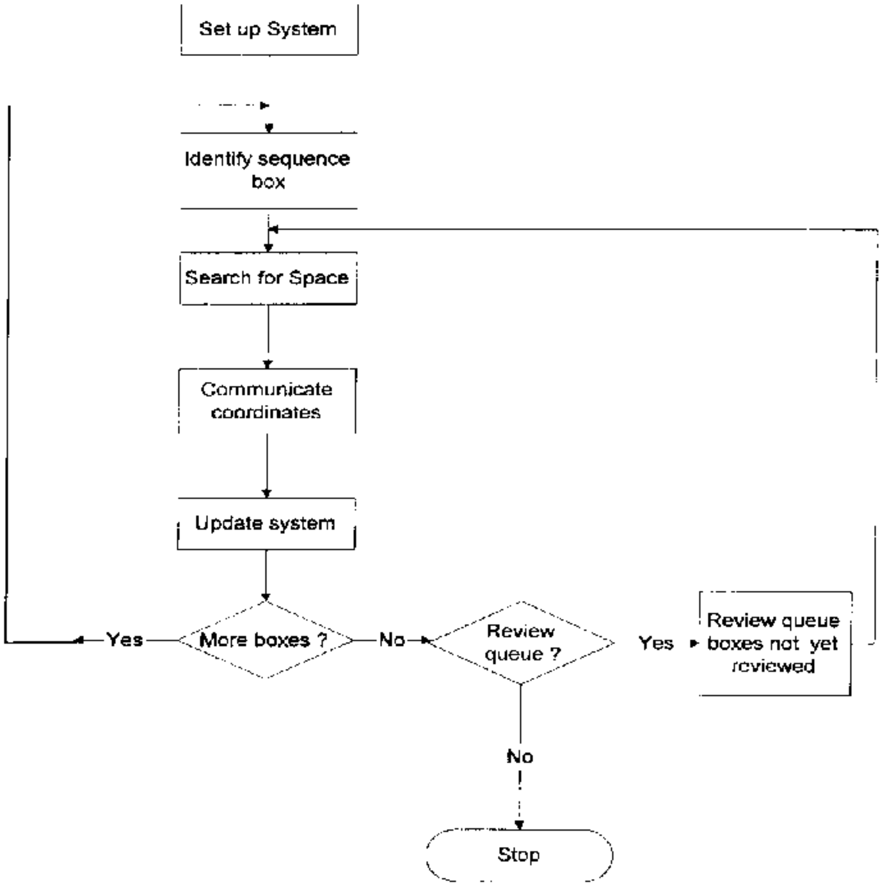


Figure 3 A flowchart of the palletizing process.

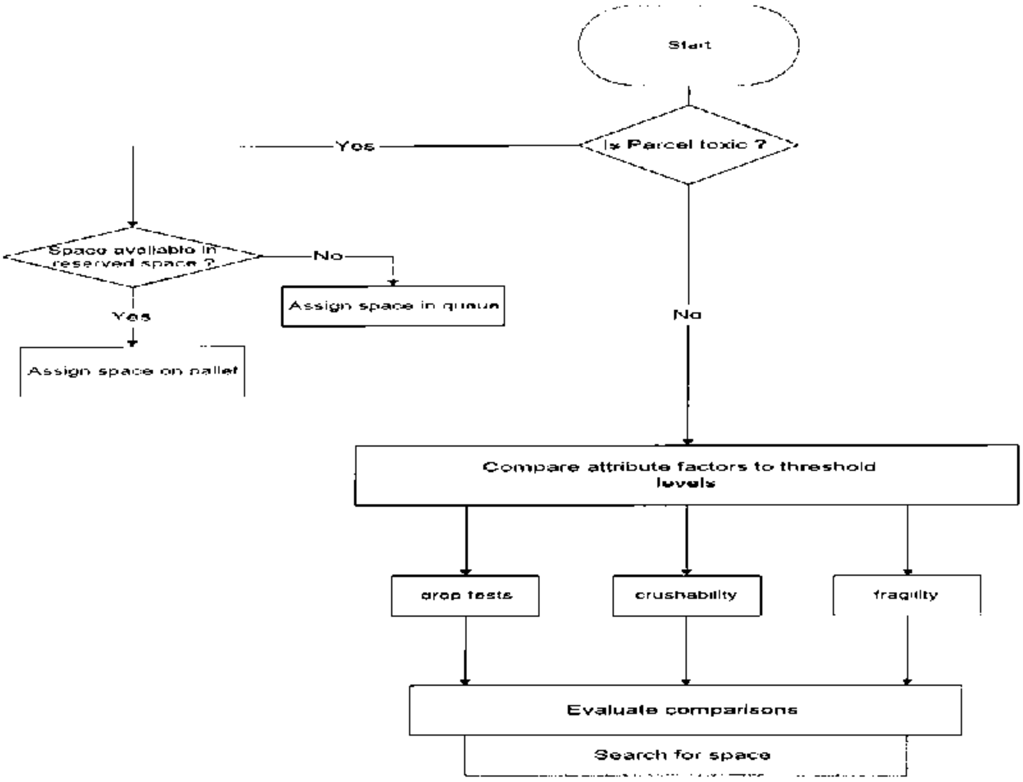


Figure 4 The overall logic of the expert system.

The second type of input is a bill of lading for a particular pallet. Orders would be processed separately to determine the number of pallet loads required for the entire order. The main emphasis for this effort was single-pallet load stacking of randomly picked SKU parcels. However, certain picking orders may be preferable and lead to faster stacking or better quality pallets. A third type of input that could be used is a pallet database.

The output of the software is the box locations for a pallet stacking arrangement. This arrangement satisfies all the rules built into the system and therefore gives an efficient pallet load. Measures of pallet quality, such as percentage utilization of the available cubic space, location of the three-dimensional centroid, etc., can be easily computed from the information available. The output file, with appropriate calibration and translation of co-ordinates, could give placement positions to a palletizing robot for each parcel on the pallet. The quality of the expert system pallet load is not "optimum" but rather "acceptable" quality, since it satisfies all the rules.

4.3.1 Establishing Set Relations

The set formation process, in which two boxes of type A are related to one box of type B in a length–width orientation, is shown in Fig. 5. Searching for box size combinations that form a stable base within the pallet size constraint forms relations.

Note: *in this section, C language code will be displayed in italics.*

In the algorithm; the `sRelat` function is called from `Main` by passing the variable `numBoxTypes`. This variable is used to determine the number of iterations to be used in the `For` loops which are used to compare different box types, scaled dimensions to form valid sets. A set relation is defined between two box types, in this case `box1` and `box2`, which can be either of the same or different types. If one or more boxes of `box1` forms the bottom layer of the set, then this bottom layer forms a stable base for the top layer comprising one or more boxes of `box2`. The type `box1` always forms the bottom layer, and the type `box2` always forms the top layer in a given set. The top layer can be either equal or one

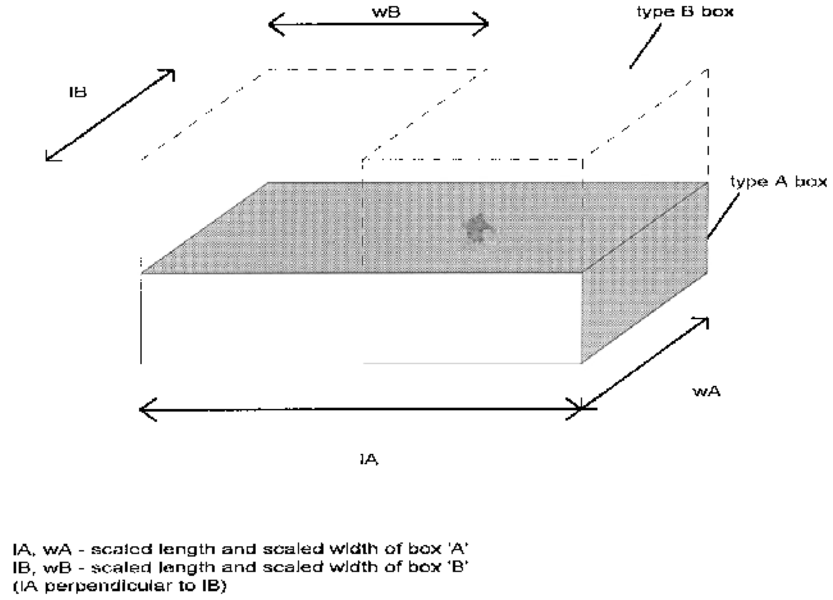

wB
type B box
lB
type A box
wA
lA
lA, wA - scalcd length and scaled width of box 'A'
lB, wB - scaled length and scaled width of box 'B'
(lA perpendicular to lB)

Figure 5 The set formation process in which two boxes of type A are related to one box of type B in a length–width orientation.

scaled-down unit smaller than the bottom layer along the length or width. A set is valid when sufficient quantities of both `box1` and `box2` types are available, and the set dimensions, defined as `setLength` and `setWidth`, are such that the `setLength` does not exceed the scaled pallet length (SPL), and the `setWidth` does not exceed the scaled pallet width (SPW). Since the code requires many `For` loops and `If` statements, to avoid confusion, only the coded used to form these sets will be discussed. Each valid set is stored in the data structure `sets_t`. All the sets formed are stored in an array of structures, `Sets[ ]`. The size of this array is defined in the header file.

```
struct sets_t{
   char box1;
   char box2;
   char b1inLength;
   char b1inWidth;
   char b2inLength;
   char b2inWidth;
   char orient;
   char setLength;
   char setWidth;
}; struct sets_tSets[MAX_SETS];
```

In the `sets_t` structure, the variable `b1inLength` is the number of boxes of type `box1` arranged along the `setLength` and `b1inwidth` is the number of boxes of type `box1` arranged along the `setWidth`. Similarly, the variables `b2inLength` and `b2inWidth` are for type `box2`. In a set, the length of `box1` is always parallel to the `setLength`, and the length of `box2` may be parallel or perpendicular to the `setLength`. If length of `box2` is parallel to `setLength`, then the variable orient is defined as `ORIENT_LL`. Otherwise, if `box2` is perpendicular to `setLength`, then `orient` is defined as `ORIENT_LW`.

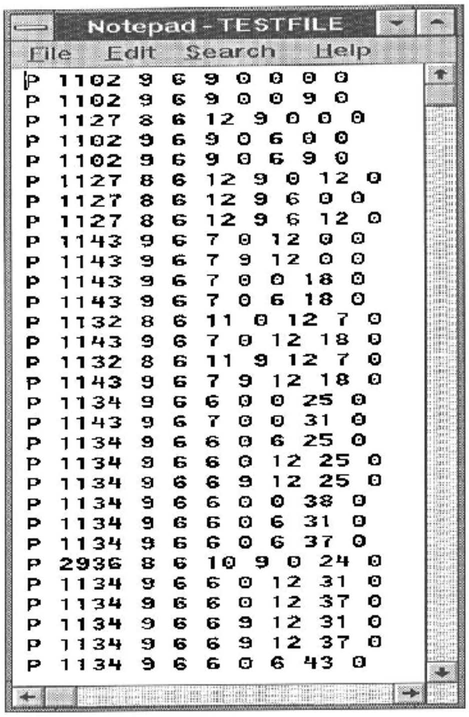
4.3.2 Search for Space

When a box is identified, a search for set relationships and quantities is made. If a set relationships is found with another box type and sufficient boxes of that type are in the bill of lading; then the box is placed, and space is reserved for the new box type. Up to this point, a framework has been constructed which will allow for the necessary user input that will enable the palletizing software to perform.

4.3.3 System Simulation and Performance

The results of this search are displayed as an output file showing box position, as shown in Fig. 6, or by an equivalent graphical output, as shown in Fig. 7. This type of output is very effective for displaying database information in a more visually pleasing and interactive form. Having both a box database and a pallet database linked to the interface also gives the user an inventory tool for the entire palletizing operation of a given distribution center/warehouse.

Several versions of the palletizing system have now been designed and constructed. A typical solution is shown in Fig. 8. The gripper is designed to lift the parcels that would be encountered in the application. The robot is selected to handle both the static load (weight) and the dynamic load of the parcels in motion. It must also have a sufficient reach to accommodate the pallets in the workcell. The operator can view both the simulation and actual box placement. In normal operation no operator is required.



```
P 1102 9 6 9 0 0 0 0
P 1102 9 6 9 0 0 9 0
P 1127 8 6 12 9 0 0 0
P 1102 9 6 9 0 6 0 0
P 1102 9 6 9 0 6 9 0
P 1127 8 6 12 9 0 12 0
P 1127 8 6 12 9 6 0 0
P 1127 8 6 12 9 6 12 0
P 1143 9 6 7 0 12 0 0
P 1143 9 6 7 9 12 0 0
P 1143 9 6 7 0 0 18 0
P 1143 9 6 7 0 6 18 0
P 1132 8 6 11 0 12 7 0
P 1143 9 6 7 0 12 18 0
P 1132 8 6 11 9 12 7 0
P 1143 9 6 7 9 12 18 0
P 1134 9 6 6 0 0 25 0
P 1143 9 6 7 0 0 31 0
P 1134 9 6 6 0 6 25 0
P 1134 9 6 6 0 12 25 0
P 1134 9 6 6 9 12 25 0
P 1134 9 6 6 0 0 38 0
P 1134 9 6 6 0 6 31 0
P 1134 9 6 6 0 6 37 0
P 2936 8 6 10 9 0 24 0
P 1134 9 6 6 0 12 31 0
P 1134 9 6 6 0 12 37 0
P 1134 9 6 6 9 12 31 0
P 1134 9 6 6 9 12 37 0
P 1134 9 6 6 0 6 43 0
```

Figure 6 Results of the search for space displayed as an output file showing box position.

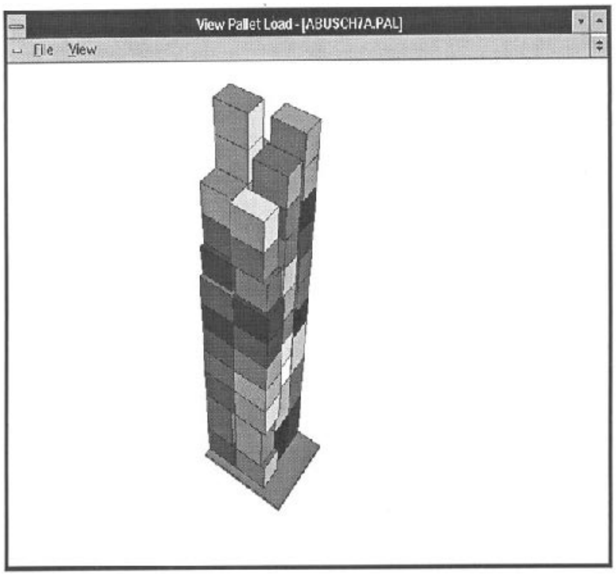


Figure 7 Equivalent graphical output of the search for space.

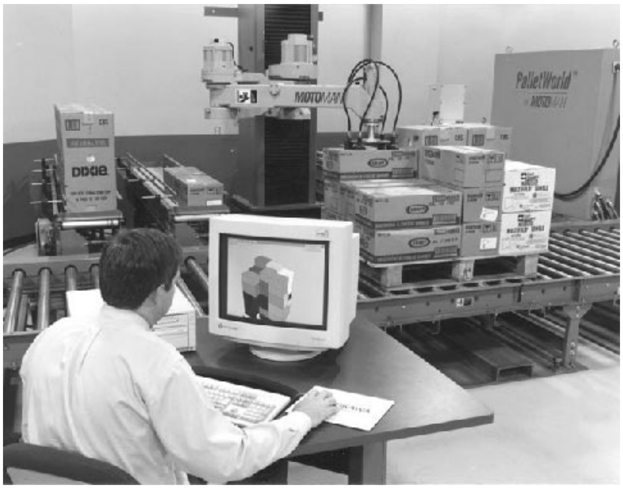


Figure 8 Motoman robotic palletizing system.

4.4 FUTURE RESEARCH ON ALGORITHMS FOR PALLETIZING

4.4.1 Expert-System Improvement

The expert-system approach has led to a solution that is practical and robust. Further rules may always be included and improvements in computer technology easily added.

4.4.2 Fuzzy Logic Approach

Fuzzy logic has received considerable attention since its introduction by Zadeh in 1965 [30]. This fundamental concept involves generalizing the traditional membership function of an element from a set of binary values $\{0, 1\}$ to continuous values on the interval $[0, 1]$. The fuzzy logic method seems appropriate for modeling several decisions encountered in palletizing. For example, in parcel placement, the amount of space used by each parcel may be modeled by a fuzzy membership function related to the volume filled by the parcel. In addition, the degree that a parcel is loaded also may be modeled by a continuous membership function. Finally, the degree of fragility of a parcel may be considered as a fuzzy set function.

To apply fuzzy logic to palletizing, the heuristic rules could be formulated in terms of imprecise propositions as well as specifications of the domains and ranges. The rules for palletizing would then be implemented using fuzzy logic. Measuring the load quality would then be performed and used to evaluate the fuzzy rules.

A promising combination of fuzzy logic and expert systems has been studied by Ralescu [31], and another interesting approach proposes the use of neural networks for computations of fuzzy logic inferences [32].

4.4.3 Neural Networks

Several faculties of neural networks make them attractive as an approach to the palletizing problem. One attractive property is the ability of a neural network to derive solutions to problems that involve "combinatorial explosion," and exponential increase in the number of possible answers. This ability was demonstrated by John Hopfield and David Tank [33] for the classic traveling salesman problem. For the palletizing problem, a three-dimensional array of parcels on a pallet could be used as the input with the requirements of a "good" pallet as the output. Various pallet configurations could be simulated from the test data to obtain a training set. Several neural network programs such as the backpropagation algorithm are available, which could be trained on the test data and tested on independent data.

Artificial neural networks (ANNs) are multilayered information processing structures consisting of large numbers of simple elements that process information in parallel [34]. These structures possess the ability to learn, associate, and generalize without rules. Artificial neural networks have been used to classify sonar data, speech, and handwriting. They have also been used to predict financial trends, to evaluate personnel data, to control robot arms, to model cognitive phenomena, and to superimpose geometrical regions. Several model ANNs have been proposed that have three things in common:

1. Distributed processing elements, or neurons
2. The connections between processing elements.
3. The rules of learning.

Artificial neural networks learn by adapting to changes in input data as the network gains experience. This learning may be categorized as supervised or unsupervised. In unsupervised learning, such as in the Kohonen net that will be discussed later, the ANN constructs internal models that capture regularities in input data. The most well-known supervised learning rules are Hebb's rule and the delta rule. Hebb theorized that biological associative memory lies in the synaptic connections between nerve cells, and that the process of learning and memory storage involved changes in the strength with which nerve signals are transmitted across individual synapses. The delta rule is a modification of Hebb's rule, stating that if there is a difference between actual output and the desired output, then the weights are adjusted to reduce the difference.

Using the above discussion of ANN adaptive learning, we can consider several model ANNs that seem to relate to the palletizing problem. Some particularly useful models include the Hopfield net, the single layer perceptron net, the multilayered perceptron net, and Kohonen's self-organizing feature-map forming net. Each of these will be briefly described.

4.4.3.1 Hopfield

The Hopfield net is primarily used with binary input. These nets are more useful when exact binary representations are possible as with ASCII text, where input values represent bits in the 8-bit ASCII of each character. However, these nets are less appropriate when input values are continuous because a fundamental representation problem must be addressed to convert the analog quantities to binary values. The Hopfield net may be used as an associative memory tool to solve optimization problems. It can also be used on problems where inputs are generated by selecting exemplar and reversing bit values randomly and independently with a given probability.

4.4.3.2 Single-Layer Perceptron

The single-layer perceptron can be used with both continuous and binary inputs. This simple net aroused much interest when it was initially developed because of its ability to learn to recognize simple patterns. The original perceptron convergence procedure was developed by Rosenblatt [35]. In this procedure, a perceptron decides whether input belongs to one of two classes by computing the weighted sum of the input elements and subtracting a threshold. The result is passed through a nonlinearity function so that the output is either $A+1$ or $A-1$. The decision rule is used to classify the output into one of the two classes. Connection weights and the threshold in a perceptron can be fixed or adapted depending on the algorithm. First connection weights and the threshold values are initialized to small random nonzero values. Then the new input with N continuous valued elements is applied to the input and the output is computed.

4.4.3.3 Multilayer Perceptrons

Multilayer perceptrons are nets where there are one or more layers of nodes between the input and output nodes, where these additional hidden nodes are not directly connected to both the input and output nodes. The strengths of multilayer perceptrons stem from the nonlinearity within the nodes. If nodes were linear elements, then a single-layer perceptron with appropriate weights could exactly duplicate the calculations performed by any multilayer net. Although multilayer perceptrons overcome many of the limitations of the single layer perceptrons, they were generally not used in the past because effective training algorithms were not available. Recently, this has been changed with the development of a new algorithm. Although it cannot be proven that these algorithms converge as the single-layer perceptrons do, they have been showm to be successful in many problems.

An interesting theorem that explains some of the capabilities of multilayer perceptrons was developed and proven by Kolmogorov. This theorem states that any continuous function of N variables can be computed using only linear summations and nonlinear, but continuously increasing, functions of only one variable. It also demonstrates that a threelayer perceptron with N^{2N+1} nodes with continuously increasing nonlinearity can compute any continuous function of N variables.

4.4.3.4 Kohonen Self-Organizing Feature Maps

One important organizing principle of sensory pathways in the brain is that the placement of neurons is orderly and often reflects some physical characteristics of the external stimulus being sensed. For example, at each level of the auditory pathway, nerve cells are arranged in relation to the frequency that elicits the greatest response in each neuron, and this organization of the auditory pathway extends up to the auditory cortex. Some of the organization is created during learning by algorithms that promote self organization. Kohonen presents one such algorithm which produces what he calls *self-organizing feature maps* similar to those that occur in the brain.

Kohonen's algorithm creates a vector quantizer by adjusting weights from common input nodes to M output nodes arranged in a two dimensional plane. Output nodes are highly linked by many local connections. Continuous input vectors are presented sequentially in time without specifying the desired output. These input vectors naturally form clusters with weights that specify the center of the vector clusters, defining the input space. After enough input vectors have been presented, the point density function of the vector center of the clusters tends to approximate the probability density function of the input vector. In addition, the weights will be organized such that close nodes are sensitive to inputs that are physically similar. Output nodes will thus be ordered in a natural manner.

This type of ordering may be important in complex systems with many layers of processing, since it can reduce lengths of interlayer connections. Kohonen [36] presents examples and proofs related to this algorithm. He also demonstrates how the algorithm can be used as a vector quantizer in a speech recognizer. The algorithm does perform relatively well in noisy systems. Since the number of classes is fixed, weights adapt slowly, and adaptation stops after training. Therefore, this algorithm is a viable sequential vector quantizer when the number of clusters desired can be specified before use, and the amount of training data is large relative to the number of clusters desired.

To apply neural network methods to the palletization problem, we start with the concept of a stable load. For a training set, parcels can be arranged in both stable and unstable stacking arrangements. Several algorithms such as the bidirectional associative memory [37], neocongnitron [38], adaptive resonant theory [39], Boltzmann and Cauchy [40], counterpropagation networks [41], and others have been studied by Hall and his students over the past several years. In many cases, both software and hardware implementations are available for this research.

4.4.4 New Architectures

Even though expert-system, fuzzy logic, and neural network approaches may be investigated and tested separately, it is possible that a new architecture or combination of these approaches will be discovered with further research. In fact, a neural network may be found to be superior for low-level decisions, while a fuzzy logic expert system may be superior for higher-level decisions.

4.5 CONCLUSIONS AND RECOMMENDATIONS

An introduction to robotic palletizing for parcels handling has been considered. Human palletizing is still commonly used even though cumulative trauma injuries may result. Robotic palletizing is similar to stock cutting, chess, and other *NP*-complete problems for which only an exhaustive search leads to an optimal solution. The expert-system approach that implements many rules seems ideal for this problem. Rules based upon relationships between the items being placed on the pallet provide useful guidance for parcel placement. Other rules regarding toxicity, crushability, fragility, etc. can also be implemented.

Other artificial approaches such as fuzzy logic and neural networks and new combinations are also briefly considered. Although the expert system solution is robust, other methods could lead to new solutions.

REFERENCES

1. KF Bloemer. A conceptual design for order pick and palletizing with robot palletizing vehicles. PhD dissertation, University of Cincinnati, 1992, p. 19.
2. HJ Steudel. Generating pallet loading patterns: a special case of the two-dimensional cutting stock problem. Manag Sci 25(10): 997–1004, 1979.
3. AP Smith, P DeCani. An algorithm to optimize the layout of boxes in pallets. J Oper Res Soc 31(7): 573–578, 1980.
4. JM Apple. Material Handling System Design. New York: Roland Press, 1972.
5. K. Dowsland. The three dimensional pallet chart: an analysis of the factors affecting the sett of feasible layouts for a class of two-dimensional packing problems. J Operat Res Soc 35(10): 895–905, 1984.
6. F Puls, J Tanchoco. Robotic implementation of pallet loading patterns. Int J Prod Res 24(3): 635–645, 1986.
7. H Carpenter, W Dowsland. Practical considerations of the pallet loading problem. J Operat Res Soc 36(6): 489–497, 1985.
8. K. Dowsland. Efficient automated pallet loading. Eur J Operat Res 44(2): 232–238, 1990.
9. A Gupta. Operations research models for design of palletization. J Inst Eng India) 57(ME 4): 183–185, 1977.
10. S Basse. Computer algorithms, introduction to design and analysis. Addison-Wesley Publishing, 1978, pp Reading, MA: pp. 268–271.
11. P Gilmore, R Gomory. A linear programming approach to the cutting stock problem. Operat Res 9(6): 849, 1961.
12. PC Gilmore, RE Gomory. Multistage cutting stock problems of two and more dimensions. Operat Res 113: 94–120, 1965.
13. JC Hertz. Recursive computational procedure for two-dimensional stock cutting. IBM J Res Develop 16(5): 462–469, 1977.
14. N Christofides, C Whitlock. An algorithm for two-dimensional cutting problems. Operat Res 25(1): 30–44, 1977.
15. T Hodgson, A combined approach to the pallet loading problem. IIE Trans 14(3): 175–182, 1982.
16. H Carlo, L. Hodgson, L Martin-Vega, E Stern. Micro-IPLS: pallet loading on a microcomputer. Computers Indust Eng 9(1): 29–34, 1985.
17. A Albano, R Orsini. A heuristic solution of the rectangular cutting stock problem. Computer J 23(4): 1980.
18. J George, D Robinson. A heuristic for packing boxes into a container. Computers Operat Res 7: 147–156, 1980.
19. AK Mazouz. Expert system for control of flexible palletizing cell for mixed size and weight parcels. PhD dissertation, University of Cincinnati, Cincannati, OH, 1987.
20. EL Hall, GD Slutzky, AK Mazouz. A final report, development and demonstration of robotic palletizing of mixed size and weight parcels for the Institute of Advanced Manufacturing Sciences, May 1987.
21. AK Mazouz, RL Shell, EL Hall. Expert system for flexible palletizing of mixed size and weight parcels. SPIE vol 848, Conference on Intelligent Robots and Computer Vision, Cambridge, MA, Nov 1–6, 1987, pp. 556–563.
22. J Gilmore, S Williams, E. Soniat du Fossat, R Bohlander, G Elling. An expert system approach to palletizing unequal sized containers. SPIE vol 1095, Applications of Artificial Intelligence VII, 1989, pp 933–942.
23. EE Witte, RD Chamberlain, MA Franklin. Task Assignment by Parallel Simulated Annealing. Procceedings of 1990 International Conference on Computer Design, Oct 1990.

24. J Kinoshita, NG Palevsky, Computing with neural networks. High Technol May: 24–31, 1987.
25. B Bavarian, Introduction to neural networks. Course Notes IEEE International Conference on Robotics and Automation, May 13–18, 1990.
26. D Redmond-Pyle, A Moore. Graphical User Interface Design and Evaluation (Guide): A Practical Process. New York: Prentice-Hall, 1995, 2.
27. A Marcus, N Smilonich, L Thompson. The Cross-GUI Handbook For Multiplatform User Interface Design. Reading, MA: Addison-Wesley, 1995, p vii.
28. T Langley. A graphical user interface for a robotic palletizing application. MS thesis, University of Cincinnati, 1996.
29. H Agha. Robotic palletizing algorithms for mixed size parcels. MS thesis, University of Cincinnati, 2000.
30. L Zadeh. Fuzzy sets. Inform Control 8: 338–353, 1965.
31. AL Ralescu. Meta-level expert system design. ACM Fuzzy Logic Conference, OUCC, Oct 1989.
32. JM Keller, RR Yager, Fuzzy logic inference neural networks. Proceedings of SPIE, vol 1192, Intelligent Robots and Computer Vision VIII, 1989, pp 582–587.
33. DW Tank, JJ Hopfield. Collective computation in neuronlike circuits. Scient Am July: 104–114, 1987.
34. JJ Hopfield. Neural networks and physical systems with emergent collective computational abilities. Proc Nat Acad Sci USA, 79: 2554–2558, 1982.
35. F Rosenblatt. Principles of Neurodynamics. New York: Spartan, 1962.
36. T. Kohonen. Self-Organization and Associative Memory. Springer-Verlag, Berlin, 1984.
37. B Kosko. Bidirectional associative memories. IEEE Trans Syst Man Cybern 18(1): 49–60, 1988.
38. K Fukushima, S Miyake. Neocognitron: A New Algorithm for Pattern Recognition Tolerant of Deformation and Shifts in Position. Pattern Recognition 15(6): 455–469, 1982.
39. G Carpenter, S Grossberg, ART2: Self-organization of stable category recognition codes or analog input patterns. Appl Optics 26(23), 4919–4939, 1987.
40. GE Hinton, TJ Sejnowski. Learning and relearning in Boltzmann machines. In: Parallel Distributed Processing, vol 1 Cambridge, MA: MIT Press, 1986, pp. 282–317.
41. PP Wasserman, Neural Computing, Theory and Practice. New York: Van Nostrand Reinhold, 1989.

Chapter 8.1

Investigation Programs

Ludwig Benner, Jr.
Events Analysis, Inc., Alexandria, Virginia

1.1 INTRODUCTION

This chapter describes what an investigation program is, what it should accomplish for an organization, how it should be created, and what investigators should do within that framework. It presents investigation fundamentals in a way that enables everyone in an organization to tailor the ideas so they satisfy their specific investigation needs. It includes models to help investigators during investigations, and references providing detailed guidance for program designers and investigators.

Accidents involving automated systems occur infrequently. However, many kinds of investigations are conducted in organizations using automated systems. Supervisors, mechanics, engineers, labor representatives, claims adjusters, safety staff, and others investigate claims, operational disruptions, equipment breakdowns, accidents, fires, injuries, outages, quality deviations, environmental insults, and other unexpected or undesired occurrences. Each type of investigation has many common tasks. These commonalties are masked by thinking about each kind of investigation as unique. In this fragmented environment nobody looks for the commonalties, or opportunities that co-ordinated thinking about all investigations might offer. Thus potential improvements in investigation programs are overlooked. This chapter addresses that oversight. It describes the overlooked opportunities and how to establish a program to take advantage of them.

1.2 WHAT IS AN INVESTIGATION PROGRAM?

An investigation program is an organization's ongoing structured activity to investigate unintended or unexpected and unwanted occurrences. This section describes the context in which such a program exists and functions, the role of the program in a dynamic organization, the nature of occurrences and investigations and the conceptual basis for an investigation program. The context provides the background that explains what an investigation program should accomplish, and what an organization should demand of an investigation program. The discussion of the role describes the relationship of an investigation program to other organizational activities. The discussion of the nature of occurrences and investigations describes useful ways to think about them within an organization. The discussion of the knowledge needed to do investigations describes essential investigation concepts and principles needed to produce the desired results

1.2.1 Investigation Program Context

Investigations take place within an organizational context and a regulatory context. The organizational context should dominate investigation programs, but must accommodate the regulatory environment.

1.2.1.1 Organizational Context

Nobody likes unpleasant surprises. Progressive managers view an investigation program broadly as a set of continuing activities designed to understand, predict, and control or prevent unpleasant and unwanted "surprises" in operations. These surprises include many kinds of occurrences, such as injuries, accidents, fires, breakdowns, outages or delays, environmental insults, operational disruptions, claims, or other kinds of undesired events. Surprises reflect deviations from expected or intended or hoped-for performance, interfering with desired outcomes. The fundamental mission of a comprehensive investigation program is to improve future performance by thoroughly understanding and acting on past occurrences of all kinds.

Recurring unpleasant surprises are in indication, in part, of investigation program shortcomings or failures, or possibly the lack of a competent investigation program.

1.2.1.2 Regulatory Context

In addition to an organization's internal interests, certain regulatory requirements affect the investigation context in most organizations employing or supplying automated systems. Most employers are subject to occupational safety and health regulations, which include investigation program requirements [1]. Briefly summarized, regulations require that:

1. All accidents should be investigated.
2. Accidents involving fatalities or hospitalization of five or more employees be investigated to determine casual factors involved and that on scene evidence be left untouched until agency inspectors can examine it.
3. Any information or evidence uncovered during accident investigations which would be of benefit in developing a new regulatory standard or in modifying or revoking an existing standard be promptly transmitted to the agency.
4. The investigative report of the accident shall include appropriate documentation on date, time, location, description of operations, description of accident, photographs, interviews of employees and witnesses, measurements, and other pertinent information, be distributed to certain people, and made available to an agency representative. The regulation does not specify explicitly the purpose of required investigations, but a standards development purpose is implied.

1.2.2 Investigation Roles

The basic functional role of investigations of all kinds is to develop a basis for and report on future action to improve future performance. The basis for action must always be a valid description and explanation of occurrences, developed promptly, efficiently, objectively, and consistently.

This requires investigators to document their description and explanation, reporting them in a way that enables managers and others to understand, accept, and want to act on this new information. Investigations should assure discovery and definition of problems or needs that require action, and of actions for addressing them. They should also provide a way to assess whether the changes introduced actually improved future performance. Investigations should also validate predictive analyses and design decisions. If these basic needs are satisfied, opportunities for additional benefits can be realized.

Investigators first look backward in time to determine and explain what happened. When they understand that, they must look forward in time to identify changes that will improve future performance. To fulfill their role, investigations must be perceived by all affected as desirable, valuable and helpful, rather than judgmental, threatening, punitive, vengeful, or accusatory. To achieve best long term results, the tone of the investigation program must encourage co-operation and support.

1.2.2.1 Desired Roles for Investigations

Competently designed and implemented investigation programs should report new understanding of occurrences in ways that help:

Reduce future surprises which interfere with desired outputs.
Resolve claims and disputes.
Satisfy regulatory requirements.

They also have the potential to:

Reduce resource needs by revealing potential process improvements.
Enhance employee capability and morale with constructive work products.
Reduce exposure to litigation.
Provide a way to audit analyses of planned functions.
Predict changes to influence future risks.
Identify shifting norms and parameters in operations.

Contribute to the organization's long term corporate memory.

One other potential role requires an executive decision. The choice is whether or not to use investigations to assess installed safety and reliability systems and their performance. Audits require special criteria and audit methods, and additional data, so it is advisable to conduct program audits as stand-alone activities rather than an element of investigations.

1.2.2.2 Traditional Views of Investigation Role

That view differs from the regulatory view of the role of investigations. Traditional investigation perceptions and assumptions in industrial settings focus narrowly on *accident investigations, failures, unsafe acts and conditions, basic, direct and indirect accident causes, and compliance*. That focus does not address or satisfy many internal needs, and limits opportunities for broader achievements. The Federal agency regulating industrial robotics safety, for example, views investigations as an element of a safety program rather than a part of a broad organizational performance improvement program. In its view investigations have a narrow goal of preventing similar accidents and incidents in the future. It holds that "thousands of accidents occur throughout the United States every day, and that the failure of people, equipment, supplies, or surroundings to behave or react as expected causes most of the accidents. Accident investigations determine how and why these failures occur" [2]. Note the negative tone of this "failure" and cause-oriented perspective.

The agency's demands of investigations are also narrow. "By using the information gained through an investigation, a *similar or perhaps more disastrous accident* may be prevented. Conduct accident investigations *with accident prevention* in mind" (emphasis added) [2].

The loss or harm threshold, rather than the surprise nature of the occurrence, narrows the field of candidates for investigation. The authority to impose penalties also influences the agency's perception of investigations, and the procedures it must follow. When it becomes involved in investigations, operating organizations must recognize and adapt to the regulatory agency's perspectives.

In summary, the role of an investigation program should be constructive, designed to develop new knowledge to support a broad range of future actions in an organization, and produce timely, efficient, objective and consistent outputs.

1.2.3 Nature of Investigation Processes

To investigate something is to examine it systematically. Any investigation should be a systematic examination process. The investigation process focuses on examining the people and objects involved in the occurrence, and everything they did that was necessary and sufficient to produce the process outcome that prompted the investigation.

Investigations involve many tasks. Most share many common investigation tasks and tools. For example, in every investigation the investigator must:

Make observations of people and objects involved in the occurrence.
Acquire, structure, document, and organize data about their interactions.
Discover, define, and describe what people and objects had to do to produce the outcomes.
Apply logic to action data to define cause-effect linkages.
Recognize, define, and act on unknowns, and frame questions to pose.
Diagnose objectively what happened to define needs for change and candidate changes.
Evaluate needs and propose actions, with ways to monitor their success.
Prepare valid and persuasive investigation work products.
Mediate differing views.

The specific nature of each task and level of effort required of the investigator differ in nature depending on the kind and level of investigation required. For example, the degree of effort required to prepare an incident report form is the least complex, and may be considered the lowest level of investigation (Level 1). The nature of that investigation is to gather data needed to complete a reporting form. That need is usually satisfied by sequencing whatever data can be acquired in a relatively brief time. Note that the data collected on forms are analyzed later by accident or claims analysts. This may mean that several similar incidents must occur before sufficient data for some analysis methods is available.

A slightly greater effort and more tasks are required to complete a logically sequenced and *tested* narrative description of what happened, or Level 2 investigation. This level requires the investigator to do some logical analysis tasks as the data are gathered. For example, understanding equipment breakdowns requires this kind of effort.

When the description of what happened must be expanded to include carefully developed explanations, a greater level of investigation is required. Level 3 investigations may involve teams, and additional analytical and testing tasks to validate the explanation and assure adequate objectivity and quality. This level is required for matters that might be involved in litigation or compliance actions, or contractual disputes over equipment performance or warranty claims.

If recommendations for actions to improve future performance are required of an investigator, the investigator must do additional forward-looking data gathering and different analytical tasks. Level 4 investigations are the most complex and demanding and usually involve an investigation team. They should be required for any major casualty, or facility or design changes driven by undesired occurrences. Thus the nature of an investigation and the knowledge and skills required to do them is dependent on the expected investigation level and outputs.

The nature of an investigation is also partially dependent on the number of investigating organizations conducting investigations of the same occurrence. The tasks where interactions occur should be reviewed with organizations which might be involved in investigations. For example, whenever fatal injuries occur, an incident might involve investigators from organizations such as a local law enforcement agency or medical examiner, a state or federal regulatory authority, an insurance representative, and an organizational team. The authority and actions of those officials should be identified before an occurrence, and general agreement reached about who would do what in an investigation. When law enforcement or regulatory investigators are involved, their interests include access to witnesses and property, and preservation of evidence until an investigation has been completed [1]. Legal rights also may affect the nature of the investigation. These interactions are complex, but planning helps everyone work together when required.

1.2.4 Investigation Knowledge Needs

Performance of investigation tasks requires knowledge about *investigation* concepts, principles and practices, and skills in applying that knowledge. Investigation knowledge is not the same as knowledge about automated or robotics systems. Every automated system expert is not intuitively an automated system investigation expert. Additionally, system experts tend to unconsciously accept assumptions and ideas on which their decisions about the system are structured. Frequently those assumptions and ideas have contributed to the occurrence. Expert investigators avoid that trap by applying their *investigation* knowledge and skills.

During the investigation process, investigators use *investigation* tools to determine, describe, and explain what happened. Sometimes they need expert help to acquire or interpret data they need from objects involved in the occurrence. These data can be acquired with the help of others by knowing how to identify the expertise needed, and how to frame the right questions for those experts. Typically, such experts have expert knowledge and experience in some specialized field of the physical sciences, and can interpret what actions were required to produce the observed postoccurrence states. Their outputs must support the investigator's concrete needs.

To discover and define needs indicated by the occurrence, investigators require data about how a specific system was intended or expected to function in its daily environment. Expert investigators get such system data from people with system knowledge, either directly or from their work products. Those system experts have knowledge of a specific system's design, manufacture, testing, programming, operational behavior, safety or failure analyses, maintenance, or other system support activities.

1.2.5 Investigation Task Knowledge

Study of investigation processes has disclosed that, to be effective, investigation process tasks must be disciplined, objective, timely, efficient, and logical, and produce demonstrably valid, credible, and readily useful outputs. Special *investigation* knowledge investigators need to perform their investigation tasks adequately includes fundamental investigation concepts, principles, and procedures. They must incorporate this knowledge into investigation program plans for all kinds of investigations.

1.2.5.1 Investigation Concepts

Concepts about occurrences and investigations guide how investigators think about what they are investigating, and what they do during an investigation [3]. Concepts needed by investigators to produce quality work products include:

A multilinear conceptual framework.
The role of change in occurrences.
An investigation data language.
Mental movies

Progressive analyses
Break down events
Energy tracing
Event pairing
Event linking
Investigation quality assurance.

Multilinear Conceptual Framework. What is the general nature of occurrences to be investigated? Research has identified at least five different perceptions of unintended and unexpected occurrences [4]. Each perception results in a different framework or model that drives what investigators think and do during investigations.

The most helpful perception of occurrences or framework for investigators is the "multilinear" events sequences concept [5a]. This framework views occurrences as a process, during which people and objects act, concurrently and in sequence, to produce successive changes resulting in the outcomes of interest. Relative timing of events in this multilinear framework is often essential to understanding and explaining what happened. The framework leads investigators to focus on developing descriptions and explanations of process interactions that produced the outcomes of interest.

Other perceptions of the nature of occurrences are often encountered. A linear "chain of events" perception of occurrences such as accidents has long been the most popular in lay circles and the legal community. It relies on experts to identify a chain of unsafe acts and conditions and accident causes "leading to the accident" or incident. Typically, it results in subjectively developed, investigator-dependent, judgment-laden and frequently controversial investigation work products. The stochastic perception is similarly investigator or analyst dependent. The tree perception is more disciplined, and helps to organize data, but lacks criteria for selecting top events and a data language, does not accommodate relative event timing and duration considerations, or show interactions among concurrent events readily. The five major perceptions are illustrated in Fig. 1.

Role of Change in Occurrences. The role of change in surprise occurrences and their analysis was defined by Johnson during research leading to the MORT safety assurance system [6]. He pointed out the congruence between change control and accidents, and the importance of examining changes during investigations.

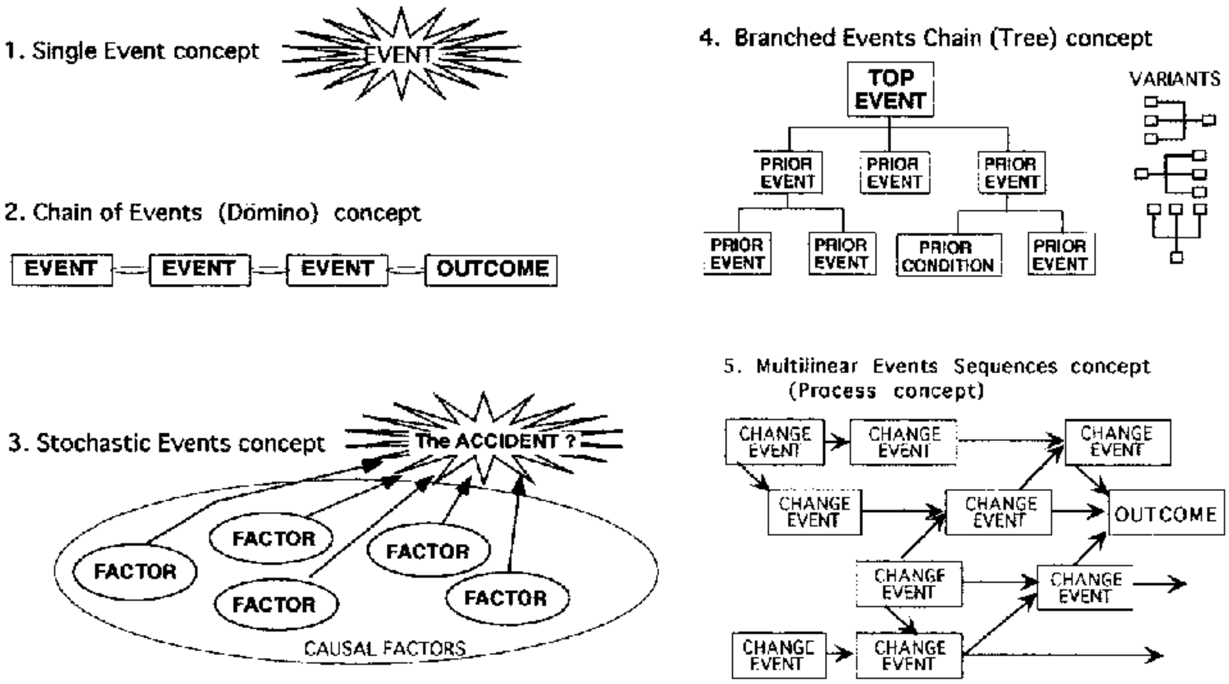


Figure 1 Perceptions of accidents: the five ways investigators perceive the nature of the accident phenomenon. Each perception influences what investigators think and do during investigations. (From Accident Investigation: Safety's Hidden Defect. Oakton, VA: Ludwig Benner & Associates, 1981.)

During the operation of a process, people or objects act on other people or objects to produce cascading changes, with resultant outputs or outcomes. When desired outputs result, change produces progress. When undesired or unintended outputs result, change produces trouble. The change concept facilitates investigations by providing a focus for investigators' data searches: look for the changes required to produce the outcome.

When people act during a process, they act to produce an intended change, to adapt to an unanticipated change to sustain the process, or to arrest undesired cascading changes. For example, if a robotic device needs adjustment, a programmer acts to reprogram the device. If a robotics device suddenly activates during maintenance, the repairman might either adapt by trying to avoid the moving parts, or arrest the progression by activating the emergency "off" control.

A useful aspect of change is the concept of *change signals*. The signal emitted by a change has consequences for investigators. For example, if the signal emitted is not detectable or detected too late, the opportunities for an adaptive response by either people or objects are foreclosed. If it is detectable, it must be detected before an adaptive response is mounted. This general adaptive subprocess has been modeled from observations during investigations (see Appendix A).

Event Data Language. Investigation data language is the language structure and terms investigators use to document, analyze, describe, and explain an occurrence. To be consistent with the process framework for occurrences, the investigation data language must be able to describe and report what people and objects did to advance the undesired process toward its outcome. The data language structure used by investigators determines what they can do during an investigation. A structure that facilitates the *verifiable reporting* of what happened and why it happened is needed. A structure and terms that undermine verifiable reporting are not helpful.

The structure should encourage investigators to focus their observations on finding and documenting data that define and permit the *value-free* reporting of what the people and objects did during the occurrence. It should steer investigators to verifiable terms, and away from terms with built-in judgments or unsupported inferences which stop thought.

The data language structure and terms that best satisfy these demands are the actor–action structure and event-related terms. The structure is simple: *one actor + one action = one event*. That is the foundation for the "think events" guidance encouraging investigators to structure their investigation thought processes. It employs the definitive power of the grammatical active voice, facilitating the visualization of specific people or objects. This "actor + action"-based structure, or "event" structure, makes possible the most economical acquisition and ordering of data. It facilitates the most concrete descriptions of what happened, the most practical approach to systematic problem discovery and remedial action selection, the implementation of objective quality controls, and timely results.

The actor + action language structure helps guide other tasks, such as facilitating *visualization* of what happened, rather than impeding visualization of what happened. It should be used while interviewing witnesses, photographing ending states of objects, or designing damaged-equipment test protocols. Documenting data with abstract, ambiguous or equivocal terms does not offer such guidance.

It is important to note that *conditions are the result of actions by someone or something*. Improving future performance requires *a change in behavior* of people or objects. A condition cannot be changed without changing the behavior of someone or something that created the condition. Thus, investigators should focus on the actor + action data language during investigations, and use observed conditions as a basis to infer the actions that produced them.

During investigations, investigators' major challenge is transforming their observations and all other information they acquire into a common format to give them building blocks for creating their description and explanation of what happened. *This task is not intuitive*. Further, it conflicts with daily language experiences. The challenge is to recast all kinds of data from all kinds of sources into a basic common format suitable for documentation, analysis, testing, reporting, and dissemination. That challenge is depicted in Fig. 2.

The exact attributes of event building blocks depend on the choice of investigation process adopted by an organization. The most basic form of event building blocks (Fig. 3) contains the following information:

Actor is *any person* or *any object* that initiates a change of state during the process required to produce the outcome achieved by the occurrence. An actor has *only one name*. Ambiguous, compound, group, or plural names will corrupt the investigation and are *unacceptable*.

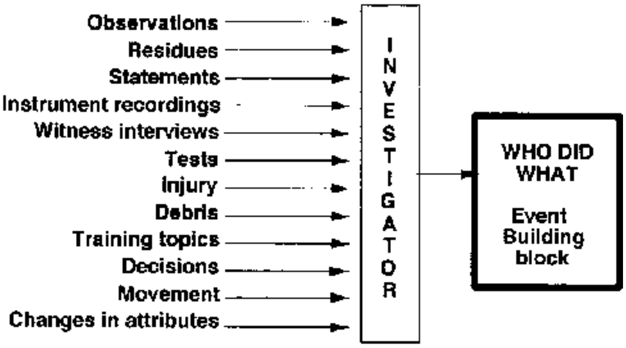


Figure 2 The investigator's data transformation challenge. The investigator must transform all kinds of data from all sources into the investigation data language format needed to describe what happened. (From 10 MES Investigation Guides. Guide 1, MES Event Building Blocks. Oakton, VA: Ludwig Benner & Associates, 1998, p. 6.)

Action is one specific act which affected another actor or action *and* helped initiate or sustain the process that produced the outcome of the occurrence.

Descriptor is used to expand the description of what the actor did, to describe what the actor acted on, or otherwise to define the act so it is uniquely described, can be visualized, and then can be related to other events.

Source is the source of the data from which the event block was formulated, noted so it can be referenced as needed to verify the event.

For more complex investigations or investigations requiring clear documentation of source data and verification of the reasoning, it is helpful to use more comprehensive building blocks, as shown in Fig. 4. The numbers refer to the sequence in which the contents are typically added.

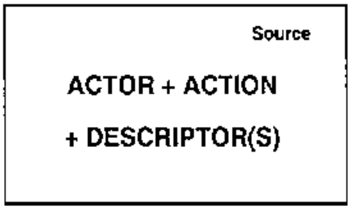


Figure 3 Minimum event building block elements. This is the minimum information required to permit investigators to arrange events into their correct sequence as they develop their description of what happened. (Adapted from K Hendrick, L Benner. Investigating Accidents with STEP. New York: Marcel Dekker, 1986, p 128.)

Without a specified data language structure to guide investigators, investigators are likely to use words that can corrupt the investigation or undermine the potential value of an investigation. Corrupting words include *ambiguous names or action descriptions, implicit conclusions*, and *words with built-in judgments*. For example, ambiguous names of actors like "they" or "she" or grouped actors like "the crew" or "the second shift" can confuse hearers or readers, because they can not visualize who did what without more data. Ambiguous actors reflecting inadvertent use of the passive voice grammatically, such as "it was decided," have the same effect. Investigators often use the passive voice to cover up their incomplete or shoddy investigation or unacknowledged unknowns. Implicit conclu-

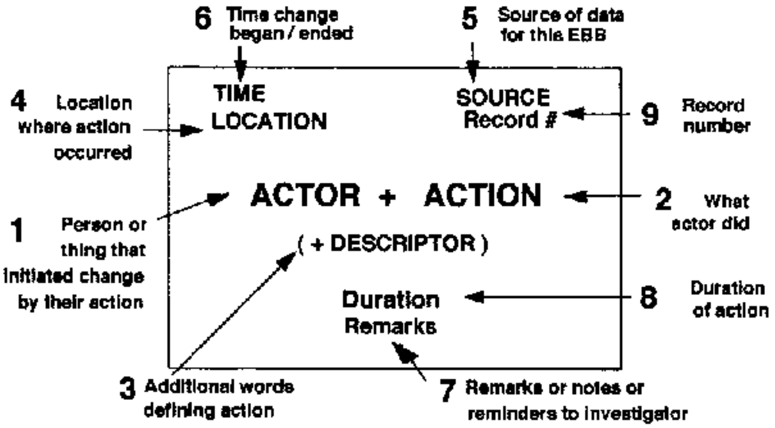

6 Time change began / ended
5 Source of data for this EBB
4 Location where action occurred
TIME
LOCATION
SOURCE
Record #
9 Record number
ACTOR + ACTION
2 What actor did
1 Person or thing that initiated change by their action
(+ DESCRIPTOR)
Duration
Remarks
8 Duration of action
3 Additional words defining action
7 Remarks or notes or reminders to investigator

Figure 4 Comprehensive event building block. This format is helpful for documenting actions during a complex occurrence, and for investigations which might be used in potentially controversial environments such as claims settlement, arbitration or litigation. (Adapted from K Hendrick, L Benner. Investigating Accidents with STEP. New York: Marcel Dekker, 1986, p 128.)

sions may be subtle, and are usually hidden in words like "did not," or "failed," or "inadequately." They should be avoided, unless the evidence and behavior standard on which the conclusion is based are also clearly defined and described.

Most corrupting are words with built-in judgments. Descriptions of occurrences should be factual, not judgmental. Frequently the judgments can not be verified, convey false certainty, rouse defensive feelings, mask differences in understanding, stifle thought, and slant viewpoints. For example, once a judgment is made that someone "failed" to act, made a "human error," or was "inadequately" prepared, the tone of what follows is set—to find out what the person did wrong and lay blame on that person. Investigators should view such words as *poison words*, and *avoid them*. A review of language pitfalls described in Hayakawa's work [7] is highly recommended. The investigator should strive to report events at the lowest rung on Hayakawa's ladder of abstraction.

Conformance to the actor + action data structure helps investigators avoid these pitfalls, economize their investigation reporting efforts, and improve investigation efficiencies.

Mental Movies. A mental movie is a sequence of visualized images of what happened, arrayed in the sequential order and approximate times they happened. Making mental pictures or a "mental movie" of what people and objects did enables investigators to cope with new data as the data are acquired. They enable investigators to integrate data gathering and analysis functions.

Mental movies serve four important investigation purposes. They force investigators to try to visualize what happened, demand concrete action data, help order the data as they are acquired, and pinpoint what they do not know about the occurrence. The mental movie construction requires investigators to visualize the specific actors and actions involved in the occurrence and the effects of their actions on others. As the data acquisition continues, the mental movie framework provides a place to order the actions relative to other data already in hand. When investigators cannot visualize what happened, each "blank frame" in the mental movie identifies unknowns, and the need for specific data about the actor or action in the time period involved. Thus blank frames define unknowns and narrow the search for additional data as the investigation progresses.

The concept also applies to witness interviews. The investigators' challenge is to transfer the mental movie from the witnesses' heads into their heads. This view helps investigators probe for concrete data from witnesses, and ask questions that generate concrete answers.

Progressive Analysis. This is the concept of integrating new data into all existing data as each new data item is acquired during the investigation. The reason for using progressive analysis methods is to integrate the data gathering and analysis functions into an efficient, effective consolidated task as the investigation progresses.

The progressive analysis concept provides a basis for establishing criteria for the selection of the investigation methods. The formulation of mental movies is an informal implementation of this concept. A more formal implementation is the multilinear events sequencing methodology and its flow charting time-events matrices, or worksheets. Using either method, investigators can achieve very efficient, real-time data gathering and analysis task integration during investigations.

The historical approach to investigation has been to gather all the facts, analyze the facts, and then draw conclusions and report findings. This approach results in separately gathering the "facts" and subsequently analyzing them to develop conclusions and findings. The approach is widely used by traditional industrial accident investigators, by litigants, and by many public investigation organizations. This process is inefficient, time consuming, and prone to overlooking relevant data. Additionally, it is more tolerant of ambiguous and irrelevant data, particularly in investigations with two or more investigators. The identification of relevant data during data gathering tasks is ill defined, and objective quality management methods are not usually viable.

Break Down Events. Breaking down or decomposing events is an old concept, but understanding how it is done is very important to investigators. When the "think events" concept is employed, unclear or grouped actors or actions can be "broken down" or decomposed into two or more actors or actions to help investigators understand what happened.

One question every investigator faces in each investigation is how long to continue breaking down events. The technical answer is "it depends"—on the need to understand what happened in sufficient detail to be able to reproduce the occurrence with a high degree of confidence. Alternatively, it may depend on the resources available for the investigation: stop when the allotted time or money is exhausted. Still another answer depends on the quality assurance task needs: stop when quality assurance tasks meet quality assurance criteria, including the degree to which uncertainties or unknowns are tolerated in work products.

Event Pairs and Sets. An event pair or event set consists of two or more events, either next to each other in the sequence, or part of a cause–effect relationship. Event pairs or sets provide the foundation for sequencing events disclosed by the investigation data, using temporal and spatial sequencing logic. After the sequential logic is satisfied, a second application of the concept is to apply cause–effect logic to determine if the events are causally related to each other. After causal relationships are established, application of necessary and sufficient logic to each related pair or set can be used to determine the completeness of the investigation or description of the occurrence.

The event pairing also enables investigators to define gaps in the occurrence description, or any uncertainties associated with those events. That in turn enables investigators to integrate each new data item into the existing event patterns and gaps as data are acquired, as shown in Fig. 5.

Event pairs are also used to compare what happened with what was expected to happen, as part of the problem discovery and definition investigative subprocess. Another use is for identifying and assessing performance improvement options, and preparing plans for monitoring implementation of new actions.

By "thinking events" and using progressive analysis methods, investigators can accelerate the investigation and reduce data-gathering burdens.

Event Linking. An event link is a representation of a cause–effect relationship between two events. The orderly sequencing of events found during the investigation generates the evolving description of what happened. To understand why events happened, the investigator needs identify and document rigorously and completely the cause–effect relationships among all the relevant the events. This task rests on the event linking concept. In practice, links are arrows on documents showing the cause–effect relationships between the earlier and later events. By convention, links lead from the triggering event to the triggered event.

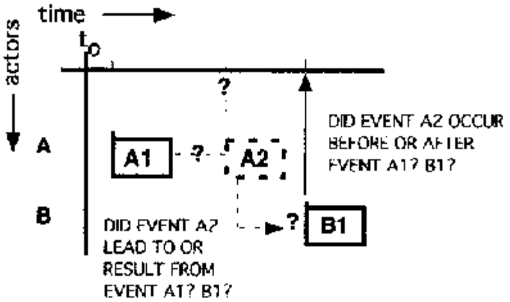

time
actors
t0
?
A
A1
?
A2
DID EVENT A2 OCCUR BEFORE OR AFTER EVENT A1? B1?
B
DID EVENT A2 LEAD TO OR RESULT FROM EVENT A1? B1?
?
B1

Figure 5 Sequencing new events. As new data defining event A2 become available, the investigator can assure its proper sequencing by determining where it should be placed on the time–actor matrix relative to other known events. (From K Hendrick, L Benner. Investigating Accidents with STEP. New York: Marcel Dekker, 1986, p 135.)

To establish links, the investigator considers each potentially relevant event in pairs or sets, to decide whether or not they have a cause–effect relationship. If one had to occur to produce the other, the investigator links the events to document that relationship. If the causal relationship is not direct but through another event, that third event (or a "?") is added to the set. If the original events in the pair have no cause–effect relationship, no link is added, and one or both of the unlinked events may be irrelevant (Fig. 6).

The linking concept provides a way to display logical cause–effect relationships for each event that is identified. It also provides a way, with the question marks, to:

Progressively incorporate relevant events into the description of the occurrence *as each is acquired.*
Identify completed data acquisition tasks.
Identify *unfinished* investigation tasks.

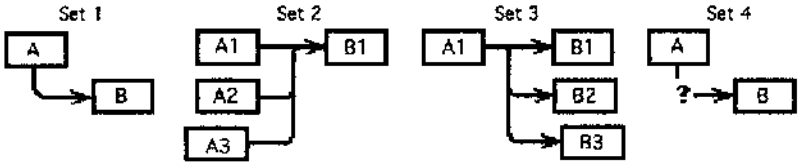


Figure 6 Linked events sets. Set 1 represents two events with a direct cause–effect relationship. Set 2 represents three events (A1, A2, A3) that will produce B1 every time they occur. Set 3 represents one event that will lead to three other events. Set 4 represents two events for which a causal relationship may exist. The "?" represents an unfinished investigation task. (From 10 MES Investigation Guides, Guide 2, Worksheets. Oakton, VA: Ludwig Benner & Associates, 1998, p 4.)

Define specific *remaining data needs* and acquisition tasks or workload.
Control expenditures of more time or money to get missing data.
Filter irrelevant or unlinked data from work products.
Show users uncertainties or unknowns at the end of an investigation.

An ideal investigation will produce a description of the occurrence that consists of all interacting or linked events, and only those which were necessary and sufficient to produce the outcomes. Anything less indicates an incomplete description of the occurrence. Anything more will almost certainly raise unnecessary questions.

Energy Tracing. This concept is also based on Johnson's MORT safety research [6]. His point was that energy is directed by barriers to do desired work. When barriers do not successfully direct the energy to its work target, the energy can do harm to vulnerable targets. These events are part of the automated system or robotics accident or incident process. Energy produces the changes investigators see in objects or people. Tracing energy paths and flows to find what produced the observed changes helps investigators explain "how did what you see come to be?"

Energy flows leave tracks of varying duration. To trace energy flows the investigator's challenge is to find those tracks or changes that resulted from the energy flow. This energy tracing can be done in a sequential way, from the time the energy enters the system until the energy has produced the work that can be observed. "Energy" should be viewed broadly, ranging from the readily identified electrical and mechanical categories to people inputs, for example [8]. It can also be a more obscure energy such as gas generated by bacterial action, temperature changes and oxygen that rusts iron. See Appendix B for a thought-starting list of energies observed by the author during investigations over a 20- year period [9,10].

Each energy form is an actor that is tracked through the system to identify any harm that it did, and any constructive work or control it brought to the system during the occurrence.

The concept also has the effect of requiring an understanding of the system in which the energy flows. Systems are designed to constructively direct energy flows with barriers. Thus the investigator needs to find out what energies might have affected the system, the barrier behaviors, and the harmful work that was done, and also to trace any amelioration work that affected the interactions or changed the potential outcome. The orderly tracing of energy flow backward from the harm produced often helps define the system, if it has not been defined before the occurrence. That is not unusual, and is why investigating minor occurrences is usually so valuable.

Witness Plates. This concept was adapted from the explosives testing field. During field tests, metal plates positioned all around an outdoor explosion bore witness to work done on them by objects and energies released when the device was exploded. Experts then interpreted the changes to the witness plates to analyze what acted on them during the explosion.

The concept defines the process for "reading" events on objects after an occurrence. It applies the energy-trace principle to investigation, in that energy which does work during occurrences leaves tracks on "witness plates." Witness plates are the keepers of the tracks left by energy exchanges. This applies to both objects and people. By viewing both as witness plates or keepers of data about events that occurred, investigators respect the sources. They recognize that their ability to access the data depends on their own skills to acquire the data, more than the witness or object's ability to communicate their data to them. Thus the concept helps investigators maintain a constructive attitude about witnesses they interview, and objects they study in investigations.

Objective Investigation Quality Assurance. Objective quality assurance is the use of nonjudgmental criteria to assess the quality of an investigation and its work products. This concept results in *displaying* events, and using rigorous *logic tests* to assess the order, relevance and completeness of the description and explanation of the occurrence. It uses time and spatial sequencing of events to assure the proper ordering of events. It then uses cause–effect logic to assure discovery of relevant interactions among events. It then uses necessary and sufficient logic to assure the completeness of the ordered and linked events which describe and explain what happened.

The display enables the investigator to invite constructive critiques of the logic flow of the events constituting the occurrence. The demand to state the data and name the sources to justify any proposed additional events or changes to a flow chart disciplines experience-based experts who want to challenge an investigator, promote their interests, redirect plans, or create uncertainty for other reasons.

1.2.5.2 Investigation Principles

Study of many investigation processes has disclosed key principles which can help investigators produce superior investigation results. These generally applicable principles should be incorporated into investigation program plans for all kinds of investigations.

If You Can't Flowchart It, You Don't Understand It. This fundamental axiom is another contribution of Johnson's MORT safety research [6]. It is especially important when occurrences are perceived and treated as processes.

Flowcharting the process interactions that produced an unexpected and unwanted outcome has many benefits. One of the most important is the discipline it imposes on investigators to produce complete, consistent, valid, and credible descriptions and explanations of what happened. They must understand the sequence, cause–effect relationships, and the necessity and sufficiency of all documented interactions during the occurrence to be able to prepare a valid flowchart.

A second and equally important reason for flowcharting occurrences is the *visibility* the flowchart documentation provides for the events and the logic of their relationships. That visibility provides a convenient mechanism to organize and analyze data as they are acquired. It enables everyone associated with an occurrence or its investigation to pool their data into objective, logical, and disciplining patterns. It helps filter out questionable or extraneous data. Additionally, flowcharts provide an abbreviated record of the occurrence to share with affected personnel for training, retraining, process or equipment design, performance monitoring, or for monitoring the effectiveness of changes recommended by the investigator. Also, flowcharts of such processes can be archived and retrieved readily from corporate memory for future applications, which is a major consideration for building corporate memories.

For investigation managers, flowcharts provide instant information about the current status of the investigation. If flow charts are developed as the data are acquired, gaps help managers pinpoint what data are still needed, and what they might gain if they get the data. Investigators have a tendency to want to eliminate every possibility to arrive at the most likely possibility. With flow charts, managers can make informed decisions about the value of expending more investigation resources.

Track Change Makers. Process outcomes result from changes introduced by people and objects during the occurrence. Therefore, investigators have to focus on the *change makers* that produced the outcomes. Some people and objects are just along for the ride, while other people or objects shape the outcomes. Investigators must look for and identify the people and objects that shaped the outcome, and show those interactions. By starting with the outcomes, and working backwards, investigators pursue the change makers in a logical sequence.

Focusing on change makers or "doers" leads to efficiencies in investigations, by minimizing the amount of time spent on irrelevant people or objects. This mind set reflects the "think events" concept. This is one of the key secrets to achieving efficient investigations.

"Do No Harm" Rule. Introducing changes to people and objects that survived the incident before you capture their data can corrupt an investigation. Thus the "do no harm" rule. Investigators must prevent any change in data sources until they have extracted the data needed from those sources.

This rule poses difficult challenges for investigators. For example, rescue workers usually must disturb some witness plates to effect their rescue. Investigators can walk on debris and change it as they try to get closer to another object to observe its condition. Investigators may try to start something or turn it on to see if it works when they get there. They shut down power to immobilize a remote controller cabinet, and lose stored data in volatile memory chips. How can essential data be preserved in these circumstances?

The answer is to make plans to prevent loss of data, and establish control over the site of the occurrence to prevent as much change as possible. Control of changes at the scene of an occurrence increases in difficulty as the size of the scene or accessibility delay increases. This is particularly important when trying to control the people and objects at the site of a large occurrence, or when the investigator may arrive at the site later. A site involving large or dispersed equipment such as a gantry robot is more difficult to control that a small single station robot site, for example. The rule reminds investigators of the importance of an on-site mental assessment of the risks to data stored in people and objects before introducing new changes that can harm the data.

Time Never Stands Still. Time is an *independent variable* during an occurrence. Every person and every object has to be someplace doing something during an incident. What they do is identifiable by when they did it, and how long it lasted. Each action during a process has a starting and an ending time. Time is used to order events data as they are acquired. Investigators should be concerned with establishing or at least roughly approximating the relative times when people and objects did something to advance the occurrence to its conclusion or outcome. Creation of a mental movie helps investigators do this.

The principle is applicable directly during interviews of people. By trying to visualize what the witness was doing from the time the witness first became aware of the occurrence, investigators can develop a "time line" of actions by the witness. Whenever a person or any object drops out of sight during the occurrence, the mental movie helps to pinpoint needed data and questions to ask.

Meeker's Law. *"Always expect everyone to act in what they perceive to be in their best interests, and you will never be disappointed"* [11]. Sometimes investigators have to deal with people who were actively engaged in the operation or process that went awry. For many reasons, they may perceive that it is their best interest to withhold some information from the investigator, or mislead an investigator, or perhaps even lie. Investigators should be aware of these perceptions of self interest, and be prepared to work around them. One way is to use the mental movie to assess the completeness and sequential logic of the actions described. Another is to document and display the events reported on a flowchart, and test their logic. Another way is to get corroborating or contradictory statements. Trust but remember the perceived interests and verify what is reported.

The Silent Witness Rule. *The witness has it, you need it, and the witness doesn't have to give it to you.* Investigators are at the mercy of people who have in their memory the data they need. A companion to the self-interest principle, this principle helps investigators adopt a helpful frame of mind for talking to witnesses about an occurrence. It reminds investigators to look for, recognize, and adapt to the perceptions, interests, and motivation of each witness. They adapt by framing the purpose of the interview and all questions in a way that encourages each witness to share data the investigator needs. Ideally, successful investigators are able to transfer the witness' mental movies of the occurrence to their minds. An investigator's challenge is to get the witness to do 95% of the talking during an interview.

Things Don't Lie. For many reasons data acquired from people are less reliable than data acquired from things. Things respond predictably to energy exchanges, according to laws of nature that enable prediction of changes in things. The value of this predictability is that investigators should rely on the most reliable data—derived from objects—to determine what happened.

While they do not lie, objects are not ardent conversationalists. Thus it is up to the investigators to extract whatever data might be stored in things. To read data from an object, it is necessary to know the state of the object both before and after an occurrence, the changes that occurred during the incident, and the energies that changed it. This means capturing and documenting the ending state promptly and efficiently is an investigation priority.

Data from objects become critically important when nobody was around during the occurrence, or when those who saw what happened did not survive the occurrence.

Experience Recycles Yesterday's Problems. Rationalizing experiences has the subtle but real capacity to normalize deviations or changes that increase risks or produce degrading performance or accidents [17]. The importance of this principle lies in the need to select investigation methods that prevent experience from leading to conclusions contrary to those demanded by the data. It also means that the selection of investigators must carefully balance their experience against their ability to subordinate it to logical thinking about the data they develop during their investigations. MORT training cautions investigators not to SLYP or solve last year's problems. Mental movies and flowcharts help prevent this.

This is another reason why primary reliance on *investigation* knowledge and skills rather than system knowledge and skills is so important in good investigation programs.

Investigations Are Remembered by Their Results. Investigations are meaningless and a waste of resources unless they contribute to timely and *enduring* change. Loss incidents have a way of bringing about temporary changes in behavior and views, even without any investigation. The challenge for any investigation program and every investigator is to produce work products leading to lasting improvements and retention of the understanding achieved. Retention is best achieved with brief, readily grasped descriptions of what happened, with obvious and broadly applicable principles that can be applied in many situations.

Given these concepts and principles, what procedures will produce the desired investigation work products?

1.2.5.3 Investigation Processes

Investigation processes traditionally reflected the intuitive understanding of investigations by individuals performing the investigations. That is changing as alternative investigation methods have become available, starting with the MORT research around 1973 [6].

When considering alternative investigation processes, several precautions are advisable. These precautions include tailoring the investigation program to the needs and capabilities of the organization. In considering a selection, it is advisable to be aware of desirable capabilities and attributes to seek, as well as attributes that may impose constraints or create problems. Selection of an investigation program methodology should match the capabilities demanded by the favored choice(s) and the capabilities that can be made available within the organization.

The following summary of criteria can assist in the task of selecting the investigation process.

Preferred Capabilities and Attributes. A preferred investigation process [12] for implementation under the program plan can:

Provide investigators with guidance about what to observe and how to frame questions.

Help investigators organize and document data they acquire promptly and efficiently.

Give investigators real-time guidance for narrowing their data searches during the investigation (progressive analysis capability).

Facilitate sequential, cause–effect and necessary and sufficient logic testing of the data documented.

Help investigators recognize and act on unknowns.

Define problems, needs, and candidate remedial actions logically and objectively.

Assist in the assessment of needs and candidate remedial actions, and prediction of their success.

Point to ways to monitor actions to evaluate their success.

Expedite preparation of valid and persuasive deliverable work products.

Mediate differing viewpoints and guide their resolution.

Adapt to the full range of occurrences likely to be encountered.

Be learned and practiced at modest cost.

Filter quickly any extraneous data during investigations, without alienating other investigators.

Prevent investigators from drawing conclusions contrary to the data.

Minimize dependence on experience and maximize dependence on logical reasoning.

Facilitate the objective assessments of the investigation process and output quality.

Attributes of Less Desirable Processes. Less attractive investigation processes also have some distinguishing attributes, including:

Informal and very experience-dependent procedures

A legally oriented facts–analysis–findings–conclusions framework

A high tolerance level for ambiguous and abstract data usage, experiential assertions, built-in judgments, and subjective interpretations and conclusions

Oversimplified descriptions and explanations of what happened, with recurring jargon such as chain of events, unsafe acts, human error, failures, fault, and the like

An emphasis on finding a single ‘‘golden bullet’’ to explain the occurrence such as ‘‘the cause’’ or the root cause or equivalent

A lack of scientific rigor or disciplining procedures demanded of investigators, such as time-disciplined demonstration of relationships

Lack of objective quality control criteria and procedures for the outputs or the process.

An understanding of these concepts and principles provides a basis for developing an investigation program plan tailored for a specific organization.

1.3 INVESTIGATION PROGRAM PLANNING

This section describes the main investigation program-planning tasks. The operation of an effective investigation program depends on the design of the program and readiness of four primary program elements: executive commitment, a sound investigation plan, adequate investigator preparations and competent investigation support. The main investigation program planning decisions and actions are summarized in Fig. 7.

Executives are responsible for an organization’s overall performance, set policies, and allocate resources to achieve desired performance. The

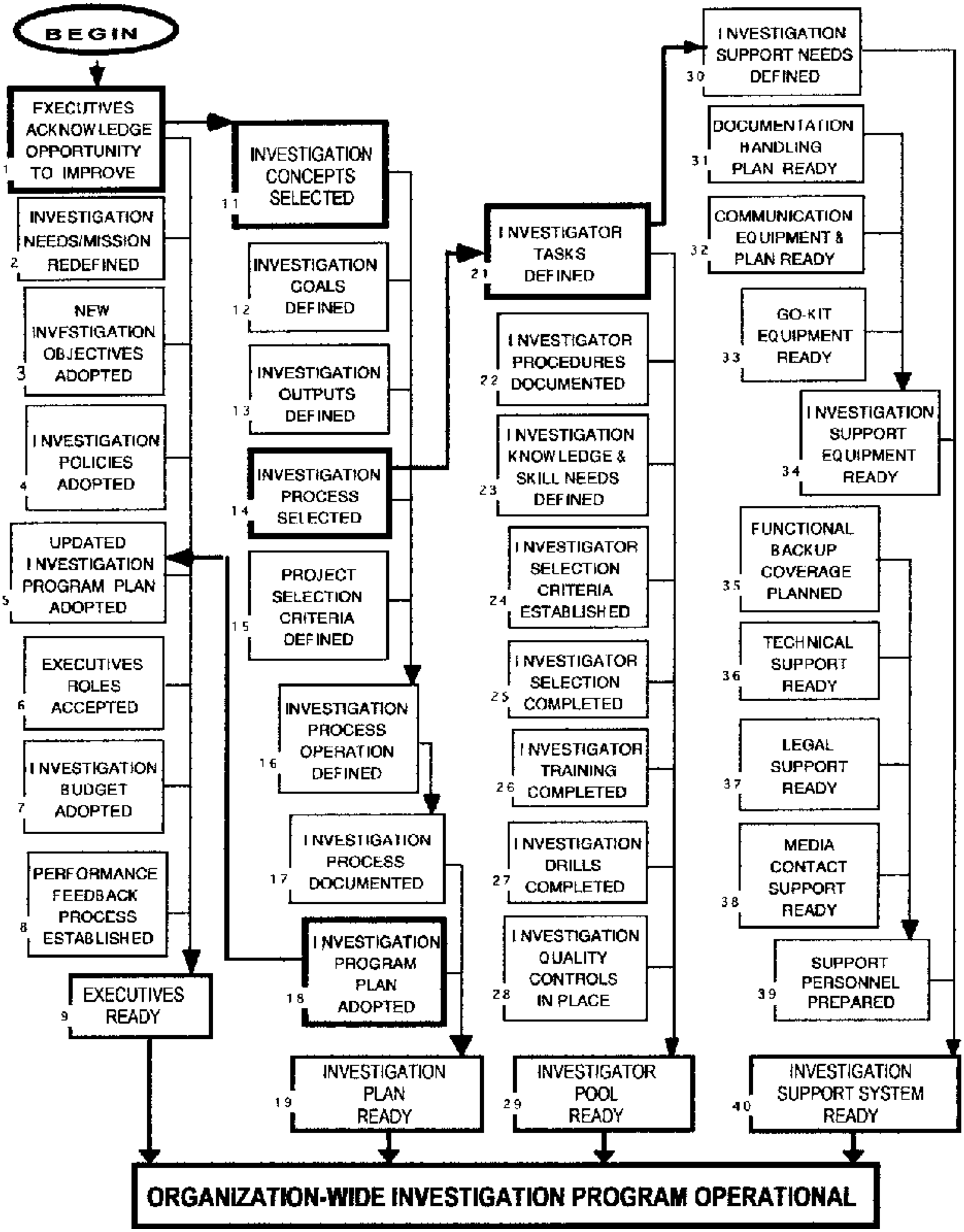

BEGIN
EXECUTIVES ACKNOWLEDGE OPPORTUNITY TO IMPROVE
1
INVESTIGATION NEEDS/MISSION REDEFINED
2
NEW INVESTIGATION OBJECTIVES ADOPTED
3
INVESTIGATION POLICIES ADOPTED
4
UPDATED INVESTIGATION PROGRAM PLAN ADOPTED
5
EXECUTIVES ROLES ACCEPTED
6
INVESTIGATION BUDGET ADOPTED
7
PERFORMANCE FEEDBACK PROCESS ESTABLISHED
8
EXECUTIVES READY
9
INVESTIGATION CONCEPTS SELECTED
11
INVESTIGATION GOALS DEFINED
12
INVESTIGATION OUTPUTS DEFINED
13
INVESTIGATION PROCESS SELECTED
14
PROJECT SELECTION CRITERIA DEFINED
15
INVESTIGATION PROCESS OPERATION DEFINED
16
INVESTIGATION PROCESS DOCUMENTED
17
INVESTIGATION PROGRAM PLAN ADOPTED
18
INVESTIGATION PLAN READY
19
INVESTIGATOR TASKS DEFINED
21
INVESTIGATOR PROCEDURES DOCUMENTED
22
INVESTIGATION KNOWLEDGE & SKILL NEEDS DEFINED
23
INVESTIGATOR SELECTION CRITERIA ESTABLISHED
24
INVESTIGATOR SELECTION COMPLETED
25
INVESTIGATOR TRAINING COMPLETED
26
INVESTIGATION DRILLS COMPLETED
27
INVESTIGATION QUALITY CONTROLS IN PLACE
28
INVESTIGATOR POOL READY
29
INVESTIGATION SUPPORT NEEDS DEFINED
30
DOCUMENTATION HANDLING PLAN READY
31
COMMUNICATION EQUIPMENT & PLAN READY
32
GO-KIT EQUIPMENT READY
33
INVESTIGATION SUPPORT EQUIPMENT READY
34
FUNCTIONAL BACKUP COVERAGE PLANNED
35
TECHNICAL SUPPORT READY
36
LEGAL SUPPORT READY
37
MEDIA CONTACT SUPPORT READY
38
SUPPORT PERSONNEL PREPARED
39
INVESTIGATION SUPPORT SYSTEM READY
40
ORGANIZATION-WIDE INVESTIGATION PROGRAM OPERATIONAL

Figure 7 Organization-wide investigation program readiness tree displaying the preparatory steps needed to assure readiness of a comprehensive investigation program.

are the program's sponsors, and must be committed to and satisfied by the program.

Investigation program planners are responsible for the determining the investigation tasks investigators will perform. They are the program's creators. Their plans must be tailored for the organization, and capable of producing the desired results.

Investigators and their supervisors are responsible for producing satisfactory deliverables within the program. They are the program implementers, and their work must satisfy their sponsor and their customers.

Persons or groups who support investigators provide knowledge, advice, and support for the investigators. They are program auxiliaries.

Investigation program readiness decisions and actions are shown for each group. Executive decisions and actions (blocks 1–9) define what the program is expected to accomplish. Investigation program planning actions (blocks 11–19) define how investigations are conducted, and what they deliver. Investigator selection training, and practice (blocks 21–29) lead to the investigation capability that will produce the desired work products. Preparation of support personnel (blocks 30–40) provides a needed "resource pool" to help investigators when they need it.

1.3.1 Executive Preparations

Executives set the direction and tone of an organization's activities. They also control the organization resources and their distribution. Investigations consume resources—sometimes twice: once when they are conducted, and a second time if the investigation is flawed and undesired surprises continue. Success of an investigation program depends on engaging executives and getting their sponsorship of the program by showing them their stake in its success.

The following actions by executives are required to get a successful investigation program underway, and to keep it going. The numbers in parentheses at the end of the task title refer to Fig. 7, the organization-wide investigation program readiness tree.

1.3.1.1 Acknowledge Opportunities (1)

This is the indispensable first step. Executives must be able to recognize the narrowness and shortcomings of conventional approaches, and why those approaches do not satisfy their efforts to continually improve performance. Upon recognizing that need, they then need to recognize that new opportunities are available to them to achieve better results. If they understand these opportunities, they will want to take advantage of them, and will be more receptive to new approaches.

1.3.1.2 Define Mission, Purpose, and Demands (2)

The opportunities enable desires for continuing improvement to become the basis for revising the investigation program mission and purposes. Rather than a narrow accident prevention mission, everyone can endorse the broader mission of facilitating continuous performance improvement. This will establish the performance demands for the investigation program. After an executive decision has been made to acknowledge and seize opportunities to improve investigation programs, the investigation program planning begins.

1.3.1.3 Establish or Update Investigation Program Objectives (3)

Establish objectives for each kind and level of investigation, such as:

Efficiently and consistently produce timely, valid, and consistent descriptions and explanations of the occurrence being investigated.

Report that new information in a form facilitating its use throughout the organization to discover and define specific needs for change, and identify and assess candidate changes to improve future performance.

Provide a basis for monitoring in real time the effectiveness of predictive analyses, and changes implemented as a result of investigations.

Do all this in a constructive, harmonious manner.

If the present investigation program has narrower objectives, establish new broader objectives for the program plan.

1.3.1.4 Adopt Investigation Policy Changes (4)

When executives are comfortable with the program objectives, they need to review the organization's investigation policy. If new investigation policies are needed, they should amend current policies. Changes should address the investigation program mission and goals, particularly regarding the tone of investigations. Determination of what happened and why it happened, and using that understanding to improve future performance should be common to all policies. Policy *changes* require executive acceptance and support.

One element of this task is to ensure that the policy is compatible with regulatory requirements. Another element is to communicate the policy and need for co-operation with investigators to everyone in the organization who might become involved in investigations.

1.3.1.5 Adopt Updated Investigation Program Plan (5)

When the investigation program plan is ready it should be considered, accepted, and advocated at the executive level of an organization. By advocating the plan, the executives show their support for it. They also become the program's sponsor. The program operation must satisfy the sponsors, or they will abandon it.

1.3.1.6 Accept Executives' Roles (6)

The investigation plan should incorporate support roles at the executive level. Support roles include participating in periodic program performance reviews, in leading high-profile investigations that might affect the public's perception of the organization, and in the resolution of difference affecting the levels of predicted residual risks accepted. These roles should be accepted by executives who will be involved in these tasks from time to time.

1.3.1.7 Ensure Investigation Budget (7)

If an initiative is worth undertaking, the organization should be prepared to pay a reasonable price to gain the benefits it expects to receive. By setting a budget for the investigation program, the value of the program is established, and one measure of its performance is put in place. The source or sources of the funds are less significant that their allocation to investigations. This can have a positive effect on investigators, who will become conscious of the need to demonstrate the value of their work. It also encourages investigation efficiencies. Caution should be exercised to avoid creating disincentives that penalize anyone via the budgeting process.

1.3.1.8 Establish Investigation Performance Feedback Process (8)

Periodic review of any function is an essential element of good management. If the broad mission for an investigation program is adopted, the suggested objectives provide a basis for assessing the program's achievements and *value*. A concomitant objective is to change or terminate the program if it is not achieving its objectives.

1.3.1.9 Executives Ready (9)

The importance of these executive-level tasks cannot be overstated. If the above actions are taken, the organization's executives will be ready to support the program and perform their role in achieving the desired benefits.

1.3.2 Investigation Process Plan

The best investigation plan for each specific organization should be identified, prepare, and implemented. Planning tasks to achieve this include selecting, adapting, and implementing an effective investigation process.

1.3.2.1 Select Investigation Concepts (11)

Selection of the conceptual framework for an investigation program is probably the second most important decision for ensuring an effective program. Criteria for program selection are applied during this task.

What governing framework should be adopted? A review of references is strongly advised [2,3,5,6,13]. Should adoption of the change-driven process model and event data language concepts be the governing framework? Or should the concept of determining cause and unsafe acts or unsafe conditions in a chain of events be chosen? Or would the energy/barrier/target MORT concept be most desirable for the organization? Use the criteria cited earlier during these deliberations, and document the reasons for the selection for later use.

1.3.2.2 Define Investigation Goals (12)

Depending on the investigation policy and the framework selected, the specific goals of an investigation are defined next. Goals of any investigation should include development of a validated understanding and explanation of what happened. Other goals are suggested by the discussion above. Document the goals selected.

1.3.2.3 Define Investigation Process Deliverables (13)

Plans should define the work products to be delivered. Plans should also include criteria by which each work product will be evaluated during the investigations, and quality assurance procedures. Deliverable work products include the description and explanation of the occurrence, source document files, visual aids and other materials required to support the description, presentations or briefings, and any reporting forms or other documentation. If desired, deliverables may also include reported needs disclosed by the investigation, and proposed changes to improve operations. Planners must consider the regulatory requirements and their effects on criteria for the deliverables. It is advisable to review the potential adverse as well as beneficial consequences of each with legal counsel before the plan is adopted.

Another important deliverable, particularly when potential injury claims or litigation are foreseeable, is the source material used to define the reported actions, consisting of documents, photos, selected debris, sketches, or other source materials. Plans should address their acquisition, handling, archiving, and disposal.

Provide specifications for deliverables expected from investigations of incidents such as breakdowns, disruptions, or similar problems. If descriptions of the occurrences are not documented, the problems will probably return. When forms are required, provide examples of entries that satisfy the form's designers.

1.3.2.4 Select Preferred Investigation Process (14)

Alternative investigation processes are available for review [2,3,6,10,13–15]. Planners should document the criteria for the investigation process selection as part of the evaluation process. They will be used in subsequent evaluations of the process. The criteria should include those described above and any additional criteria needed to tailor the process to the organization's capabilities. Each candidate process should be evaluated against each criterion and the results compared, and documented for later review.

1.3.2.5 Define Case Selection Process (15)

What occurrences should be investigated? In a narrowly focused program, case selection is limited to those required by regulatory authorities, usually involving. a reportable or lost-time injury or fatality. In a broad program, the emphasis is on investigating any surprises, especially those during which a worse outcome was averted by successful intervention of people or object controls.

Case selection can be made automatic. For example, when an automated system breaks down or functions erratically, the troubleshooting and repair are a form of investigation that is initiated automatically. When disruption of a production process occurs, investigations are also preuthorized. Preparations should include identification of the kinds of occurrence that are investigated automatically and those for which a case selection decision will be required.

Criteria for determining the scope of each kind of investigation should also be prepared to guide investigators in specific cases. These specifications should address at least the extent of the scenario to be developed, hours to be committed, and the investigation deliverables, among other specifications.

1.3.2.6 Define Investigation Operations (16)

This aspect of the investigation plan should address administrative matters. That includes guidance for the assignment of investigators, record keeping, and administrative requirements, including time and expense record keeping; also notification of and co-ordination with others who have requested notification, including regulatory agencies if applicable. Develop requirements for others in the organization to co-operate with investigators, and output reviews, distribution procedures, and any other specifications needed to tailor the plan for the organization. Do not overlook guidance for investigators when they suspect a crime rather than an unintended occurrence.

1.3.2.7 Document Investigation Process (17)

This section of the plan documents the selected investigation process. It outlines guidance for acquiring data from people and objects; for handling those data after they are acquired; for validating interactions or bridging gaps; for problem discovery and definition; for assessing support needs and capabilities; for quality assurance procedures; for distribution of work products; for media contacts; for consultations with counsel; for self or peer assessment of investigation performance; and any other tailored elements of the selected process. References provide detailed descriptions of investigation tasks [10,16].

1.3.2.8 Adopt Investigation Plan (18)

This step is the final co-ordination step. Each person affected by the plan reviews it, confirms the intended operation and benefits, demonstrates any difficulties it might present, helps modify it if so required, and commits to it with a sign-off indicating concurrence. This step typically involves selected executives, managers of activities that might be investigated, and the likely investigation program manager or senior or lead investigators who will implement the plan.

1.3.2.9 Investigation Plan Ready (20)

Plan future tasks needed to maintain plan readiness in the future. This may require periodic review and tune-up of the plan, by using performance indicators to identify the needed modifications or updates.

1.3.3 Investigator Preparation

1.3.3.1 Define Investigator Tasks (21)

What are investigators expected to do during investigations? The investigation process selection decision determines investigators' tasks. If a formalized process is adopted, descriptions of various useful techniques can be identified from the literature [2,3,6,9,10,13–20]. Each reference offers different ideas about investigation processes and techniques. Review those ideas and techniques to determine if they should be incorporated into the plan. For example, tailoring might be needed because of special classes of robotics equipment. Tailoring might also be needed because of special regulatory considerations, or because of the nature of the organization or its policies. If so, descriptions of techniques in the literature should, of course, be modified to accommodate those special needs. Task definitions need not be complex but should provide sufficient detail to prevent inconsistent overall performance for different levels of investigation.

Consider investigation task definitions to be dynamic, subject to change whenever circumstances change. For example, introduction of new products or new materials, expanded processes or any other changes affecting investigation performance might precipitate a review and modification of the investigation procedures.

1.3.3.2 Document Investigator Procedures (22)

This planning task requires documentation of the orderly sequencing and execution of investigation tasks to provide guidance for investigators, and to accommodate the nature and levels of anticipated investigations. Investigation procedures might range from troubleshooting to major accidents. Planning for investigations of serious accidents should have priority. Include a "walk through" of each kind of investigation during planning, to determine personnel involved in the investigation, task interactions of these personnel, and the timing of the investigation tasks relative to each other. Project planning software can be a useful aid to this documentation process. Be sure to designate a supervisory process to accommodate decision flows during the investigations.

Pay special technical attention to procedures for acquiring data after an incident from digital electronic devices and applications which control automated systems. In transport systems, recording devices are used to keep track of selected operating system parameters during operations, providing the ability to determine their behavior if an accident occurs. Without such recorders, operators should identify proposed means to examine such devices for whatever data they might hold. If this situation arises, information about available investigation techniques is available through sources such as the International Society of Air Safety investigators [21].

1.3.3.3 Define Investigator Knowledge and Skill Needs (23)

Define knowledge and skills required to perform investigation tasks next, to ensure that investigators are capable of performing their assigned tasks. The knowledge needs to include an understanding of tasks and outputs required, and methods that will equip investigators to perform those tasks. The knowledge needs are outlined above, but a few are worth repeating. Absolutely essential are knowledge of observation processes and how to transform data into investigation data language, practical knowledge of logical reasoning procedures, an understanding of problems that certain kinds of words and language can create, and awareness of investigation concepts and principles described earlier. These apply to all levels of investigation.

1.3.3.4 Establish Investigator Selection Criteria (24)

Selection of personnel to qualify as investigators must be responsive to the needs dictated by the investigation process used. Generally, the more rigorously disciplined the investigation process, the more stringent the criteria for investigator selection. The level of investigations required also has a bearing on those criteria.

Basic criteria for investigators must include the ability to make observations without distortion or bias; transform and document investigation data with minimal interjection of personal views, experiences or personal values; order and present data sequentially; visualize patterns and scenarios; reason logically; have a reasonable memory for detail; and be self-critical without personal constraints. Physical capabilities should be consistent with the tasks and environment in which the investigator will be working. If special knowledge or skills will be required because of the nature of the systems being investigated or the levels of investigation, include these criteria.

1.3.3.5 Complete Investigator Selection (25)

The interviewing and selection of investigators contribute to the subsequent tone of the activities and expectations of both the employee and supervisor. When describing the duties of the position, discuss the mission and policies governing investigations, particularly with respect to regulatory agencies. Also, resolve any questions about the levels of investigations, their authority and responsibilities, acceptability of their work products, and how their performance will be judged.

Investigator selection might include selection of outside experts to assist in or direct Level 4 or 3 investigations under contract. The anticipated frequency of major investigation cases may be too low to justify development and maintenance of an investigation team within the organization. If so, the selection of an outside team to perform the investigations under contract might be desirable. Select all contractors based on their ability to perform the investigations as prescribed by the adopted investigation program plan criteria.

1.3.3.6 Train Investigators (26)

Typically, selected investigators will have systems expertise in some form. They will probably not have formal investigation training, so plans should include training in the selected investigation process and techniques before they are assigned a case. For level 1 investigations, an apprenticeship to trained investigators might offer sufficient training. For other levels, consider providing classroom investigation training. Design the training to accommodate the mission, policies and plans, the tools provided for investigations, and tasks required to achieve the desired outputs. Ensure training in quality assurance tasks.

1.3.3.7 Complete Investigation Drills (27)

Include investigation drills for investigators in the program. During training, drills simulating hypothetical investigations or case studies can be developed for specific operations to strengthen the investigator thought processes. It is also good practice to have newly trained investigators do an independent quality assurance check of a recently completed case report. Have them use flowcharting tools and the quality assurance checks in Appendix C to build their skills in applying the desired thought processes. Use feedback about the results to reinforce any investigation process elements that need attention.

A more sophisticated drill to give new investigators insights into and experience with the kinds of problems that arise during an actual investigation is desirable. Have them investigate the breakdown of an automated system to determine what happened and why it happened. This has the added advantage of introducing operational personnel to the thought processes used in investigations.

A third kind of drill is to rotate the quality assurance of work products among investigators. This enables investigators to keep abreast of findings by others, while continually improving their data gathering, analysis, and logic skills.

1.3.3.8 Put Quality Controls in Place (28)

Quality control procedures should be put in place to assure the adequacy of the deliverables, the quality of actions implemented as a result of the investigations, and the quality of the investigation process itself.

The quality of the deliverables and findings can be screened before they are distributed. Use the quality assurance check lists described in Appendix C. References provide detailed quality assurance instructions [9,22]. Tracking acceptance of the findings provides another indicator of deliverable quality [4d]. The greater the controversy or reluctance to accept the findings, the greater the likelihood that the quality of the investigation deliverables needs to be improved.

The quality of the actions proposed as a result of the investigation is also assessed by the results they produce. Before they are released, any recommended actions should predict the intended effects in a way that can be verified. Quality control plans should address the monitoring and verification tasks.

The quality of the investigation process is assessed by occasional spot auditing of the procedures at the conclusions of an investigation. It is also indicated by the performance against budgets, nature of any delays, handling of uncertainties and unknowns, complaints about the investigators or their actions, and the work products produced. Quality control plans should address this need.

1.3.4 Investigation Support Preparations

Investigators may need various kinds of help, particularly if the investigation involves casualties.

1.3.4.1 Define Support Needs for Organization (30)

Plans should identify the kind of help that may be needed during investigations, and assure that it will be available promptly when needed. The help required may include functional backup for collateral duty investigators, document handling, communications, go-kit maintenance, investigation equipment, readiness monitoring, technical expertise, legal expertise, and media expertise. Technical expertise may include electronics, mechanical, chemical, medical, or design experts.

1.3.4.2 Establish Data Source Handling Plan (31)

During investigations, sources from which investigators extract data increase in number as the investigation progresses. Address how data sources are to be documented, processed and preserved during and after investigations, including possible chain-of-custody implementation tasks in specified cases. If willful harm is involved, consult counsel for planning guidance.

1.3.4.3 Communications Protocols and Equipment Ready (32)

Depending on the occurrence, voice or data communications links may be required at the site of an occurrence. Preparations should be responsive to any anticipated special requirements for such communications. For example, consider communications protocols, security concerns, interfacility electronic data transfers or exchanges, external data source access with suppliers, and incompatibility with facility communications or control equipment.

1.3.4.4 Complete Go Kit Preparations (33)

Go-kits are the investigators' transportable tool kits, containing the equipment an investigator will need on arrival at the site of an occurrence. At a minimum, preparations should specify the equipment to be provided, and its regular maintenance or updating. A camera and film, sketch pad, voice recorder, personnel protective equipment, and note-taking equipment are absolute minimum equipment. A list of special instruments and special experts, with contact information, is also a must. Planning should keep in mind that the investigator must be able to transport the go-kit, usually without help.

1.3.4.5 Confirm Support Equipment Readiness (34)

Surprises occur irregularly and hopefully infrequently. Thus preparations should address how support equipment will be maintained in a continuing state of operational readiness, including assignment of task responsibility and perhaps checklists and inspection schedules.

1.3.4.6 Arrange for Functional Backup (35)

Investigations interrupt the normal duties assigned to investigators except in large organizations where investigation might be a full-time job. When investigations occur, preparations should provide for backup personnel to take over the investigator's normal duties when the investigator is at the occurrence site or elsewhere doing investigation tasks.

1.3.4.7 Prepare Technical Support Personnel (36)

Technical support personnel are likely to be required when an investigator encounters a need to understand how something was designed to work, or how it actually worked. For example, if a tear down of an automated system or component is needed to examine internal parts after an occurrence, the investigator may need help in planning the tear down sequence and methods to minimize data loss. A tear down will probably require skilled mechanics or engineers. Expertise may be required to recover data or settings from electronic control components. Planners should identify and provide for access to in-house expertise, supplier expertise and contractor expertise.

1.3.4.8 Prepare Legal Support (37)

When people are harmed in any way, and sometimes when warranties are involved, investigators may need help from a legal expert or counsel. Planners should work with counsel to identify when investigators should seek legal advice, and how to access and use it.

1.3.4.9 Prepare Media Support (38)

Occasionally occurrences may precipitate media interest, particularly if fatalities occurred, or regulatory agencies become involved. Recognize that the media tend to focus on controversy. Plans should establish procedures and contact points for responding to media inquiries, and for adequate guidance and compliance with those procedures.

1.3.4.10 Support Personnel Prepared (39)

The planning should include some scheduled feedback arrangement, to accommodate changes in personnel, personnel assignments, housekeeping, equipment obsolescence, expired supply dates, etc. Plans should assign feedback and review tasks to whoever manages or directs the investigation program.

1.3.5 Monitor Startup of Investigation Processes

Any change requires planning before implementation, and monitoring after implementation. If a new program is initiated, or changes in investigation practices are instituted, performance monitoring and feedback to the executive in charge of the investigation program should be provided frequently during startup, and periodically during subsequent operations.

After a predetermined time, monitor the program for its achievements, and distribute periodic reports of its achievements.

1.4 CONDUCTING INVESTIGATIONS

This section outlines how to conduct investigations involving automated systems, and the basic tasks common to all kinds of investigations. Each occurrence is different, but this general guidance is applicable in all investigations. For detailed investigation task guidance, consult the references.

Before beginning, remember a basic axiom of investigation: *if you can't flowchart it, you don't understand it*. Thus, as an investigation progresses it is good practice to work toward developing a flowchart of events constituting the occurrence. With automated systems, the relative timing of events involved with the system controls often become critical to understanding what happened [23].

1.4.1 Initial Response to Notification

The first information about an occurrence is usually very sketchy, especially if any casualties are involved. However sketchy, it is necessarily the basis on which the decision is made to launch an investigation based on case selection guidance. Delaying a response often raises the risk of losing data or increasing the loss. Preauthorized automatic launches can begin immediately. If the responsible manager decides to launch an investigation, the manager assigns the investigator, with initial instructions about its specific goals, concerns, resources available, and deliverable schedule. The manager also implements the plans for backup and support services, including site preservation assignments pending arrival of the investigator. The manager then initiates the planned contacts, to provide them the information requested or required by regulation or local law. The investigator or team goes to the site of the occurrence.

Early communications should also consider directions to retrieve data about the equipment or processes involved, or any earlier analyses of the operations, addressed to the custodians of those data. If deemed appropriate or necessary, those contacts can include directions to protect any data sources.

1.4.2 On-Site Tasks

On arrival at the site, the investigator has four priorities. They are to preserve the data at the site; to overview the occurrence setting and get a rough idea about what happened; to set priorities for data gathering; and, frequently, to restart the system.

1.4.2.1 Data Protection

Generally stated, the first task is to prevent inadvertent or deliberate changes to the ending state of objects surviving the incident, or to the memories of people who acquired data during the occurrence. This can be done by roping off or otherwise isolating the area or the equipment or the people. Alternatively, post guards until the sources have been examined and the data they offer has been acquired by the investigator.

1.4.2.2 Data Acquisition and Processing

The next task is to begin to acquire data. The task challenge is to develop data in a format that supports efficient and timely development of a description and explanation of what happened. Further details are provided in the investigation plan discussion and references.

Start with a "walkaround" at the site to get a general overview of what is there and what might have happened. During the walkaround task, the investigator begins to plan the order of the data acquisition tasks, setting priorities to examine or record the condition of any perishable data sources before they change. Priorities may also be required for examination or documentation of equipment that is to be returned to service to restore operations. If potential witnesses to the occurrence are going to be leaving the site, arrangements for acquiring their data also require priority attention.

A part of the walkaround task is to document the ending state of the objects found at the site. This can be accomplished with cameras, video camcorders, sketches, or drawings, or maps if debris is scattered. Notes describing what was photographed or sketched should be part of the documentation.

Investigators should be prepared to delegate some of these tasks, such as the witness contacts and scheduling, or photographing damaged or changed equipment, for example. Use support personnel freely when needed.

The acquisition of data from people and objects continues after the walkaround. The investigator's challenge is to identify the people and objects that contributed to the outcome, identify what they did, and document those actions. The order in which this is done varies with the occurrence, but generally it begins with asking the first people found at the scene for their observations. They can help the investigator identify other data sources like witnesses and victims, or equipment that they observed doing something. They can also describe changes to debris or other objects which they introduced or saw occurring.

When physical conditions of objects can be observed, investigators have to determine "how what you see came to be"—or who or what did what to produce the condition(s) they see. Detailed guidance for extracting action data from physical objects can be found in some references [3,19,16,20]. Sometimes investigators want to conduct tests or simulations to determine what produced the condition or attributes of an object. Before *any* testing is initiated, the investigator should insist on a test plan, which incorporates criteria for creating actor–action formatted data as a test output. A test plan should specify who will do what and in what sequence, to ensure needed data are not lost inadvertently [24,25].

The acquisition of stored data from digital or analog electronic devices associated with automated or robotics systems poses different challenges. The plans for acquiring these data should be followed. Exercise caution to avoid changing such data before it has been retrieved.

The next step is transforming observations of the data sources into descriptions of the actions that can be sequentially organized and analyzed. The best way to do this is to use the actor–action data structure. When listening to people describe what they observed, the investigator should be listening for and documenting data that describe who or what did what and when it happened, and documenting the data as events.

1.4.2.3 Restarts

Sometimes the desire to restart equipment or processes involved in the occurrence with minimal delay influences the investigation. When this desire exists, the need to identify what happened and identify the problems and new controls needed quickly influences the priorities for the investigation, and the resources devoted to the investigation. The need to take shortcuts and the consequences of doing so should be discussed with the operational and investigation managers, who are responsible for accepting the risks.

1.4.3 Data Handling Tasks

As data about interactions are acquired, the investigator can add newly discovered actions to the developing scenario, as "frames" to the mental movie, or events to the events worksheets. The data handling goal is to narrow the focus of the investigation continuously. A concurrent goal may be to assure compliance with chain of custody requirements. When other organizations are involved, ensure that the investigation plan is observed.

1.4.3.1 Event Linking and Testing

The technical goal of this task is to prepare a completed event flowchart that describes what happened and why it happened during the occurrence. Gaps in the mental movie or sequential layout of events point to specific questions that the investigator should pursue. Working the gaps narrows the focus of the data acquisition effort, and separates the relevant from the irrelevant events. As data are added to the mental movie or analysis worksheets, relevance is determined by examining events in pairs, and identifying cause–effect relationships between them. When they exist, the events should be linked to show the sequence and where important the relative timing of related interactions. The examination can begin with the process outcome, and proceeds backward in time toward the change event(s) that initiated the process. Alternatively, it can start wherever an event is displayed; work in both directions. Events that do not have a cause–effect relationship with other events should be considered irrelevant, and tentatively set aside; they can be recalled if additional information shows a need to do so.

When available data have been exhausted, the events with a cause–effect relationship should describe what happened in the proper sequence, from the first deviant event to the last harmful event. In cases where adaptive actions prevented a loss outcome, the last event would be restoration of the original activity or a shutdown.

The testing task checks the description of the events for completeness. The investigator looks at each "causing" event to determine if the "causing" event was sufficient to produce the "effect" event every time it occurred. If so, that part of the description is complete. If not, the investigator needs to determine what additional actions were needed to produce the "effect" event each and every time the "causing" events occurred. When this task is finished the investigator has identified all the events necessary to produce the outcome. Remaining gaps in the description identify remaining unknowns.

Methods for hypothesizing events to fill gaps in the scenarios exposed by the testing task use bounded logic trees to display possible alternative scenarios. A bounded logic tree has both the top and bottom events defined. Alternative hypothesized scenarios are developed to link the bottom event to the top event. The most likely alternative supported by data recovered after the occurrence can be used to complete the description, provided their source is noted, and uncertainties acknowledged.

1.4.4 Work Product Development Tasks

These tasks depend on the deliverables specified in the investigation plan. An investigation is remembered by the work products it produces. The description and explanation documentation are the main work product common to all investigations.

An investigator's main task is to produce the documentation describing the flow of interactions that produced the outcome. The mental movies or events flowcharts provide the basis for describing and explaining what happened. If acceptable in the organization, the flowchart satisfies the basic documentation needs. If not, a narrative description prepared from the flowchart may be needed to satisfy other needs. If another reporting structure is required, such as the facts/analysis/findings/conclusions reporting format, the events flowcharts with their supporting data enable investigators to produce that work product.

Regardless of the format, the description must provide sufficient information to enable the user to visualize what happened, and why it happened. If illustrations are needed to achieve this, they should be added to the description. An additional element of the investigation work product is the group of supporting source documents or objects. Each source of data used should be identified and archived and retained according to the investigation plan.

1.4.4.1 Problem Identification

One constructive use of the descriptions is problem discovery and definition. The problem discovery task requires a supplemental data-gathering effort. To define a problem, an investigator must identify what a person or object was expected to do, or the norm for an action. Then the investigator must define the difference between what a person or object did and what they were expected to do, and examine why that occurred. This comparative approach is the basis for defining problems. Investigators sometimes fold this task into the description or explanation development tasks. That is not recommended unless restarting is urgent.

Alternatively, if a flowchart is available, a problem can be defined as a subset of the events constituting the scenario. For example, if all expectations were satisfied, but a person or object did not have adequate time to adapt to prior events, that subset of events identifies a problem. See Appendix A for guidance about the kinds of events for which timing may be critical. However, then the investigator must pursue the reason the problem came into being. That pursuit can lead to design, administrative, supervisory, training, programming, or other less direct decisions, assumptions, or actions.

1.4.4.2 Recommendation Development

Another constructive use of the descriptions is in the development of recommendations for future actions [26]. If a flowchart of events is available, the events sets used to define problems provide the basis for examining potential changes that might be introduced to change the future flow of events. For each event in the set, every actor, action, or link can be examined as a candidate for change. Changes might include different sequencing or timing of events, changing the magnitude of events, or substitution of components, energies, or barriers, for example. Then, the consequences of each change can be estimated by studying what effect the change might have on the subsequent events involved in the occurrence.

Comparing the predicted consequences of each candidate change provides a basis for evaluating and ranking the desirability of the alternative choices, in terms of their relative efficacy and efficiency. This comparison and evaluation should include a discussion of the costs of implementation and value of predicted performance improvements [27].

1.4.4.3 Success Monitoring

Another constructive use of the descriptions is to develop a monitoring plan with which the predicted success of proposed actions can be monitored if they are implemented. The approach is to look for the recurrence of *problem events sets* during future operations. Thus by identifying and monitoring those events sets the effectiveness can be identified. If they recur, the change may be unsuccessful. If they do not recur, the change is probably successful.

1.4.4.4 Other Uses of Occurrence Descriptions

Action-based flowcharts of occurrences can be used for efficient personnel training or retraining, to identify operational improvements, for design reviews or new design guidance, or to support changes in standards, codes, or regulations. The information on the flowcharts is easy for individuals to assimilate because they can *quickly relate actions shown on the flowchart to their own daily tasks*. They can also see their relationship to others' tasks, and use that as a basis for doing their tasks more efficiently.

Other kinds of investigation outputs have more limited uses, usually take longer to absorb, and are more difficult for individuals to assimilate.

1.4.5 Quality Assurance Tasks

Quality assurance tasks involve examining the quality of the investigation work products, the investigation process, and the investigation program.

1.4.5.1 Investigation Work Product Quality

The quality assurance task varies with the investigation process chosen. If the actor–action-based process is used, the quality assurance task consists of having another investigator review the flowchart and supporting data for their logic and sufficiency. This helps identify and remove conjecture, speculation, and unsupported conclusions. Other indicators of quality problems are the number of questions raised by users, or the degree of controversy that follows release of a report.

If another process is chosen, the quickest way to assess the quality of the work products is to flowchart the reported actions and links showing relationships, and look for gaps or logic errors to identify problems with the work product or perhaps the investigation process [21].

1.4.5.2 Investigation Process Quality

Problems with the quality of the work products provide one indication of problems with the investigation process. If work products are found to have problems during the quality assurance tasks, the investigation process should be re-examined as one possible reason for the problems. The problem may also result from the process chosen, or its execution.

Another indicator of problems with the investigation process is the level of complaints about investigator actions or behavior during the investigation.

1.4.5.3 Investigation Program Quality

A third and broader indicator of problems is in the value of the results produced by investigations over time. A measure of performance is the comparison of the resources allocated to the investigation program and the cost of investigations with the value of the improvements produced. This quality assurance task is more subjective, but still requires attention. How that is done should reflect the assessment practices applied to other kinds of activities in the organization.

1.4.6 Deliverables

The specifications in the investigation plan define the content of the deliverables produced by the investigation, and their distribution. The investigator's task is to produce the required deliverables. From time to time, the investigators are called upon to make oral presentations to explain or defend their deliverables. To do this well, investigators should ensure that the logic of their reasoning has been checked carefully.

Another important deliverable, particularly when potential injury claims or litigation are foreseeable, is the source material. That consists of documents, photos, selected debris, sketches, or whatever source materials were used to define and document the reported actions. Chain-of-custody documents may be an essential element of these deliverables.

When regulatory agencies prepare a report, ensure that internal deliverables are compatible with those reports, or explain any inconsistencies. When accidents requiring investigation under regulations are investigated, it is desirable to prepare reports containing the descriptive data required by regulations to meet their requirements. They require "the date, time, description of operations, description of the accident, photographs, interviews of employees and witnesses, measurements, and other pertinent information." A separate report, with "information or evidence uncovered during the investigation" that would be of benefit in developing a new or changed standard is also required to be submitted to the agency [1]. It is good practice to review reports prepared for regulatory agencies to remove subjective opinions or built-in judgments from the reports. Cause statements can be contentious. To minimize arguments about causes, specify what action(s) caused what specific effects, using event terms.

1.4.7 Postinvestigation Tasks

Postinvestigation tasks involve distributing and using the investigation work products for additional purposes. These uses range from resolution of claims and disputes to long-term enhancement of corporate memory. Some uses are obvious, like resolution of claims, where the description of what happened provides a basis for negotiating settlements. Other uses include their incorporation into training or recurrent training of personnel as case studies. Charts can be used as guidance for operating personnel to help them to understand interactions among system components, and to do their jobs more efficiently. Distribution to designers helps them identify design options or changes that could improve future performance. In multifacility organizations, the charts are convenient to exchange among facilities for the same purposes.

When consistency among all investigation is achieved, it becomes possible to combine the outputs into a growing body of process flowcharts covering any aspect of operations that may have been investigated. Analogous to mapping DNA, use of investigation work products to build a map of operations increases understanding of the processes. That also provides new opportunities for developing improvement recommendations beyond those made in a single occurrence. This is a new frontier in industrial operations.

Another important use of the work products is to update predictive job, hazard, or operational analyses of the systems. Investigations provide the main tool for assessing the quality of or verifying prior risk and hazard analyses of the system, and showing where changes might be needed [23].

1.5 SUMMARY

For an investigation program to be of constructive value to an organization, a positive approach transcending conventional compliance and prevention perspectives needs to be in place. The program must rest on progressive policies, concepts, principles, plans, and preparations. The material presented helps organizations accomplish this, while enabling investigators to satisfy regulatory agency and other narrower demands of investigations.

Investigation technology is expanding at an accelerating rate. Investigators are urged to use the Internet resources and newly published information to keep abreast of new developments. They present opportunities to learn more about the investigation process and emerging developments, and the investigation requirements imposed by regulations. The Internet is in flux, but the web sites noted with the references are useful starting points as this is written.

1.6 APPENDICES: INVESTIGATION TOOLS

The models and checklists in the appendices offer guidance to help investigators during investigations.

A. General Human Decision Model for Investigators

This model was developed from observations of human behaviors in many transportation accident investigations (Fig. A1). It was developed by tracking what happened during the accident, and by using the event-based data language and event matrices to show the events flows found.

A.1 Application of the General Human Decision Model for Investigators

To apply this model during investigations or interviews, identify people who appear to have had a role in the incident process. For each relevant action:

1. Begin by finding the change or changes in the activity that created a need for action by that person to keep the activity progressing toward its undesired outcome.
2. When identifying that change, determine if the change emitted some kind of signal that a person *could* detect or observe. If it did not,

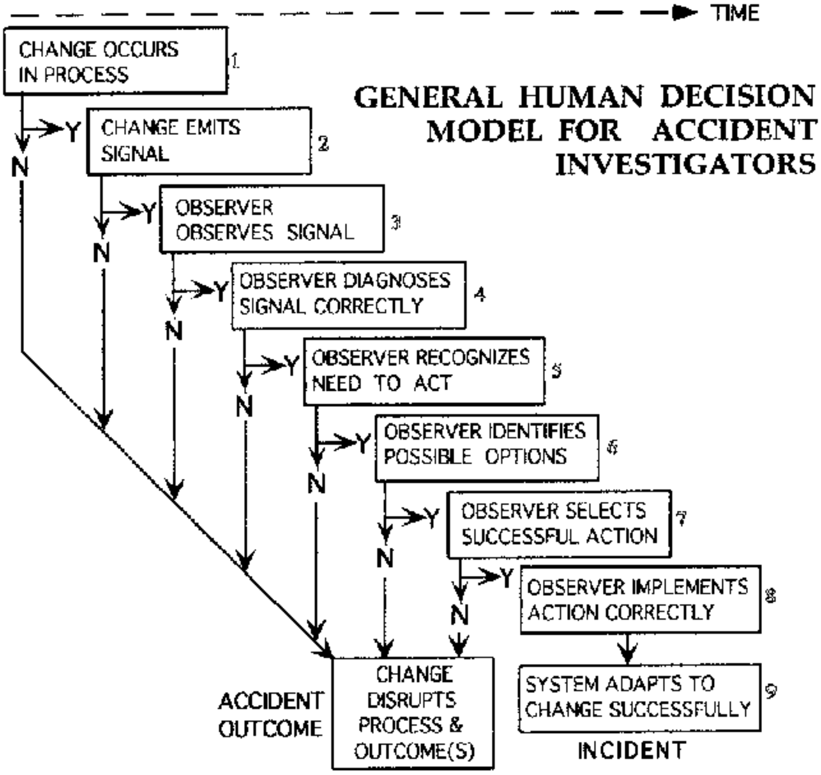

TIME
GENERAL HUMAN DECISION MODEL FOR ACCIDENT INVESTIGATORS
CHANGE OCCURS IN PROCESS 1
CHANGE EMITS SIGNAL 2
OBSERVER OBSERVES SIGNAL 3
OBSERVER DIAGNOSES SIGNAL CORRECTLY 4
OBSERVER RECOGNIZES NEED TO ACT 5
OBSERVER IDENTIFIES POSSIBLE OPTIONS 6
OBSERVER SELECTS SUCCESSFUL ACTION 7
OBSERVER IMPLEMENTS ACTION CORRECTLY 8
SYSTEM ADAPTS TO CHANGE SUCCESSFULLY 9
INCIDENT
ACCIDENT OUTCOME
CHANGE DISRUPTS PROCESS & OUTCOME(S)
Y
N

Figure A1 This model represents decisions by people in response to changes. It helps investigators understand the roles of stimuli, sensory inputs, communications, diagnostics, decision making and implementation, training, design, procedural, supervisory, and many other "human factors" issues in occurrences. (Adapted from Four Accident Investigation Games, Simulations of the Accident Investigation Process. Appendix V-F, General Human Decision Model for Accident Investigation. Oakton, VA: Lufred Industries, Inc., 1982.)

explore why it did not and what effect that had on the outcome.

3. If it did emit a signal, explore whether the person saw, heard, felt, or otherwise "*observed*" the signal. If not, explore why not, and what effect that had on the outcome.
4. If the person observed the signal, was the signal *diagnosed* correctly? Was the person able to predict the consequence(s) of the change from the signal, and knowledge of the system and its operation? If not, explore why not, and its effects.
5. If the predicted consequences of the change were correctly identified, did the person *recognize* that action was needed to counter those consequences? If not, explore why not, and its effects.
6. If so, did the person identify the *choices* for action that were available for successful intervention? If not, explore why not, and its effects. Was this a new situation where the action had to be invented, or was this something that prior training anticipated and provided the responses to implement? In other words, was the person confronted by demand for an *adaptive* or *habituated* response? (Here, you start to get into the person's decision-making process, and potential personal judgment issues, so explore this area empathetically with the witness, particularly for adaptive responses.)
7. If any response actions were identified, did the person *choose* a successful response to implement? If a successful response was not chosen, explore why not.
8. If a successful response was chosen, did the person successfully *implement* the desired action? If not, explore why not.
9. If a suitable response was implemented, the system adapted to the change without an unintended loss or harm. If the response did not

achieve a no-accident outcome, explore why it did not. Often this leads to discovery of invalid system design assumptions or other design problems.

After working with this model, you will be in a much better position to discover, define, describe, and explain what happened when a so-called "human error" or "failure" is alleged. You will also be able to identify more than one possible action to improve future performance of that system.

B. Energy Trace and Barrier Analysis

This model (Fig. B1) describes generally the steps to follow to examine events with the energy trace and barrier analysis (ETBA) method during investigations. It is a generalized model. Investigators can apply it to understand energy flows that produced an observed changed condition or reaction, or when an action needs to be understood in greater detail.

The kinds of energy that might be involved in a specific investigation are listed below. Look for possible energy sources when the energy source is not obvious.

Man-made energy sources:

1. Electrical.
2. Mass/gravity/height (MGH).
3. Rotational kinetic.
4. Pressure/volume/kinetic displacement (P/V/KD)
5. Linear kinetic.
6. Noise/vibration.
7. Dust.
8. Chemical (acute or chronic sources).
9. Thermal.
10. Etiological agents.
11. Radiation.
12. Magnetic fields.
13. Living creatures or things.
14. Moisture humidity.

Natural energy sources:

15. Terrestrial.
16. Atmospheric.

(Condensed from Ref. 9, Guide 5, p. 4. For an exhaustive listing of energy types see Ref. 14.)

C. Investigator Checklists

C.1 Quality Assurance Checklists

Description of Occurrence. After a quality check of the description of what happened, check it against the following checklist one last time.

Event form and content in description okay?
Event names consistent in description?
Abstractions and ambiguities removed?
Sources referenced okay?
Scope of description adequate?
Flowchart causal links complete and valid?
Uncertainties clearly indicated.

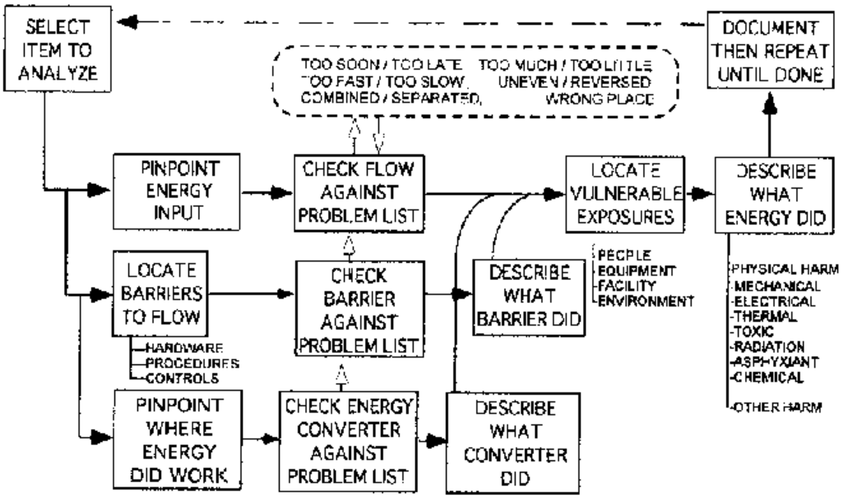


Figure B1 Energy trace and barrier analysis process model for accident investigation. (Adapted from 10 MES Investigation Guides, Guide 5, ETBA. Oakton, VA: Ludwig Benner & Associates 1998, p 4.)

Mental movie (visualization) supported?
Editorial adjectives and adverbs removed?
Unsupported opinions or judgments removed?
QA checks completed okay?
Referenced sources in archives?

C.2 Recommendation Checklist

Recommendations. When including recommendations in a report, use the following checklist to review each recommendation [27].

Does the recommendation simply and concretely describe the problem?
Does the recommendation clearly identify who will have to do what is proposed?
Does the report state what specific improvement is expected to occur if the recommendation is implemented?
Does that person have adequate authority and resources available to implement the proposed action?
Is there enough uncertainty to indicate that a field test of the action is needed before making the recommendation, or before it is widely implemented? If so, is the required testing, defined?
Are appropriate implementation milestones needed? If so, are they included and reasonable?
If more than one recommendation results from the investigation, are priorities for implementation necessary or provided?
Do you know how the people who have to implement the recommendations will respond to them?
Have you determined how both you and the recipient will know when your recommendation has produced successful results?
Have you defined the follow-up process that is required to ensure implementation and verify that predicted performance was achieved?
If you had to implement the recommendation, would you be willing to do so? Good rule: don't ask anyone to do something you wouldn't be willing to do yourself if you received the recommendation.

C.3 Problem Words Checklist

Listed below are words which are known to create problems with descriptions and explanations of occurrences developed by investigators. Use the list to locate problem words, and replace them in final reports if possible.

And
Or
He
It
She
They
Blame
Not
Nor
Was
Were
Did not
Failed to
Fault
Plural actor names (*firefighters, crew*)
Compound actor names (*third shift, crowd*)
Verbs with wrong tense (*drive*).
Passive voice verbs (*was struck*)
Built-in judgments (*too, failed, erred, misjudged, violated, should have, ignored,* etc.)
Editorial "ly" adjectives and adverbs (*badly, improperly, inadequately, poorly*, etc.)
Words conveying what did *not*, happen (e.g., *did not replace*). Say what *did* happen!

REFERENCES

Developing Descriptions and Explanations of What Happened

1. U.S. Department of Labor, Occupational Safety and Health Administration. Title 29, Code of Federal Regulations, 1960.29 Accident Investigations (http://www.osha-slc.gov/OshStd_data/1960_0029.html).
2. U.S. Department of Labor, Occupational Safety and Health Administration. OSHA's Small Business Outreach Training Program. Instructional Guide, Accident Investigation, May 1997 (http://www.osha-slc.gov/SLTC/smallbusiness/sec6.html). OSHA's view of investigation programs for small businesses.
3. K Hendrick, L Benner. Thinking about accidents and their investigation. In: Investigating Accidents with STEP. New York: Marcel Dekker, 1986. chap 2.
4. For a description of these perceptions and how they influence investigations, see L Benner. 5 Accident Perceptions: Their Implications For Accident Investigators. ASSE Professional Safety, February 1982, or L Benner. Accident Perceptions: Their Implication For Investigators. International Society of Air Safety Investigators Forum, 14:1, 1981 (http://www.iprr.org).

5. K Hendrick, L Benner. Investigation concepts. In: Investigating Accidents with STEP. New York: Marcel Dekker, 1986, pp 30, 235.
6. WG Johnson. The Management Oversight and Risk Tree-MORT. SAN 821-2 UC4l, prepared for the U.S. Atomic Energy Commission, Division of Operational Safety under contract AT(04-3)-821, submitted February 12, 1973, p 59. See also WG Johnson. MORT Safety Assurance Systems. New York: Marcel Dekker, 1980, Chapter 5. Many useful ideas about nature of mishaps and investigations. See also U.S. Department of Energy. Accident/Incident Investigation Manual, DOE/SSDC 76-45/27. 2nd ed. 1985. Built upon Johnson's research, and contains some additional useful investigation techniques; based on legal framework.
7. SI Hayakawa, AR Hayakawa. Language in Thought and Action. 5th ed. San Diego, CA: Harcourt Brace & Company, 1960, chap 3, pp 48, 85. Essential reading for investigators who want to use factual language to develop their work products.
8. N Leveson. Safeware: System Safety and Computers. Reading MA: Addison-Wesley 1995. Appendix A, Medical devices: The Therac-25 Story describes interactions between personnel input timing and software design, for example.
9. L Benner. 10 MES Investigation Guides. Oakton, VA: Ludwig Benner & Associates, 1998. Guides cover wide range of investigation procedures and practices.
10. D Vaughan. The Challenger Launch Decision, Chicago, II: University of Chicago Press, 1996, chap 5. Describes how deviations become normalized.
11. K Hendrick, L Benner. Meekers law and perceived interests, In: Investigating Accidents with STEP. New York: Marcel Dekker, NY, 1986, pp 149, 235.
12. Adapted from L Benner. Rating accident models and investigation methodologies. J Safety Res 16(3): 105-126, 1985. This work can also be found in ref 3.
13. Handbook P88-I-1. Investigation of Mining Accidents and Other Occurrences Relating to Health and Safety. U.S. Department of Labor, Mine Safety and Health Administration, 6/21 /94 Release 3. (http://www.msha.gov/READROOM/HANDBOOK/PH88-I-l.pdf.)
14. Introduction to investigation. (Companion guide to a video film with the same title), Stillwell OK: International Fire Service Training Association, Fire Protection Publications, 1997. Detailed investigation guidance is consistent with multilinear events sequencing concepts.
15. NFPA 921. Traditional guidance for fire investigations. Quincy, MA: National Fire Protection Association, 1995, Traditional guidance for fire investigations.
16. Accident Investigation. (Companion guide to a video film with the same title), Stillwell OK: International Fire Service Training Association, Fire Protection Publications, 1997. Describes detailed investigation tasks, with checklist form of summary of tasks in chap 4.
17. Guidelines for Investigating Chemical Process Incidents, New York: Center for Chemical Process Safety, American Institute of Chemical Engineers, 1992, p 50. Lists different kinds of investigation approaches, with subjective thoughts of their attributes.
18. "U.S. Department of Labor, Occupational Safety and Health Administration, OSHA Field Inspection Reference Manual CPL 2.103, Section 6-Chapter II. Inspection Procedures; OSHA Instruction CPL 2.94 July 22, 1991 OSHA Response to Significant Events of Potentially Catastrophic Consequences http://www.osha.gov/(do site search)"
19. Fatality Inspection Procedures. OSHA Instruction CPL 2.113, U.S. Department of Labor, Directorate of Compliance Programs, Washington, DC April 1, 1996. Regulatory agency investigator guidance. (http://www.osha-slc.gov/OshDoc/Directive_data/CPL 2_113.html.)
20. RH Wood, W Sweginnis. Aircraft Accident Investigation. Casper, WY: Endeavor Books, 1995. Chapter 34 helpful for "reading" events from witness plates.
21. Proceedings with technical papers from annual seminars and individual papers are available through the International Society of Air Safety Investigators, Sterling, VA 20164.

Assuring Quality

22. L Benner, I Rimson. Quality Management For Accident Investigations (in two parts). International Society of Air Safety Investigators Forum, 24(3), October 1991; 25(1): February 1992.(http://www.patriot.net/users/luben/5IRRQC.html.)

Developing Recommendations

23. I Rimson, L Benner. Mishap Investigations: Tools For Evaluating The Ouality Of System Safety Program Performance. Proceedings of 14th System Safety Conference, Albuquerque, NM. (www.iprr.org/LIB/QMA_P1.html from http:/www.patriot.net/5IRRQC.html.)
24. K Hendrick, L Benner. Appendix E. In: Investigating Accidents with STEP. New York: Marcel Dekker, 1986. Presents instructions for preparing a test plan.
25. L Benner. 10 MES Investigation Guides. Oakton, VA: Ludwig Benner & Associates, 1998, Investigation Test Plan Guide.

26. L Benner, 10 MES Investigation Guides 2nd ed. Oakton, VA: Ludwig Benner & Associates, 1998, Guide 8. Helps investigators develop recommendations.

27. K Hendrick, L Benner. Investigating Accidents with STEP, New York: Marcel Dekker, 1986, pp 361–365. Describes recommendation quality control problems to avoid.

Internet Site References

Several of the references provide Internet web addresses. Internet sites change frequently, and more are coming on line daily. Sites of interest to investigators can be located by doing searches on accident investigation using one or more of the many internet search engines available. The searches will disclose the online sites when the search is conducted. References to additional investigation manuals, research reports, papers, books, videos, and links to investigation sites and investigation reports, publications, and other resources of interest can be found at www.iprr.org.

Chapter 8.2

Government Regulation and the Occupational Safety and Health Administration

C. Ray Asfahl
University of Arkansas, Fayetteville, Arkansas

2.1 INTRODUCTION

A comprehensive program of industrial automation must take into consideration the impact that government regulation and the Occupational Safety and Health Administration (OSHA) have had upon the workplace. Automated systems sometimes introduce hazards that are either not present in manual systems, or they do not pose a threat in manual setups. The introduction of these hazards may put the firm in violation of OSHA standards. Even more important to automation than the negative impact that automation can sometimes have upon hazards and violations of OSHA standards is the positive role automation can play in removing workers from hazardous workplaces or environments. In this respect, OSHA and federal safety and health standards can act as a driving force for the introduction of automated systems and robots to do jobs that are considered unsafe for humans or in violation of OSHA safety and health standards. Whether the impetus is to bring automated systems into compliance by removing hazards that they might introduce or to determine where automated systems can remove worker exposures to OSHA violations, the automated system designer needs to know what critical role the OSHA plays in today's workplace.

2.1.1 OSHA Enforcement Powers

When the OSHA was brought forth by Congress in 1970 it had a stunning impact upon industry nationwide. Made a part of the U.S. Department of Labor, the new agency was given enforcement powers never before seen in the American workplace. For the first time, federal inspectors were authorized to enter virtually any industry without advance notice at any "reasonable time," including the night shift, and look around for violations of federal safety and health standards and then write citations with monetary penalties. As the law was originally written, the OSHA inspector had sweeping powers to enter the workplace without having to show cause or a court-issued search warrant, but later, in the landmark Barlow Decision, the U.S. Supreme Court ruled unconstitutional this broad power and required the federal OSHA inspector to seek and obtain a court order to enter a plant, if the proprietor demanded it [1].

Industry was caught by surprise by such sweeping powers. Congress specified a mere four-month period between passage of the law and its effective date. This unusually short period for familiarization of the law was all the more remarkable for the fact that the force of the law was to be so generally applied. At the time of passage it was estimated that the law would affect some 57 million workers in more than four million workplaces across the nation [2]. It would soon be found that the OSHA would affect virtually every U.S. citizen, nonworkers as well as workers. But before that was to happen, many industries had to become aware of the agency and what to do about it. Millions of firms and millions of workers were still uninformed of the OSHA's powers at the time that the law went into effect.

When U.S. industry did find out what the OSHA was all about, it reacted quickly and sometimes humorously. As one industry representative put it, "If you think that 'OSHA' is a small town in Wisconsin, you are in big trouble!" The new agency's powers were likened to Hitler's Gestapo. The OSHA quickly grew to become one of the most hated government agencies ever created. There is a suggestion that Congress had some premonition of the public indignation that the agency would generate, because in the wording of the law itself Congress specified a penalty of up to life imprisonment for anyone convicted of killing an OSHA inspector.

The vehemence with which industry attacked the new OSHA agency suggested that industry did not care about worker safety and health. The ostensible purpose of the law—to punish employers with monetary fines if they violated established and published standards for safety and health—seemed honorable and difficult to question. So why else would industry be so upset, unless they were indeed guilty of gross neglect of worker safety and health? The reason behind the public outcry becomes more apparent when the structure of standards promulgated by the OSHA is examined, in the section which follows.

2.1.2 Federal Standards

Congress set another quick deadline on the OSHA agency as soon as the agency went into effect. The agency was given two years in which the agency had the right to put into effect, without public comment, "national consensus standards" for occupational safety and health. The OSHA was faced with the task of quickly assembling appropriate existing standards for establishment as mandatory federal standards enforceable by law. To wait beyond the two-year deadline would require the OSHA to justify to the public the necessity for each standard using a lengthy, legal promulgation process. There were several problems associated with accomplishing the task of adopting a set of national consensus standards.

The first problem was to decide which standards should be included. Congress specified criteria for adoption under the classification "national consensus." However, many standards met the general criteria specified by Congress, such as "adopted and promulgated by a nationally recognized standards-producing organization" or "formulated in a manner which afforded an opportunity for diverse views to be considered." But the real problem was to determine whether the standards under consideration for adoption were ever intended to be mandatory, with inspections and penalties for failure to comply.

In case of conflicts between existing standards, Congress expressly took the more aggressive approach and specified that OSHA shall "promulgate the standard which assures the greatest protection of the safety or health of the affected employees." The OSHA could easily interpret the spirit of this mandate as "when in doubt, go ahead and adopt a given standard."

Another problem was to decide which standards applied to which industries. When standards were enforced by states, the state often published standards for each industry covered. While not all industries were covered, the states had the opportunity to focus upon industries that had poor safety records. But the OSHA law was so general in scope that it applied to virtually all industries. A very complex structure would be needed to determine exactly to which category each industry belonged and many decisions would have to be made to determine which "national consensus standards" should be applied to which industry. Since the OSHA law generalized upon a more specific structure for standards and enforcement, it became difficult for the agency to classify industries and make judgments to determine which industries would be required to abide by which standards. It is even questionable whether the new OSHA agency had been extended the authority by Congress to make such judgments. Therefore, OSHA departed from the standards-by-industry approach and adopted a strategy that became known as the "horizontal standards approach." Therefore, early on, the OSHA published a "national consensus standard" consisting of standards to be applied in general, in all industries in which the specific circumstances described in the standard could be found. The new standard was designated "General Industry Standards." The OSHA retained the traditional approach for certain industries, such as construction, sawmills, textile mills, and maritime industries and others for which it could identify specific industry standards. These standards became known as “vertical standards” to contrast them with the horizontal types contained in the General Industry Standards.

The General Industry Standards met the Congressional deadline for adoption within two years of the effective date of the OSHA law, but the agency was soon to regret that it had not used more discretion in selecting which standards to adopt. Many of the standards had been written in advisory language because, as was stated earlier, they had never been intended to be mandatory with inspections and penalties for failure to comply. To get around the problem of advisory language, the OSHA in some instances simply substituted the word “shall” for the word “should” when it adopted the language of the national consensus standards. The OSHA’s right to do this was later challenged, as Congress had not given the OSHA the authority to change the wording of standards that met the definition of “national consensus” without going through the lengthy process of legal promulgation of the new standard, allowing the public to comment on whether the standard should be adopted.

Another indiscretion on the part of the OSHA was its failure to screen standards to eliminate antiquated provisions that were no longer adhered to by anyone. For example, the OSHA was subjected to ridicule for adopting a standard that prohibited ice from coming into contact with drinking water. This ancient provision dated back to the days when ice was cut from ponds and was not considered safe for drinking.

But the most critical problem generated by the new national consensus standards was their sheer volume. Since the OSHA had been unable to grapple with the problem of sorting out the standards into vertical standards for each industry, the “horizontal standards” strategy adopted by the OSHA left to the public the responsibility for reading the rules and determining which standards applied to their own operations. The standards were, and still are, so voluminous that industries large and small have had difficulty finding which ones are sufficiently relevant to their own operations to study in detail and formulate plans for coming into compliance. The volume of the standards has become a powerful weapon in the hands of the federal inspector, who has the latitude to dig deeply into the standards to find violations, if he chooses to do so in a given inspection, or to take a more lenient posture overlooking minor violations. It is the inspector’s option, because, as one industry representative put it, “if you are alive, you’re wrong,” in any careful, literal reading of the entire volume of the OSHA standards applied “wall-to-wall” throughout an industrial plant of any reasonable size.

2.2 GENERAL REQUIREMENTS

Most OSHA standards have survived the public controversy over their volume, their origins, and their relevance to current industrial operations. While it may seem that nearly every topic is covered by standards, in fact OSHA inspection authorities are often unable to identify a particular standard to apply to a given hazardous situation. In these situations, one solution is to turn to the General Duty Clause of the OSHA law.

2.2.1 General Duty Clause

Congress anticipated that federal standards would not be in place to deal with every situation so they drafted a general requirement stated right in the text of the law, quoted here as follows [3]:

> Public Law 91-596
> Section 5(a) Each employer . . .
> (1) shall furnish to each of his employees employment and a place of employment which are free from recognized hazards that are causing or are likely to cause death or serious physical harm to his employees . . .

Particularly in the field of automation, the OSHA applies the General Duty Clause because there are very few specific standards that apply to the specific areas covered by this book, i.e., programmable logic controllers, robots, etc.

Another way that the OSHA applies the General Duty Clause is to find a specific rule that applies to some other operation and then generalize the rule to other applications, using Section 5(a)(1). For example, types of point of operation guarding is well-specified for certain machines, such as mechanical power presses. For some other machines, such as hydraulic presses or forging machines, the standards are less specific about how the machine should be guarded. Since it would be inappropriate for OSHA to cite a rule for a different machine than the type specified in the standard, the OSHA can turn to the Section 5(a)(1) General Duty Clause and cite the dangerous situation anyway. Of course, to be upheld by the courts, the OSHA will have to provide justification that the hazardous condition is “likely to cause to death or serious physical harm” to employees. In the example studied here, lack of point of operation guarding could lead to an amputation, a condition that would meet the criterion of "serious physical harm."

There is also a General Duty Clause for employees, as follows:

> Public Law 91-596
> Section 5(b) Each employee shall comply with occupational safety and health standards and all rules, regulations, and orders issued pursuant to this Act which are applicable to his own actions and conduct.

The General Duty Clause for employees is rarely, if ever, cited by the OSHA. Indeed there are no penalties specified in the law that are to be applied to employee conduct under Section 5(b). The General Duty Clause for employers, however, is very frequently cited and citations of this clause usually carry heavy penalties.

2.2.2 Enforcement Priorities

The OSHA agency, without doubt, must operate with limited resources, resources that have become even more limited in recent years. The OSHA must set priorities, and the result is that some categories of workplace are rarely visited. The OSHA law recognized the presence and importance of "imminent dangers" that are of such a serious nature as to threaten immediate and serious harm to employees. Accordingly, the OSHA places the highest priority for inspection on "imminent dangers" in the allocation of its inspection resources. Imminent dangers are rare, so rare that inspectors are not regularly scheduled to handle these situations. On the rare occasion that one does arise, usually at a construction site, the OSHA has been given authority to seek a court order to close down the worksite. Although this inspection priority is high, automation planners and designers should not expect to experience an imminent danger inspection.

The second highest inspection priority is the investigation of reported fatalities or major incidents involving the hospitalization of five or more employees. The OSHA law originally required the employer to notify the OSHA agency within 24 hr of occurrence of workplace fatalities or major incidents involving hospitalization of five or more employees. The 24 hr deadline has since been shortened to 8 hr. Telephone, telegraph, or other suitable means are acceptable means of informing OSHA of these major incidents. Fortunately, tragedies of this magnitude are rare for applications of automation, but unlike the imminent danger category, fatalities or major accidents are definitely a possibility when working around automatic machines, robots, or other automated systems. Accounts of human fatalities at the hands of industrial robots make big headlines in newspapers.

Besides robots, other automated systems and automatic machines have caused fatalities, sometimes in surprising ways. Consider an automatic screw machine, for example. Although there are many moving parts in the highly mechanized operation of a multispindle automatic screw machine, most of these moving parts are enclosed or shielded in the operational head of the machine. For both safety and general containment of chips and cutting fluid, automatic screw machines usually have enclosed heads. It is difficult to see how a person could be injured seriously by an automatic screw machine, much less killed by one. But, as the following account reveals, fatalities around automatic screw machines are certainly a possibility.

Case study. Worker Fatality—Multispindle Automatic Screw Machine: *A six-spindle automatic screw machine was operating with exposed rotating bar stock in a manufacturing facility for the production of socket wrenches. The bar stock had recently been changed and the bars were still very long, extending from the backside of the machine head as shown in Fig. 1. An awareness barrier, in the form of a stanchion and chain arrangement, warned personnel not to enter the area of the rotating bars. However, one of the bars was slightly bent when the machine was placed in operation. The situation was exacerbated when the machine was turned on and the bent bar began to rotate rapidly. The long length of the slightly bent bar also added to the problem. The combination of factors caused the bar to make contact with one of the other five rotating bars, bending it also. Almost instantly, the entire bundle of rotating steel bars became a tangled mass, swinging wildly outside their usual work envelope. A female worker standing nearby was struck by the rotating tangle of steel bars and was fatally injured.*

This case study illustrates the need for awareness of the ways in which automated systems can become very dangerous in surprising ways. The OSHA was summoned in this incident in response to the high priority placed upon fatalities and major accidents.

The next highest category for OSHA enforcement actions are complaints filed by employees or others. The OSHA law provides for a procedure by which an employee can complain about a suspected violation of OSHA standards and request an inspection. To protect the employee from employer retaliation or from the fear of potential employer retaliation, the law

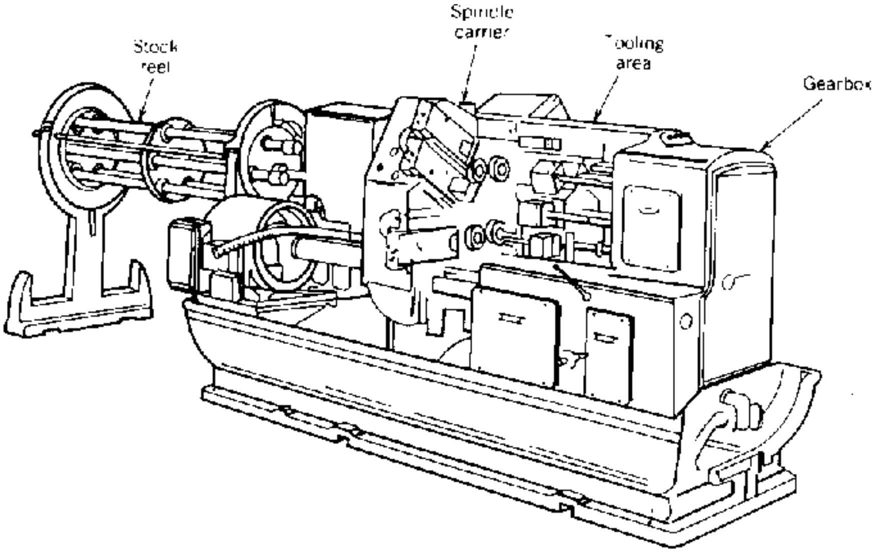


Figure 1 Multiple-spindle automatic screw machine. (From Ref. 4, copyright 1979 by John Wiley & Sons, Inc., New York, NY. Reprinted by permission.)

provides for concealing the identity of the employee making the complaint. Before undertaking any automation initiative, management needs to think about this provision of the OSHA law. Automation projects often generate hostility on the part of employees who fear that they may lose their jobs or otherwise be adversely affected by the new automated system. People experience anxiety when they are confronted with change beyond their control. Management would do well to consider carefully the impact of any automation project upon the attitudes of workers who will be affected. More specifically, management should expect workers to be intolerant toward any safety or health hazards that might be introduced by an automation project and should not be surprised if workers get the idea to call the OSHA and file a complaint. The OSHA can be expected to place a high priority on such complaints. Since the two higher priority categories (imminent danger and fatalities/major hospitalizations) are very rare, an inspection is almost a certainty if an employee files a complaint against the firm.

The foregoing account seems to condemn automation as a generator of OSHA problems. The opposite is often closer to the truth, because on the other side of the complaint issue is the fact that a primary motivation for automation is to eliminate hazards to employees. Indeed, an employee complaint to the OSHA regarding a workplace hazard may ultimately wind up as an automation project to eliminate the hazard. The OSHA itself places a high premium on "engineering controls" as a solution to workplace hazards, instead of relying on "work practice controls" or personal protective equipment.

2.2.3 Compliance Procedures

By whatever priority it is selected, once a firm is selected for inspection, the OSHA inspector has authority to search for violations of any standards contained in the general industry (29 CFR 1910) standards, provided the firm is not a special industry such as maritime, for which other volumes of standards apply. During the walkaround inspection of the facilities, the OSHA may request a representative of employees to accompany the inspector and a representative of management during the inspection. If violations are alleged, the firm is afforded the opportunity to request an informal conference in which the alleged violations and corrective actions are discussed by company management and the OSHA. If the OSHA believes that an observed violation "has no direct or immediate relation to safety or health," it may choose to designate an alleged violation as "de minimus" and propose no penalty to be paid by the firm. Sometimes citation of an alleged violation is waived completely by the OSHA, especially if the violation is corrected immediately.

If a citation is issued, it will be received by the company within six months, the statutory limit for issuing a citation of alleged violations. After a citation is issued, time is of the essence, because from the date a citation is issued, the firm has 15 working days in which to decide whether to contest the citation. If left uncontested, the citation becomes a final order and the employer is obliged to pay whatever monetary penalties have been proposed by the OSHA. If the citation is contested, the employer has an opportunity for the citation to be reviewed by the Occupational Safety and Health Review Commission (OSHRC), an independent reviewing body appointed by the President. Once the citation comes under review, the OSHRC is authorized to lower or to eliminate the penalty for the violation, but the firm should also be aware of the fact that the OSHRC also has the authority to *increase* the penalty to a level higher than was proposed by the OSHA. Congress extended to the OSHRC, not OSHA, the authority to set final penalty levels, and although the OSHRC usually lowers penalty levels, it sometimes exercises its authority to set higher penalties.

A more serious concern than the payment of fines for OSHA violations is the correction of the violation within whatever reasonable time for abatement is set by the OSHA. The OSHA usually discusses a proposed abatement period with the firm during the informal conference after the inspection. The abatement period is very important, and the firm should be sure that it is reasonable, and once the period is agreed upon, the firm should be diligent in taking whatever actions are necessary to correct the violation. If during the correction period if becomes evident that deadlines will not be met, the firm should get in touch with the OSHA and request an extension of the abatement period. If no extension is requested and the abatement period expires, the OSHA has authority to re-enter the plant and check to see whether violations have been corrected. If a violation is found to remain uncorrected, the firm becomes subject to severe penalties that are assessed for each day the violation remains uncorrected.

2.2.4 Recordkeeping

One of the first things the OSHA inspector examines during an inspection is the company's records for injuries and illnesses. The Log/Summary of Occupational Injuries/Illnesses (OSHA form 200) is now well recognized by industries as mandatory, except for small firms having fewer than 10 employees. At the end of each year the right side of the two-page form acts as a summary, to be posted by February 1 for employees to review the previous year's record. The required posting duration is 30 days. One might expect that recordkeeping violations are considered minor, even de minimus, but the OSHA takes seriously whether a firm is diligent in maintaining the log. Recordkeeping fines have exceeded $1000, and in the early 1990s, recordkeeping violations were among the most frequently cited violations on record.

2.2.5 Performance Measures

One reason recordkeeping is so important to the OSHA is that recordkeeping is the basis for computing injury/illness performance measures. Standard incidence performance measures are computed ratios of the number of injuries or illnesses per man-hours worked, the ratio being multiplied by a constant (200,000) to make the ratio more meaningful. The figure 200,000 is approximately the number of man-hours worked per year by a 100-employee firm. The injury incidence rate is considered more meaningful than the illness rate, because it is difficult to define a time of occurrence for illnesses. The most popular incidence rate is the "Lost Workday Incidence Rate" (LWDI) which includes injuries, but not fatalities. It should be recognized that the OSHA considers a workday "lost" if the worker is unable to do his or her *regular* job after the injury or illness. If the worker remains at work, but is transferred to another job after the injury or illness, the days spent in the other job are considered "lost workdays" by the OSHA. The LWDI is a measure of the number of incidents that result in one or more lost workdays, but it does not consider how many workdays are lost in each incident. The OSHA has used this popular index as a criterion for whether to continue an inspection of a given firm, and it has also used the LWDI as a criterion for targeting a given industry group using the Standard Industrial Classification (SIC). The average LWDI for all manufacturing industries toward the end of the twentieth century was around 3 or 4. Some of the transportation industries have LWDIs of 10 or more. The next section examines ways in which automation can be used to keep incidence rates below the national averages and to avoid OSHA citation.

2.3 MACHINE GUARDING

It should not come as a surprise that the OSHA finds thousands of violations of machine guarding standards every year, as machine guarding seems to typify the subject of industrial safety. The automated-system designer should consider the safety aspects of manufacturing systems in the design phase, so that hazards of the human/machine interface can be guarded or engineered out before the system is built. It is particularly appropriate to consider machine guarding in this handbook, because the technology of automation can be quite successfully brought to bear upon the problem of guarding machines in general. Before examining the solutions this technology can bring forth, the various mechanical hazards introduced by machines should be explored.

2.3.1 Mechanical Hazards

The OSHA uses general requirements to cover all machines before proceeding to specialized requirements for such dangerous machines as punch presses, saws, and grinding machines. One of the most frequently cited general industry OSHA standards is as follows [5]:

> 29 CFR 1910. General Industry
> 212 General requirements for all machines.
> (a) Machine guarding.
> (1) Types of guarding. One or more methods of machine guarding shall be provided to protect the operator and other employees in the machine area from hazards such as those created by point of operation, ingoing nip points, rotating parts, flying chips and sparks. Examples of guarding methods are: barrier guards, two-hand tripping devices, electronic safety devices, etc.

Note that this standard leaves latitude for choice on the part of the process planner or machine designer with regard to the particular types of guards or guarding methods. Included are "electronic safety devices," which include a wide variety of automatic safeguarding mechanisms, such as photoelectric or infrared sensing barriers and other sophisticated control apparatus. In the context of this standard, the term "electronic" should be construed to include "electromechanical" devices. The general machine guarding standard is one of the most frequently cited of all OSHA standards. Even more frequently cited, though, is the general standard for point of operation guarding, detailed in the section which follows.

2.3.2 Point of Operation Safeguarding

The point of operation of a machine is defined as the "area on a machine where work is actually performed upon the material being processed." The frequently cited general OSHA standard for guarding the point of operation is quoted as follows [5]:

> 29 CFR 1910. General Industry
> 212. General requirements for all machines.
> (a) Machine guarding.
> (3) Point of operation guarding. (ii) The point of operation of machines whose operation exposes an employee to injury, shall be guarded. The guarding device shall be in conformity with any appropriate standards therefor, or, in the absence of applicable specific standards, shall be so designed and constructed as to prevent the operator from having any part of his body in the danger zone during the operating cycle.

The point of operation is generally the most dangerous part of the machine and at the same time it is usually the most difficult to guard. Machines usually must have an opening to admit material to be processed, and the same opening that admits material also often exposes the operator to injury. Often the solution is to devise some means of alternating the opening and closure of the area around the danger zone so that the operator can have access to the point of operation during the loading and unloading portion of the cycle, and then the guarding system can close the danger zone during the processing portion of the machine cycle. If such a system can be made to be automatic, an increased level of safety is usually possible, because the system does not depend upon the judgment or discipline of the operator to be effective. The following sections describe some of the automated systems used to control access to the danger zone of machines.

2.3.2.1 Gates

The OSHA standards recognize a device (known as a "gate") for protecting the danger zone of punch presses (see Fig. 2). Gates are actually of two types. Type A closes and remains closed during the entire cycle of the ram as it closes the upper die upon the lower die and then reopens and rises until it comes to rest. The Type B gate closes for the downward portion of the cycle and then reopens as the ram returns to its up position. Type A is a little more conservative in that it provides more protection in the case of a "repeat," in

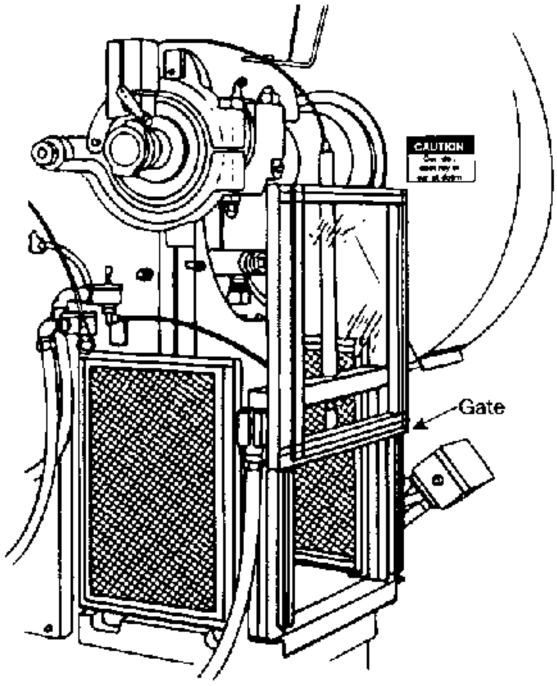


Figure 2 Gate device for projection of the point of operation danger zone. (From Ref. 6.)

which the ram unexpectedly returns down and recloses the die after the cycle has finished. Although the Type B gate is not quite as safe, it affords higher production by shortening the closure of the danger zone to the more dangerous portion, the downward stroke. Control of gates of either Type A or Type B is possible using automated control systems such as are described in this book. The design and control of sophisticated gate devices for punch presses serve as an example of the ways that automated systems can be used to eliminate point of operation hazards on other types of machines as well.

2.3.2.2 Presence-Sensing Devices

Figure 3 illustrates a system for sensing objects that penetrate an infrared screen and enter the danger zone. Such systems are recognized by the OSHA standards as a valid means of safeguarding the point of operation of some types of mechanical power presses. They can also be used very effectively for some machines other than presses.

The principle behind the infrared "light curtain," as these systems are sometimes called, is that the actua-

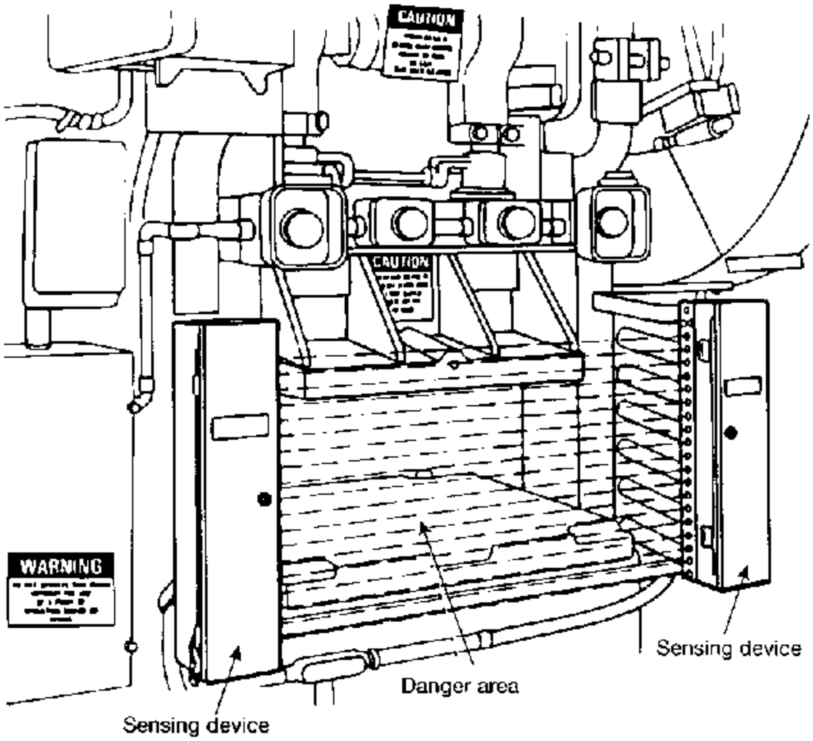


Figure 3 Photoelectric presence sensing screen. (From Ref. 6.)

tion of the machine is halted once the infrared beam is broken. For this principle to work the machine must be able to stop actuation mid-cycle. Some types of mechanical presses, especially the older types, are actuated by a heavy flywheel that is engaged with the press ram in some positive way, such as with a mechanical dog, during each cycle. The inertia of the flywheel is so great that once engaged the ram cannot be stopped. This seriously limits the effectiveness of the infrared sensing barrier, and indeed OSHA standards disqualify the presence-sensing barrier as a means of safeguarding the point of operation on such presses. Fortunately, many modern presses are equipped with clutches and brakes so that the flywheel can be disengaged midstroke.

Visible light can also be used instead of infrared, but infrared has several advantages. The main advantage is that ambient visible light does not trigger a sense condition in an infrared sensor. Besides using infrared radiation, the technology now permits these sensors to be LED-based and to be pulsed. A pulsed LED source of infrared radiation can be programmed in some unique way to produce a characteristic infrared radiation pattern that is not likely to be misinterpreted as ambient radiation. Another way that the pulsed LED infrared source can be programmed is to make a criss-cross pattern across the field so that an object is unable to slip through the barrier without detection.

Other types of presence sensing systems rely upon radio-frequency field disturbances introduced by a person or object penetrating the danger zone. Radio-frequency detection systems can vary in their sensitivities from object to object, depending upon their size and physical composition.

2.3.2.3 Control Systems

Where presence sensing systems are used to protect operators from danger zones, there are ample opportunities to employ automation technology using digital logic to control what takes place whenever the barrier is violated. The same can be said of sophisticated two-hand control systems as shown in Fig. 4. The automated control system can detect when the operator has not depressed the palm buttons concurrently. In other words, the press should not operate if the operator depresses one palm button and then the other.

2.3.2.4 Automatic Brake Monitoring

Another consideration is the monitoring of the clutch and brake arrangement for presses designed to stop in midstroke. If the press is dependent upon the clutch release and brake engagement for safety, the automatic control system can be designed to monitor the condition of the brake and clutch to alert the operator or shut down the machine if stopping time deteriorates beyond established limits. The OSHA is very specific about brake monitoring and control systems as applied to mechanical power presses, as stated in the following provision of the OSHA standard [5]:

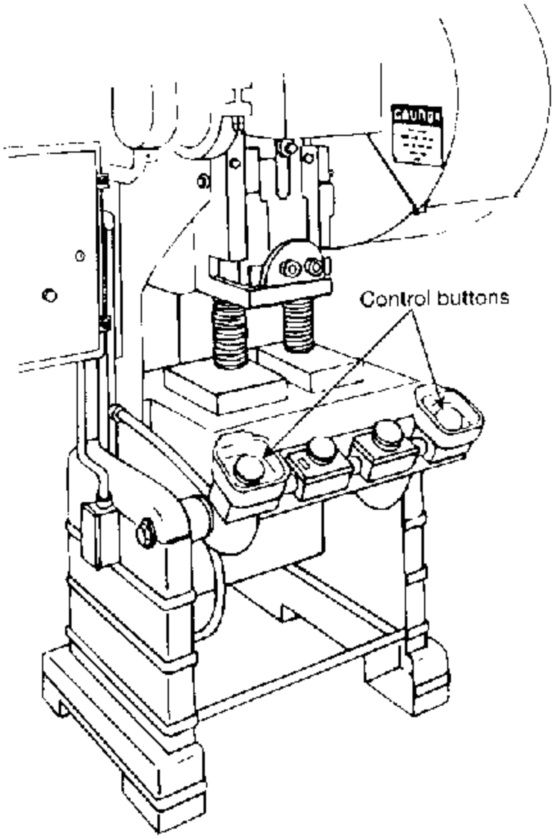


Figure 4 Two-hand palm button control for keeping hands out of the danger zone. (From Ref. 6.)

> 29 CFR 1910. General Industry
> 217. Mechanical power presses
> (c) Safeguarding the point of operation.
> (5) Additional requirements for safe-guarding. Where the operator feeds or removes parts by placing one or both hands in the point of operation, and a two hand control, presence sensing device of Type B gate or movable barrier (on a part revolution clutch) is used for safeguarding:
> (i) The employer shall use a control system and a brake monitor which comply...

For machines other than mechanical power presses, the automated system designer can borrow from the technology required for mechanical power presses.

2.3.3 Interlock Mechanisms

Related to brake monitoring s and control systems are interlockmechanisms for preventing a machine from being operated if any part of the machine is in an unsafe mode. In a very simple example, an ordinary clothes dryer will not operate if the door is open. More complex machines may need to be interlocked dependent upon a complex logical relationship among several variables. The automated system designer is an ideal candidate to delineate the logic that such systems require on a case-by-case basis. A programmable logic controller is an effective device for housing and executing that logic.

2.3.4 Failsafe Systems

The OSHA standards reflect the principles of failsafe operation in a number of the detailed provisions. Automated system design should take into consideration the need to build failsafe logic into the system as it is being developed. For instance, if a fault is detected by the automated system that requires logical action, the control system can disable the ability of the machine to perform additional cycles, but the system can preserve the ability of the machine to keep itself in a safe mode. Example OSHA provisions or portions of provisions that recognize failsafe principles follows:

> 29 CFR 1910. General Industry
> 217. Mechanical power presses
> b. Mechanical power press guarding and construction, general—
> 8. Electrical.
> (iii) All mechanical power press controls shall incorporate a type of drive motor starter that will disconnect the drive motor from the power source in event of control voltage or power source failure, and require operation of the motor start button to restart the motor when voltage conditions are restored to normal.
> (vi) Electrical clutch/brake control circuits shall incorporate features to minimize the possibility of an unintended stroke in the event of the failure of a control component to function properly, including relays, limit switches, and static output circuits. (13) Control reliability.... the control system shall be constructed so that a failure within the system does not prevent the normal stopping action from being applied to the press when required, but does prevent initiation of a successive stroke until the failure is corrected.
> (i) Be so constructed as to automatically prevent the activation of a successive stroke if the stopping time or braking distance deteriorates to a point where...

One can see from the foregoing that the OSHA standards specify the recognition of failsafe principles in the construction of the automatic control systems that are specified to control presses. The principles apply to other machines as well. To an extent, then, it can be said that the OSHA requires automated systems to control the safety of certain machines.

2.4 REPETITIVE MOTION HAZARDS

Many jobs in the manufacturing environment require highly repetitious motion that can lead to trauma. The most publicized of these types of trauma is carpal tunnel syndrome (CTS), a disability of the wrist due to repetitive hand and finger motion. The OSHA has taken note of the high incidence of CTS in industrial exposures, especially in the last two decades of the twentieth century. Various designations have arisen to represent the general class of injuries typified by CTS. One such designation is cumulative trauma disorders (CTD); another is repetitive strain injuries (RSI). The high rate of Workers Compensation claims for these types of injuries during the last two decades of the century led the OSHA to seek a general standard for ergonomics that addressed repetitive motion hazards in a systematic way. At the time of this writing, no general OSHA standard for ergonomics had been finalized. One difficulty is the wide variety of workstation conditions that can lead to repetitive motion injuries. The OSHA attempted to specify a general review process whereby repetitive motion hazards could be identified and removed through redesign of the workplace or by revision of the production process.

Certain industries have proven to be more susceptible to repetitive motion hazards. Among these is the meatpacking industry, especially poultry processing. Any processing or manufacturing activity that is labor intensive is a potential candidate for repetitive motion hazards. In the absence of a specific OSHA standard for ergonomics or for repetitive motion injuries, the OSHA has turned to the General Duty Clause for citations of this hazard, with associated heavy monetary penalties. An example is a snack food manufacturer, cited in the early 1990s for a total of $1,384,000 in penalties, of which more than half was for "repetitive motion illnesses." Each instance of citation for repetitive motion illness resulted in a proposed penalty of $5000, for a total of $875,000 for repetitive motion illnesses alone [7].

Repetitive motion illnesses have been one of the biggest drivers of automation research in the meatpacking industry. By transferring jobs that are at risk from people to machines, the company often removes hazards, avoids OSHA penalties, and reduces production costs all at the same time. It is for this reason, more than any other, that this chapter has been included in this handbook.

2.5 HAZARDOUS ENVIRONMENTS

Another driver for automation in the workplace is the presence of hot, toxic, radioactive, flammable, or simply unpleasant work environments. There are many ways automation can be brought to bear on this problem. Even when workers remain in the hazardous environment, automated systems can monitor exposures and control ventilation and other systems for keeping hazards in check. In other situations, too hazardous for human exposure, robotic solutions to production operations may be the answer.

2.5.1 Environmental Monitoring

The OSHA sets exposure limits for air contaminants in the workplace. These limits are designated permissible exposure levels (PELs). For most air contaminants these limits are expressed as 8 hr time-weighted averages (TWAs). This procedure recognizes that even very harmful toxic contaminants are tolerable for short durations. For most toxic substances, what is important is the overall average exposure for a full shift. The OSHA PELs are based upon findings of the American Conference of Governmental Industrial Hygienists (ACGIH) Committee on Industrial Ventilation (CIV). ACGIH publishes a list of threshold limit values (TLVs) for each toxic air contaminant. The TLV is considered the level above which the contaminant can do significant, permanent harm to the e individual exposed. The OSHA attempts to set PELs at the TLV levels, whenever possible, however it is difficult to promulgate a new standard every time the TLV is adjusted by ACGIH. The OSHA also recognizes an action level (AL) which is generally one-half the concentration of the PEL. The idea behind the action level is to provide a buffer to permit actions to be taken to control harmful air contaminants before they reach illegal levels.

Without some sort of automated monitoring system it is difficult to respond to a change in concentration of toxic air contaminants. This suggests that automated systems can be employed to not only control the hazardous environment, but also to provide documentation of compliance with OSHA standards for air contaminants. Such systems can be controlled by a computer or a programmable logic controller. Figure 5 diagrams an environmental control system that suggests an automated logic control system to monitor filter conditions, control valves, and log compliance with air contamination limits.

2.5.2 Robotic Solutions

Some environments are too hazardous for employee exposure. Even some environments that in the past have been marginal are now considered intolerable in light of increased awareness of long-term effects of toxic air contaminants and the technological and economic price of bringing these environments into compliance with OSHA standards. The presence of federal regulation and the OSHA has also made management and worker sensitive to the safety hazards associated worker exposures to some hazardous environments. All of these factors are supporting increased research into robotic solutions to make possible the removal of the human worker from direct exposure to the hazards.

On the other side of the robotics issue is the fact that robots themselves present new hazards. The intermittent, unpredictable nature of their very flexible operation, when programmed to perform a complex task, increases the hazard when personnel are exposed to the robot's work movement envelope.

The OSHA has been sensitive to the new hazards the industrial robot brings to the workplace. However, the development of standards to deal with these hazards is a slow process. Most of the OSHA General Industry Standards were put into effect as national consensus standards in the OSHA's opening years. In the early 1970s industrial robots were scarcely in existence, much less were national consensus standards to address their safety. As of this writing there still exists no specific standard addressing the safety of industrial robots.

Whenever a serious accident or fatality results from human exposure to industrial robots, the OSHA

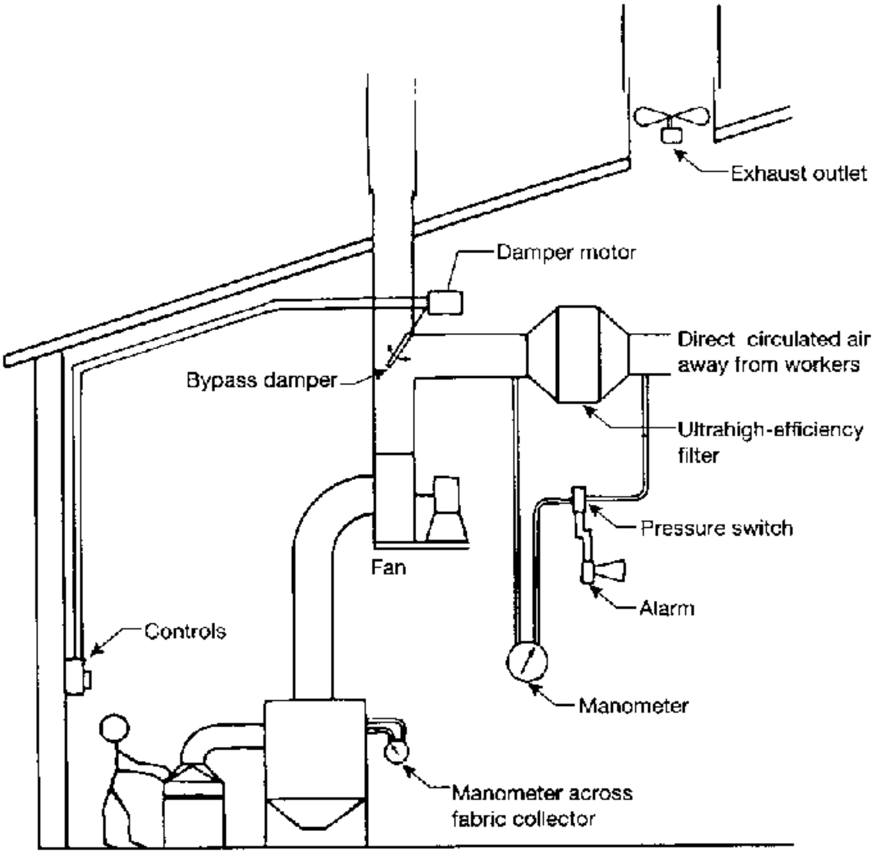

Exhaust outlet
Damper motor
Direct circulated air away from workers
Bypass damper
Ultrahigh-efficiency filter
Pressure switch
Fan
Alarm
Controls
Manometer
Manometer across fabric collector

Figure 5 An automatic control system for exhaust ventilation for removing particulate air contaminants. (From Ref. 8, used with permission.)

generally cites Section 5.a.l—General Duty Clause of the OSHA law. Penalties are typically severe. Even without the presence of safety hazards, robots in the workplace generate animosity on the part of some employees. When an accident occurs, that animosity is amplified, not only by workers, but by the press as well.

If a firm is cited for an alleged hazardous condition caused by the motion of a robotic system, the Occupational Safety and Health Review Commission (OSHRC) may recognize the benefit provided by an automatic interlock system that controls access to the danger zone, as the following excerpt from an OSHRC decision demonstrates [9]:

> It is undisputed that the machines had extensive precautions to protect servicing and maintenance employees. An electronically interlocked gate surrounded the machine area in each case. Once an employee opened that gate or pushed an emergency stop button, a time-consuming series of eight to twelve steps were required before any hazardous movement of the machine could occur. The evidence indicated that the restart procedures would provide plenty of warning to the employees, in the form of alarms and visible motions, so that they could avoid any hazardous movement of the machinery.

The advantage of an automated system to control access by means of an interlock is obvious in the foregoing description of an alleged hazardous condition at General Motors Corporation that was vacated by the reviewing body.

2.6 SUMMARY

This chapter began with an introduction to OSHA and the origin of federal standards that have an impact upon the field of industrial automation. The OSHA has been granted powers by Congress to make inspections and write citations for alleged violations of detailed standards from a large volume of general industry standards. The OSHA has changed the way companies have dealt with safety and health standards. One of these changes is that issues that were formerly advisory in nature have become mandatory with penalties for violation. Another way that standards have changed is that a more horizontal structure predominates the OSHA approach, which has replaced a more vertical, or industry-by-industry approach used by the various states prior to the OSHA.

Despite the detail of the OSHA standards, many situations believed to be hazardous do not seem to fit any of the described conditions spelled out in existing national consensus standards. This is especially true of industrial automation systems, such as robots, material handling systems, and process control systems. Many modern industrial automation systems, such as those described in this book, were not invented at the time of the writing of the general body of national consensus standards that represent the bulk of what the OSHA enforces. Therefore, the OSHA has turned to a general provision of the OSHA law itself, the so-called "General Duty Clause" of the OSHA Act. Many citations of industrial automation systems are citations of this general duty clause, and it is a provision that the automated system designer should understand and heed. The OSHA does have a priority structure for inspections, placing imminent danger highest, followed by fatalities and major accidents, and employee complaints, respectively. Because of the resentment that industrial automation systems sometimes engender on the part of employees, system designers should prepare for possible OSHA inspection as a result of employee complaints. If an employer does receive a citation of alleged violations of OSHA standards, there is an appeals process available, but timely action is essential. Recordkeeping is another aspect of the safety system that OSHA takes very seriously, as is evidenced by significant fines for recordkeeping violations. One reason for the importance placed upon recordkeeping is that it forms the database for the computation of performance measures that are used to evaluate both companies and entire industries and to be used as decision criteria for OSHA inspection frequency.

Machine guarding is one of the most relevant areas of federal standards as they apply to industrial automation systems. Within the machine guarding standards, the most significant application area for automated systems is guarding the point of operation. The OSHA has some very specific rules for guarding the point of operation on certain machines, such as mechanical power presses. The principle behind these specific rules can be used in the design of industrial automation systems to be used on a wide variety of machines in general. Typical point-of-operation safeguarding strategies that employ automated control systems are gates, presence sensing devices, and two-hand controls. To facilitate the operation of these systems, overall machine control systems and automatic brake monitoring systems are also appropriate in many cases. The concepts of interlock mechanisms and failsafe principles also apply to many of the OSHA standards and are relevant to the application of the principles contained in this book.

An industrial hazard category of increasing significance is the collection of disorders known as "repetitive motion hazards." The OSHA has responded to this increasing significance by vigorous enforcement in industries reporting a high incidence of this type of hazard. Since there is no specific standard for this class of hazard, OSHA is citing the General Duty Clause and is proposing severe penalties, especially in the meatpacking industry. The presence of repetitive motion hazards and OSHA penalties is a strong motivation to implement a program of industrial automation using the concepts and principles contained in this book.

Another strong motivation for automation is the presence of hazardous environments, such as atmospheres contaminated with toxic substances. The OSHA has had an impact in this area, too, by setting standard "permissible exposure levels" (PELs). For the most part these PELs are time-weighted averages over an 8 hr shift. Such standards form a rational basis for implementing automatic monitoring and control systems for protecting employees and documenting compliance with OSHA standards.

Finally, some environments are so hazardous that they support development and investment in industrial robots to replace humans in these undesirable workstations. The presence of federal standards and the OSHA enforcement agency have a significant impact upon the present and future development of robots and industrial automation systems. At the same time, industrial automation systems designers should remain aware of the fact that public sentiment is not very tolerant of machines that hurt people. The sentiment is particularly compelling if that machine is an industrial robot, which is often perceived as a replacement for human workers. The profession of the industrial automation system designer, then, is challenged with the task of providing robots and automated systems that eliminate hazards and heed OSHA standards, while at the same time taking care not to introduce new hazards that run afoul of other OSHA standards.

REFERENCES

1. CR Asfahl. Industrial Safety and Health Management. 3rd ed. Upper Saddle River, NJ: Prentice-Hall, 1995.
2. CR Asfahl. Robots and Manufacturing Automation. Rev ed. New York: John Wiley, 1992.
3. Job Safety and Health Act of 1970. Washington, DC: The Bureau of National Affairs, 1971.
4. Williams-Steiger Occupational Safety and Health Act of 1970 (Public Law 91-596). December 29, 1970.
5. BH Amstead, F Ostwald, ML Begeman. Manufacturing Processes. 7th ed. New York: John Wiley, 1979.
6. General Industry OSHA Safety and Health Standards (29CFR1910, OSHA 2206 (rev). Washington DC: U.S. Department of Labor, Occupational Safety and Health Administration, 1989.
7. Concepts and Techniques of Machine Safeguarding (OSHA 3067). Washington, DC: U.S. Department of Labor, Occupational Safety and Health Administration, 1980.
8. OSHA proposes fines for ergonomics-related injuries. Mater Handling Eng 44(2): 42, 1989.
9. Industrial Ventilation. 15th ed. Lansing, MI. American Conference of Governmental Industrial Hygienists, Committee on Industrial Ventilation, 1978.
10. OSHRC Docket Nos. 91-2973, 91-3116, and 91-3117. Secretary of Labor, Complainant, vs. General Motors Corporation, Delco Chassis Division, Respondent, April 26, 1995.

Chapter 8.3

Standards

Verna Fitzsimmons and Ron Collier
University of Cincinnati, Cincinnati, Ohio

3.1 INTRODUCTION

Standards and codes are part of our everyday life. We are raised with a set of standards taught to us by our parents. We abide by a set of standard rules and codes at school and in sports. We are considered "law-abiding citizens" if we follow the social rules and regulations established in our communities. It should not be a surprise, then, that we also have standards, codes, regulations, and guidelines in our work. All professionals have a "Code of Ethics" usually developed by the professional organization.

A standard or code is simply a rule, test or, requirement that is deemed necessary by someone in authority. There are domestic or national standards and codes (defines as applicable in originating country). There are international rules. Some of these rules are required by law (mandatory) in the form of regulations, while others are recommended practices or guidelines which can be interpreted as voluntary. There are rules for product design, compatibility, safety, and reliability. There are rules for construction, manufacturing, and processes. There are rules of management, operations, and for selling goods and services.

This is just the beginning! A word of caution: do not underestimate the importance of standards, codes, regulations, recommended practices, and guidelines. Every person, every process, and every product is, in some way, affected by these. It is the duty of the engineer to understand the applicable standards, codes, regulations, design practices, and guidelines and to be able to apply them in his or her work.

There are literally hundreds of standards, codes, design practices, and guidelines that come into play within the engineering professions. Regulations are equally voluminous. Unfortunately, being able to determine or understand exactly what standards, codes, or regulations may apply to any given engineering design or project can be daunting to say the least, especially for the engineer just starting his or her career. The labyrinth of standards, codes, regulations, recommended practices, and guidelines that could apply, and in many cases are required, can best be described as a bowl of spaghetti. These crucial documents are published by an array of organizations (government and private) and, as often as not, are intertwined and interrelated.

Just where does one start? This chapter is intended to help the engineer answer this question, and find some path through the spaghetti bowl called standards, codes, regulations, design practices, and guidelines.

3.2 TERMINOLOGY

First, there is much confusion derived from terminology. Many organizations that develop and publish standards, codes, regulations, recommended practices, and guidelines toss these terms around lavishly and in many cases use them interchangeably. This can add to the confusion and the complexity of when and where to apply a standard, code, regulation, practice, or guideline.

In order to reduce this confusion a review of the definition of each of these terms is recommended. *Webster's New Collegiate Dictionary* defines each of these terms as follows:

Standard: "Something established by authority, custom, or general consent as a model or example: CRITERION: something set-up and established by authority as a rule for measure of quantity, weight, extent, value, or quality."

Code: "A systematic statement of a body of law; one given statutory force. A system of principles or rule."

Practice: "To do or perform often, customarily, or habitually: to do something customarily: a repeated or customary action: the usual way of doing something: systematic exercise for proficiency."

Regulation: "An authoritative rule dealing with details or procedure: a rule or order having the force of law issued by an executive authority of a government."

Guideline: "A line by which one is guided: an indication or outline (as by a government) of policy or conduct."

As can be seen from these definitions, there are some distinct differences between standards, codes, regulations, practices, and guidelines. For example, in the definition of a standard, notice the phrase "established by authority." This authority is usually an appointed body, association or committee of interested experts from a particular field of interest.

This convened body of experts does not have legislative authority. The committee develops a "consensus" standard, so named because a majority of the committee members must agree. Unanimity is not necessary, and in theory, not preferred, since unanimity generally results in a standard of lower quality. Their primary objective is to formulate a standard set of rules that meets the due process requirements of their industry and the intent of the standard they are to develop. The committee is also considered a duly constituted review authority for the standard and subsequent revisions to it. Therefore, the standards that are developed, although established by an authoritative group of recognized experts, do not have the force of law. Standards are documents that require some degree of consent [1] by the users; standards generally have no criminal penalty attached to them for noncompliance. They are strictly voluntary.

Again, there are exceptions to the "voluntary" standard definition. If a consensus standard is referenced in a federal, state, or local law, then that consensus standard can become a part of that law and made mandatory by de facto. An example of this is the listing of certain National Fire Prevention Association (NFPA) standards and guidelines in many state building codes. The NFPA standards and guidelines are documents that establish a reasonable level of fire safety for life and property. The NFPA standards and codes themselves do not carry the force of law. They are recommended practices based on the fire prevention experience that builders of industrial, commercial, and residential properties, and their insurance underwriters have gained over many years of study. They are purely advisory documents as far as the National Fire Protection Associate is concerned [2]. When listed in a local building code, which does have the force of law, these referenced standards and guidelines become a part of the building code and, therefore, state law.

Another example of where NFPA standards and guidelines have been adopted into law is the inclusion of many of them in the Occupational Safety and Health Act (OSHA) of 1970. The OSHA uses the NFPA standards and codes either by citing them as references in the OSHA regulation, or by actually extracting the written text from the NFPA standard and guideline and including it directly within the body of the OSHA rule or regulation. In either case, the NFPA standard or guideline cited becomes part of the federal law.

International standards, such as those developed by the International Standards Organization (ISO) and the International Electrotechnical Committee (IEC), are consensus standards. In these cases, interested experts from different countries come together. The process involves the standard being approved and accepted by the member countries.

There is also a blurring of definitions between governmental and nongovernmental organizations. The federal government promulgates standards that have the force of law. These standards are usually consensus standards that are first developed by a committee of nongovernmental, voluntary experts and reviewed by the interested industries as well as government agencies before being incorporated into the law as a regulation published in the Code of Federal Register.

All of this simply means that a document called a "standard" may or may not be required by law, carrying noncompliance and, possibly, criminal penalties. The engineer must determine the origin and authority of any standard that may apply to the project. Ignoring the history and legal authority of a standard can prove to be detrimental and costly to the engineer and the company. As a rule of thumb, it is best to comply with any standard related to the project work, in particular, if the standard deals with public or worker safety.

Codes, on the other hand, do have the force of law. Paraphrasing *Webster's* definition will help to understand the difference between a standard and a code. A code is a system of principles or rules that have statutory force by their inclusion or reference in a systematic statement of a body of law [i.e., Code of Federal Register (CFR), or State Administrative Code (SAC)]. This body of law is generally established by a group of individuals who has legislative authority (i.e., congress, state legislators, etc.). Noncompliance with a code can carry criminal penalties and can be punishable by fines, and in some cases prosecution and imprisonment.

In the United States, codes and regulations are generally recognized as requirements that have the force of law. An example of this is the three-model Building Code: the Uniform Building Code (UBC), the National Building Code (NBC), and the Southern Building Code (SBC). These documents detail the design and occupancy requirements for all buildings in the United States. They become law by adoption and reference in the state and local statutes and regulations. The individual states and localities can pass tighter codes, but the model Building Codes are the minimum requirement in the jurisdictions adopting them.

Another example of a standard that became law through codification is the American Society of Mechanical Engineers (ASME) Boiler and Pressure Vessel Code. This code is recognized throughout the world as the primary document for the design and construction of boilers and pressure vessels used in industry today. It not only establishes design and construction parameters for all boilers and pressure vessels used in the United States, Canada, and other parts of the world, but also details maintenance, and inspection criteria for this equipment.

The Boiler and Pressure Vessels Code was originally developed as a recommended consensus standard after industry had experienced a significant number of boiler explosions that resulted in loss of life and property during the early 1900s. Since then, the original standard has been legally adopted (codified), in whole or in part, by most countries using boilers and pressure vessels. The standard has also become an American National Standards Institute (ANSI) standard. Many of the state building codes and the federal Environment Protection Agency (EPA) and OSHA regulations reference ASME Boiler and Pressure Vessel Code in the body of their text, thus making the Boiler and Pressure Vessel Code part of the law by inclusion. The ASME Boiler and Pressure Vessel Code is a living document that is constantly undergoing reviews and revision as the state of the art in technology changes.

In addition to standards, codes, and regulations there are recommended design practices and guidelines. These are usually written methods or procedures for development of a product or process. Generally, they are developed within a particular industry and are accepted as the ''best'' method for conducting business. These documents are often referred to as ''Good Engineering Practices'' or ''Good Manufacturing Practices (GMP).'' In all cases the development of design practices or guidelines is based on experience of the particular industry to which the practice or guideline applies, and has developed over time. Usually, this experience factor is documented and the guidelines are continuously updated by the industry as further experience is gained.

The key point to remember about recommended practices and guidelines is that they are not fixed, firm methods that must be followed without question. They are rarely, if ever, considered mandatory. Instead, they are procedures or principles that are based on experience factors and should be considered, built upon, and improved. They do not depict the only way of doing something. They are but one possible solution. Recommended practices and guidelines allow the engineer to include independent thoughts on how something can be designed or built.

Recommended practices and guidelines are established much in the same way as standards. These generally tend to be more corporate specific, as opposed to industry-wide acceptance. They are generally used for standardization of procedures and methods for performing a particular function. Penalty for noncompliance is internal to the company and is left to the discretion of management.

However, there is another exception. The Federal Drug Administration (FDA) requires GMP to be followed in drug and medical device manufacturing. Normally, one would consider GMP as an administrative management decision, not a requirement imposed by a government agency.

From the above discussion one hopefully begins to appreciate why the term ''bowl of spaghetti'' is used when considering the subject of standards, codes, regulations, practices, and guidelines.

Since many organizations use the terms standard, code, regulation, practice, and guideline interchangeably in their documents, how does one determine a particular document's legal or mandatory requirement, In particular, if the document has applicability to the project's work? An easy way for the engineer to do this is to look for key words internal to the standard, code, and regulation, practice, or guideline.

The NFPA's definition for each of the above terms demonstrates this "word association" technique. The NFPA explains:

> A *standard* is a document containing both mandatory provisions (using the word "shall" to indicate requirements) and advisory provisions (using the word "should" in indicate recommendations [3].
>
> A *code* or regulation is a document containing only mandatory provisions using the word "shall" in indicate requirements. Explanatory material may be included only in the form of "fine print notes," in footnotes, or in an appendix [3].
>
> A *recommended practice* is a document containing only advisory provisions (using the word "should" to indicate recommendations) in the body of the text [3].
>
> A *guideline* is a document which is informative in nature and does not contain requirements or recommendations [3].

This word association technique is not 100% foolproof, and should not be substituted for a thorough verification of the document's mandatory or statutory requirements. It is, however, a consistent method to use to get started in determining what standards, codes, regulations, and guidelines apply to a given project. It should also be noted that this word association technique works equally as well with international standards.

In this chapter, the terms standard, code, regulation, recommended practice, and guideline will be used as the parent or originating organization has called it. This is done so that as the engineer looks for the needed document, the published term can be used.

3.3 WHY DO WE HAVE STANDARDS?

Standards have been developed because of some perceived need. There are a variety of driving forces behind standards. Governments and consumers demand safety, quality, and performance. Marketing departments demand products that can be sold to the targeted market and the ability to compete in specific markets. The military and aerospace industries require high reliability and operation in adverse environmental conditions.

Some newly emerging markets are fragile and need protection. Some markets are very sensitive and need protection for national interest reasons. There is an increase in imports and exports requiring standards to protect the environment, agriculture, and consumers. By having rules, governments and consumers can have some degree of confidence that what they are allowing into the country and buying is not going to be harmful to the people or environment.

In some industries, there has been a converging of product designs and materials that have proven to be the best practice. For example, videocassettes have converged to the VHS format in the United States, leaving people with beta machines out in the cold. The same has happened with computer floppy disks. A person would be hard pressed to find a 5.25 in. disk drive as standard equipment in today's PC.

Standardization helps to advance society. The Romans standardized the width of roads. Without standardization railways could not share tracks. Transportation standards incorporating seat belts backed by the force of law in their use has reduced deaths as a result of car accidents.

3.4 WHY DO WE NEED STANDARDS?

The only reason for standards is protection! To protect public safety, the federal government passed regulations such as the Clean Air Act (CAA), Resource Conservation and Recovery Act (RCRA), Clean Water Act (CWA), and the Occupational Safety and Health Act of 1970. The government has created agencies such as the OSHA and the EPA to oversee compliance with these rules. The air and water quality is regulated by the EPA. To protect the health and safety of workers, the OSHA has regulatory authority.

Depending on what part of the United States you are in, building construction and occupancy safety is governed by one of three model building codes; the Uniform Building Code, the National Building Code, or the Southern Building Code. These building codes have the force of law through their adoption by either state or, in some cases, local administrative codes. By the year 2000 all three of these codes will be combined into what will be known as the International Building Code and will be used by the United States and Canada for all building construction. The International Mechanical Code and the International Plumbing Code are already in use by both countries.

The United States protects consumers by federal agencies such as the FDA, and the Consumer Product and Safety Commission (CPSC). Governments protect fragile markets with such rules as those developed by the Securities and Exchange Commission § and the various banking laws.

In order to ensure product usability and compatibility, professional organizations such as the Institute of Electrical and Electronic Engineers (IEEE) have issued standards for electrical and electronic devices. Many of the IEEE standards have been accepted worldwide.

As consumers, we want to feel confident that when we purchase a product it will work or perform and that it continues in operation for a reasonable length of time. The standards world is present in this arena. Underwriters Laboratory has been active in developing standards and testing methods for electrical and electronic equipment for decades. The ISO has developed a set of standards known as ISO 9000, which is a quality management standard. This particular set of documents has been accepted worldwide and many customers are demanding compliance to it.

The ISO has continued in its standard writing efforts for the general protection of the environment; ISO 14000 is the first international environmental management standard.

3.5 ROLE OF STANDARDS ORGANIZATIONS

Standard writing organizations give consideration to safety, product usefulness and service life, as well as advancements in technology, and past experiences. They assist interested parties in identifying the common basis for behavior for a product, whether the product is a toy or a chemical refinery or an airplane. The interested parties include suppliers (manufacturers, sellers, designers, etc.), consumers (end users, purchasers, etc.), experts (technicians, researchers, lawyers, etc.), advocates (politicians, moralists, etc.), government agencies, and professional associations.

The common base of behavior has several components. The minimum acceptable level of design or performance is studied. An established method of testing is needed to ensure compliance. This leads to the need for measurements and a way of grading the variations. Processes are also a piece of the common base of behavior. Standardize design criteria is also considered. This includes possible interchangeability and the actual use of anything.

3.6 WHEN DO WE APPLY STANDARDS, CODES, RECOMMENDED PRACTICES, AND GUIDELINES?

In general, we would like to say that standards, codes, and guidelines are used everyday to protect property and life. However, in reality, standards, codes, and guidelines are used when it is in the best interest of the business to do so. This means that if a business can sell its goods and services only if the goods and services are in compliance with some set of standards, codes, or guidelines, then businesses will conform. The threat of criminal or civil penalties for noncompliance also provides motivation for businesses to comply.

3.7 PROVING COMPLIANCE

Now that we know the standards, codes, and guidelines associated with the project, determination of compliance is needed. In the standards and codes that are dictated by law, a means for verifying compliance is also dictated. The means of verifying compliance can vary from self-certification to a third-party audit by a notified body.

Self-certification is the process whereby the company with the legal responsibility for the project performs the required testing for the project and signs a document detailing compliance. Self-certification can be economically attractive, but may not be available in all situations.

When self-certification is not an option, independent testing houses are available. There are many organizations around the world that will test a product, process or operations for compliance with standards. These groups are generally known as "third party" because they are not directly involved with the design, manufacturing or sale of the items that they test. In more and more situations consumers and governments are requiring this type of third party testing or auditing as proof of compliance. In the United States, the OSHA maintains a list of what are called the National Recognized Testing Laboratories (NRTL).

Third-party testing is evident all around us. Just look at the back of any computer monitor or power tool tags. There are labels stating compliance with various standards and codes.

3.8 WHERE TO GO TO FIND STANDARDS

When starting a new project the question of what standards, codes, design practices, and guidelines apply, and where to go to find them can be a nagging concern, and impact schedule. It would be nice to have a convenient list or "road map" to get started, but unfortunately standards are not that simple.

The first place for any engineer to start is right inside the company. Most medium- to large-sized companies and many small companies have developed their own set of standards called policies and procedures. These rules can dictate as well as guide the engineer through product design, manufacturing, operations, sales and marketing, maintenance, and decommission. Company standards are usually reviewed by the company legal representative or an attorney in that industry. The complex product liability and malpractice law suits have made this review advisable and, in some cases, a business necessity.

Within the company, there may be another resource. The very lucky engineer might find an "in-house guru," that is, someone who has significant experience with standards and their application. This resident expert may be another engineer or legal counsel. These resources may be able to provide a historical look at the product, process or operation evolution, including standards, codes, regulations, practices, and guidelines.

If the company is a consulting firm, the client will usually specify the appropriate standards in the contract. However, it may be up to the consultant to discover standards, codes, regulations, and guidelines that apply that the client is not aware of.

The NRTLs can also provide assistance in determining what standards may be required. Some of the NRTLs in the United States have developed agreements with testing houses in other countries. This provides additional assistance in requirements for sale of a product anywhere in the world.

Another resource that the engineer should use is industry-related societies and associations. These organizations frequently have libraries with standards, codes, regulations, practices, and guidelines available for member use. They often support lobbying efforts on behalf of the industry as well.

Also available to the engineer are outside information service organizations that can be accessed through the internet. Most of these organizations have a wide range of technical databases, which include US and International Industrial Standards, Military and Federal specifications, Electronic Component Databases, Medical Device Standards, Regulatory Data, vendor equipment catalog information, and more. These service organizations usually have the most complete sets of data available under one source library, and they regularly update this information to keep it current. Their databases can be delivered right to the engineer's desktop computer in what ever format is needed. Such access to critical engineering data, standards, codes, and regulatory data is very cost effective, and in most cases reduces the time spent researching for the required project information. Most of these information services organizations can provide their information on CD-ROM, or through online databases and internet websites. When project deadlines or product-to-market schedules are critical, efficient and effective access to key standards, codes, and regulations can be of the utmost importance if a successful project is to be achieved.

3.9 WHAT IF NO STANDARD EXISTS?

From time to time engineers are presented with the opportunity to work on a project that uses "cutting-edge technology" to produce a new, never-before-seen product or process. The engineer assigned to such a job often discovers that existing standards do not adequately cover the new design. They may in some cases find that adequate standards simply do not exist. The latter case is rare. Because of the vast arena of standards, codes, regulations, practices, and guidelines existing in industry and the military today, it is hard to develop any new product or process not yet addressed to some degree. For those cases where existing standards are not adequate, what does the engineer do? Where does one start?

The first reaction the engineer may have when finding themselves in this situation may be: "Great! I don't have any standards, codes, or regulations that need considering, so I have a wide-open field of design opportunities." To some degree this is true. However, the engineer should not forget why there are standards, codes, regulations, recommended practices, and guidelines in the first place. For protection! The engineer has a direct and vital impact on the quality of life of many people and their "Code of Ethics" requires them to be dedicated to the protection of the public health, safety, and welfare [4]. Thus, the real opportunity comes with being the first to set the pace for technology yet to come. Being the first one to develop and establish an industry standard provides the opportunity for the engineer to become the expert, the technical guru of this new field of interest.

When developing a new standard, practice or guideline, two key points to remember are to use common sense and simplify the language of the document. Eliminating ‘‘technospeak’’ in the body of text will add clarity and simplicity to the document. This in turn improves the understanding of the subject matter and usability of the document.

The first step in developing a standard for the new work is to organize a committee of interested experts from within the company. This team may include key members of the product or process design team, manufacturing, health and safety representative, etc. Once assembled, determine the scope of the effort. A review committee that includes corporate counsel should also be established.

The next step is to ask the questions, ‘‘Is a standard necessary or would a recommended practice or guideline be more appropriate?’’ A recommended practice or guideline provides more latitude in the product, process, or operation development. If the team of experts determine a standard is necessary, then a development subcommittee should be appointed to prepare a draft text of the proposed standard.

The first thing the development subcommittee should consider is using a technical data or information service organization to assemble any related standards, codes, regulations, practices, and guidelines for there assignment. These documents although not directly applicable, often as not can be used as resources and possible templates for the drafting of the proposed standard.

After the development subcommittee has prepare a draft of the proposed standard, then the entire team of experts should review the draft text, resolve differences, provide clarification where needed, and develop the document details for a final presentation to the review committee.

The next step, and usually a very wise one, is to get a legal review of the new proposed standard by the company’s legal department or through an independent attorney who specializes in corporate liability affairs.

After the draft of the proposed standard is reviewed and all legal questions answered, a final draft can be prepared and circulated among all team members along with the review committee’s comments and recommended changes.

A deadline for receipt of all advice and comments from team members on the final draft of the proposed standard should be established and rigidly held to.

A consensus vote of all team members should then be called on the final draft of the proposed standard, with the understanding that a majority vote of team members determines acceptance of the final draft as written.

A test of the proposed standard’s validity by benchmarking it against an existing standard for a comparable product or process, even if a competitor’s, should be performed. Under certain cases where public safety may be impacted by the new process or product, it may be wise to solicit public comment as well.

After performing a benchmark test of the proposed standard, further meetings should be held so that the team of experts can reconsider the standard in regards to the comments or data received from the test, or public comment if that be the case. The review team should also provide their input on these results and comments. The standard should be amended accordingly if necessary, and a reballoting of all team members with two-thirds approval required to proceed with publication.

The official standard should be published and submitted to the corporate board of directors or the industrial association within which the standard is to be applied, for adoption and signature.

Following its official adoption the new standard should be issued by its parenting organizations for implementation. Where appropriate the standard should be presented to the American National Standards Institute (ANSI) for approval as an ANSI standard in accordance with the institute’s by-laws and regulations.

Although the above steps are not inclusive of the total process to develop and implement a new standard, they do cover the basics and are indicative of the same process used by many of the larger standards writing organization. The key point that the engineer must remember when developing any new standard, practice, or guideline, is that the document must address any process or product parameter that could presently or in the future have an impact upon the health, safety, and welfare of the workers, consumer, and general public. If the engineer can achieve this, then the standard will be a credit to the profession and the process of developing it will be a rewarding experience. It is not an easy task, but it can be a rewarding one.

3.10 SUMMARY

The engineer has a professional obligation to protect the general health and welfare of the public. This encompasses any, and all, work done while working for a corporation, or as an independent consultant. The engineer is assisted in this endeavor by using the appropriate standards, codes, regulations, practices, and guidelines as required. This may sound relatively simple, but it is not. Just locating the appropriate documents can be an awesome and time-consuming experience.

The first place the engineer should start in the quest for the applicable rules and guides is within the company's policies and procedures. Another starting point for the engineer would be to contract a technical data or information service organization to do the search, or provide the databases by which the research can be expedited, especially if the engineer's project schedule is critical. Such database services can assist the engineer in reducing the time required to locate project pertinent standards, codes, regulations, practices and guidelines. The less time the engineer spends searching for critical documents the more time he or she will have to apply to the design, which in most cases, is schedule driven.

Maneuvering through the maze of standards, codes, regulations, design practices, and guidelines that are required for projects in today's global economy can be an engineer's nightmare. The efficiency and effectiveness of how quick this documentation can be assessed and absorbed often dictates the difference between a successful project and one that fails. A solid understanding of standards, codes, regulations, practices, and guidelines by the engineer cannot be emphasized more strongly when considering the litigious society in which we live and work. "Remember, always design per codes and standards. Just because you meet the codes and standards established by your profession, does not mean you or your company is off the liability hook. However, if you don't meet the codes and standards, you can rest assured you will be held negligent should someone suffer personal harm or property damage" (from a personal conversation with Dr Rodney Simmons, University of Cincinnati, May 1996).

A word of encouragement: the engineer cannot hope to be a marvel at determining what standards, codes, regulations, practices, or guidelines apply or where to find the documents when first entering this strange world. This skill is learned and developed through practice and time spent in the maze. It does become easier. Good luck!

3.11 ORGANIZATION LISTINGS

This section lists the names, addresses, phone numbers, and websites of parenting organizations for many of the standards, codes, regulations, practices, and guidelines used throughout the engineering community. It is a generalized list and is not specific to any one engineering discipline. Many corporate engineering departments and consulting firms have similar listings to aid their staff in reaching the required documents. Also, it is not a comprehensive list of all organizations who establish and maintain such documents. It is an attempt at providing the engineer with a list of the more commonly known organizations as a starting point to build upon and develop their own.

One point of caution must be noted: as with any list of addresses, phone numbers, and even names of organizations, this information is subject to change! Organizations tend to move, merge, change their name, and in some cases cease to exist. The following list of organizations, and the information provided for each one, is current at the time of writing this chapter. The engineer is encouraged to add organizations to this list as experience demands, and to review and update the information on a regular basis.

3.11.1 Regulatory Organizations

CA Congressional Acts
Superintendent of Documents
U.S. Government Printing Office
809 Channing Place, NE
Washington, DC 20018
Phone: 202/576-6693
Website: www.explore.gpo.gov

CFR Code of Federal Regulations
Superintendent of Documents
U.S. Government Printing Office
809 Channing Place, NE
Washington, DC 20018
Phone: 202/576-6693
Website: www.explore.gpo.gov

DOE U.S. Department of Energy
1000 Independence Avenue, SW
Washington, DC 20585
Phone: 202/586-9642
Website: www.explore.doe.gov

EO Executive Orders
National Archives and Records Administration
8th Street and Pennsylvania Avenue, NW
Washington, DC 20408

Phone: 202/523-5230
Website: www.nara.gov/fedreg

ERDA (Energy Research and Development
Administration, see U.S. Department of Energy,
DOE)

FR Federal Register
Superintendent of Documents
U.S. Government Printing Office
809 Channing Place, NE
Washington, DC 20018
Phone: 202/576-6693
Website: www.explore.gpo.gov

GSA General Services Administration
Public Buildings Service
Office of Government-wide Real Property Policy
and Oversight
19th and F Streets, NW
Washington, DC 20405
Phone: 202/501-0398
Website: www.explore.gsa.gov

OMB Office of Management and Budget
Executive Office of the President
Washington, DC 20503
Phone: 202/395-3000
Website: www.whitehouse.gov/wh/eop/omb

3.11.2 Reference Standards and Guides Organizations

AA Aluminum Association
900 19th Street, NW, Suite 300
Washington, DC 20006
Phone: 202/862-5100
Website: www.aluminum.org

AABC Associated Air Balance Council
1518 K Street, NW
Washington, DC 20005
Phone: 202/737-0202
Website: www.aabchq.com

AAMA American Architectural Manufacturers
Association
1827 Walden Office Square, Suite 104
Schaumberg, IL 60173
Phone: 847/303-5664
Website: www.aamanet.org

AAMI Association for the Advancement of
Medical Instrumentation
3330 Washington Blvd.
Arlington, VA 22201-4598
Phone: 703-525-4890
Website: www.aami.org

AASHTO American Association of State
Highway and Transportation Officials
444 N. Capitol St. NW, Suite 249
Washington, DC 20001
Phone: 202/624-5800
Website: www.aashto.org

AATCC American Association of Textile
Chemists and Colorists
1 Davis Drive
P.O. Box 12215
Research Triangle Park, NC 27709
Phone: 919/549-8141
Website: www.aatcc.org

ABMA American Boiler Manufacturers
Association
950 North Glebe Road
Suite 160
Arlington, VA 22203
Phone: 703/522-7350
Website: www.abma.com

ABMA (DC) American Bearing Manufacturers
Association
12001 19th Street, NW
Washington, DC 20036
Phone: 202/429-5155
Website: www.abma-dc.org

ACGIH American Conference of Governmental
Industrial Hygienist
1330 Kemper Meadow Drive
Cincinnati, OH 45240
Phone: 513/742-2040
Website: www.acgih.org

ACI American Concrete Institute International
38800 Country Club Drive
Farmington Hills, MI 48331
Phone: 248/848-3700
Website: www.aci-int.org

ACTS (See Congressional Acts, CA, Regulatory
Organizations)

AGA (American Gas Association, see
International Approval Services, Inc., IAS)

AGMA American Gear Manufacturers
Association
1500 King Street, Suite 201
Alexandria, VA 22314
Phone: 703/684-0211
Website: www.agma.org

AI Asphalt Institute
Research Park Drive
P.O. Box 14052
Lexington, KY 40512
Website: www.asphaltinstitute.org

AIA/NAS Aerospace Industries Association of America, Inc.
1250 Eye Street, NW
Washington, DC 20005
Phone: 202/371-8400
Website: www.aia-aerospace.org

AIA American Institute of Architects
1735 New York, NW
Washington, DC 20006
Phone: 202/626-7300
Website: www.aiaonline.com

AICHE American Institute of Chemical Engineers
345 East 47th Street
New York, NY 10017
Phone: 212/705-7335
Website: www.aiche.org

AIIM Association for Information and Image Management
1100 Wayne Avenue, Suite 1100
Silver Springs, MD 20910
Phone: 301/587-8202
Website: www.aiim.org

AISC American Institute of Steel Construction
1 East Wacker Drive, Suite 3100
Chicago, IL 60601-2001
Phone: 312/670-2400
Website: www.aisc.org

AISI American Iron and Steel Institute
1101 17th Street, NW, Suite 1300
Washington, DC 20036
Phone: 202/452-7100
Website: www.steel.org

AMCA Air Movement and Control Association
30 West University Drive
Arlington Heights, IL 60004
Phone: 847/394-0150
Website: www.amca.org

ANL Argonne National Laboratory
9800 South Cass Avenue
Argonne, IL 60439
Phone: 630/252-2000
Website: www.anl.gov

ANS American Nuclear Society
555 North Kensington Avenue
LaGrange Park, IL 60526
Phone: 708/352-6611
Website: www.ans.org

ANSI American National Standards Institute
11 West 42nd Street, 13th Floor
New York, NY 10036
Phone: 212/642-4900
Website: www.ansi.org

API American Petroleum Institute
1220 L Street, NW
Washington, DC 20037
Phone: 202/682-8000
Website: www.api.org

AREA American Railway Engineering Association
8201 Corporate Drive, Suite 1125
Landover, Md. 20785
Phone: 301/459-3200
Website: None

ARI Air Conditioning and Refrigeration Institute
4301 North Fairfax Drive, Suite 425
Arlington, VA 22203
Phone: 703/524-8800
Website: www.ari.org

ARMA Asphalt Roofing Manufacturers Association
4041 Powder Mill Road
Centerpark Suite 404
Calverton Md. 20705
Phone: 301/231-9050
Website: www.asphaltroofing.org

ARMY U.S. Department of the Army
National Technical Information Services
5285 Port Royal Road
Springfield, VA 22161
Phone: 703/605-6000
Website: www.ntis.gov

ASA Acoustical Society of America
120 Wall Street
32nd Floor
New York, NY 10005-3993
Phone: 212/248-0373
Website: www.asastds@ait.org

ASCE American Society of Civil Engineers
1801 Alexander Bell Drive
Reston VA. 20191-4400
Phone: 703/295-6000
Website: www.asce.org

ASHRAE American Society of Heating, Refrigerating, and Air-Conditioning Engineers
1791 Tullie Circle, NE
Atlanta, GA 30329
Phone: 404/636-8400
Website: www.ashrae.org

ASME American Society of Mechanical Engineers
22 Law Drive
P.O. Box 2900
Fairfield, NJ 07007
Phone: 973/882-1167
Website: www.asme.org

ASNT The American Society For Nondestructive Testing
1711 Arlingate Lane
P.O. Box 28518
Columbus, OH 43228-0518
Phone: 614/274-6003
Website: www.asnt.org

ASQ American Society for Quality
3611 East Wisconsin Avenue
Milwaukee, WI 53203-4606
Phone: 414/272-8575
Website: www.asq.org

ASTM American Society for Testing and Materials
100 Harbor Drive
W. Conshohocken, PA 19428-2959
Phone: 610/832-9500
Website: www.astm.org

AWS American Welding Society
550 NW LeJune Road
Miami, FL 33126
Phone: 305/443-9353
Website: www.aws.org

AWWA American Water Works Association
6666 West Quincy Avenue
Denver, CO 80235
Phone: 303/794-7711
Website: www.awwa.org

BIA Brick Industry Association
11490 Commerce Park Drive, Suite 300
Reston, VA 20191
Phone: 703/620-0010
Website: www.bia.org

BOCA Building Officials and Code Administrators International, Inc.
4051 West Flossmoor Road
Country Club Hills, IL 60487
Phone: 708/799-2300
Website: www.bocai.org

CAA (Clean Air Act, see Congressional Acts, Regulatory Organizations)

CABO (Council of American Building Offcials, See International Conference of Building Officials, ICBO)

CERC Coastal Engineering Research Center & Hydraulics Laboratory
U.S. Army Corps of Engineers 3909 Halls Ferry Rd.
Vicksburg, MO; 39180-6199
Phone: 601/634-3339
Website: www.chl.wes.army.mil

CERCLA (Comprehensive Environmental Response, Compensation and Liability Act, see Congressional Acts, Regulatory Organizations)

CFR (Code of Federal Regulations, see Regulatory Organizations)

CGA Compressed Gas Association
1725 Jefferson Davis Highway, Suite 1004
Arlington, VA 22202-4102
Phone: 703/412-0900
Website: www.cganet.com

CMAA Crane Manufacturers Association of America
8720 Red Oak Blvd. Suite 201
Charlotte, NC 28217
Phone: 704/676-1190
Website: www.mhia.org

CRI Carpet and Rug Institute
310 Holiday Avenue
Box 2048
Dafton, GA 30720
Phone: 706/278-3176
Website: www.carpet-rug.com

CSA Canadian Standards Association
178 Rexdale Blvd.
Etobicoke (Toronto), Ontario, Canada
M9W1R3
Phone: 416/747-4000
Website: www.csa.ca

CTI Cooling Tower Institute
530 Wells Fargo, Suite 218
Houston, TX 77090
Phone: 281/583-4087
Website: www.cti.org

CWA (Clean Water Act, see Congressional Acts, Regulatory Organizations)

DOD U.S. Department of Defense
Defense Technical Information Center, Suite 0944
Fort Belvoir, VA 22060-6218
Phone: 703/767-8274
Website: www.dtic.mil

DOE (U.S. Department of Energy, see Regulatory Organizations)

DOE/OSTI DOE/Offce of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831
Phone: 423/574-1000
Website: www.explore.doe.gov

DSWA Defense Special Weapons Agency
6801 Telegraph Road
Alexandria, VA 22310
Phone: 703/325-8775
Website: www.dswa.mil

EIA Electronics Industries Alliance
2500 Wilson Blvd.

Arlington, VA 22201-3834
Phone: 703/907-7500
Website: www.eia.org

EO (Executive Orders, see Regulatory Organizations)

EPA (Environmental Protection Agency, see U.S. Environmental Protection Agency)

EPRI Electric Power Research Institute
P.O. Box 10412
Palto Alto, CA 94303
Phone: 650/855-2000
Website: www.epri.com

ERDA (Energy Research and Development Administration, see U.S. Department of Energy, DOE)

FAA Federal Aviation Administration
U.S. Department of Transportation
800 Independence SW
Washington, DC 20591
Phone: 202/267-3111
Website: www.faa.gov

FAI Fauske & Associates, Inc.
16W070 West 83rd Street
Burr Ridge, IL 60521
Phone: 630/323-8750
Website: www.fauske.com

FEMA Federal Emergency Management Agency
Federal Center Plaza
500 C Street, SW
Washington, DC 20472
Phone: 202/646-4600
Website: www.fema.gov

FGMA Glass Association of North America
White Lakes Professional Building
3310 SW Harrison Street
Topeka, KS 66611
Phone: 913/266-7013
Website: www.gana.org

FIPS Federal Information Processing Standards
National Bureau of Standards
Bldg. 820, Room 562
Gaithersburg, MD 20899
Phone: 301/975-2816
Website: none

FM Factory Mutual Engineering and Research
1151 Boston Providence Turnpike
Norwood, MA 02062
Phone: 781/762-4300
Website: www.factorymutual.com

FR (Federal Register, see Regulatory Organizations)

FS Federal Specifications
Naval Inventory Control Point
700 Robbins Avenue
Philadelphia, PA 19111
Phone: 215/697-4374
Website: None

GA Gypsum Association
810 First Street, NE, Suite 510
Washington, DC 20002
Phone: 202/289-5440
Website: www.gypsum.org

HES Health Education Services
P.O. Box 7126
Albany, NY 12224
Phone: 518/439-7286
Website: www.hes.org

IAPMO International Association of Plumbing and Mechanical Officials
20001 Walnut Drive
Walnut, CA 91789-2825
Phone: 909/595-8449
Website: www.iapmo.org

IAS International Approval Services, Inc.
(formerly American Gas Association)
8501 East Pleasant Valley Road
Cleveland, OH 44131
Phone: 216/524-4990
Website: www.iasapproval.org

ICBO International Conference of Building Officials
5360 South Workman Mill Road
Whittier, CA 90601
Phone: 562/699-0541
Website: www.icbo.org

ICC (International Code Council, see International Conference of Building Officials, ICBO)

IEC (International Electrotechnical Committee, see American National Standards Institute, ANSI)

IEEE Institute of Electrical and Electronics Engineers
United Engineering Center
345 East 47th Street
New York, NY 10017
Phone: 212/705-7000
Website: www.ieee.org

IESNA Illuminating Engineering Society of North America
120 Wall Street
17th Floor

New York, NY 10005
Phone: 212/248-5000
Website: www.iesna.org

IMC (International Mechanical Code, see
International Conference of Building Officials,
ICBO)

IPC The Institute for Interconnecting and
Packaging Electronic Circuits
2215 Sanders Road
Northbrook, IL 60062
Phone: 847/509-9700
Website: www.ipc.org

IPC (International Plumbing Code, see
International Conference of Building Officials,
ICBO)

IPSDC (International Private Sewage Disposal
Code, see International Conference of Building
Officials, ICBO)

ISA Instrument Society of America
P.O. Box 12277
Research Triangle Park, NC 27709
Phone: 919/549-8411
Website: www.isa.org

ISDSI Insulated Steel Door Systems Institute
30200 Detroit Road
Cleveland, OH 44145
Phone: 440/899-0010
Website: www.isdi.org

ISHM International Society for Hybrid
Microelectronics
1861 Wiehle
Reston, VA 22090

ISIAQ International Society of Indoor Air
Quality and Climate
P.O. Box 22038
Ottawa, ON, Canada K1V 0W2
Phone: 613/731-2559
Website: www.isiaq.org

ISO (International Standards Organization, see
American National Standards Institute, ANSI)

LANL Los Alamos National Laboratory
P.O. Box 1663
Los Alamos, NM 87545
Phone: 505/667-7000
Website: www.lanl.gov

LBL Lawrence Berkeley Laboratory
1 Cyclotron Road
Berkeley, CA 94720
Phone: 510/486-4000
Website: www.lbl.gov

LLNL Lawrence Livermore National Laboratory
7000 East Avenue
Livermore, CA 94550
Phone: 925/422-1100
Website: www.llnl.gov

MBMA Metal Building Manufacturing
Association
1300 Summer Avenue
Cleveland, OH 44115-2180
Phone: 216/241-7333
Website: www.taol.com/mbma

MHMS Material Handling and Management
Society
8720 Red Oak Blvd.
Charlotte, NC 28217
Phone: 704/676-1183
Website: www.mhia.org/mhms

MIL (Millitary Specifications, see Department of
Defense, DOD)

NACE National Association of Corrosion
Engineers
1440 South Creek Drive
Houston, TX 77084
Phone: 281/492-0535
Website: www.nace.org

NADCA North American Die Casting
Association
9701 W. Higgins Road, Suite 880
Rosemont, IL 60018-4721
Phone: 847/292-3600
Website: www.diecasting.org

NAPHCC National Association of Plumbing-
Heating-Cooling Contractors
180 South Washington Street
Falls Church, VA 22046
Phone: 703/237-8100
Website: www.naphcc.org

NASA National Aeronautics and Space
Administration NASA Headquarters
300 East Street, SW
Washington, DC 20546
Phone: 202/453-2928
Website: www.nasa.gov

NAVFAC (U.S. Naval Facilities Engineering
Command; see U.S. Department of the Navy,
US NAVY)

NBC (National Building Code, see BOCA)

NBS National Bureau of Standards (currently
National Institute of Standards and
Technology)
Clopper & Quints Orchard
Gaithersburg, MD 20899
Phone: 301/975-2000
Website: www.nist.gov

NCMA National Concrete Masonry Association
2302 Horse Pen Road
Herndon, VA 20171
Phone: 703/713-1900
Website: www.ncma.org

NEC (National Electric Code, see National Fire Protection Association STD-70)

NEMA National Electrical Manufacturers Association
1300 North 17th Street, Suite 1847
Rosslyn, VA 22209
Phone: 703/841-3200
Website: www.nema.org

NEPA (National Environmental Policy Act, see Congressional Acts, Regulatory Organizations)

NFPA National Fire Protection Association
11 Tracy Drive
Avon, MA 02322
Phone: 800/344-3555
Website: www.nfpa.org

NIJ National Institute of Justice
810 7th Street, NW
Washington, DC 20531
Phone: 202/307-2942
Website: www.usdoj.gov/nij

NOAA National Oceanic and Atmospheric Administration
1315 East West Highway
Silver Springs, MD 20910
Phone: 301/713-4000
Website: www.noaa.gov

NPDES (National Pollution Discharge Elimination System, see 40 CFR 125, Regulatory Organizations)

NRC U.S. Nuclear Regulatory Commission
1155 Rockville Pike
Rockville, MD 20852
Phone: 301/415-7000
Website: www.nrc.gov

NRCA National Roofing Contractors Association
10255 West Higgins Road
Suite 600
Rosemont, IL 60018
Phone: 847/299-9070
Website: www.roofonline.org

NSA National Security Agency/Central Security Service
9800 Savage Road
Fort Meade, MD 20755
Phone: 301/688-7111
Website: www.nsa.gov

NTIS National Technical Information Services
5285 Port Royal Road
Springfield, VA 22161
Phone: 703/605-6000
Website: www.ntis.gov

NWWDA National Wood Window and Door Association
1400 East Touchy Avenue, Suite 470
Des Plaines, IL 60018
Phone: 847/299-5200
Website: www.nwwda.org

OMB (Office of Management and Budget, see Regulatory Organizations)

OSHA (Occupational Safety and Health Act, see Congressional Acts, Regulatory Organizations)

OSHA Occupational Safety and Health Administration
200 Constitution Avenue NW
Washington, DC 20210
Phone: 202/219-8151
Website: www.osha.gov

PCA Portland Cement Association
5420 Old Orchard Road
Skokie, IL 60077-1083
Phone: 847/966-6200
Website: www.portcement.org

PCI Prestressed Concrete Institute
175 West Jackson Boulevard
Chicago, IL 60604
Phone: 312/786-0300
Website: www.pci.org

PCMI Photo-Chemical Machining Institute
810 Knott Place
Phone: 215/825-2506
Website: www.pcmi.org

PDCA Painting and Decorating Contractors of America
3913 Old Lee Highway, Suite 33B
Fairtax, VA 22030
Phone: 703/359-0826
Website: www.pdca.com

PMA Precision Metalforming Association
27027 Chardon Road
Richmond Heights, OH 44143
Phone: 440/585-8800
Website: www.metalforming.com

PTI Post-Tensioning Institute
1717 West Northern Avenue
Phoenix, AZ 85021
Phone: 602/870-7540
Website: www.pti-usa.org

RCRA (Resource conservation and Recovery Act, see Congressional Acts, Regulatory Organizations)

RFCI Resilient Floor Covering Institute
966 Hungerford Drive
Suite 12-B
Rockville, MD 20850
Phone: 301/340-8580
Website: www.rfci.org

SACMA Suppliers of Advanced Composite Materials Association
1600 Wilson Blvd.
Arlington, VA 22209
Phone: 703/841-1556
Website: www.sacma.org

SAE Society of Automotive Engineers, Inc.
400 Commonwealth Drive
Warrendale, PA 15096
Phone: 724/776-4841
Website: www.sae.org

SBC (Standard Building Code, see Southern Building Code Congress International, Inc., SBCCI)

SBCCI Southern Building Code Congress International, Inc.
900 Montclair Road
Birmingham, ALA 35213-1206
Phone: 205/591-1853
Website: www.sbcci.org

SCS Soil Conservation Service, U.S. Department of Agriculture
14th and Independence Avenue, SW
Washington, DC 20250
Phone: 202/205-0026
Website: www.doa.gov

SDI Steel Door Institute
30200 Detroit Road
Cleveland, OH 44145
Phone: 440/899-0010
Website: www.steeldoor.org

SDWA (Safe Drinking Water Act, see Congressional Acts, Regulatory Organizations)

SEMI Semiconductor Equipment and Materials International
805 East Middlefield Road
Mountain View, CA 94043
Phone: 650/964-5111
Website: www.semi.org

SJI Steel Joist Institute
3127 10th Avenue North
Myrtle Beach, SC 29577
Phone: 843/626-1995
Website: www.steeljoist.org

SMA Screen Manufacturers Association
2850 South Ocean Blvd., No. 114
Palm Beach, FLA 33480-5535
Phone: 561/533-0991
Website: none

SMACNA Sheet Metal and Air Conditioning Contractors National Association
4201 Lafayette Center Drive
Chantilly, VA 20151
Phone: 703/803-2980
Website: www.smacna.org

SNL Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185
Phone: 505/284-3958
Website: www.irnsandia.gov

SPI Society of the Plastics Industry
1801 K Street N.W., Suite 600K
Washington, DC 20006
Phone: 202/974-5251
Website: www.socplas.org

SPICI (SPI Composites Institute, see Society of the Plastics Industry, SPI)

SSFI Scaffolding, Shoring, and Framing Institute
1300 Sumner Avenue
Cleveland, OH 44115
Phone: 216/241-7333
Website: www.taol.com/ssfi

SWI Steel Window Institute
1300 Sumner Avenue
Cleveland, OH 44115-2851
Phone: 216/241-7333
Website: www.taol.com/swi

TFA The Ferroalloys Association
900 2nd Street N.E., Suite 201
Washington, DC 20002
Phone: 202/842-0292
Website: www.tfa.org

TSCA (Toxic Substance Control Act, see Congressional Acts, Regulatory Organizations)

UBC (Uniform Building Code, see International Conference of Building Officials, ICBO)

UL Underwriters Laboratories
333 Pfingsten Road
Northbroook, IL 60062
Phone: 847/272-8800
Website: www.ul.com

UPC (Uniform Plumbing Code, see IAPMO)

USAF U.S. Department of the Air Force
National Technical Information Services
5285 Port Royal Road
Springfield, VA 22161

Phone: 703/605-6000
Website: www.ntis.gov

USARMY U.S. Department of the Army
National Technical Information Services
5285 Port Royal Road
Springfield, VA 22161
Phone: 703/605-6000
Website: www.ntis.gov

USEPA U.S. Environmental Protection Agency
401 M. Street, SW
Washington, DC 20460
Phone: 202/260-2090
Website: www.epa.gov

USNAVY U.S. Department of the Navy
National Technical Information Services
5285 Port Royal Road
Springfield, VA 22161
Phone: 703/605-6000
Website: www.ntis.gov

USNRC (U.S. Nuclear Regulatory Commission, see NRC)

WEF Water Environment Federation (formerly Water Pollution Control Federation)
601 Wythe Street
Alexandria, VA 22308
Phone: 703/684-2400
Website: www.wef.org

WQA (Water Quality Act, see Congressional Acts, Regulatory Organizations)

WRC Water Resources Council, Hydrology Committee
U.S. Department of the Interior
1849 C Street, NW
Washington, DC 20240
Phone: 202/208-3100
Website: www.doi.gov

REFERENCES

1. JF Thorpe, WH Middendorf. What Every Engineer Should Know About Product Liability. New York: Marcel Dekker, 1979, p 42.
2. National Fire Protection Association. Guide to OSHA Fire Protection Regulations. Boston, MA: National Fire Protection Association, 1971.
3. National Fire Protection Association. Guide to OSHA Fire Protection Regulations. Boston, MA: National Fire Protection Association, 1971, p 7.
4. National Society of Professional Engineers. Code of Ethics. Preamble, p 1.

ADDITIONAL READING

AL Batik. The Engineering Standard, A Most Useful Tool. Ashland, OH: Book Master/EI Rancho, 1992.

Chapter 9.1

Perspectives on Designing Human Interfaces for Automated Systems

Anil Mital
University of Cincinnati, Cincinnati, Ohio

Arunkumar Pennathur
University of Texas at El Paso, El Paso, Texas

1.1 INTRODUCTION

1.1.1 Importance and Relevance of Human Factors Considerations in Manufacturing Systems Design

The design and operation of manufacturing systems continue to have great significance in countries with large and moderate manufacturing base, such as the United States, Germany, Japan, South Korea, Taiwan, and Singapore. It was widely believed in the 1980s that complete automation of manufacturing activities through design concepts such as "lights-out factories," would completely eliminate human influence from manufacturing, and make manufacturing more productive [1]. However, we now see that complete automation of manufacturing activities has not happened, except in a few isolated cases. We see three basic types of manufacturing systems present and emerging—the still somewhat prevalent traditional manual manufacturing mode with heavy human involvement in physical tasks, the predominant hybrid manufacturing scenario (also referred to traditionally as the mechanical or the semiautomatic systems) with powered machinery sharing tasks with humans, and the few isolated cases of what are called computer-integrated manufacturing (CIM) systems with very little human involvement, primarily in supervisory capacities. Indeed, human operators are playing, and will continue to play, important roles in manufacturing operations [2].

Another important factor that prompts due consideration of human factors in a manufacturing system, during its design, is the recent and continuous upward trend in nonfatal occupational injuries that has been observed in the manufacturing industry in the United States [3]. While these injuries may not be as severe and grave as the ones due to accidents such as the Chernobyl Nuclear Reactor accident (the Three Mile Island nuclear accident prompted an upswing in human factors research, especially in nuclear power plants and in process industry settings), the increasing trend in injuries leaves the claim that "automation" of manufacturing has resulted in softer jobs for manufacturing workers questionable. In fact, many manufacturing researchers and practitioners believe that an increase in severe injuries in manufacturing is primarily due to the automation of simpler tasks, leaving the difficult ones for the humans to perform. This belief is logical as the technology to automate difficult tasks is either unavailable or expensive.

The factors discussed suggest that manufacturing systems (our definition of a system is broad; a system may thus be a combination of a number of equipment/machines and/or humans) be designed with human limitations and capabilities in mind, if the system is to be productive, error-free, and safe, and result in quality goods and services, all vital goals for manufacturing organizations.

1.1.2 The Human–Machine System Framework for Interface Design

Traditionally, system designers have accounted for human limitations and capabilities by considering the human operator as an information processor having sensory and motor capabilities and limitations (Fig. 1). It can be readily seen from Fig. 1 that the key elements to the efficient and error-free functioning of a human–machine system are the provision of information to human operators in the system, and the provision for control of the system by humans.

Displays provide information about the machine or the system to human operators, and controls enable human operators to take actions and change machine or system states (conditions). Operator feedback is obtained through interaction with the controls (tactile sensing, for instance). Thus, in the classical view, human interaction with automation is mediated through displays and controls for a two-way exchange of information.

The recent view of the human–machine system, resulting out of advances in computerized information systems, sees the human operator as a supervisory controller [4] responsible for supervisory functions such as planning, teaching, monitoring, intervening, learning, etc. (Fig. 2). Even though, in such a view, the human operator has a changed role, displays and controls still provide the fundamental medium for human interaction with the system.

Indeed, properly designed displays and controls are fundamental to the efficient and error-free functioning

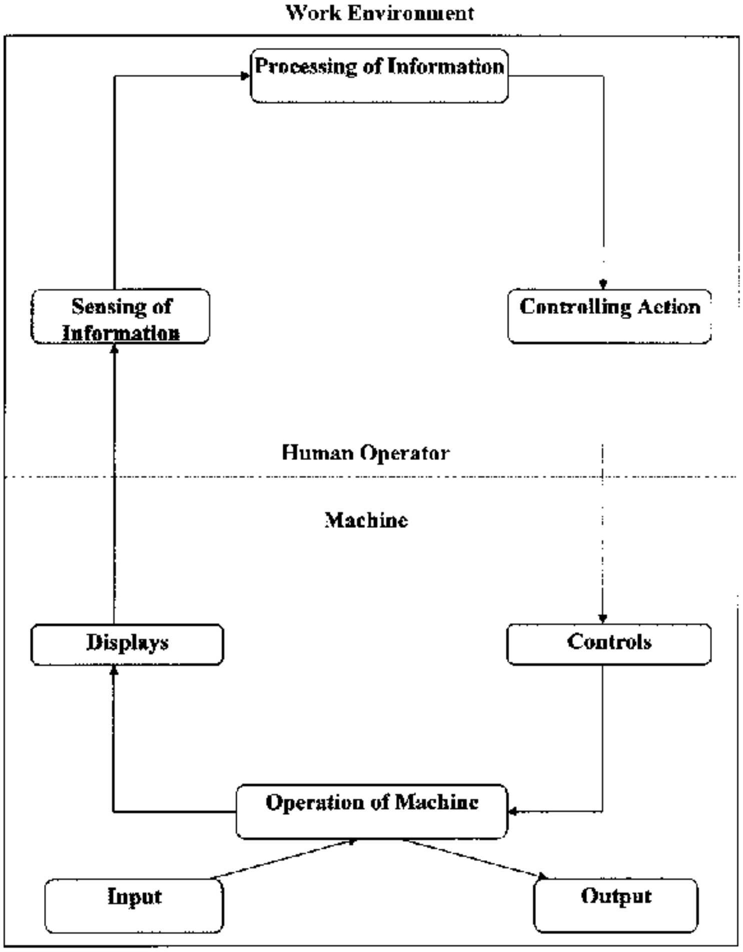


Figure 1 Traditional representation of human interaction with machine.

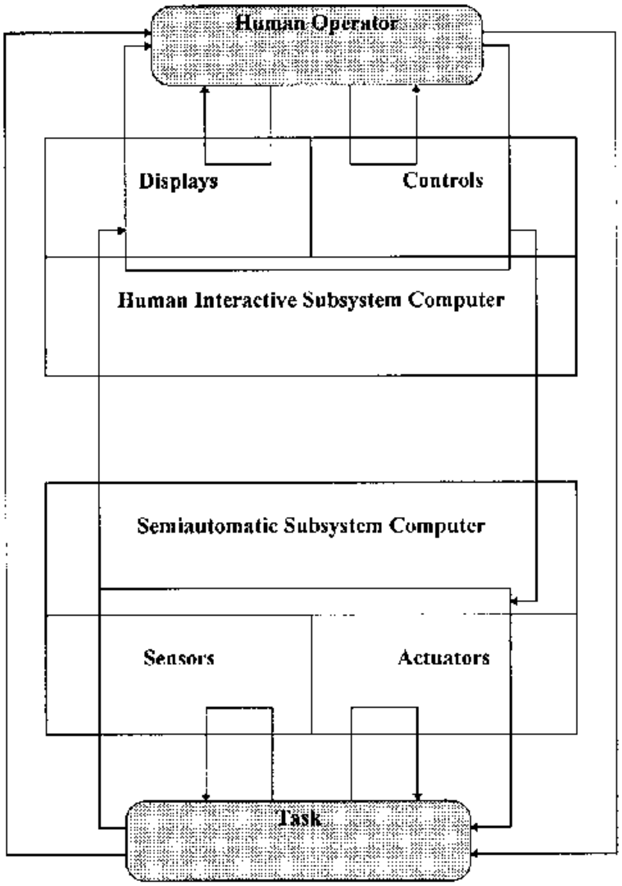

Human Operator
Displays
Controls
Human Interactive Subsystem Computer
Semiautomatic Subsystem Computer
Sensors
Actuators
Task

Figure 2 The latest notion of human as a supervisory controller.

of manufacturing systems. Ergonomics, which we define as the study of issues involved in the application of technology to an appropriate degree to assist the human element in work and in the workplace, provides recommendations for interface design based on research in human sensory and motor capabilities and limitations.

1.1.3 Scope of This Chapter

Even though displays and controls, and their effective design, are fundamental to the efficient and error-free operation of the system, a number of important activities need to be carried out before one can think of displays and controls. These activities stem from the central need to build systems to suit human limitations and capabilities. Some of these activities, such as "user needs analysis," are relatively new concepts and form the core of what is called the "usability engineering approach" to design. Techniques associated with other activities, such as task analysis and function allocation between humans and automated equipment, are an integral part of designing "good" jobs, and have been in existence for some time. We present some of these techniques and methods.

Inherent throughout our presentation is the essence of the "human-centered interface design approach." We first present elements of this approach and contrast it with the "system-centered interface design approach." It is recommended that this concept of human-centered design guide the designer at both the system, as well as at the nuts-and-bolts, design levels.

Displays and controls, the selection, design, and evaluation of which will be the theme for the remainder of the chapter, form a part of aids, equipment, tools, devices, etc., that are necessary for a system to operate satisfactorily. Due to the wide variety of available technologies, and due to the fact that most ergonomics recommendations for the design of displays and controls remain the same regardless of the technology used (e.g., recommendations on the design of lettering remain the same whether the recommendation is for a conventional hand-held meter, a visual display unit, or printed material), we provide only general recommendations for different types of displays and controls, without reference to commercial products and equipment.

A few other notes about the scope of this chapter: due to the vast literature available in the area of design of human–machine systems, our emphasis in this chapter is on the breadth of coverage rather than depth in any area. This emphasis is deliberate, and is motivated, in addition, by our intention to provide the reader a taste of the *process* of design and evaluation of a modern human–machine system. Readers interested in more detail in any one area or technique should refer to our recommended reading list. Also, even though the recommendations and guidelines summarized in this chapter come from research in human–machine settings other than hard-core manufacturing settings, they are equally applicable to manufacturing systems—the general framework and the specific recommendations we have collected and provided in this chapter for design of human–machine systems are applicable across systems.

1.2 APPROACHES TO DESIGNING SYSTEMS FOR HUMAN–MACHINE INTERFACE

1.2.1 The System-Centered Design Approach

The system-centered design approach, as the name suggests, analyzes the system currently in use, designs and specifies the new system based on this analysis, builds and tests the new system, and delivers the system and makes minor changes to the system (Fig. 3). The focus is on the goals of the *system* and the goals of the *organization* within which the system is to perform. Designers following this approach fail to consider the users before designing the system. As a result, users of such systems are required to remember too much information. Also, typically, these systems are intolerant of minor user errors, and are confusing to new users. More often than not, such systems do not provide the functions users want, and force the users to perform tasks in undesirable ways. New systems designed the system-centered way have also been shown to cause unacceptable changes to the structure and practices in entire organizations [5].

1.2.2 The Human-Centered Design Approach

The human-centered design approach to human–machine interaction, unlike the system-centered approach, puts the human attributes in the system ahead of system goals. In other words, the entire system is built around the user of the system—the human in the system. This approach has been variously called the "usability engineering approach," the "user-centered approach" or the "anthropocentric approach to production systems," etc. Figure 4 provides our conception of the human-centered approach to interface design. The first step in this design approach is information collection. Information about user needs, information about user cognitive and mental models, information on task demands, information on the environment in which the users have to perform, information on the existing interface between the human operator (the user of the system) and the machine(s), requirements of the design, etc., are some of the more important variables about which information is collected. This information is then used in the detailed design of the new interface. The design is then evaluated. Prototype development and testing of the prototype are then performed just as in any other design process. User testing and evaluation of the prototype, the other important characteristic of this design process which calls for input from the end user, is then carried out. This results in new input to the design of the interface, making the entire design process iterative in nature.

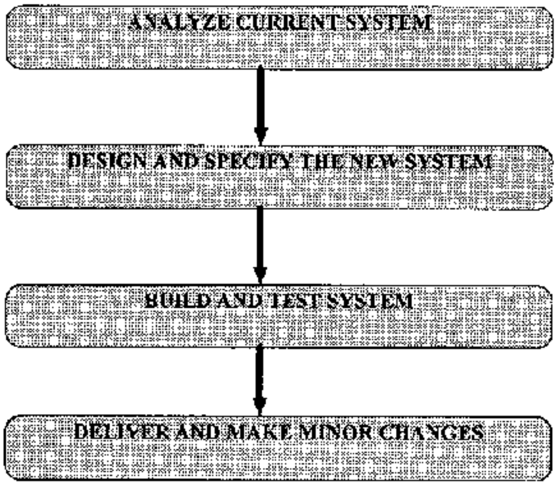

ANALYZE CURRENT SYSTEM
DESIGN AND SPECIFY THE NEW SYSTEM
BUILD AND TEST SYSTEM
DELIVER AND MAKE MINOR CHANGES

Figure 3 System-centered approach to design.

Even though the human-centered design approach is intended to take human capabilities and limitations into account in system design and make the system *usable*, there are a number of difficulties with this approach. The usability of the system is only as good as its usability goals. Thus, if the input from the users about the usability goals of the system are inappropriate, the system will be unusable. One approach to overcome this problem is to include users when setting usability goals; not just when measuring the usability goals. Another common difficulty with this approach is the lack of provision to take into account qualitative data for designing and refining the design. This is due to the deficiency inherent in the definition of usability which calls for quantitative data to accurately assess the usability of a system. There is also the drawback that this approach is best suited for designing new systems, and that it is not as effective for redesign of existing systems.

Despite these limitations, the human-centered design approach merits consideration from designers because it proactively takes the user of the product (displays and controls with which we are concerned, and which make up the interfaces for human–machine interaction, are products) into the system design process, and as a result, *engineers* usability, into the product.

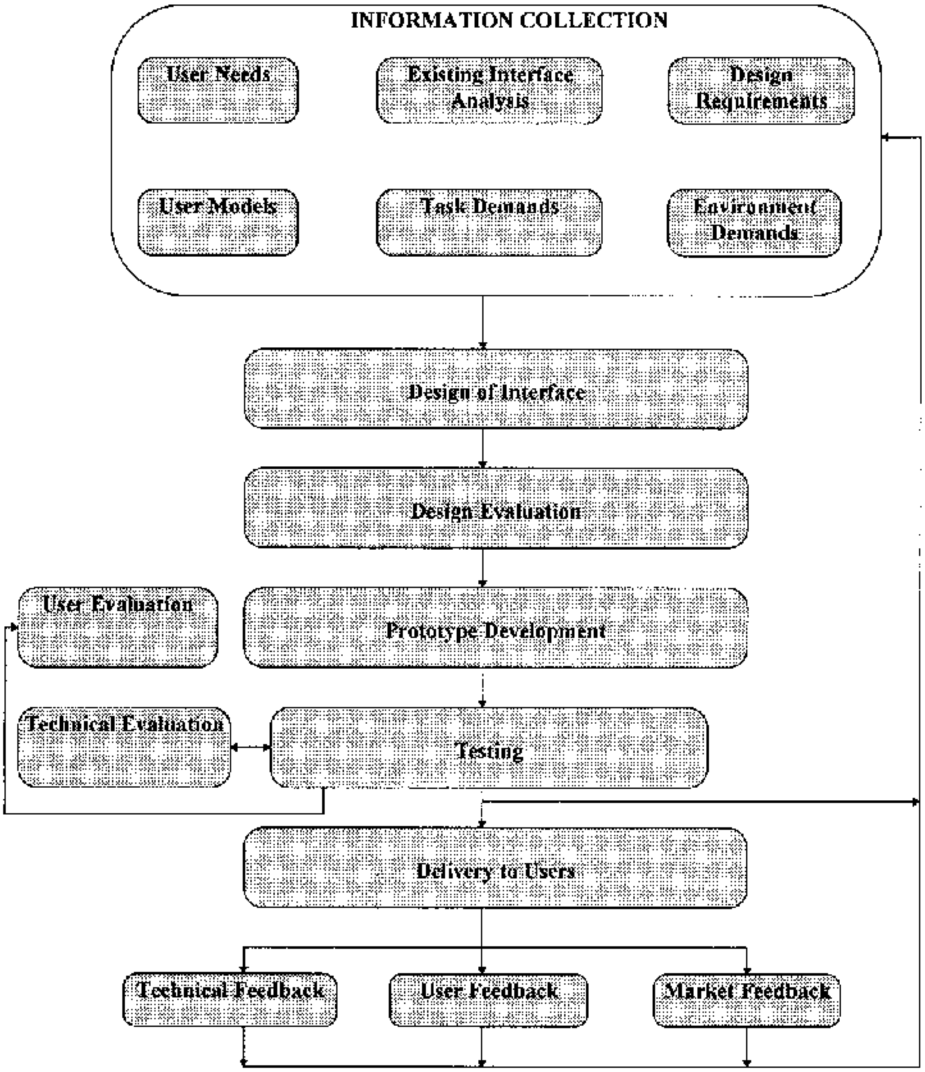


Figure 4 Human-centered approach.

1.3 THE PROCESS OF SOLVING HUMAN–MACHINE INTERFACE PROBLEMS

Even though displays and controls are the final means of information exchange between humans and machines in a system, the actual design of the hardware and software for displays and controls comes only last in order, in the *process* of solving human–machine interface problems. The other key steps in this process include user-needs analysis, task analysis, situation analysis, and function allocation decisions, after which the modes of information presentation and control can be decided. In the following sections, we discuss each of these steps.

1.3.1 User-Needs Analysis

The goal of user-needs analysis is to collect information about users and incorporate it into the design process for better design of the human–machine interface. User-needs analysis typically involves the following activities: characterization of the user, characterization of the task the user performs, and characterization of the situation under which the user has to perform the task. What follows are guidelines and methods for performing each of these three activities prior to designing the system.

1.3.1.1 Characterization of the User

Table 1 provides a user characterization checklist. Included in this checklist are questions to elicit information about the users, information about users' jobs, information about users' backgrounds, information about usage constraints, and information about the personal preferences and traits of the users.

As is obvious from the nature of the questions in the checklist, the goal of collecting such information is to use the information in designing a *usable* system.

1.3.1.2 Characterization of the Task

Characterization of the tasks users have to perform to attain system goals is done through task analysis. Task analysis is defined as the formal study of what a human operator (or a team of operators) is required to do to achieve a system goal [6]. This study is conducted in terms of the actions and/or the cognitive processes involved in achieving the system goal. Task analysis is a methodology supported by a number of techniques to help the analyst collect information about a system, organize this information, and use this information to make system design decisions. Task analysis is an essential part of system design to ensure efficient and effective integration of the human element into the system by taking into account the limitations and capabilities in human performance and behavior. This integration is key to the safe and productive operation of the system.

The key questions to ask when performing task analysis activities are shown in Table 2. The task analysis methodology finds use at all stages in the life cycle of a system—from initial conception through the preliminary and detailed design phases, to the prototype and actual product development, to the storage and demolition stage. Task analysis is also useful for system evaluation, especially in situations involving system safety issues, and in solving specific problems that may arise during the daily operations of a system. Task analysis can be carried out by system designers or by the operations managers who run the system on a day-to-day basis.

Table 1 User Characteristics Checklist

Data about users	What is the target user group? What proportion of users are male and what proportion are female? What is average age/age range of users? What are the cultural characteristics of users?
Data about job	What is the role of the user (job description)? What are the main activities in the job? What are the main responsibilities of the user? What is the reporting structure for the user? What is the reward structure for the user? What are the user schedules? What is the quality of output from the user? What is the turnover rate of the user?
Data about user background	What is the education/knowledge/experience of the user relevant to the job? What are the relevant skills possessed by the user? What relevant training have the users undergone?
Data about usage constrains	Is the current equipment use by users voluntary or mandatory? What are the motivators and demotivators for use?
Data about user personal preferences and traits	What is the learning style of the user? What is the interaction style of the user? What is the aesthetic preference of the user? What are the personality traits of the user? What are the physical traits of the user?

Adapted from Ref. 5.

Table 2 Checklist for Task Analysis Activities

Goals	What are the important goals and supporting tasks?
For every important task:	
Intrinsics of the task	What is the task?
	What are the inputs and outputs for the task?
	What is the transformation process (inputs to outputs)?
	What are the operational procedures?
	What are the operational patterns?
	What are the decision points?
	What problems need solving?
	What planning is needed?
	What is the terminology used for task specification?
	What is the equipment used?
Task dependency and criticality	What are the dependency relationships between the current task and the other tasks and systems?
	What are the concurrently occurring effects?
	What is the criticality of the task?
Current user problems	What are the current user problems in performing this task?
Performance criteria	What is the speed?
	What is the accuracy?
	What is the quality
Task criteria	What is the sequence of actions?
	What is the frequency of actions?
	What is the importance of actions?
	What are the functional relationships between actions?
	What is the availability of functions?
	What is the flexibility of operations?
User discretion	Can the user control or determine pace?
	Can the user control or determine priority?
	Can the user control or determine procedure?
Task demands	What are the physical demands?
	What are the perceptual demands?
	What are the cognitive demands?
	What are the environmental demands?
	What are the health and safety requirements?

Adapted from Ref. 5.

While many different task analysis techniques exist to suit the different design requirements in systems, our primary focus here is on techniques that help in designing the interface. The key issues involved in designing a human interface with automated equipment are assessing what will be needed to do a job (the types of information that human operators will need to understand the current system status and requirements; the types of output that human operators will have to make to control the system), and deciding how this will be provided. Table 3 provides a summary of the important activities involved in the process of interface design and the corresponding task analysis technique to use in designing this activity. We present brief summaries of each of these techniques in the following sections. The reader should refer to Kirwan and Ainsworth [6], or other articles on task analysis, for a detailed discussion of the different task analysis techniques.

Hierarchical Task Analysis. This enables the analyst to describe tasks in terms of operations performed by the human operator to attain specific goals, and "plans" or "statements of conditions" when each of a set of operations has to be carried out to attain an operating goal. Goals are defined as "desired states of

Table 3 Summary of Task Analysis Activities and Methods Involved in Interface Design

Activity	Task analysis method
Gathering task information representing the activities within the task	Hierarchical task analysis Activity sampling
Stating required information, actions, and feedback	Work study Task decomposition Decision/action diagrams
Checking adequacy of provisions for information flows for successful completion of the task	Table-top analysis Simulation Walk-through/talk-through Operator modifications surveys Coding consistency surveys
Identifying links between attributes (total system check) to ensure system success	Link analysis Petri nets Mock-ups Simulator trials
Provide detailed design recommendations	Person specification Ergonomics checklists

Modified from Ref. 6.

systems under control or supervision'' (e.g., maximum system productivity). Tasks are the elements in the method to obtain the goals in the presence of constraints (e.g., material availability). Operations are what humans actually do to attain the goals. Thus, hierarchical task analysis is ``the process of critically examining the task factors, i.e., the human operator's resources, constraints and preferences—in order to establish how these influence human operations in the attainment of system goals.'' System goals can be described at various levels of detail (or subgoals), and hence the term ``hierarchical.'' The hierarchical task analysis process begins with the statement of overall goal, followed by statements of the subordinate operations, and the plans to achieve the goal. The subordinate operations and the plans are then checked for adequacy of redescription (of the goal into suboperations and plans). The level of detail necessary to adequately describe a goal in terms of its task elements determines the ``stopping rule'' to use when redescribing. A possible stopping rule could be when the probability of inadequate performance multiplied by the costs involved if further redescription is not carried out, is acceptable to the analyst.

Activity Sampling. This is another commonly used task analysis method for collecting information about the type and the frequency of activities making up a task. Figure 5 shows the steps involved in activity sampling.

Samples of the human operator's behavior at specified intervals are collected to determine the proportion of time the operator spends performing the identified activities. Two key factors for the activity sampling method to work include the requirements that the task elements be observable and distinct from one another, and that the sampling keep pace with the performance of the task. Typically, the analyst performing activity sampling, classifies the activities involved, develops a sampling schedule (these two aspects form the core of the design of activity sampling), collects and records information about activities, and analyzes the collected activity samples. Activity sampling has its advantages and disadvantages. Objectivity in data recording and collection, ease of administering the technique, and the ability of the technique to reveal task-unrelated activities that need analysis, are some of the advantages of the method. Requirements of a skilled analyst (for proper identification and description of the task elements), and the inability of the technique to provide for analysis of cognitive activities are the main disadvantages of the technique.

Task Decomposition. This is a method used to exactly state the tasks involved in terms of information con-

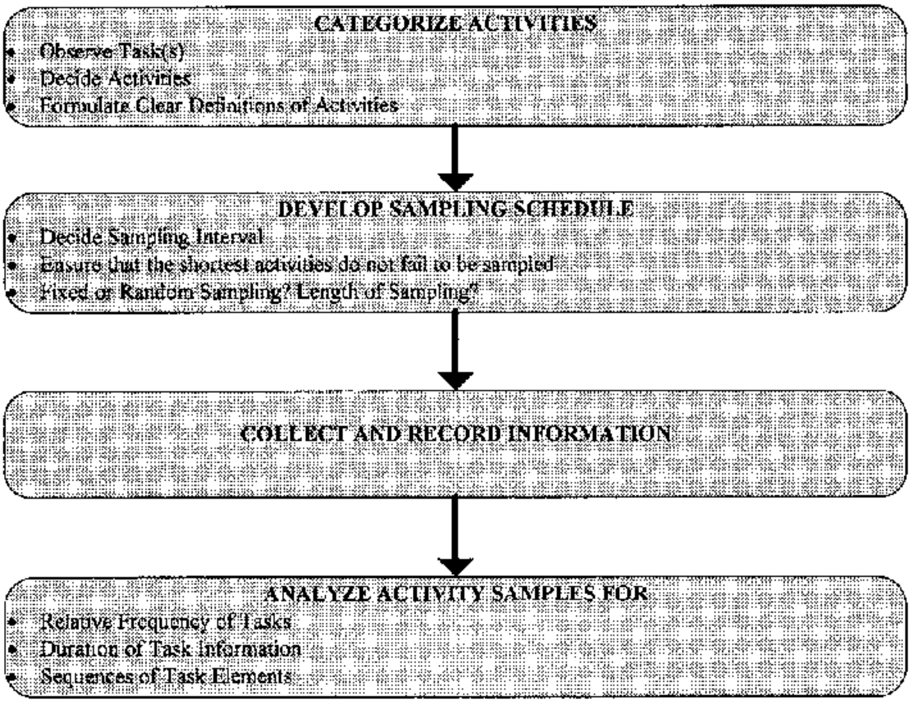

CATEGORIZE ACTIVITIES
Observe Task(s)
Decide Activities
Formulate Clear Definitions of Activities
DEVELOP SAMPLING SCHEDULE
Decide Sampling Interval
Ensure that the shortest activities do not fail to be sampled
Fixed or Random Sampling? Length of Sampling?
COLLECT AND RECORD INFORMATION
ANALYZE ACTIVITY SAMPLES FOR
Relative Frequency of Tasks
Duration of Task Information
Sequences of Task Elements

Figure 5 Activities involved in activity sampling.

tent, and actions and feedback required of the operator. Once a broad list of activities and the tasks involved have been generated using either hierarchical task analysis or activity sampling, task decomposition can be used to systematically expand on the task descriptions. The various steps involved in task decomposition are presented in Fig. 6.

Decision–Action Diagram. This is one of the most commonly used tools for decision making. Figure 7 is an example of a decision–action diagram [7]. The decision–action diagram sequentially proceeds through a series of questions (representing decisions) and possible yes/no answers (representing actions that can be taken). The questions are represented as diamonds, and the possible alternatives are labeled on the exit lines from the diamond. A thorough knowledge of the system components, and the possible outcomes of making decisions about system components is essential for constructing complete and representative decision–action diagrams.

Table-Top Analysis. As the name implies, this is a technique through which experts knowledgeable about a system discuss specific system characteristics. In the context of interface design, this task analysis methodology is used for checking if the information flows identified during the initial task analysis and task description, is adequate for successful task completion. Table-top analysis, hence, typically follows the initial hierarchical or other forms of task analysis which yield task descriptions, and provides information input for the decomposition of the tasks. A number of group discussion techniques exist in practice, including the Delphi method, the group consensus approach, the nominal group technique, etc., for conducting table-top analysis, each with its own merits and demerits.

Walk-Through/Talk-Through Analysis. These analyses involve operators and other individuals having operational experience with the system, walking and talking the analyst through observable task components of a system in real time. Walk-through is normally achieved in a completely operational system or in a simulated setting or even in a mock-up setting. Talk-through can be performed even without a simulation of the system—the only requirements are drawing and other system specific documentation to enable the analysts to set system and task boundaries while performing the talk-through analysis. For more information on walk-through and talk-through analyses, refer to Meister [8].

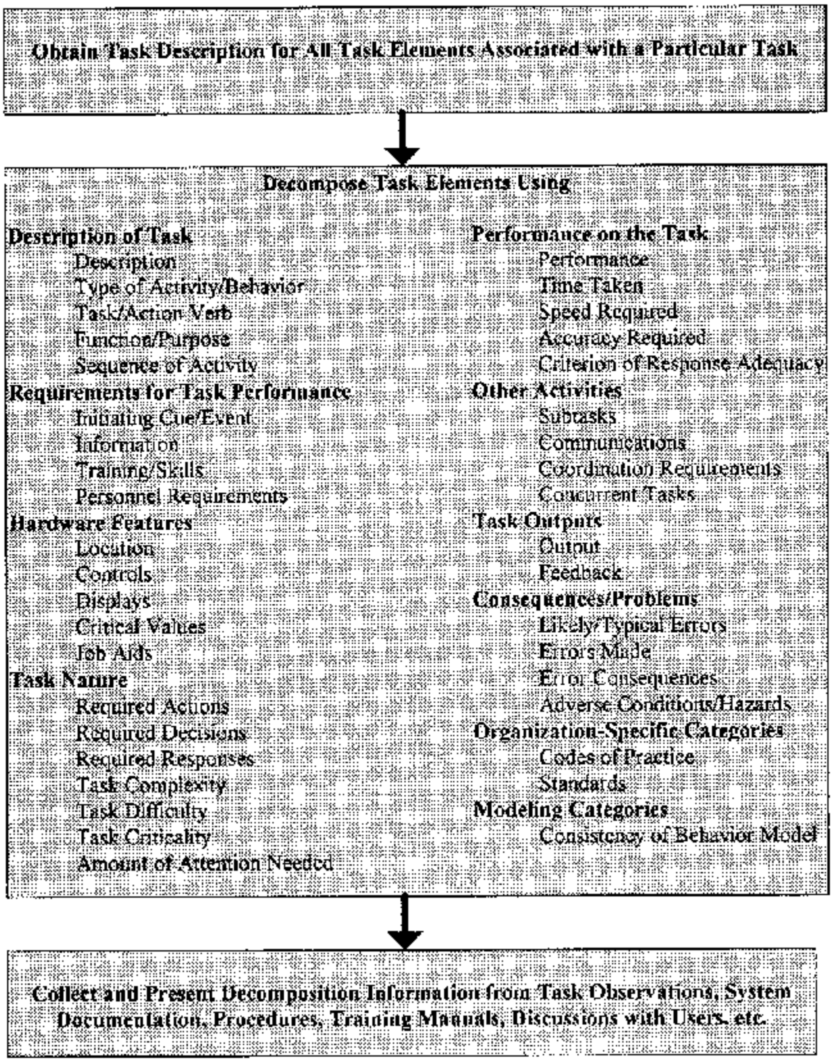


Figure 6 The task decomposition process.

Operator Modification Surveys. These surveys are performed to gather input from the actual users, (i.e., the operators) of the system, to check if there will be difficulties in using the system, and of what types. This checking of the adequacy of the interface design of the system from the users' perspective is done through surveys conducted on similar already operational systems. In general, operators and other users of systems maintain and provide information on design inadequacies through *memory aids*, such as their own labels on displays to mark safe limits, *perceptual cues*, such as makeshift pointers, and *organizational cues*, such as grouping instruments through the use of lines. These makeshift modifications done by the operators indicate design deficiencies in the system, and can be planned for and included in the redesign of the existing system or in the design of a new system.

Coding Consistency Surveys. These surveys are used to determine if the coding schemes in use in the system are consistent with the associated meanings, and if and where additional coding is needed. The recommendation when performing coding consistency surveys is to record the description of the location of the item, a description of the coding used for that item (intermittent siren sound), a description of any other coding schemes used for that item (inter-

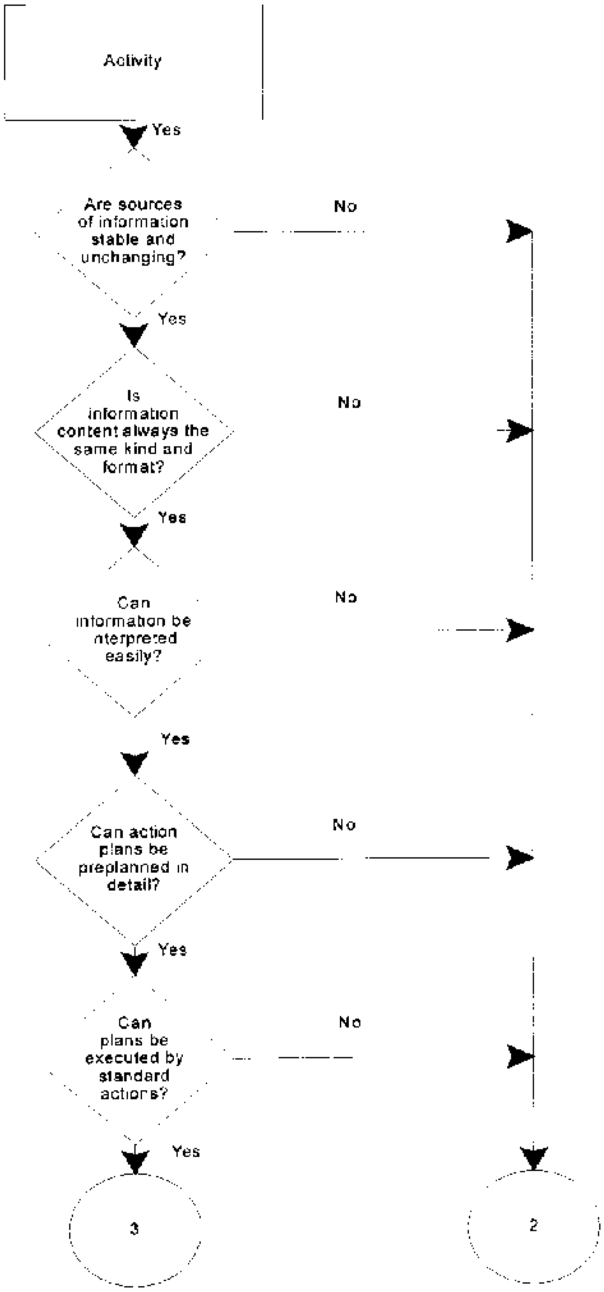

Activity
Yes
Are sources of information stable and unchanging?
No
Yes
Is information content always the same kind and format?
No
Yes
Can information be nterpre:ed easily?
No
Yes
Can action plans be preplanned in detail?
No
Yes
Can plans be executed by standard actions?
No
Yes
3
2

Figure 7 Generic function allocation analysis flowchart.

mittent siren sound accompanied by a yellow flashing light), and a complete description of the function being coded.

Link Analysis. This is a technique used to identify and represent the nature, frequency, and/or the importance of relationships or *links* existing between individual operators and some portion of the system [9]. Link analysis has been found to be particularly useful in applications where the physical layout of equipment, instruments, etc., is important to optimize the interaction of the human operator with the system. Link analysis does not require extensive resources to perform (in fact, paper and pencil are the only resources required to perform a link analysis). Link analysis proceeds by first collecting information about the system components used during task performance. This information is then used to develop a complete list of links between individual system elements. The links thus established are then diagramed and ranked for importance. The order of importance may be determined based on the frequency of activity between two links, or based on other appropriate measures decided by the system expert. The nature of the links to be studied (is it a movement of attention or position between parts of the system?), and the level of detail to include in defining each link are important factors that determine the overall structure and usefulness of the links established. Link analysis does not need observational data collection; a mere description of the procedures in the form of a technical manual is sufficient for identifying and establishing the links. The extensive graphical and tabular representations involved in link analysis, however, limits the use of this technique for large systems with involved linkages in the system.

Simulator Analysis. The goal of simulation studies is to replicate, and observe, system (including operator and operating environment) performance while making the performance environment as representative and close to the real-time environment as possible. Different forms of simulations exist depending on the platform or the simulator used for the simulation: a simple paper-and-pencil simulation, to a mock-up of a system that may or may not be dynamic, to a dynamic simulation which will respond in real time. Whatever the method of simulation used, the key consideration in simulation studies is the trade-off between the fidelity of simulation (deciding the features of the system that need fidelity is an issue too), and the cost of involved in building high-fidelity simulations. Despite this limitation, simulation analysis can be useful when designing task situations that are dangerous for humans to perform, or difficult to observe.

Person Specification. The goal of person specification is to detail the key physical and mental capabilities, the key qualifcations and personality traits, and experience, required of the operator to perform specif ed tasks. Person specification is similar to the user char-

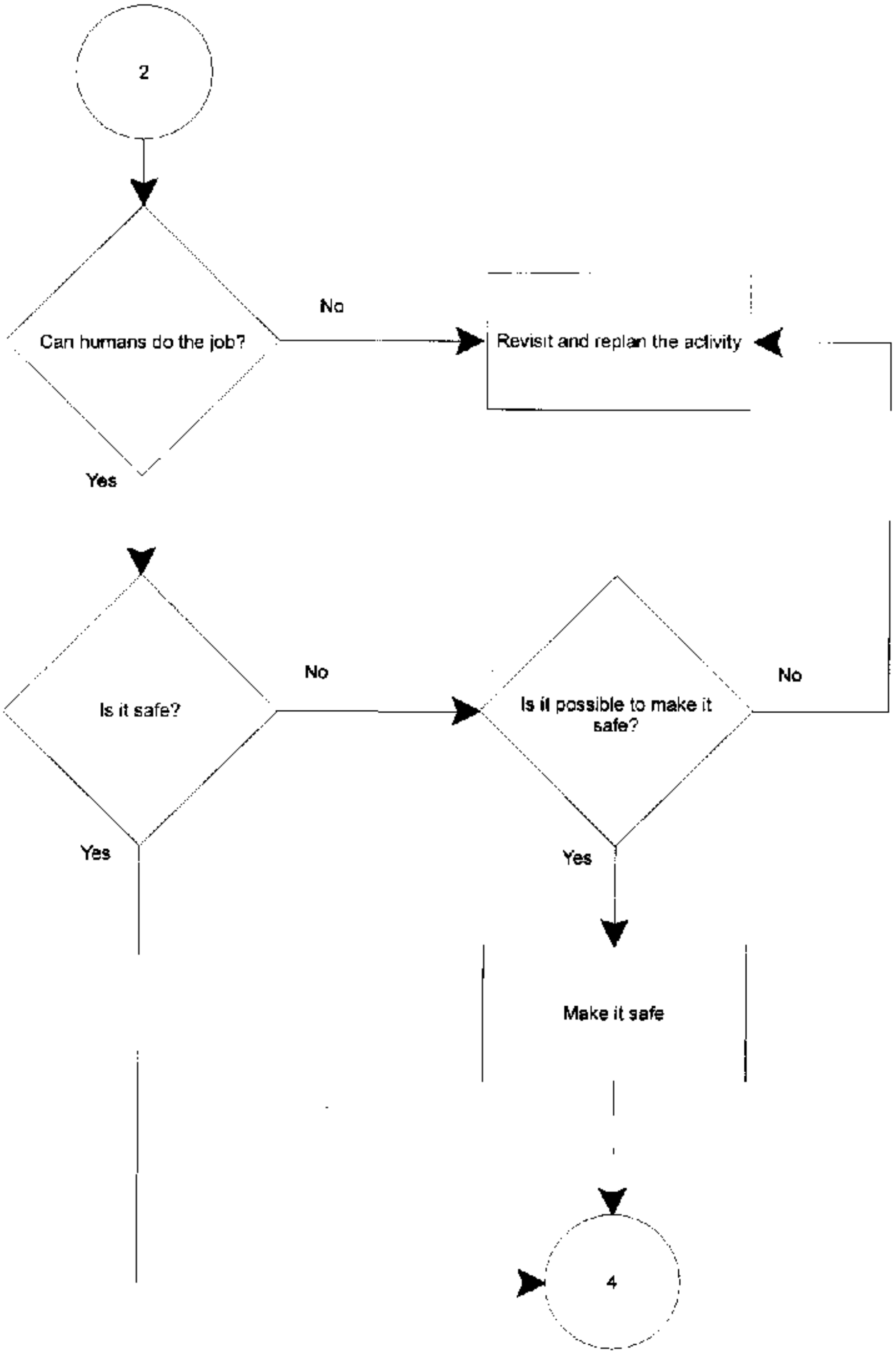

2
Can humans do the job?
No
Revisit and replan the activity
Yes
Is it safe?
No
Is it possible to make it safe?
No
Yes
Yes
Make it safe
4

Figure 7 (continued)

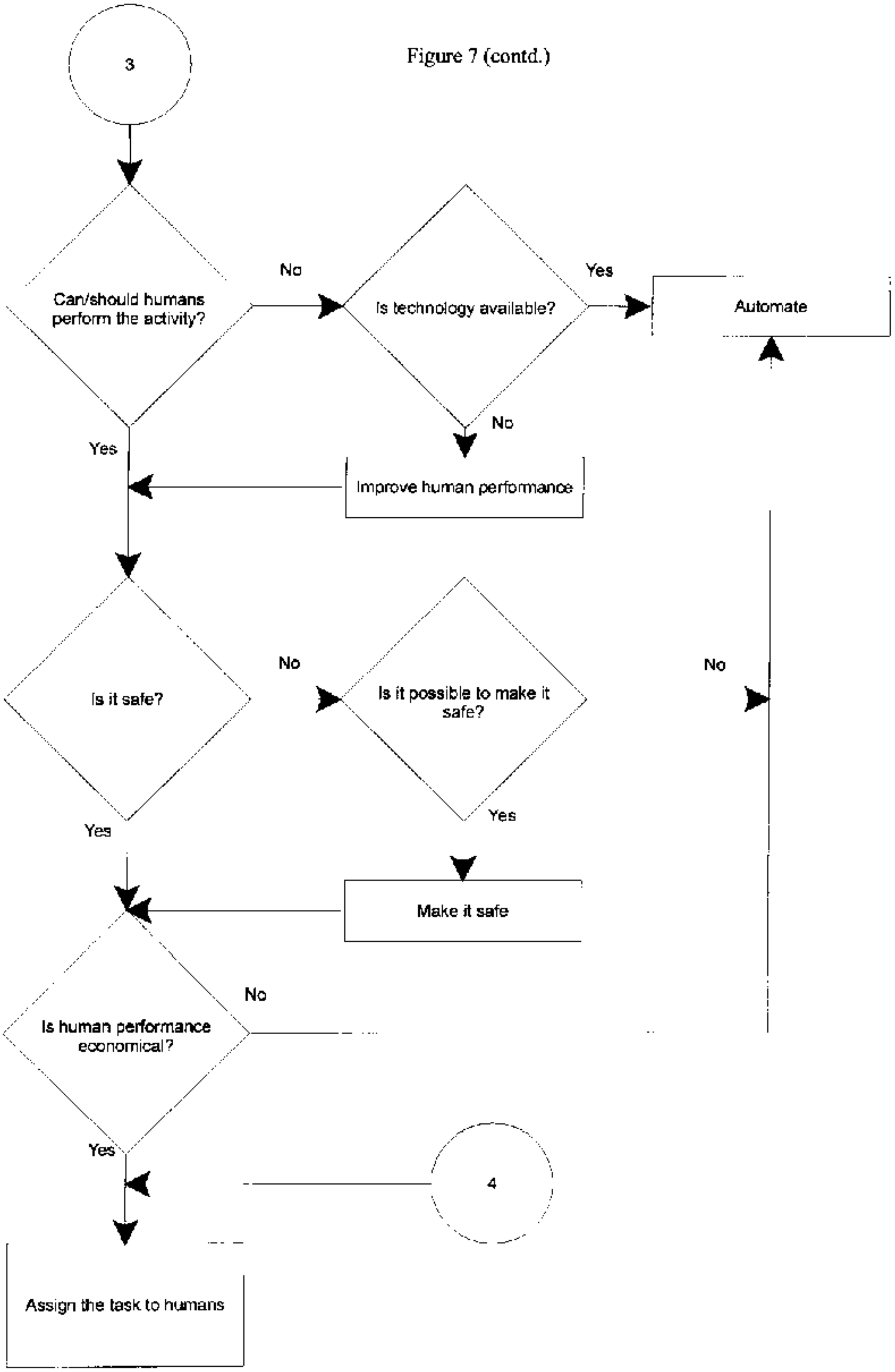

3
Figure 7 (contd.)
Can/should humans perform the activity?
No
Is technology available?
Yes
Automate
No
Yes
Improve human performance
Is it safe?
No
Is it possible to make it safe?
No
Yes
Yes
Make it safe
Is human performance economical?
No
Yes
4
Assign the task to humans

Figure 7 (continued)

acterization exercise described in Sec. 1.3.1.1; the checklist used for user characterization can be used for person specification also. One of the widely used techniques for person specification is the position analysis questionnaire. Broadly, position analysis questionnaires require the operator to identify for their specified tasks andjobs, the information input, the mental processes, the work output, the context of the job, the relationship with other personnel in the system, and any other relevant job characteristics. Using the responses from the operators, the skill content of tasks and jobs can be determined, and can help in designing personnel selection and training programs to ensure optimal human–machine interaction.

Ergonomics Checklists. These checklists are generally used to ascertain if a particular system meets ergonomic principles and criteria. Ergonomics checklists can check for subjective or objective information and can cover issues ranging from overall system design to the design of individual equipment. Checklists can also range in detail from the broad ergonomic aspects to the minute detail. Table 4 provides an example of a checklist for equipment operation. A number of other standard checklists have also been developed by the ergonomics community. Important among these are the widely used and comprehensive set of checklists for different ergonomics issues by Woodson [10,11], MIL-STD 1472C [12] which covers equipment design (written primarily for military equipment, but can be used as a guide to develop checklists), EPRI NP-2360 [13] which is a checklist for maintenance activities in any large-scale system, NUREG-0700 [14] which is a comprehensive checklist for control room design, the HSE checklist [15] which deals with industrial safety and human error, and the numerous checklists for CRT displays and VDUs [16,17].

1.3.1.3 Characterization of the Situation

Apart from the user and the task variables that could affect system performance, the external environment in which the system functions can also influence the human–system interaction performance. Table 5 provides a representative checklist for the most commonly encountered situations for which the system analyst must obtain answers, and attempt to provide for these situations in design.

1.3.2 Allocation of Functions

In designing the human–machine interface, once comprehensive information about the users and the activities/tasks these users will perform is known (through the use of tools presented in the earlier sections), the specific activities and tasks need to be assigned either to the humans or to the machines. The allocation of functions is a necessary first step before any further design of the interface in the human–machine system can be carried out.

The need for solving the function allocation problem directly stems from the need to decide the extent of automation of manufacturing activities. This is so because, in the present day manufacturing scenario, the decision to make is no longer whether or not to automate functions in manufacturing, but to what extent and how.

The function allocation problem is perhaps as old as the industrial revolution itself. Fitts' list, conceived in 1951 (Table 6), was the first major effort to resolve the function allocation problem.

However, while Fitts' list provided fundamental and generic principles that researchers still follow for studying function allocation problems, its failure to provide quantitative criteria for function allocation resulted in its having little impact on engineering design practices. The development of practical and quantitative criteria for allocating functions is compounded by an important issue: unless one can describe functions in engineering terms, it is impossible to ascertain if a machine can perform the function; and, if one can describe human behavior in engineering terms, it may be possible to design a machine to do the job better (than the human). But many functions cannot be completely specified in engineering (numerical) terms. This implies that those functions that cannot be specified in engineering terms should be allocated to humans, with the rest allocated to the machines. In addition, for the practitioner, function allocation considerations have been limited due to the lack of [19]:

1. Systematic and step-by-step approaches to decision making during function allocation
2. Systematic and concise data for addressing issues such as the capability and limitations of humans and automated equipment, and under what circumstances one option is preferable over the other
3. Methodology for symbiotic agents such as manufacturing engineers and ergonomists, to integrate human and machine behaviors
4. Unified theory addressing domain issues such as roles, authorities, etc
5. Integration of other decision-making criteria (such as the economics of the situation) so

Table 4 Example of an Ergonomics Checklist for Equipment Operation

Characteristic	Satisfactory	Compromise but acceptable	Unsatisfactory
Console shape/size			
Desk height, area			
Control reach			
Display view			
Body, limb clearance			
Panel location			
Frequency of use			
Sequence of use			
Emergency response			
Multioperator use			
Panel layout			
Functional grouping			
Sequential organization			
Identification			
Clearance spacing			
Displays			
Functional compatibility for intended purposes			
Intelligibility of information content			
Control interaction			
Legibility; figures, pointers, scales			
Visibility; illumination, parallax			
Location			
Identification			
Controls			
Functional compatibility for intended purposes			
Location, motion excursion, and force			
Display interaction			
Spacing, clearance, size			
Identification			

Adapted from Ref. 10.

Table 5 Checklist for Situation Analysis

	What are the likely situations that could arise during system use and how will these affect use of the system?
Equipment	Falls short of target performance Falls short of specification Fails
Availability	Data is missing Materials are missing Personnel are missing Support is missing
Overloads	Of people/machines Of data, information, materials, etc.
Interruptions	The process breaks down Complete restart of process required
Environment	Changes: in physical or social environment
Policy changes	Changes in laws, rules, standards and guidelines

Adapted from Ref. 5.

that the function allocation decision is not made in isolation

6. Easily usable tools to simulate different configurations of humans and machines.

In spite of these shortcomings, research on function allocation has permitted the following general inferences for the practitioner:

1. Function allocation cannot be accomplished by a formula—or example, rules which may apply in one situation may be irrelevant in another.
2. Function allocation is not a one-shot decision—the final assignment depends on activities at the levels of the tasks, the conflation of tasks into jobs, the relationships of jobs within a larger workgroup, and the likely changes in the higher level manufacturing processes themselves.
3. Function allocation can be systematized—it is clear that there are a number of sequential steps that can be taken to best allocate functions.
4. Both humans and machines can be good or bad at certain tasks.
5. Using analogies can facilitate the function allocation process.
6. Function allocation can be targeted to a specific time frame.
7. Function allocation depends on the nature of the task—it varies based on whether the task is perceptual, cognitive, or psychomotor.
8. Function allocation decisions must be based on sound economic analyses of options as well as the capabilities and limitations of humans and machines.
9. Human and machine performances are not always antithetical.
10. Function allocation decisions must consider technology advances within a given time frame.
11. In cases where both humans and machines can perform a function, the system should be designed in such a way so that humans can delegate the function to machines, or can

Table 6 Fitts' List

Humans appear to surpass present-day machines with respect to the following:
- Ability to detect small amounts of visual or acoustic energy
- Ability to perceive patterns of light or sound
- Ability to improvise and use flexible procedures
- Ability to store very large amounts of information for long periods and to recall relevant facts at the appropriate time
- Ability to reason inductively
- Ability to exercise judgment

Present-day machines appear to surpass humans with respect to the following:
- Ability to respond quickly to control signals, and to apply great force smoothly and precisely
- Ability to perform repetitive, routine tasks
- Ability to store information briefly and then to erase it completely
- Ability to reason inductively, including computational ability
- Ability to handle complex operations, i.e., to do many different things at once

Adapted from Ref. 18.

take over the function when circumstances demand it.

A number of approaches have been suggested in the literature for solving the function allocation problem. Some of the promising approaches include function allocation criteria based on specific performance measures (time required to complete tasks, for example) [20–24], criteria based on comparison of capabilities and limitations of humans with particular attention given to knowledge, skills, and information sources and channels [25–34] criteria based on economics (allocate the function to the less expensive option), [21,35,36], and criteria based on safety (to the human operator in the system) [37–39].

Experiments with these approaches suggest that functions that are well-proceduralized permitting algorithmic analysis, and requiring little creative input, are prime candidates for automation. On the other hand, functions requiring cognitive skills of a higher order, such as design, planning, monitoring, exception handling, etc., are functions that are better performed by humans. The prime requirements for automating any function are the availability of a model of the activities necessary for that function, the ability to quantify that model, and a clear understanding of the associated control and information requirements. Clearly, there are some functions that should be performed by machines because of:

1. Design accuracy and tolerance requirements.
2. The nature of the activity is such that it cannot be performed by humans.
3. Speed and high production volume requirements.
4. Size, force, weight, and volume requirement.
5. Hazardous nature of the activity.

Equally clearly, there are some activities that should be performed by humans because of:

1. Information-acquisition and decision-making needs
2. Higher level skill needs such as programming
3. Specialized manipulation, dexterity, and sensing needs
4. Space limitations (e.g., work that must be done in narrow and confined spaces)
5. Situations involving poor equipment reliability or where equipment failure could prove catastrophic
6. Activities for which technology is lacking.

Mital et al. [7] provide a generic methodology in the form of decision-making flowcharts for the systematic allocation of functions between humans and machines. Figure 7, presented earlier is a part of these flowcharts. These flowcharts are based on the requirements of complex decision making, on a detailed safety analysis, and on a comprehensive economic analysis of the alternatives. These function allocation flowcharts are available for different manufacturing functions such as assembly, inspection, packaging, shipping, etc., and should be consulted for a detailed analysis of the question of manufacturing function allocation.

1.3.3 Information Presentation and Control

1.3.3.1 The Scientific Basis for Information Input and Processing

Reduced to a fundamental level, human interaction with automation can be said to be dependent upon the information processing ability of the human, and upon the exchange of information among the different elements in a system. Over the years, behavioral scientists have attempted to explain human information processing through various conceptual models and theories. One such theory is the information theory [40] *Information*, according to information theory, is defined as the reduction of uncertainty. Implicit in this definition is the tenet that events that are highly certain to occur provide little information; events that are highly unlikely to occur, on the other hand, provide more information. Rather than emphasize the importance of a message in defining information, information theory considers the probability of occurrence of a certain event in determining if there is information worth considering. For instance, the "no-smoking" sign that appears in airplanes before takeoff, while being an important message, does not convey much information due to the high likelihood of its appearance every time an aircraft takes off. On the other hand, according to information theory, messages from the crew about emergency landing procedures when the plane is about to perform an emergency landing convey more information due to the small likelihood of such an event. Information is measured in bits (denoted by H). One *bit* is defined as the amount of information required to decide between two equally likely alternatives.

When the different alternatives all have the same probability, the amount of information (H) is given by

$$H = \log_2 N$$

where N is the number of alternatives. For example, when an event only has two alternatives associated with it, and when the two alternatives are equally likely, by the above equation, the amount of information, in bits, is 1.0.

When the alternatives are not equally likely (i.e., the alternatives have different probabilities of occurrence), the information conveyed by an event is given by

$$h_i = \log_2 (1/p_i)$$

where h_i is the information associated with event i, and p_i is the probability of occurrence of event i.

The average information (H_{av}) conveyed by a series of events having different probabilities is given by

$$H_{\text{av}} = \sum p_i(\log_2 (1/p_i))$$

where p_i is the probability of the event i.

Just as a *bit* is the amount of information, *redundancy* is the amount of reduction in information from the maximum due to the unequal probabilities of occurrence of events. Redundancy is expressed as a percentage, and is given by

$$\% \text{ Redundancy} = (1 - H_{\text{av}}/H_{\text{max}}) \times 100$$

Information theory, while providing insight into measuring information, has major limitations when applied to human beings. It is valid only for simple situations which can split into units of information and coded signals [41]. It does not fully explain the stimulus-carrying information in situations where there are more than two alternatives, with different probabilities.

The *channel capacity theory*, another theory explaining information uptake by humans, is based on the premise that human sense organs deliver a certain quantity of information to the input end of a channel, and that the output from the channel depends upon the capacity of the channel. It has been determined that if the input is small, there is very little absorption of it by the channel, but that if the input rises, it reaches the threshold channel capacity, beyond which the output from the channel is no longer a linear function of the input [41]. Experimental investigations have shown that humans have a large channel capacity for information conveyed to them through the spoken word than through any other medium. A vocabulary of 2500 words requires a channel capacity of 34 to 42 bits per second [42]. Designers must keep in mind that in this day and age of information technology, the central nervous system of humans is subjected to more information than the information channel can handle, and that a considerable reduction in the amount of information must be carried out before humans process the information.

In addition to theories such as the information theory and the channel capacity theory that explain information uptake, many conceptual models of human information processing have been proposed by researchers over the last four decades. Figure 8 shows one such fundamental model (most other models contain elements of this basic model) depicting the stages involved in information processing [43]. The key elements of the model are perception, memory, decision making, attention, response execution, and feedback. The following is a brief discussion of each of these elements.

Perception may involve *detection* (determining whether or not a signal is present), or *identification and detection* (involving detection and classification). The theory of signal detection [43–45] through the concept of *noise* in signals, attempts to explain the process of perception and response to the perceived signals. Four possible outcomes are recognized in signal detection theory: (1) hit (correctly concluding that there is a signal when there is one), (2) false alarm (concluding that there is a signal when, in actuality, there is none), (3) miss (concluding that there is no signal when, in actuality, there is one and (4) correction rejection (correctly concluding that there is no signal when there is none). The fundamental postulate of signal detection theory is that humans tend to make decisions based on criteria whose probabilities depend upon the probabilities of the outcomes above. The probability of observing a signal, and the costs and benefits associated with the four possible outcomes above, determine the responses of the human to the signal. The resolution of the human sensory activities (ability to separate the noise distribution from the distribution of the signal) has also been found to affect the signal detection capability of the human.

Memory, in humans, has been conceptualized as consisting of three processes, namely, sensory storage, working memory, and long-term memory [43]. According to this conception, information from sensory storage must pass through working memory before it can be stored in long-term memory. Sensory storage refers to the short-term memory of the stimulus. Two types of short-term memory storage are well known—*iconic storage* associated with visual senses, and *echoic storage* associated with the auditory senses [46]. Sensory storage or short-term memory has been shown to be nearly automatic requiring no sustained attention on the part of the human to retain it.

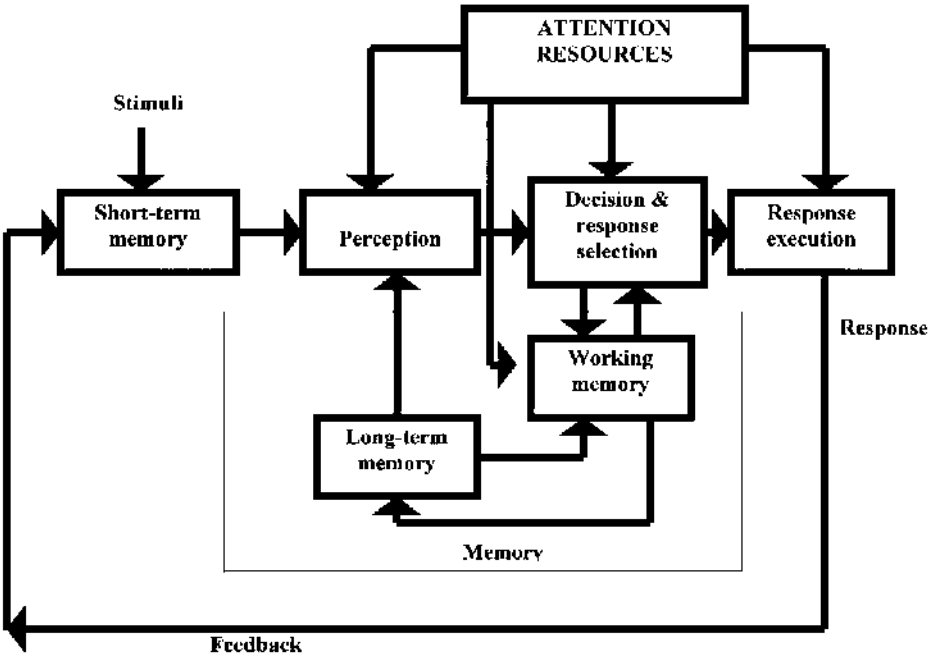


Figure 8 Fundamental model of human information processing.

Information transfer from sensory storage to working memory is brought about through *attention* (to the process). Information from stimuli is believed to be stored in the working memory primarily in the form of either visual, phonetic, or semantic codes. It is also believed that the capacity of working memory is five to nine chunks of information (similar units regardless of the size) [47]. Researchers recommend presenting five to nine meaningful and distinct chunks of information for improved recall. It has also been determined that there is a linear relationship between the number of items in a memorized list and the time required to search the list of items in the working memory [48]. Also, all items in the working memory are searched one at a time, even if a match is found early in the search process. The transfer of information from working memory to the long-term memory is believed to take place through semantic coding, i.e., by analyzing, comparing, and relating information in the working memory to past stores of knowledge in the long-term memory [46]. The extent to which information can be retrieved from long-term memory depends on the extent of organization of the information in the long-term memory.

Rational decision making is defined as the process that involves seeking information relevant to the decision at hand, estimating the probabilities of various alternatives, and attaching values to the anticipated alternatives. A number of *biases*, however, have been identified to exist among humans that often makes decision making irrational. Table 7 lists some of these biases.

Attention is another key factor influencing human information input and processing. Research has identified four types of tasks or situations requiring attention. These are selective attention, focused attention, divided attention, and sustained attention. When several information sources are to be monitored to perform a single task, attention is said to be *selective* (e.g., a process control operator scanning several instrument panels before detecting a deviant value). Table 8 provides guidelines for improving performances in tasks requiring selective attention. When a human has to focus attention on one source of information and exclude all other sources of information for task performance, attention is said to be *focused.* Task performance under focused attention is affected by the physical proximity of the sources of information. While physical proximity enhances performance in tasks requiring selective attention, it impedes performance in tasks requiring focused attention. Table 9 provides guidelines for improving performances in tasks requiring focused attention. When humans do more than one task at a time, their atten-

Table 7 Common Human Biases

Humans attach more importance to early information than subsequent information.
Humans generally do not optimally extract information from sources.
Humans do not optimally assess subjective odds of alternative scenarios.
Humans have a tendency to become more confident in their decisions with more information, but do not necessarily become more accurate.
Humans tend to seek more information than they can absorb.
Humans generally treat all information as equally reliable.
Humans seem to have a limited ability to evaluate a maximum of more than three or four hypotheses at a time.
Humans tend to focus only on a few critical factors at a time and consider only a few possible choices related to these critical factors.
Humans tend to seek information that confirms their choice of action than information that contradicts or disconfirms their action.
Human view a potential loss more seriously than a potential gain.
Humans tend to believe that mildly positive outcomes are more likely than mildly negative or highly positive outcomes.
Humans tend to believe that highly negative outcomes are less likely than mildly negative outcomes.

Adapted from Ref. 43.

tion is said to be *divided* (among the tasks). While much of the theoretical base for explaining performance of tasks requiring divided attention is still evolving [43,49], some guidelines for designing tasks that require divided attention are available, and are provided in Table 10. When humans maintain attention and remain vigilant to external stimuli over prolonged periods of time, attention is said to be *sustained.* Nearly four decades of research in vigilance and vigilance decrement [50–53] has provided guidelines for improving performance in tasks requiring sustained attention (Table 11).

In addition to the factors discussed above, considerable attention is being paid to the concept of mental workload (which is but an extension of divided attention). Reviews of mental workload measurement techniques are available [54–56], and should be consulted for discussions of the methodologies involved in mental workload assessment.

1.3.3.2 The Scientific Basis for Human Control of Systems

Humans respond to information and take controlling actions. The controlling actions of the human are mediated through the motor system in the human body. The human skeletal system, the muscles, and the nervous system help bring into play *motor skills* that enable the human to respond to stimuli. Motor skill is defned as ‘‘ability to use the correct muscles with the exact force necessary to perform the desired response with proper sequence and timing’’ [57]. In addition, *skilled* performance requires adjusting to changing environmental conditions, and acting consistently from situation to situation [58]. A number of different types of human movements have been recognized in the literature [46]. These include discrete movements (involving a single reaching movement to a target that is stationary), repetitive movements (a single movement is repeated), sequential movements

Table 8 Recommendations for Designing Tasks Requiring Selective Attention

Use as few signal channels as possible, even if it means increasing the signal rate per channel.
Inform the human the relative importance of various channels for effective direction of attention.
Reduce stress levels on human so more channels can be monitored.
Inform the human beforehand where signals will occur in future.
Train the human to develop optimal scan patterns.
Reduce scanning requirements on the human by putting multiple visual information sources close to each other, and by making sure that multiple sources of auditory information do not mask each other.
Provide signal for a sufficient length of time for individual to respond; where possible, provide for human control of signal rate.

Adapted from Ref. 46.

Table 9 Recommendations for Designing Tasks Requiring Focused Attention

Make the different channels of information as distinct as possible from the channel to which the human must attend.
Physically separate the channel of interest from the other channels.
Reduce the number of competing channels.
Make the channel of interest prominent by making it larger in size, or brighter, or louder, or by locating it centrally.

Adapted from Ref. 46.

(a number of discrete movements to stationary targets), continuous movements (involving muscular control adjustments during movement), and static positioning (maintaining a specific position of a body member for a specified period of time). In addition, certain theoretical models of human motor responses explain the control aspects of human responses based on only two fundamental types of movements—fast and slow. Closed-loop theories [59,60], whether the movement be fast or slow, use the concept of sensory feedback (sensory information available during or after the motor response) to explain motor responses (to correct/reduce errors obtained through feedback). The sensory receptors for feedback and feedforward (sensory information available prior to the action that regulates and triggers responses), are believed to be located in the muscle spindles (for sensing the muscle length and the rate of change of length) [58,61], tendons (the Golgi tendons inhibit muscle contraction and regulate muscle action), joints (the tension in the joints influences the generation of nerve impulses), cutaneous tissue (skin is believed to have receptors that affect joint movement), and the eyes (important for timing of responses) [62]. Open-loop theories, on the other hand, are based on the belief that there are higher-level structured motor programs containing information necessary for patterning the different movements [63,64]. Different deficiencies, such as the *error of selection* (where a person calls the wrong motor program for a controlling action) and the *error of execution* (where the correct motor program fails during execution of controlling actions) have been identified with motor programs [65]. Much of the development in understanding human controlling actions in response to stimuli is still in its infancy, but has important practical consequences (how to improve skilled performance, for example).

Table 10 Recommendations for Designing Tasks Requiring Divided Attention

Minimize the potential sources of information.
Provide human with a relative priority of tasks to optimize the strategy of divided attention.
Keep the level of difficulty of tasks low.
Make tasks as dissimilar as possible in terms of task demands on the human.

Adapted from Ref. 46.

Table 11 Recommendations for Designing Tasks Requiring Sustained Attention

Provide appropriate work–rest schedules.
Provide task variation by interpolating different activities.
Make the signal larger, and/or more intense, and/or longer in duration, and/or distinct.
Reduce uncertainty in time and place of occurrence of signal.
Use artificial signals and provide feedback to humans on their performance.
Reduce the rate of presentation of stimuli if it is high.
Provide optimal environmental conditions such as lighting, noise level, etc.
Provide adequate training to humans to clarify the nature of signals to be identified.

The *time* it takes for the human *to respond* to stimuli is another critical factor that has been studied extensively in the literature [46]. An understanding of response time of the human is essential for good design of the tasks involved in human interaction with automated systems. *Response time* is, in general, composed of reaction time, and movement time. *Reaction time* is defined as the time from the signal onset calling for a response, to the beginning of the response. *Simple reaction time* (reaction time in the presence of a single source of stimulus) has been shown to be between 0.15 sec and 0.20 sec. The mode through which the single stimulus occurs (visual, auditory etc.,) the detectability of the stimulus (intensity, duration, and size), the frequency, the preparedness (of the human for the stimulus), the age, and the location of the stimulus (location in the peripheral field of view, for instance) are among the factors that have been shown to affect simple reaction time. *Choice reaction time* (reaction time in the presence of one of several possible stimuli each with different possible responses), is a function of the probability of a stimulus occurring, i.e., the reaction time is faster for events with greater probability. It has been shown to increase by about 0.15 sec for each doubling of the number of possible alternative stimuli [66]. Choice reaction time has been shown to be influenced by a numerous factors, including the degree of compatibility between stimuli and responses, practice, presence or absence of a warning signal, the type and complexity of the movement involved in the responses, and whether or not more than one stimulus is present in the signal. *Movement time* is defned as the time from the beginning of the response to its completion. It is the time required to physically make the response to the stimulus. Movements based on pivoting about the elbow have been shown to take less time, and have more accuracy, than movements based on upper-arm and shoulder action. Also, it has been determined that movement time is a logarithmic function of distance of movement, when target size is a constant, and further that movement time is a logarithmic function of target size, when the distance of movement is constant. This finding is popularly known as Fitts' law [67], and is represented as

$$\mathrm{MT} = a + b \log_2(2D/W)$$

where MT is the movement time, a and b are empirical constants dependent upon the type of movement, D is the distance of movement from start to the center of the target, and W is the width of the target.

Human response to stimuli is not only dependent upon the speed of the response, but also on the accuracy of the response. The accuracy of the human response assumes special importance when the response has to be made in situations where there is no visual feedback (a situation referred to as "blind positioning"). Movements that take place in a blind positioning situation have been determined to be more accurate when the target is located dead-ahead than when located on the sides. Also, targets below the shoulder height and the waist level are more readily reachable than targets located above the shoulder or the head [68]. The distance and speed of movement have also been found to influence the accuracy of the response [69,70].

1.3.3.3 Displays

Types of Displays. A display is defined as any indirect means of presenting information. Displays are generally one of the following four types: visual, auditory, tactual, and olfactory. The visual and the auditory modes of displaying information are the most common. Displays based on tactile and olfactory senses are mostly used for special task or user situations (e.g., for the hearing impaired).

Selecting the mode of display whether it should be visual or auditory in nature) is an important factor due to the relative advantages and disadvantages certain modes of display may have over other modes, for specific types of task situations (auditory mode is better than visual displays in vigilance), environment (lighting conditions), or user characteristics (person's information handling capacity). Table 12 provides general guidelines for deciding between two common modes of information presentation, namely, auditory and visual.

The types of displays to use to present information also depend on the type of information to present. Different types of information can be presented using displays when the sensing mode is indirect. Information can either be dynamic or static. Dynamic information is categorized by changes occurnng in time (e.g., fuel gage). Static information,

Table 12 Guidelines for Deciding When to Use Visual Displays and When to Use Auditory Displays

Characteristics	Visual displays	Auditory displays
Message characteristics		
Simple message		√
Complex message	√	
Short message		√
Long message	√	
Potential reference value of message		
High	√	
Low		√
Immediacy of action requirement of message		
High		√
Low	√	
Message deals with events in time		√
Message deals with locations in space	√	
Human capability		
Auditory system overburdened	√	
Visual system overburdened		√
Environmental factors		
Location too bright or too dark requiring significant adaptation		√
Location too noisy	√	

Adapted for Ref. 71.

on the other hand, does not change with time (e.g., printed safety signs). A number of other types of information are also recognized in the literature. Table 13 provides a list of these types along with a brief description of the characteristics of these types of information.

In the following sections, we discuss recommendations for the design of different types of visual and auditory displays (we restrict our attention in this chapter only to these two common modes). We first provide a brief discussion of the different factors affecting human visual and auditory capabilities. We then present specific display design issues and recommendations for these two broad types of displays.

Visual displays: factors affecting design. *Accommodation* refers to the ability of the lens in the eye to focus the light rays on the retina. The distance (of the target object from the eye) at which the image of the object becomes blurred, and the eye is not able to focus the image any further, is called the near point. There is also a far point (infinity, in normal vision) beyond which the eye cannot clearly focus. Focal distances are measured in diopters. One diopter is 1/(distance of the target in meters). Inadequate accommodation capacity of the eyes result either in nearsightedness (the far point is too close) or in farsightedness (the near point is too close). Literature recommends an average focusing distance of 800 mm at the resting position of the eye (also known as the resting accommodation) [72]. Due to changes in the iris (which controls the shape of the lens), aging results in substantial receding of the near point, the far point remaining unchanged or becoming shorter. Figure 9 shows how the mean near point recedes with age. It is recommended that the designer use this information when designing visual displays.

Visual acuity is defined as the ability of the eye to separate fine detail. The minimum separable acuity refers to the smallest feature that the eye can detect. Visual acuity is measured by the reciprocal of the visual angle subtended at the eye by the smallest detail that the eye can distinguish. Visual angle (for angles less than 10°) is given by

$$\text{Visual angle (in minutes)} = (3438H)/D$$

where H is the height of the stimulus detail, and D is the distance from the eye, both H and D measured in the same units of distance. Besides minimum separable visual acuity, there are other types of visual acuity measure, such as vernier acuity (ability to differentiate lateral displacements), minimum perceptible acuity (ability to detect a spot from its background), and stereoscopic acuity (ability to differentiate depth in a single object). In general, an individual is considered to have normal visual acuity if he or she is able to resolve a separation between two signs 1′ of arc wide. Visual acuity has been found to increase with increasing levels of illumination. Luckiesh and Moss [73] showed that increasing the illumination level from approximately 10 l× to 100 l× increased the visual acuity from 100 to 130%, and increasing the illumination level from approximately 10 l× to 1000 l× increased the visual acuity from 100 to 170%. For provision of maximum visual acuity, it is recommended that the illumination level in the work area be 1000 l×. Providing adequate contrast between the object being viewed and the immediate background, and making the signs and

Table 13 Commonly Found Types of Information and Their Characteristics

Type of information	Characteristics
Quantitative information	Information on the quantitative value of a variable
Qualitative information	Information on the approximate value, trend, rate of change, direction of change, or other similar aspects of a changeable variable
Status information	Information on the status of a system, information on a one of a limited number of conditions, and information on independent conditions of some class
Warning and signal information	Information on emergency or unsafe conditions, information on presence or absence of some conditions
Representational information	Pictorial or graphic representations of objects, areas, or other configurations
Identification information	Information in coded form to identify static condition, situation, or object
Alphanumeric and Symbolic information	Information of verbal, numerical, and related coded information in other forms such as signs, labels, placards, instructions, etc.
Time-phased information	Information about pulsed or time-phased signals

Adapted from Ref. 46.

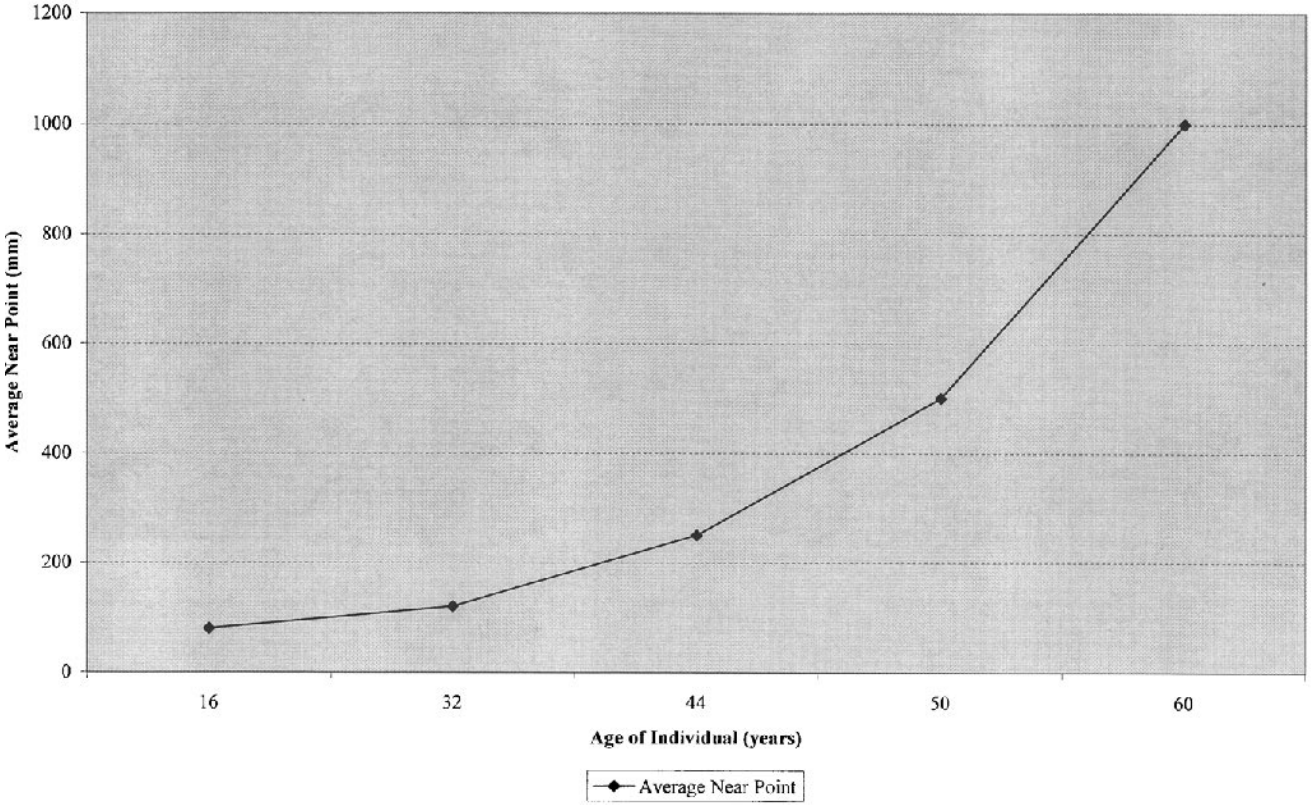


Figure 9 Effect of age on near point for visual accomodation.

characters (in the object being viewed) sharp, will also increase visual acuity. The general recommendation is to use dark symbols and characters on a bright background than vice versa, as the former increases the visual acuity. Visual acuity has also been shown to decrease with age [74]. Figure 10 illustrates how visual acuity decreases with age.

Contrast sensitivity is another factor that has implications for design of the interface. It is the ability of the eye to differentiate lightness between black and white. Contrast sensitivity is generally expressed as the reciprocal of the threshold contrast, where the threshold contrast is the level of contrast that just stops short of making the colors appear homogeneous. Other measures for contrast sensitivity include modulation contrast computed as

$$C = (L_{\max} - L_{\min})/(L_{\max} + L_{\min})$$

where $L_{\max}$ and $L_{\min}$ are the maximum and the minimum luminances in the pattern. The literature provides certain general rules to follow when designing displays in order to provide the best possible contrast sensitivity. Since contrast sensitivity is greater for larger areas, it is recommended that the viewing area be made as large as possible. Also, making the object boundaries sharper will increase contrast sensitivity. The surrounding luminance, and the intensity of light (or the level of illumination), have been shown to have an effect on contrast sensitivity. Contrast sensitivity

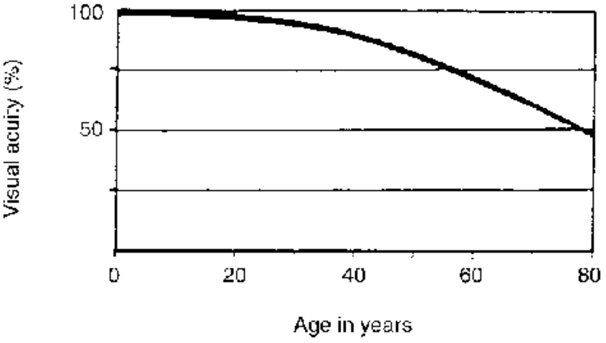


Figure 10 Effect of age on visual acuity.

has been determined to be the largest when the surrounding luminance is within the range of 70 cd/m^2, and more than 1000 cd/m^2 [75]. Also, Luckiesh and Moss [73] showed that increasing the illumination level from approximately 10 l× to 100 l× increased the contrast sensitivity from 100 to 280%, and increasing the illumination level from approximately 10 l× to 1000 l× increased the contrast sensitivity from 100 to 450%. The literature [41] also recommends that the background be at least 2% brighter or darker than the target for optimal contrast sensitivity. As briefly described above, visual acuity and contrast sensitivity are affected by a number of factor, such as luminance level (in general, the higher the luminance, the more the visual acuity and contrast sensitivity), contrast, exposure time, motion of the target, age (there is a decline in both visual acuity and contrast sensitivity with age), and training (through surgery of the eye or through corrective lenses, etc.).

Adaptation is another factor that affects the visual capability of the human eye. It is defined as the changes in the sensitivity of the eye to light. A measure of adaptation is the time it takes for the eye to adapt to light or dark. It has been found that, in general, adaptation to light occurs more quickly than adaptation to the dark. Darkness adaptation has been found to be quick in the first 5 min of exposure; nearly 80% of the adaptation to darkness has been shown to take about 25 min with full adaptation taking as much as one full hour [41]. Adaptation can also be partial (depending on whether the visual field contains a dark or a bright area), and can affect the sensitivity of the retina and the vision. For optimal adaptation, the overall recommendation is to provide the same order of brightness on all important surfaces, and provide a stable and nonfluctuating levels of illumination. It is also important to avoid the effects of glare (which is a process of overloading the adaptation processes of the eye). This can be achieved by avoiding excessive brightness contrasts, avoiding excessive brightness in the light source, and providing for transient adaptation.

The ability of the eye to discriminate between different colors is called *color discrimination*. Color discrimination deficiency is due to the reduced sensitivity of the particular (to a color) cone receptors. While it is difficult to measure precisely the type and degree of a person's color deficiency, it is important from the perspective of designing tasks which require perception of colored targets for task performance.

The ability to *read*, and the ability to *perceive* meaning, are the other key factors that have to be accounted for when designing visual displays.

Design recommendations for visual displays. As already mentioned, visual displays are classified on the basis of the type of information they present to the user. Information presented to the user can be *static* or *dynamic* in nature. Display of dynamic information will require capture of the changing nature of the information (for example, continuous changes in speed indicated by the tachometer in the car). Static displays do not display, in real time, the changes in the information content in time. (Note that, in static displays, the displays themselves do not change with time. However, static displays can be used to present, in the form of graphs, for example, changes in information content over time, after the event has occurred; static displays do not provide information in real time.) Almost all dynamic visual displays contain elements of one of the more fundamental forms of static information displays, namely, textual information, information in the form of graphical displays, information in some coded form, or symbolic information. In the following sections, we first briefly present recommendations on design of these four forms of static visual displays. We then provide guidelines on designing dynamic information displays.

Static visual displays. The literature distinguishes between two forms of *textual displays*—textual displays in hardcopy format, and textual displays in visual display terminals or computer screens [46]. While there are differences in performance based on whether the display is in hardcopy form or in a visual display unit, there are three essential characteristics of any display in the form of text; the textual display should be visible, legible, and readable. *Visibility* of the text refers to the characteristic that makes a character or a symbol distinguishable and separate from its surroundings. *Legibility* of the text refers to the characteristic of alphanumeric characters that makes it possible to identify one character from the other. The stroke width, the character format, contrast, illumination etc., influence the legibility of the text. *Readability* of the text refers to the characteristic of alphanumeric characters that enables organization of the content into meaningful groups (of information) such as words and sentences.

Various factors influence the visibility, the legibility, and the readability of textual information presented in hardcopy form. They are typography, size, case, layout, and reading ease.

The typography has been found to be especially important when the viewing conditions are unfavorable, or when the information is critical (such as a sign warning of danger). Typography depends on factors such as the stroke width of the alphanumeric character (ratio of thickness of the stroke to height of the character), the width-to-height ratio of the character, and the type style.

Table 14 also provides accepted guidelines, based on research, for size of characters, case, layout of characters, and for reading ease of alphanumeric characters. Some examples of type style and other aspects in typography of text are given in Fig. 11.

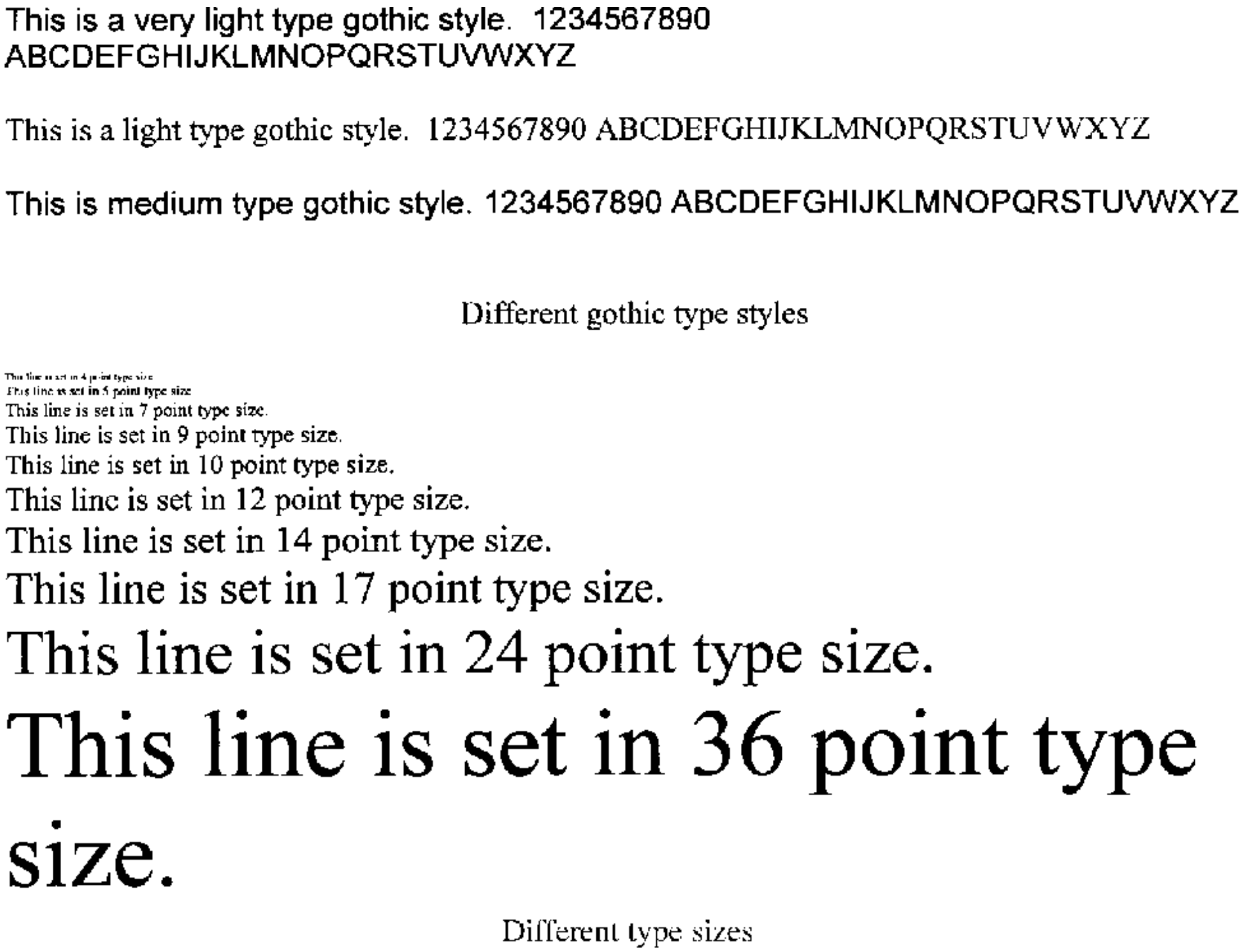
Numerical text can also be represented in graphical forms. Graphs can be in different forms such as line graphs, bar and column graphs, pie charts, etc. Pictorial information such as in the form of graphs improves the speed of reading, but the general recommendation in the literature is to combine pictorial information with information in the form of plain text, improve the accuracy of the information presented [76,77].

The visibility, readability, and legibility of visual-display-terminal-based text has been found to depend upon the typography, the reading distance, the size of characters, and hardware considerations, such as the

Table 14 Recommendations for Design of Hardcopy Text

Characteristic	Recommendations
Typography	
Stroke width	When the illumination is reasonable, use 1 : 6 to 1 : 8 for black on white and 1 : 8 to 1 : 10 for white on black When the illumination is reduced, use thick letters than thin letters for greater readability When illumination is low or with low background contrast, use boldface characters with a low stroke width–height ratio When letters are highly luminous, use 1 : 12 to 1 : 20 ratio When letters are black on a highly luminous background, use thick strokes
Width–height ratio	Use a 3 : 5 ratio for most practical applications; for transluminated or engraved legends, use 1:1
Size of character	
For close reading	*When the reading distance is 12 to 16 in.:* Use 0.09–0.11 in. or 22 to 27′ of visual angle for normal use of alphanumeric characters *When the viewing distance is 28 in.:* For critical use under 0.03 fL luminance, and variable position of character, use 020–0.30 in. height For critical use over 1.0 fL luminance, and variable position of character, use 0.12–0.20 in. height For critical use under 0.03 fL luminance, and fixed position of character, use 0.15 to 0.30 in. height For critical use over 1.0 fL luminance, and fixed position of character, use 0.10–0.20 in. height For noncritical use 0.05–0.20 in. height
For distant reading	Use $W_s = 1.45 \times 10^{-5} \times S \times d$, and $H_L = W_s/R$, where W_s is the stroke width, S is the denominator of the Snellen acuity score (20/20, 20/40 etc.), d is the reading distance, H_L is the height of the letter, and R is the stroke width-to-height ratio of the font
Case	In general, use lowercase letters than uppercase letters for better readability Use initial uppercase for search tasks
Layout	
Interletter spacing	Provide close-set type than regular-set type where possible for easier readability
Interline spacing	Increase spacing between lines for better clarity
Reading ease	
Type of sentence	Use simple, affirmative, active sentences where possible
Order of words	Match order of words in sentence to the order of actions to be taken

Adapted from Ref. 46.



This is a very light type gothic style. 1234567890
ABCDEFGHIJKLMNOPQRSTUVWXYZ

This is a light type gothic style. 1234567890 ABCDEFGHIJKLMNOPQRSTUVWXYZ

This is medium type gothic style. 1234567890 ABCDEFGHIJKLMNOPQRSTUVWXYZ

Different gothic type styles

This line is set in 4 point type size
This line is set in 5 point type size
This line is set in 7 point typc size.
This line is set in 9 point type size.
This line is set in 10 point type size.
This linc is set in 12 point type size.
This line is set in 14 point type size.
This line is set in 17 point type size.
This line is set in 24 point type size.
This line is set in 36 point type size.

Different type sizes

Figure 11 Examples of different type styles and type size.

polarity of the screen, and screen color. It is generally recommended that the size of the dot matrix for alphanumeric characters used in visual display terminals be at least 7 × 9 for continuous reading of the text. The ANSI recommendation for reading distance is 18–20 in. This distance denotes the distance from the eye to the screen, and is based on the assumption that the user is seated in an upright position. The ANSI specification for the minimum character height for capital letters is 16′ of visual angle for reading tasks where legibility is important. The maximum character height according to ANSI should be 24′ of visual angle, with the preferred character height set at 20–22′ of visual angle. As regards polarity of the screen, since the sensitivity to flicker is greater when the screen background is brighter, the literature recommends that display units with light backgrounds have a higher refresh rate than display units with dark backgrounds.

Information, in the form of stimuli, can be sensed either through direct observation of the object, or through the use of a indirect mediating device. During indirect sensing, the stimuli themselves mostly come in a coded form (such as a visual or an auditory display), and sometimes in the form of exact or modified (in size) reproductions of the original stimulus (such as a picture on a television screen). Coding of information can be along different stimulus dimensions; for example, coding can be done based on color, shape, size, etc. Research has shown that the success of coding in conveying the necessary information depends on people's ability to distinguish between two or more stimuli which vary along a dimension (e.g., which of the two stimuli is smaller in size), and on the ability to identify a single stimulus based on the measure of that stimulus on the dimension scale (e.g., whether the target is bright or dim) [46]. These abilities, respectively, are said to be dependent on the relative (comparing more than one stimulus) and the absolute judgments (identification of stimulus without the opportunity to compare) of people. It has also been shown that humans, in general, have the ability to make better relative judgments than absolute judgments [47,78]. This being the case, the orthogonality or independence of the coding schemes determines how unique the information provided by a code is, and results in an increase in the number of stimuli that can be identifed on an absolute basis [46]; for example, if size (large and small) and color (black and white) were orthogonal dimensions of coding, then each of the possible codes namely, large-black, large-white, small-black, and small-white, would provide unique information. A number of different guidelines also exist in the literature that can help good coding practices. Table 15 summarizes general guidelines for designing a good visual coding system. In addition, different visual coding methods have their own specific design features that can be exploited by the designer for specific tasks and work situations. Using alphanumeric characters (which are $0, 1, 2, 3, \ldots, 9$ and $a, b, c, \ldots, z$ in the English language), singly and in combination, for instance, has been found to be useful for identification purposes, and for situations with space constraints. Color coding of surfaces (24 or more combinations of hues, saturation, and brightness are possible, though research recommends use of no more than nine combinations) are useful for industrial tasks requiring searching and counting. Color-coding surfaces can, however, be ineffective if the worker population is color deficient [79,80]. Color coding any lights used in the workplace has been shown to be effective for qualitative reading [81]. The recommendation is to limit the number of lights coded to three. Coding using geomerical shapes (there are a total of 15 or more geometrical shapes), has been found to be useful in situations using symbolic representation of an action or an event. The literature recommends the use of no more than five geometrical shapes, as using more than five will lead to difficulty in discrimination of the different shapes [81]. While a total of 24 different angles of inclination (of characters) are available if coding is to be done by using angles of inclination, the recommended limit is 12 [82]. Using this form of coding has been found to be useful for indicating direction, angle, or position on round instruments. Other commonly used forms of visual coding include differing brightness of lights (recommended limit is two levels) [81], and differing flash rates of lights (recommended limit is two levels).

Table 15 General Recommendations for Designing a Good Coding System

Make codes *detectable* by the human sensory mechanisms under the given environmental conditions,
Make codes *discriminable* from each other by providing for a difference threshold or a just-noticeable difference.
Make codes *meaningful* to the user by providing for conceptual compatibility.
Where possible, *standardize* codes from situation to situation.
Use *multidimensional* codes to increase the number and discriminability of coding stimuli used.

Adapted from Ref. 46.

Using symbols for coding information is another important means of representing visual information. The effectiveness of symbolic coding depends on how strongly the symbol is associated with the concept or objects it is intended to represent. The strength of this association has been shown to depend on any existing and established association [83], and on the ease of leaming any new associations. The normal procedure in setting up symbolic coding systems in the workplace should involve considerable experimentation with existing and any new proposed symbolic codes. The experimentation should involve the worker population for which the symbolic coding system is intended, and coding system should be evaluated on the basis of the ease of recognition, on matching symbols with what they represent (based on reaction time of participants), and based on the preferences and opinions of the users. Figure 12 provides examples of good and bad symbol designs. The symbol labeled "bad design" in the figure has too much detail and is not simple in design. The symbol labeled "good design" in the figure has all the

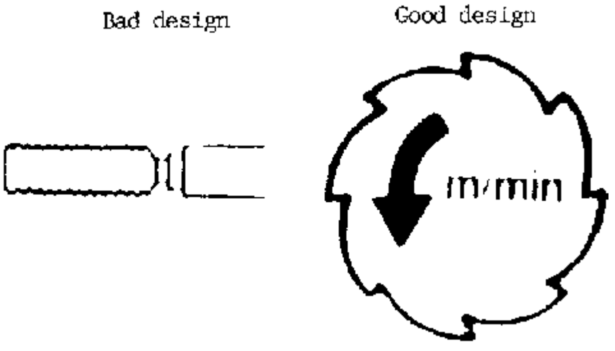
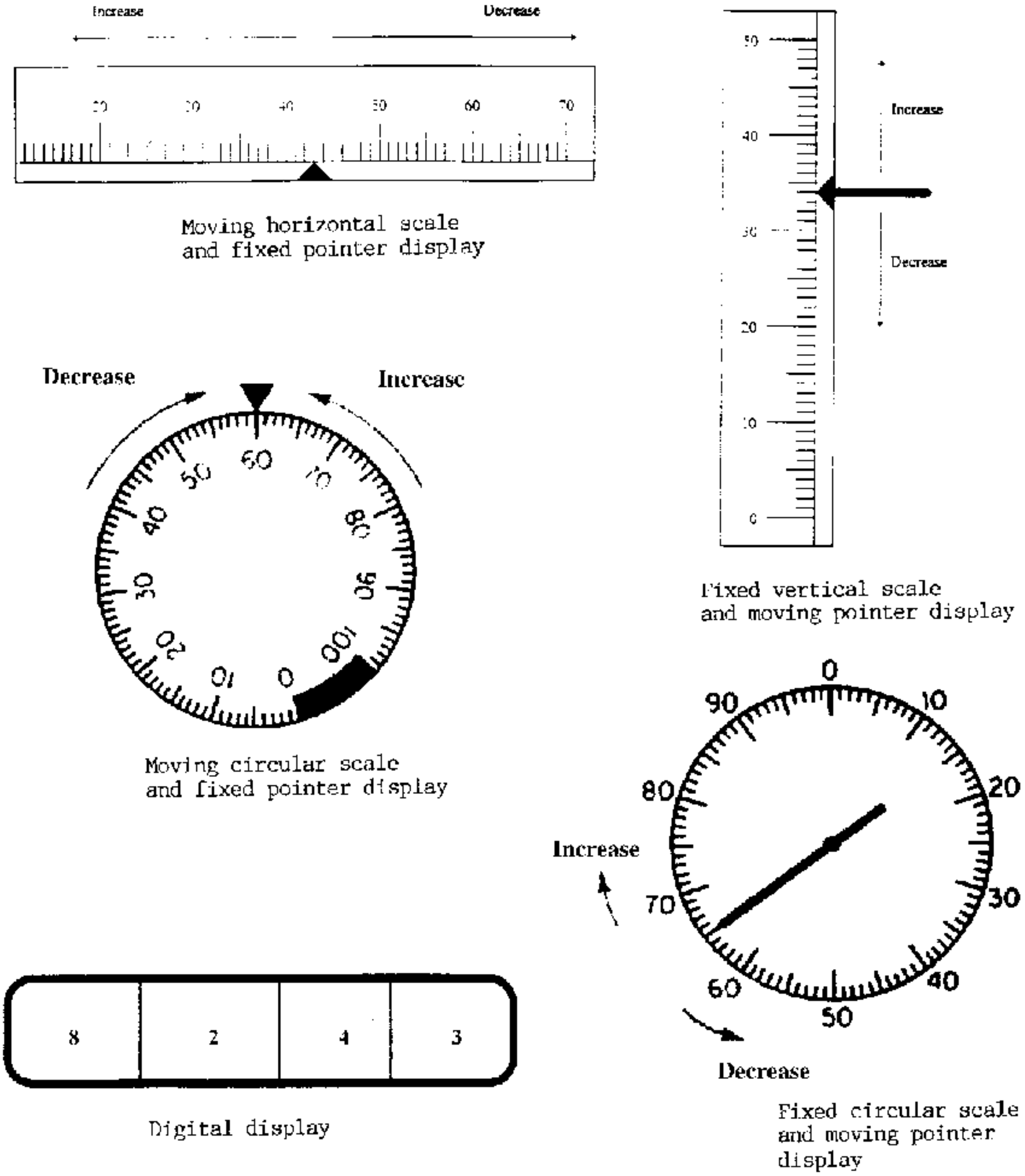


Figure 12 Examples of good and bad symbol design.

details identifying the symbol within the boundary of the symbol, and not outside the symbol.

Dynamic information displays. These displays are used to present information about variables that are subject to change in time. Depending on the type of information presented by the display, dynamic displays can provide quantitative information, qualitative information, check readings, and information on situation awareness measures.

Quantitative visual displays provide information about the quantitative value of a variable of interest. The conventional types of displays used to convey quantitative information include analog displays (fixed scale and moving pointer, moving scale and fixed pointer) and digital displays (mechanical type counters). Figure 13 provides some examples of the three conventional types of displays. Research in analog displays has provided certain general guidelines for



Moving horizontal scale and fixed pointer display

Fixed vertical scale and moving pointer display

Moving circular scale and fixed pointer display

Digital display

Fixed circular scale and moving pointer display

Figure 13 Commonly used quantitative displays.

designing such displays [84]. Fixed scale and moving pointer displays are preferable to moving scale and fixed pointer displays in most cases. This is more so especially when manual control is used to control the moving element in the display (since it is better to control the pointer rather than the scale). Also, any small variations are better apparent when using a moving pointer, fixed scale device. However, when the range of numerical values is too large to be accommodated within the scale, the recommendation is to use a fixed pointer, moving scale display with rectangular open windows in the scale for easier reference. In general, it has been determined that digital displays perform better than analog displays where precise numerical values are needed, and when the presented numerical values are not continuously changing.

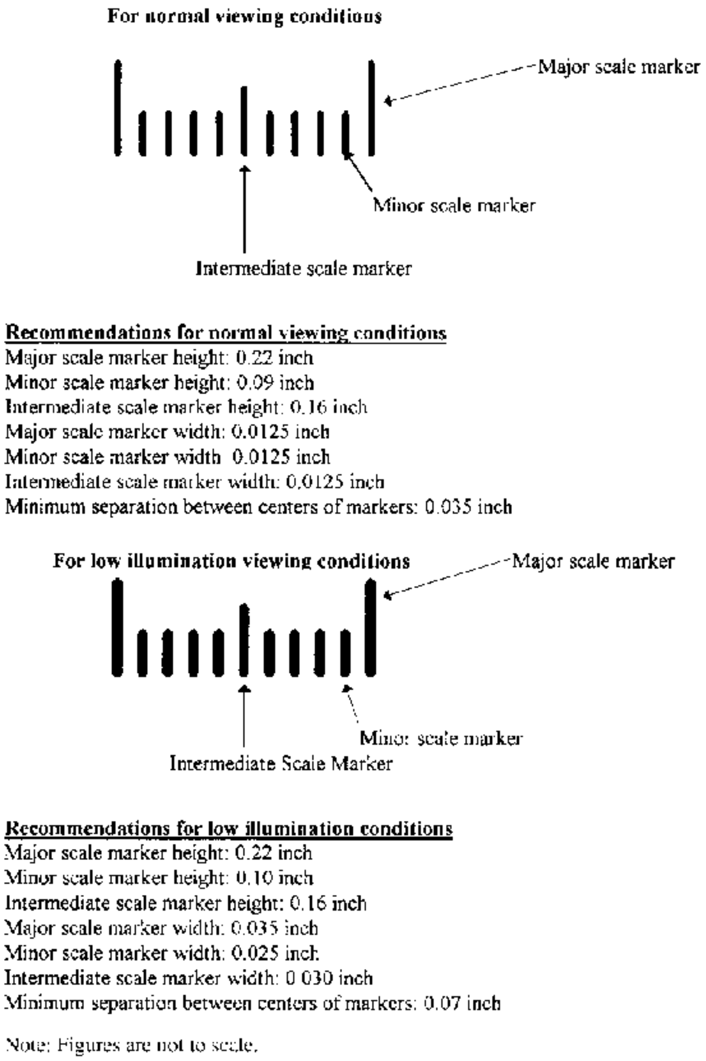
In addition to these guidelines for the design of quantitative displays, research has identified numerous characteristics that contribute towards making design of quantitative displays effective and efficient. Some of these characteristics include the design of the scale range (difference between the largest and the smallest scale values), the design of the numbered interval in the scale (numerical difference between adj acent scale numbers), the design of the graduation interval (the difference between the smallest scale points), the design of the scale unit (smallest unit to which the scale can be read), the numerical progressions used in scales, the design of scale markers, and the design of scale pointers [46]. The numerical progression by 1's $(0, 1, 2, 3, \ldots)$ has been found to be the easiest to use. Decimals in scales, and scales with unusual numerical progressions such as by 6's and 7's are discouraged. The most common recommendation for the length of the scale unit is to use values ranging from 0.05 to 0.07 in. The key factor in deciding the length of the scale unit is that the values should be as distinct as possible to permit easy human reading. Recommendations [81] are also available for design of scale markers (see Fig. 14 for a summary of these recommendations). Some common recommendations for design of pointers include having a pointed (about 20° tip angle) pointers, and having the tip of the pointer meet the smallest of the scale markers in the scale. Also, to avoid parallax between the scale and the pointer, it is recommended to have the pointer as close as possible to the surface of the scale [46].

Qualitative visual displays are used to present information on a changing variable based on quantitative information about a variable. The information presented could be indicative of a trend in the variable, or a rate of change of the variable. Also, qualitative displays can be used to determine the status of a variable in terms of predetermined ranges (whether the fuel tank is empty, full, or half-full), or for maintaining a desirable range of values of a variable (such as speed). The most common forms of presenting qualitative information through displays is by color coding or by using shapes (or areas to represent variables of specific interest, such as "danger") to code the information. Figure 15 provides an example of a color- and area-coded qualitative display.

Research [85] on check-reading displays (used to determine if a particular reading is normal or not) has provided the following conclusions about the design of such displays:

1. In general, males make fewer errors in check-reading tasks than females.
2. The accuracy of check reading is a function of viewing time; fewer errors will be made if the exposure time is relatively long (greater than 0.5 sec); also, check-reading performance differences between males and females become insignificant when exposure time is increased to 0.75 sec.
3. The selection of background color is important for check-reading tasks; for exposure time less than 0.75 sec, black dials and pointers with a white background lead to fewer errors in check reading than with white dials and pointers on a black background; however, for exposure times greater than 0.75 sec, fewer errors result with a black dial background; the final selection of the background color should be based on the time routinely available for check reading (Fig. 16).
4. Both the 9 o'clock and the 12 o'clock pointer positions in the dial yield acceptable performances; the actual design has then to be based on user preferences.
5. Check-reading performance is not affected by the presence of between 1% and 3% deviant dials.
6. The normal reading must be coded clearly; if many instruments are used in concert, the displays must be configured clearly so that the deviant reading stands out. Figure 16 provides examples of good and bad check reading displays.

One other type of qualitative display is status indicator displays. These indicators are usually representative of discrete pieces of information such as whether the condition is normal or dangerous, or


For normal viewing conditions
Major scale marker
Minor scale marker
Intermediate scale marker

Recommendations for normal viewing conditions
Major scale marker height: 0.22 inch
Minor scale marker height: 0.09 inch
Intermediate scale marker height: 0.16 inch
Major scale marker width: 0.0125 inch
Minor scale marker width 0.0125 inch
Intermediate scale marker width: 0.0125 inch
Minimum separation between centers of markers: 0.035 inch

For low illumination viewing conditions
Major scale marker
Minor scale marker
Intermediate Scale Marker

Recommendations for low illumination conditions
Major scale marker height: 0.22 inch
Minor scale marker height: 0.10 inch
Intermediate scale marker height: 0.16 inch
Major scale marker width: 0.035 inch
Minor scale marker width: 0.025 inch
Intermediate scale marker width: 0.030 inch
Minimum separation between centers of markers: 0.07 inch

Note: Figures are not to scale.

Figure 14 Recommendations on scale marker design for normal and low illumination viewing conditions.

if the working surface is hot or cold. Colored lights are the most commonly used form of status indicators.

Signal and warning lights are the types of dynamic information displays that have relevance in the context of industrial settings. Detectability (of such lights) is the most important design issue related to signal and warning lights. The detectability of signal and warning lights is influenced by factors such as the size of the light, the luminance, the exposure time, the color of lights, and the flash rate of lights. Table 16 provides widely accepted guidelines for the design of signal and warning lights.

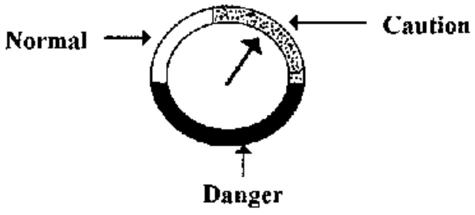


Figure 15 Example of area coded display.

Auditory displays: factors affecting design. Literature identifies four different types of tasks involved in detection of auditory signals [46]: *detection* (determining whether or not a signal is present), *relative discrimination* (differentiating between two or more close signals), *absolute identification* (identifying a particular signal when only one signal is present),

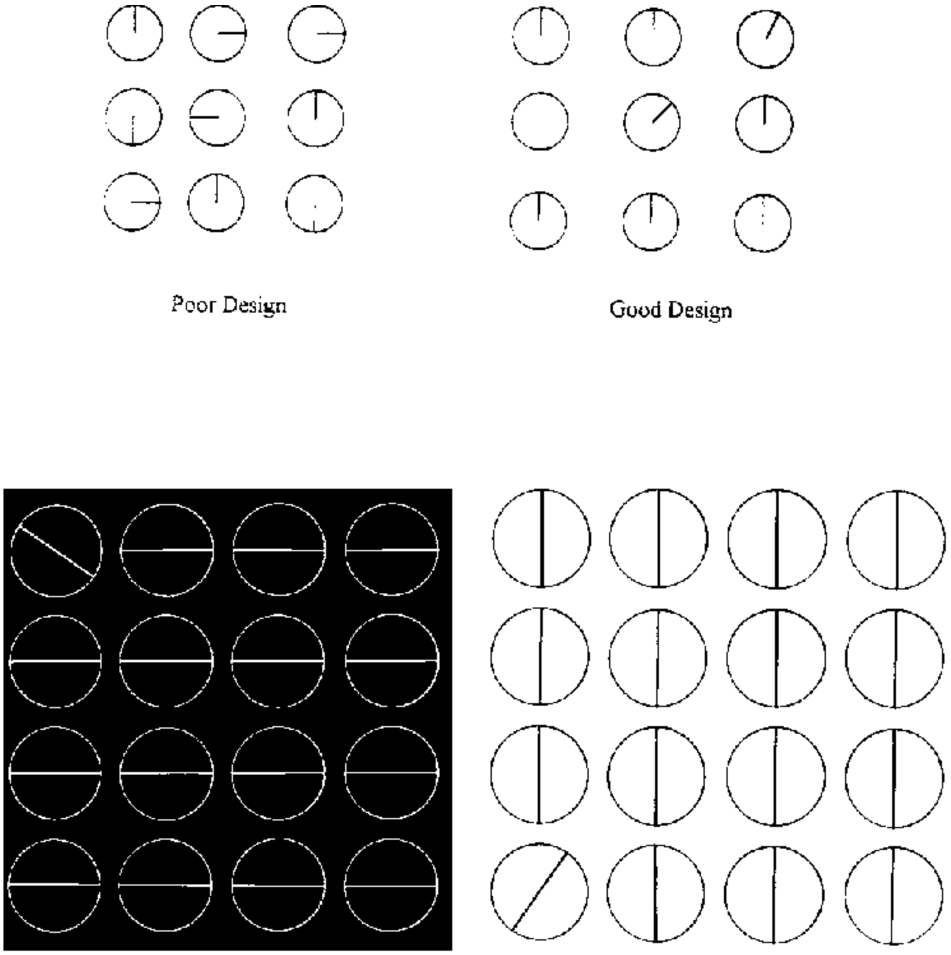


Figure 16 Check reading displays.

and *localization* (determining the direction of the signal source). These functions are based upon fundamental attributes in sound energy propagation, namely, the frequency of sound, and the intensity of sound. The number of cycles of sound waves produced in one second is called frequency. Frequency of sound is expressed in hertz (Hz). It is generally true that the human ear can detect frequencies ranging from 20 to 20,000 Hz. A related concept is the pitch of the sound (pitch denotes the highness or lowness of a sound; high frequencies result in high pitched tones, and low frequencies result in low-pitched tones). The intensity of sound is defined as the sound power in one square meter of area (W/m^2). Since it is difficult to measure sound power level directly, the intensity of sound is measured in terms of the sound pressure level. Sound pressure level, in decibels, is given by $20 \log(P_1/P_0)$, where P_1 is the sound power level corresponding to the sound to be measured, and P_0 is the sound power level corresponding to 0 dB. The sound pressure levels can be directly measured using commercially available sound level meters.

The *detectability* of auditory signals depends upon the environmental influences (noise) present in the signal. In the presence of noise in the surroundings, the threshold of detectability of the signal is increased, i.e., the signal intensity must exceed this threshold if it is to be detected. A rule of thumb pertaining to auditory signal detection in the presence of noise or masking states that the signal intensity (at the outer ear) should be midway between the masked threshold of the signal in the presence of noise and 110 dB. In quiet surroundings, the detectability of the signal depends upon the frequency and the duration of the

Table 16 Recommendations for Design of Signal and Warning Lights

Use signal and warning lights to warn of an actual or potential danger.
Use only one light in normal circumstances; if several lights are used, have a master warning light to indicate specific danger.
For commonly encountered danger or warning situations, do not use a flashing light; use only a steady light. For situations that are new or occasional, use flashing warning lights.
Use four flashes per second when using flashing warning lights. When using different flashing rates to indicate different levels of some variable, do not use more than three such rates with one light.
Have the signal or warning light at least twice as bright as the background.
Use red color for these lights and differentiate danger lights from other signal lights in the immediate environment.
Ensure that the warning lights subtend at least a visual angle of 1°.

Adapted from Ref. 84.

signal. The standard recommendation is that these signals should be at least 500 ms in duration; if they are shorter than this, the recommendation is to increase the intensity of the signal. Different recommendations have been made by researchers to improve the detectability of auditory signals. A summary of these recommendations is provided in Table 17.

The *relative discriminability* of auditory signals also depends upon the intensity and the frequency of sound, and the interaction between these two factors. Relative discriminability is usually measured in terms of the *just-noticeable difference*, which is the smallest change in the intensity or frequency that can be noticed by humans 50% of the time. The smaller the just-noticeable difference, the easier it is to detect the differences in either intensity or frequency of sound. Research has shown that it is easier to detect the smallest differences when the intensity of sound is higher (at least 60 dB above the threshold level). Also, with respect to frequency, it is recommended that signals use lower frequencies for higher discriminability. However, since ambient noise is also a low-frequency sound, it is advisable to use signals in the 500–1000 Hz range. Also, it is good to keep signals 30 dB or more above the threshold level for efficient frequency discrimination.

Table 17 Recommendations for Increasing the Detectability of Auditory Signals

Reduce the intensity of noise near the frequency of the signal of interest.
Increase the intensity of the signal.
Present the signal for at least 0.5–1 sec.
Determine the frequency where noise is low, and change the signal frequency to correspond this frequency.
Present noise to both ears and the signal to one ear only.
Introduce a phase shift in the signal and present the unshifted signal to one ear and the shifted signal to the other.

Adapted from Ref. 86.

It has also been determined that, on an absolute basis (identification of an individual stimulus presented by itself), it is possible for the human ear to identify four to five levels of intensity, four to seven levels of frequency, two to three levels of duration, and about nine levels of intensity and frequency combined.

Sound localization is the ability to determine and localize the direction of the sound. The differences in the intensity of sounds, and the differences in the phase of sounds are the primary measures by which the human auditory system determines the direction of the sound source. It has been shown that for frequencies below 1500 Hz, if the source of the auditory signal is directly to one side of the head, the signal reaches the nearer ear approximately 0.8 msec before it reaches the other ear. Also, localization is difficult at low frequencies, since there is very little difference in the time it takes for the signal to reach both ears simultaneously. However, at high frequencies (generally above 3000 Hz), the presence of the head between the ears makes intensity differences more pronounced resulting in effective localization of the sound source.

Design recommendations for auditory displays. A summary of recommendations for the design of auditory displays is provided in Table 18. This is in addition to the recommendations in the table on when to use auditory displays, as opposed to visual displays.

1.3.3.4 Controls

General Considerations in Control Design. Controls are the primary means of transmitting the controlling action to devices and systems. Numerous factors affect the design of control devices. These factors include the ease of identification, the size of the control, control–response ratio, resistance of the control, lag, backlash, deadspace, and location. In the following paragraphs,

Table 18 Checklist for Designing Auditory Displays

Design elements	Questions to ask
Compatibility	Are the signal dimensions and the coded displays compatible with user excectations?
Approximation	If the information presented is complex, are the signals attention-getting, and providing precise information as well?
Dissociability	Are the auditory signals of interest clearly discernible from other signals?
Parsimony	Do the signals provide the correct amount of information?
Invariance	Is a particular signal used for providing the same information every time?
Presentation	Are the signals moderate and not extreme?
	Is the signal intensity level such that it is not masked by noise?
	Has care been taken not to overload the auditory system of operator by presenting too many signals at the same time?
Installation	Has the signal been tested with the target user group?
	Are the new signals really new (are they noncontradictory to the existing signals?
	If auditory displays are entirely new to the setting, have the operators been given enough time to adjust to the new type of display?

Note: Answering "Yes" to all the above questions in the checklist is the desirable scenario.
Adapted from Refs. 46, 87, and 88.

we will summarize and present recommendations from research for each of these factors.

The *ease of identification* of controls depends upon how well the controls have been coded. The effcacy of the coding used can be determined using measures mentioned in an earlier section, namely, using detectability, discriminability, compatibility, meaningfulness, and the extent of standardization. Controls can be coded using shape, texture, size, location, operational methods, color, and labels. Shape coding uses tactual sensitivity of the human for discriminating between the different shapes of the controls. Figure 17 provides examples of different shapes that are commonly used in controls. Three different types of textures have been identified as being suitable for coding control devices: smooth surface, fluted surface, and knurled surface.

The most important consideration when coding by size is to provide adequate discriminability between the different sizes used. For coding based on location of controls, the recommendation is to use at least 2.5 in. between adjacent vertical controls, and at least 4 in. between adjacent horizontal controls. In addition, it is recommended that the general guidelines provided in table be followed when coding controls based on location. There are instances when coding is based on the method of operation of the control (push-button controls, for example). Table 19 provides the recommended minimum separation distances when this is the case. Such operational coding, is undesirable, however, when operation time or potential operator errors are considerations. Another way to code controls is by color. Meaningful colors (such as red for a danger button), combined with other coding dimensions such as shape and size, have been shown to be effective in enhancing the discriminability of the controls. Color coding, however, cannot effective in situations with poor illumination or in dirty environments. One of the most commonly used methods of coding controls is by labels. In fact, labels are considered a minimum requirement in many situations as they do not place extensive learning demands on the operators. Labels, however, have the disadvantage in that they take time to read and are not useful as a coding method in situations that have a high operation speed. Also, the placement of the label on the control has been shown to pose accessibility problems to the reader. Control devices can have unique combinations of codes, or even redundant codes. Considerations such as the illumination and the potential visual handicaps of the operator, maintenance of mechanical controls, and the speed and the accuracy with which the controls have to be operated, are other factors to consider in designing controls for ease of identification.

Control–response ratio (denoted by C/R) is the ratio of the movement of the control device to the movement of the system response. By this definition, a sensitive control will have a low C/R ratio (i.e., the response will be large even for a slight change in the control). It is believed that human motor actions take place at two levels—at a gross-adjustment level, and at a fine-adjustment level. Hence the optimal level of C/R ratio to use in a control device, is generally decided as a

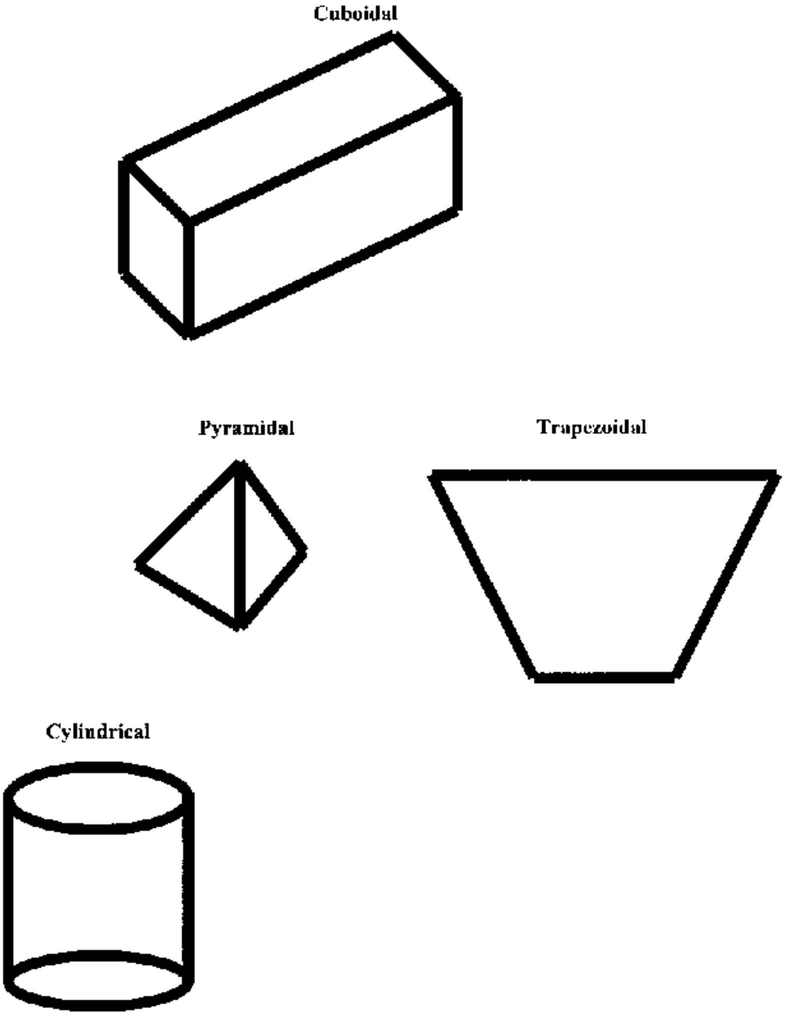


Figure 17 Different shapes that have been commonly used and demonstrated to be effective for coding controls.

tradeoff between the time it takes to accomplish the gross movement and the time it takes for the fine adjustment involved in a controlling action. It has been shown that an optimum C/R ratio is dependent upon factors including the type of control (lever, crank, wheel, etc.), the size of the display, and the tolerance permitted in setting the control.

Resistance in a control is responsible for providing feedback about the controlling action to the operator. In essence, the resistance offered by the control is made up of two fundamental elements: the force applied to the control, and the distance to which this force is applied (or the distance to which the control moves). Free-position or isotonic controls offer no resistance to movement, and feedback to the operator is based on the displacement that occurs. Isometric controls, on the other hand, provide feedback, based only on the force or the pressure applied to the control. Most controls use a combination of both pure displacement and pure force mechanisms for providing operator feedback. Control resistance can significantly affect operator performance by affecting the speed and precision of control operations, by changing the *feel* in the control, by changing the smoothness of the control movement, and by subjecting the control to the effect of shock and vibration. It is therefore vital to consider control resistance when designing or selecting controls for a specific task. Some design guidelines regarding control resistance are provided in Table 20.

Deadspace is defined as the amount of movement near the null position of the control. The amount of deadspace in a control device has been shown to affect

Table 19 Recommended Minimum and Maximum Separation for Different Control Devices

		Recommended separation (in inches)	
Control	Use	Minimum	Desired
Push button	Randomly with one finger	0.5	2
	Sequentially with one finger	0.25	1
	Randomly or Sequentially with different fingers	0.5	0.5
Toggle switch	Randomly with one finger	0.75	2
	Sequentially with one finger	0.5	1
	Randomly or Sequentially with different fingers	0.625	0.75
Crank and lever	Randomly with one hand	2	4
	Simultaneously with two hands	3	5
Knob	Randomly with one hand	1	2
	Simultaneously with two hands	3	5
Pedal	Randomly with one foot	4 (between the inner sides of the pedal)	6 (between the inner sides of the pedal)
	Randomly with one foot	8 (between the outer sides of the pedal)	10 (between the outer sides of the pedal)
	Sequentially with one foot	2 (between the inner sides of the pedal)	4 (between the inner sides of the pedal)
	Sequentially with one foot	6 (between the outer sides of the pedal)	8 (between the outer sides of the pedal)

Adapted from Ref. 89.

Table 20 Recommendations on Control Resistance and Control Operation

Control movements should be as short as possible

Positive indication of control activation must be provided to the operator.

Feedback on system response to control activation must be provided to the operator.

Control surfaces should be designed to prevent slippage when activating.

Arm or foot support should be provided to the operator if precise, sustained positioning of the controls is required.

Controls must be provided with enough resistance to avoid accidental activation due to the weight of hands or feet.

If a seated operator has to push a force more than 5 lbf on a one-hand control, a backrest must be provided to the operator.

The operator has to be able to move the trunk and entire body if both hands are required to exert more than 30 lbf through more than 15 in. in the fore-and-aft plan.

The speed, force, and accuracy of controls should fit most people, not just the most capable.

Adapted from Ref. 90.

the sensitivity, and hence the performance, of the control system. It has been shown by researchers that deadspace in a control device can be compensated to a certain extent by making the control-ratio relationships less sensitive.

Backlash in a control device is defined as the deadspace at any control position. Research on backlash shows that systems with high control gain need to have minimum backlash to reduce system errors. If the control system design makes it impossible to reduce the backlash, the recommendation is to make the control gain as low as possible, since humans have been shown to cope badly with backlash errors.

Types of Control Devices. Controls can be classified as being discrete or continuous controls based on whether they transmit discrete (on and off) or continuous (machine speed increase from 0 to 100 km/hr) information. Controls are also classified based on the amount of force required to operate them (small or large). The most common types of control devices used to transmit discrete information and requiring a small force to operate include push buttons, keyboards, toggle switches, rotary selector switches, and detent thumb wheels. Common control devices used to transmit discrete information and requiring a large amount of force include detent levers, large hand push buttons, and foot push buttons. For transmitting continuous information, the traditional control devices such as rotary knobs, multirotational knobs, thumb wheels, levers or joysticks, and small cranks, require only a small amount force to operate them. On the other hand, other traditional control devices used to impart continuous information, such as handwheels, foot pedals, large levers, and large cranks, need large amounts of force to manipulate and operate. In general, control selection for common controls, such as toggle switches, rocker switches, knobs, cranks, handwheels, etc., is based on operational factors such as speed, accuracy, space requirements, and ease of operation. With the advent of information technology, control devices such as joysticks, trackballs, mice, touch tablets, light pens, touch screens, etc., are becoming popular devices for transmitting continuous information to the system. Technology has advanced to such an extent that these modern devices demand only a small amount of physical force from the human operator. Given the variety of both traditional and modern control devices in use in industry (see Fig. 18 for examples of some of these control devices), it is beyond the scope of this chapter to explain the design of each of these devices in detail. Besides, many excellent design tables and recommendations already exist in the literature for design and selection of control devices, and are widely available. The interested reader is referred to these design guidelines. Such guidelines can be found in Sanders and McCormick [46], Woodson et al. [11], Chapanis and Kinkade [91] Salvendy [92], Eastman Kodak [90], etc.

1.3.3.5 Other Design Considerations in Information Presentation and Control

Besides the individual design factors affecting the design and operation of displays and controls, there are other general considerations in display and control design that affect the overall effectiveness of the information presentation and control system as a whole. We have chosen to present two such important factors. They are compatibility, and grouping and location of controls.

Compatibility. This the relationship between the expectations of the human and the input stimuli and responses of the system with which the human is interacting. Any system with human users should be compatible with the human expectations. In general, good compatibility will result in fewer user errors, and better

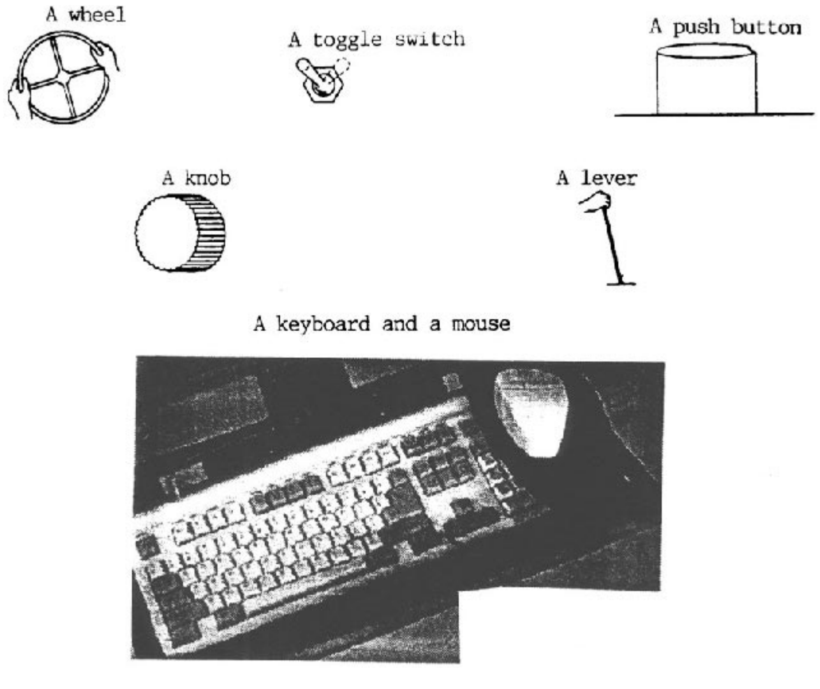


Figure 18 Examples of common control devices.

human and overall system performance. Literature identifies four types of compatibility [47] conceptual, movement, spatial and modality compatibilities.

Conceptual compatibility refers to the matching that should exist between certain forms of stimuli such as symbols, and the conceptual associations humans make with such stimuli. *Movement compatibility* (also commonly referred to as population stereotypes) denotes the relationship between the movement of the displays and controls and the output response of the system being controlled. Numerous types of movement compatibilities have been studied by researchers. The most important types of movement compatibilities include the movement of a control to follow the movement of a display, the movement of a control to control the movement of a display, the movement of a control to produce a specific system response, and the movement of a display without any related response. The common principles of movement compatibility for various types of displays and control devices are presented in Table 21.

Spatial compatibility refers to the relationship that should exist between, the physical features, and arrangement, of the controls and their associated displays. A good example of compatibility in physical features between the displays and the controls is the design of the function keys on a keyboard, and the corresponding labels for these function keys. In a number of experiments with household stove tops, human factors researchers have demonstrated conclusively the need for physically arranging displays and the associated controls in a corresponding and compatible way.

Modality compatibility is a fairly new addition to the list, and refers to certain stimulus-response combinations being more compatible with some tasks than with others.

Principles of Control-Display Arrangement in a Workspace. The physical location and arrangement of the displays and controls in a given workspace also has to be based on the human sensory capabilities, and the anthropometric, biomechanical, and other characteristics of the human user. Table 22 provides general guidelines for locating controls in a workspace. The ideal goal of placing each and every display and control at an optimal location and in an optimal arrangement with respect to the human user, is difficult, if not impossible, to achieve in practice. A few general principles of control-display location and arrangement are useful in setting priorities and in determining tradeoffs for *good* design, if not the optimal.

According to the *importance principle*, components that are vital to system goals should be placed in convenient locations. System experts determine what these vital goals are. According to the *frequency-of-use principle*, components that are frequently used should be placed in convenient locations. According to the *functional principle,* components that are functionally related in the operation of the overall system should be grouped and placed together. Figures 19a (before redesign) and 19b (after redesign) illustrate the use of the principle of functional grouping in the redesign of the machining controller of a Dynamite DM2400 bench-top programmable machining center. According to the *sequence-of-use principle*, components should be arranged in the sequence in which they find frequent use in the operation of the system or in the performance of a task. Use of one or a combination of these principles requires that the system designer collect information about the human users involved (the user characterization step described in Sec. 1.3.1 as the first step in the process of solving human–machine interaction problems), the tasks involved (the task characterization step using task analysis techniques also described in Sec. 1.3.1 as the second step in the process), and the environment in which the user has to perform the task (characterization of the situation, again mentioned in Sec. 1.3.1 as the third step in the process). Based on extensive research, the recommendations that have been suggested for designing workspaces with various forms of displays and controls are presented in Table 23.

1.4 SUMMARY

This chapter presented the overall ‘‘process’’ of designing and evaluating systems involving humans and automated devices. The key elements involved in this process were briefly described, and the *essentials* of these elements were presented in the form of guidelines and recommendations for practice.

ACKNOWLEDGMENTS

We thank Mr. Sampath Damodarasamy, doctoral student in industrial engineering at the University of Cincinnati, for help with the evaluation of the DYNAMYTE 2400 machine controller example, and the two figures he generated for this example.

Table 21 Common Principles and Recommendations for Movement Compatibility for Different Displays and Controls

Type of display–control relationship	Principles of movement compatibility
Rotary displays and rotary controls in same plane	For fixed scale/rotary pointers, ensure that clockwise turn of the pointer is associated with clockwise tum of the control. For fixed scale/rotary pointers, clockwise rotation of pointer/display should indicate increase in value and vice versa. For moving scale/fixed pointer, ensure scale rotates in the same direction as control knob. Ensure scale numbers increase from left to right. Ensure clockwise turn of control increases value.
Linear displays and rotary controls in same plane	When the control is located to the side of the display, the common expectation is the display pointer will move in the same direction of that side of the control which is nearest to it. The common expectation is pointer will move in the same direction as the side of the control knob on the same side as the scale markings on the display. The common expectation is a clockwise turn of a rotary control will increase the value on the display no matter where the control is located relative to the display.
Movement of displays and controls in differnt planes	For rotary controls, the common expectation is a clockwise rotation results in an increase in value. For rotary controls, the common expectation is a clockwise rotation results in movement away from individual and vice versa. For stick-type controls (both horizontally mounted on vertical plane and vertically mounted on horizontal plane), the common expectation is an upward movement of control results in an increase in value and an upward movement of display.
Movement of power switches	U.S. system is switch-up is for on, and switch-down is for off. British system is switch-up is for off switch-down is for on.
Directional compatibility of operator movement (when operator is not directly facing the control)	The common expectation is that a control movement in a certain direction produces a parallel movement of the indicator on the display, irrespective of the position of the operator. The direction of movement of the display indicator when the indicator is in the visual field of the subject, is the same as the direction of movement of the controlling limb. The direction of movement of the display indicator when the indicator is in the visual field of the subject, is the same as the direction of movement of the control relative to the subject's trunk.

Adapted from Refs. 41, 89, 91, and 93–97

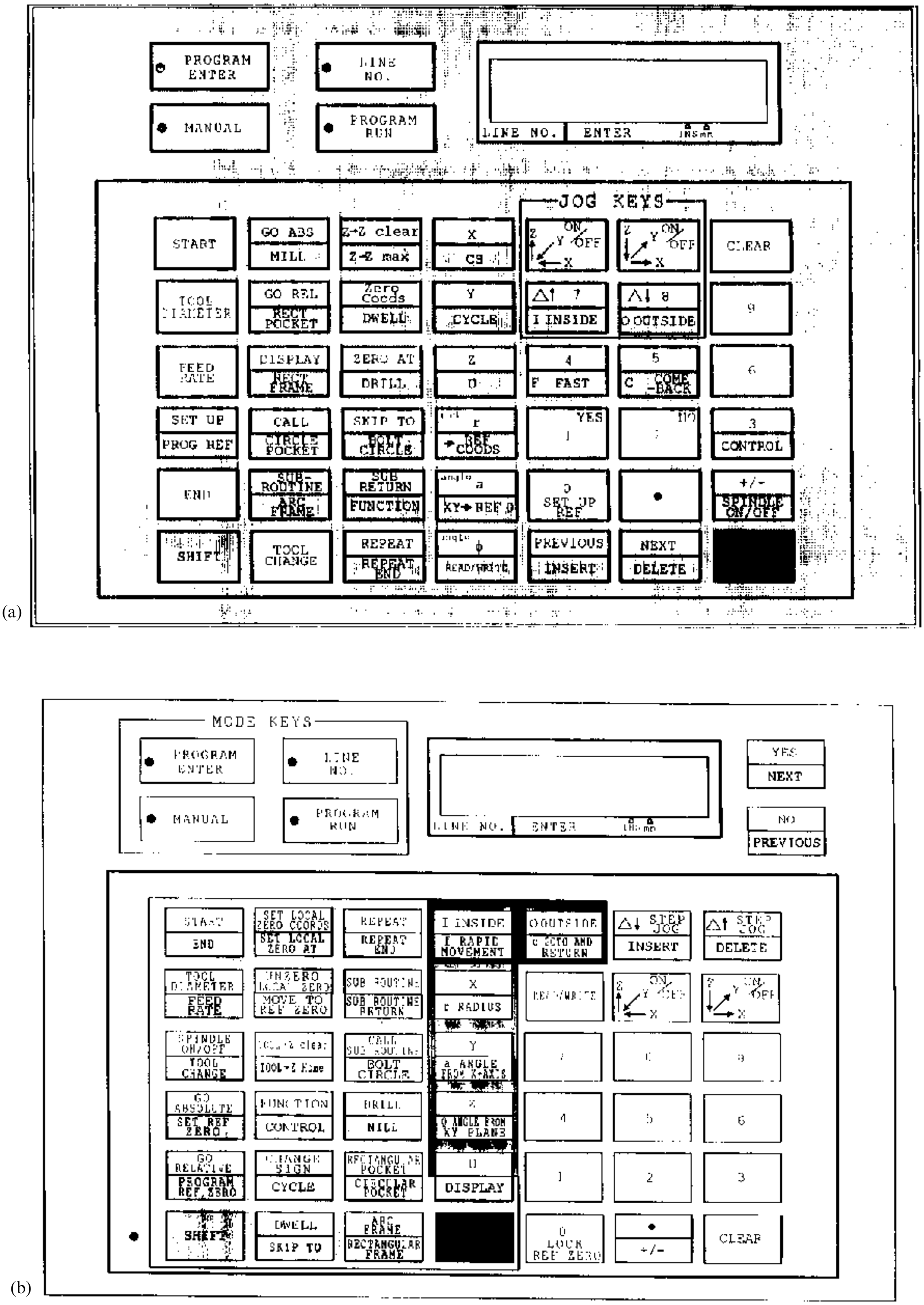


Figure 19 (a) The machining controller before redesign. (b) Functional grouping in a machining controller—After redesign.

Table 22 Guidelines for Location of Controls

Keep the number of controls minimum.
Ensure easy and simple activation of controls except when designing to avoid accidental activation.
Arrange controls to allow for operator posture adjustments.
For one-hand or one-foot operation of several controls in sequence, arrange controls to allow continuous movement through an arc.
Controls requiring high-speed or high-precision operations should be assigned to the hands.
If there is only one major control device, place it in front of the operator midway between hands.
If a control requires a precise movement, place it on the right as a majority of the population are right handed.
Controls requiring high forces for operation should be assigned to the feet.
Emergency controls and displays must be distinguished from controls and displays used in normal working situations.
Emergency controls should be placed within 30 degrees of the operator's normal line of sight.
Keep the same relative groupings for major controls and displays; if not make the exception conspicuous.
To prevent accidental activation of a control, place it away from frequently used controls, or make it conspicuous.

Adapted from Ref. 90.

Table 23 List of Priorities When Making Tradeoffs for Optimum Design of Workspace with Displays and Controls

Priority	Elements in design
First	Primary visual tasks
Second	Primary controls that interact with primary visual tasks
Third	Control–display relationships
Fourth	Arrangement of elements to be used in sequence
Fifth	Convenient location of elements to be used in sequence
Sixth	Consistency with other layouts within the system or in other systems

Adapted from Ref. 98.

REFERENCES

1. HJ Bullinger, HJ Warnecke, Factory of the Future. Heidelberg, Germany: Springer-Verlag, 1985.
2. A Mital, What role for humans in computer integrated manufacturing? Int J Computer Integ Manuf 10: 190–198, 1997.
3. A Mital, A Pennathur, Musculoskeletal overexertion injuries in the United State: an industrial profile. Int J Ind Ergon: 25: 109–129, 1999.
4. TB Sheridan, Supervisory control. In: G Salvendy, ed. Handbook of Human Factors, New York: John Wiley & Sons, 1987.
5. PA Booth, An Introduction to Human–Computer Interaction. Hove and London: Lawrence Erlbaum Associates, 1989.
6. B Kirwan, LK Ainsworth, eds. A Guide to task Analysis. London: Taylor & Francis, 1992.
7. A Mital, A Morotwala, M Kulkarni, M Sinclair, C Siemieniuch, Allocation of functions to humans and machines in a manufacturing environment: Part II—The scientific basis (knowledge base) for the guide. Int J Ind Ergon 14: 33–49, 1994.
8. D Meister, Human Factors Testing and Evaluation. New York: Elsevier Science Publishers, 1986.
9. A Chapanis, Research Techniques in Human Engineering. Baltimore: John Hopkins, 1976.
10. WE Woodson, Human Factors Design Handbook. New York: McGraw-Hill, 1981.
11. WE Woodson, B Tillman, P Tillman, Human Factors Design Handbook: Information and Guidelines for the Design of Sciences, Facilities, Equipment, and Products for Human Use. New York: McGraw-Hill, 1991.
12. Department of Defense, Human Engineering Criteria for Military Systems, Equipments and Facilities. MIL-STD 1472C, Washington, DC: U.S. Department of Labor, 1981.
13. J Seminara, Human Factors Methods for Assessing and Enhancing Power Plant Maintainability. Report EPRINP-2360. Electric Power Research Institute, Palo Alto, CA. 1949.
14. Nuclear Regulatory Commission, Guidelines for Control Room Design Reviews. Report no. NUREG-0700, U.S. Nuclear Regulatory Commission, Washington, DC. 1981.
15. HSE, Human Factors in Industrial Safety. HS(G). London: HMSO, 1989.
16. A Cakir, DJ Hart, TFM Stewart, Visual Units. Chichester: John Wiley, 1980.
17. HS Blackman, DI Gertman, WE Gilmore. CRT Display Evaluation: The Checklist Evaluation of CRT-Generated Displays, Report No. NUREG/CR-3557, Washington DC: Nuclear Regulatory Commission, 1983.
18. P Fitts, Huyman Engineering for An Effective Air-Navigation and Traffic-Control System. Washington, DC: National Research Council, 1951.
19. A Mital, A Motorwala, M Kulkarni, M Singlair, C Siemieniuch, Allocation of functions to humans and machines in a manufacturing environment: Part I–Guidelines for the practitioner. Int J Ind Ergon 14: 3–31, 1994.
20. RP Paul, S Nof, Work methods measurements—a comparison between robot and human task performance. Int J Prod Res, 17: 277–303, 1979.

21. A Mital, A Mahajan, ML Brown, A Comparison of manual and automated assembly methods, In: Proceedings of the IIE Integrated Systems Conference. Norcross, GA: Institute of Industrial Engineers, 1988, pp 206–211.
22. A Mital, A Mahahjan, Impact of production volume, wage, and interest rates on economic decision making: the case of automated assembly. Proceedings of the Conference of the Society for Integrated manufacturing Conference. Norcross, GA: 1989, pp 558–563.
23. A Mital, Manual versus flexible assembly: a cross-comparison of performance and cost in four different countries. In: M Pridham, C O'Brien, eds. Production Research: Approaching the 21st Century. London: Taylor & Francis, 1991.
24. A. Mital, Economics of flexible assembly automation: influence of production and market factors. In: HR Parsaei, A Mital eds. Economics of Advanced Manufacturing Systems. London: Chapman & Hall, pp 45–72, 1992.
25. H Andersson, P Back, J Wirstad, Job Analysis for Training Design and Evaluation—Description of a Job Analysis Method for Process Industries. Report no. 6, Ergonomrad, Karlstad, Sweden, 1979.
26. J Badaracco, The Knowledge Link. Cambridge, MA: Harvard Business School Press, 1990.
27. P Ehn, The Work Oriented Design of Computer Artifacts. Stockholm: Arbetsmiljo, Arbelistratum, 1988.
28. T Engstrom, Future assembly work—natural grouping. In: Designing for Everyone—Proceedings of the XIth Congress of the International Ergonomics Association, vol 2, London: Taylor & Francis, 1991, pp 1317–1319.
29. AM Genaidy, T Gupta, Robot and human performance evaluation. In: M Rahimi, W Karwowski, eds. Human–Robot Interaction. London: Taylor & Francis, 1992, pp 4–15.
30. LS Bainbridge, SAR Quintanilla. Developing Skills with Information Technology. Chichester: John Wiley, 1989.
31. CK Prahalad, G Hamel, The core competence of the corporation. Harv Bus Rev 68: 79–91.
32. J Rasmussen, Some Trends in Man–Machine Interface Design for Industrial Process Plants. Report number Riso-M-2228. Riso National Laboratory, Roskilde, Denmark. 1980.
33. P Shipley. The analysis of organizations as an aid for ergonomics practice. In: JR Wilson, EN Corlett, eds. Evaluation of Human Work: A Practical Ergonomics Methodology. London: Taylor & Francis, 1995.
34. J Wirstad, On knowledge structures for process operators. In: LP Goodstein, HB Ansderson, SE Olsen, eds. Tasks, Errors and Mental Models. London: Taylor & Francis, 1988.
35. A Mital, R Vinayagamoorthy, Case study: economic feasibility of a robot installation. Eng Economist 32: 173–196, 1987.
36. A Mital, LJ George, Economic feasibility of a product line assembly: a case study. Eng Economist 35: 25–38, 1989.
37. BC Jiang OSH Cheng, Six severity level design for robotic cell safety. In: M Rahimi, W Karwowski eds. Human-Robot Interaction, 1992.
38. J. Hartley, Robots at Work: A Practical Guide for Engineers and Managers. Bedford: IFS; Amsterdam: North-Holland, 1983.
39. A Mital, LJ George, Human issues in automated (hybrid) factories. In: F Aghazadeh, ed. Trends in Ergonomics/Human Factors V. Amsterdam: North-Holland, 1988, pp 373–378.
40. CE Shannon, W Weaver. The Mathematical Theory of Communication. Urbana, IL: University of Illinois Press, 1949.
41. E Grandjean, Fitting the Task to the Man. 4th ed. London: Taylor & Francis, 1988.
42. JR Pierce, JE Karlin, Reading rates and the information rate of a human channel. Bell Teleph J 36: 497–516.
43. C Wickens, Engineering Psychology and Human Performance, Merrill, Columbus, Ohio, 1984.
44. J Swets ed. Signal detection and recognition by human observers: contemporary readings. Los Altos, CA: Peninsula Publishing, 1988.
45. D Green, J Swets, Signal Detection Theory and Psychophysics. Los Altos, CA: Peninsula Publishing, 1988.
46. MS Sanders, EJ McCormick, Human Factors in Engineering and Design. New York: McGraw-Hill, 1993.
47. G. Miller, The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol Rev 63: 81–97, 1956.
48. S Sternberg, High-speed scanning in human memory. Science 153: 652–654, 1966.
49. D Lane, Limited capacity, attention allocation, and productivity. In: W Howell, E Fleishman, eds. Human Performance and Productivity: Information Processing and Decision Making. Hillsdale, NJ: Lawrence Erlbaum Associates, 1982.
50. A Graig, Vigilance: theories and laboratory studies. In: S Folkard, T Monk, eds. Chichester: Wiley, 1985.
51. R Parasuraman, Vigilance, monitoring, and search. In: K Boff, L Kaufmann, J Thomas eds. Handbook of Perception and Human Performance: Cognitive Process and Performance. New York: Wiley 1986.
52. D Davies R Parasuraman, The Psychology of Vigilance. London: Academic Press, 1982.
53. J Warm, ed. Sustaining attention in human performance. Chichester: Wiley, 1984.
54. FT Eggemeier, Properties of workload assessment techniques. In: P hancock, N Meshkati, eds. Human Mental Workload. Amsterdam: North-Holland, 1988.
55. N Moray, Mental workload since 1979. Int Rev Ergon 2: 123–150, 1988.

56. R O'Donnell, FT Eggemeier, Workload assessment methodology. In: K Boff, L Kaufman, J Thomas, eds. Handbook of Perception and Human Performance, New York: Wiley, 1986.
57. C Jensen, G Schultz, B Bangerter, Applied Kinesiology and Biomechanics. New York: McGraw-Hill. 1983.
58. J Kelso, Human Motor Behavior: An Introduction. Hillsdale, NJ: Lawrence Erlbaum Associates, 1982.
59. J Adams, A closed-loop theory of motor learning. J Motor Behav 3: 111–150, 1981.
60. J Adams, Issues for a closed-loop theory of motor learning. In: G Stelmach Motor Control: Issues and Trends. New York: Academic Press, 1976.
61. C Winstein, R Schmidt, Sensorimotor feedback. In: H Holding ed. Human Skills, 2nd ed. Chichester: Wiley, 1989.
62. B Bridgeman, M Kirch, A Sperling, Segregation of cognitive and motor aspects of visual information using induced motion. Percept Psychophys 29: 336–342, 1981.
63. S Grillner, Neurobiological bases of rhythmic motor acts in vertebrates. Science 228: 143–149, 1985.
64. S Klapp, W Anderson, R Berrian, Implicit speech in reading, reconsidered. J Exper Psychol 100: 368–374, 1973.
65. R Schmidt, Motor Control and Learning: A Behavioral Emphasis, Champaign, IL: Human Kinetics. 1982.
66. R Woodworth, Experimental Psychology. New York: Henry Holt, 1938.
67. P Fitts, The information capacity of the human motor system in controlling the amplitude of movement. J Exper Psychol 47: 381–391, 1954.
68. P Fitts, A study of location discrimination ability. In: P Fitts, ed. Psychological Research on Equipment Design. Research Report 19, Army, Air Force, Aviation Psychology Program, Ohio State University, Columbus, OH, 1947.
69. J Brown, E Knauft, G Rosenbaum, The accuracy of positioning reactions as a function of their direction and extent. Am J Psychol 61: 167–182, 1947.
70. R Schmidt, H Zelaznik, B Hawkins, J Frank, J Quinn, Jr. Motor output variability: a theory for the accuracy of rapid motor acts. Psychol Rev 86: 415–451, 1979.
71. BH Deatherage, Auditory and other sensory forms of iformation presentation. In: HP Van Cott, R Kinkade, eds. Human Engineering Guide to Equipment Design, Washington, DC: Government Printing Office, 1972.
72. H Krueger, J Hessen, Obkective kontinuierliche Messung der Refraktion des Auges. Biomed Tech 27: 142–147, 1982.
73. H Luckiesh, FK Moss, The Science of Seeing. New York: Van Nostrand, 1937.
74. H Krueger, W Muller-Limmroth, Arbeiten mit dem Bildschirm-aber richtig! Bayerisches Staatsministerium für Arbeit und Sozialordnung, Winzererstr 9, 8000 Munuch 40, 1979.
75. IBM, Human Factors of Workstations with Visual Displays. IBM Human factors Center, Dept. P15, Bldg. 078, 5600 Cottle Road, San Jose, CA. 1984.
76. H Booher, Relative comprehensibility of pictorial information and printed words in proceduralized instructions. Hum Factor 17: 266–277, 1975.
77. A Fisk, M Scerbo, R Kobylak, Relative value of pictures and text in conveying information: performance and memory evaluations. Proceedings of the Human Factors Society 30th Anual Meeting, Santa Monica, CA, 1986, pp 1269–1271.
78. G Mowbray, J Gebhard, Man's senses vs. information channels. In: W Sinaiko ed. Selected Papers on Human Factors in Design and Use of Control Systems, New York: Dover, 1961.
79. J Feallock, J Southard, M Kobayashi, W Howell, Absolute judgements of colors in the gederal standards system. J Appl Psychol 50: 266–272, 1966.
80. M Jones, Color Coding. Hum Factors: 4: 355–365, 1962.
81. W Grether, C Baker, Visual presentation of information. In: HP Van Cott, R Kinkade, eds. Human Engineering Guide to Equipment Design. Washington, DC: Government Printing Office, 1972.
82. P. Muller, R Sidorsky, A. Slivinske, E. Alluisi, P Fitts, The Symbolic Coding of Informationa on Cathode Ray Tubes and Similar Displays. TR-55-375. Wright-Patterson Air Force base, OH. 1955.
83. P Cairney, D Seiss, Communication effectiveness of symbolic safety signs with different user groups. App Ergon 13: 91–97, 1982.
84. H. Heglin, NAVSHIPS Display Illumination Design Guide, vol. 2. NELC-TD223. Naval Electronics Laboratory Center, San Diego, CA: 1972.
85. A Mital, S Ramanan, Results of the simulation of a qualitative information display. Hum Factors, 28: 341–346, 1986.
86. B Mulligan, D McBride, L Goodman, A Design Guide for Non-Speech Auditory Displays. SR-84-1. Naval Aerospace Medical Research Laboratory, Pensacola, FL. 1984.
87. SA Mudd, The scaling and experimental investigation of four dimensions of pure tone and their use in an audio-visual monitoring problem. Unpublished doctoral dissertation, Purdue University, Lafayatte, IN. 1961.
88. JCR Licklider, Audio Warning Signals for Air Force Weapon Systems. TR-60-814, USAF, Wright Air Development Division, Wright-Patterson Air Force Base, OH, 1961.
89. JV Bradley, Desirable Control-Display Relationship For Moving-Scale Instruments. TR-54-423, USAF, Wright Air Development Center, Wright-Patterson Air Force base, OH, 1954.
90. Eastman Kodak Company, 1983, Ergonomic Design for People at Work. Belmont, CA: Lifetime Learning Publications, 1983.

91. A Chapanis, R Kinkade, Design of controls. In: HP Van Cott, R Kinkade eds. Human Engineering Guide to Equipment Design. Washington, DC: Government Printing Office 1972.
92. G Salvendy, Handbook of Human Factors and Ergonomics. New York: John Wiley & Sons. 1997.
93. MJ Warrick, Direction of movement in the use of control knobs to position visual indicators. In: PM Fitts, ed. Psychological Research on Equipment Design.
94. J Brebner, B Sandow, The effect of scale side on popuation stereotype. Ergonomics 19: 471–580, 1976.
95. H Petropoulos, J Brebner, Stereotypes for direction-of-movement of rotary controls associated with linear displays: the effect of scale presence and position, of poiter direction, and distances between the controls and the display. Ergonomics, 24: 143–151, 1981.
96. DH Holding, Direction of motion relationships between controls and displays in different planes, J Appl Psychol 41: 93–97, 1957.
97. C Worringham, D Beringer, Operator orientation and compatibility in visual-motor task performance. Ergonomics, 32: 387–400, 1989.
98. HP Van Cott, R Kinkade, Human Engineering Guide to Equipment Design. Washington, DC: Government Printing Office, 1972.

RECOMMENDED READING LIST

A Guide to Task Analysis (B Kirwan, LK Ainsworth, eds.). London: Taylor & Francis, 1992.

Applied Ergonomics Handbook (B Shakel, ed.) London: Butterworth Scientific, 1982.

Barnes RM. Motion and Time Study: Design and Measurement of Work. New York: John Wiley & Sons, 1980.

Booth PA. An Introduction to Human–Computer Interaction. Hove and London: Lawrence Erlbaum Associates, 1989.

Eastman Kodal Company. Ergonomic Design for People at Work, London: Lifetime Learning Publications, 1983.

Evaluation of Human Work: A Practical Ergonomics Methodology (JR Wilson, EN Corlett, eds.) London: Taylor & Francis, 1995.

Grandjean E. Fitting the Task to the Man. 4th ed. London: Taylor & Francis, 1988.

Handbook of Human Factors and Ergonomics (G Salvendy, ed.). New York: John Wiley & Sons, 1997.

Helander M. Handbook of Human–Computer Interaction. Amsterdam: North-Holland, 1988.

Human Engineering Guide to Equipment Design (HP Van Cott, R Kinkade, eds.), Washington, DC: Government Pringtin Office, 1972.

Nielsen J. Coordinating User Interfaces for Consistency. New York: Academic Press, NY, 1989.

Ravden S, Johnson G. Evaluating Usability of Human–Computer Interfaces. New York: Ellis Horwood, 1989.

Sanders MS, McCormick EJ. Human Factors in Engineering and Design. New York: McGraw-Hill, 1993.

Woodson, WE, Tillman B, Tillman P. Human Factors Design Handbook: Information and Guidelines for the Design Systems, Facilities, Equipment, and Products for Human Use. New York: McGraw-Hill, 1991.

Chapter 9.2

Workstation Design

Christin Shoaf and Ashraf M. Genaidy
University of Cincinnati, Cincinnati, Ohio

2.1 INTRODUCTION

Workstation design, including consideration of work methods, can be used to address several problems facing the contemporary workplace. With the spread of video display terminal (VDT) use in the workplace, cumulative trauma disorders are being reported with increasing frequency [1]. This new technology has increased the incidence of health disorders due to its physical requirements manifested in terms of repetitive motion and static constrained posture demands [2]. Workstation design principles can be used to lessen the stress demands imposed by these postures and motions and therefore reduce the risk of injury. Secondly, as companies continue to cut costs and strive to achieve more with fewer people, workstation design can also be used as an effective tool to optimize human effectiveness, thus resulting in increased efficiency and productivity. In a competitive industrial environment with health treatment, litigation and disability costs all rising, workstation design has become a significant factor not only in determining the health of the employee but the success of the business as well.

When considering the design of workstations, work methods, tools and handles, three factors account for the majority of ergonomic problems across a variety of industries. Therefore, the design principles guiding the biomechanical solution of these problems is based on the control of these factors. The three general methods [3] of reducing stress requirements are the reduction of extreme joint movement, excessive forces, and highly repetitive jobs. While this chapter is intended to provide general principles and guidelines for ergonomic workstation design, detailed specifications are available from several sources [4–6].

2.2 PHYSICAL LAYOUT CONSIDERATIONS

Safe work results when the job fits the worker. The contemporary workplace is composed of an increasing number of women, elderly, and minorities. Although recommended workstation dimensions based on anthropometric data are available [5], this data may adequately describe the varied population in today's work environment. Therefore, it is very important that workstations be designed to allow the maximum degree of flexibility in order to accommodate the contemporary worker population.

The ideal work situation is to alternate between sitting and standing at regular intervals [7]. Frequently changing body postures serves to minimize the discomfort and fatigue associated with maintaining the same posture for a long period of time. However, if a job cannot be designed to include tasks which include both sitting and standing postures, the seated position is preferable as it provides:

Stability required for tasks with high visual and motor control requirements
Less energy consumption than standing
Less stress on the leg joints

Lower hydrostatic pressure on leg circulation [5, 7].

2.2.1 Chair Design Guidelines

Prolonged work in the seated position can result in pain, fatigue, or injury in the lower back, shoulders, legs, arms, and neck. However, careful consideration to chair design can reduce the likelihood of these problems. Table 1 provides a list of chair parameters.

2.2.2 Height of Work Table/Activity

The work table should be adjustable to allow for work to be performed in the seated or standing position or to accommodate various seated tasks. Easy adjustability ensures that a large population of workers can be accommodated and awkward postures can be avoided. For most jobs, the work area should be designed around elbow height when standing or sitting in an erect posture. For precise work, the working height should be 2–4 in. above the elbow; for heavy manual work, the working height should be about 4–5 in. below the elbow [3,6].

2.2.3 Materials, Controls, Tools, and Equipment

All materials, tools, and equipment should be easily accessible to the worker to prevent awkward postures. All reaching should be below and in front of the shoulder and frequent work should be kept in the area that can be conveniently reached by the sweep of the arm with the upper arm hanging in a natural position at the side of the trunk [3].

Table 1 Chair Design Parameters

Parameter	Requirements
Backrests	Should be adjustable for height and angle of tilt, and provide continuous lumbar region support; should be independent from the seat pan
Height	Should be adjustable
Footrests	Should be provided and adjustable
Seat pan	Front edge should roll forward to prevent compression of the leg
Arm rests	Should be provided when feasible
Casters	Provide "safety" caster chairs (these do not roll easily with no weight in the chair) when lateral movements are required within the work area

2.2.4 Lighting

Adequate overhead lighting should be provided for all work areas. Task-specific localized lighting should also be provided if detailed work is required. Where documents are read, illumination levels of 500–700 l are recommended. Inadequate lighting may cause employees to work in awkward postures.

2.3 WORK METHOD CONSIDERATIONS

Although physical workstation design is the primary mechanism contributing to a healthful workplace, work method considerations can be employed as a temporary solution when workstation design changes are not immediately possible, or can be used as a complement to the design changes. Training in how to perform the work task as well as how to use all tools and equipment is mandatory for new employees and should be available as needed to experienced employees. Self pacing of work, especially for new employees, is recommended to alleviate mental and physical stresses. Also, frequent rest breaks should be allowed.

In addition to these measures, the design of the work situation can be altered to lessen stress effects. Highly repetitive jobs may be automated or a job consisting of few repeated tasks can be enlarged by combining varying tasks. Job enrichment is another job redesign technique which may be employed to increase job satisfaction. Job enrichment increases the amount of control and meaningfulness the employee experiences. When job enlargement or job enrichment is not feasible, rotating job tasks among employees is an alternative method of relieving stress due to repetition.

2.4 VIDEO DISPLAY TERMINAL GUIDELINES

The VDT work environment has become commonplace in recent years and has provoked a significant amount of discomfort and health complaints. Visual problems as well as musculoskeletal injuries are two frequently reported concerns which can be lessened by workstation design and work methods changes.

Visual fatigue can result from viewing objects on the VDT screen at a close range for an extended period of time as well as from excessive reflected glare. A brightly lit office environment, often found in the conventional office setting, can create a risk in VDT work as screen reflections occur. Several measures, including reorienting the VDT screen, selective removal of light sources or use of partitions or blinds, can aid in controlling light in VDT work areas [8]. If these solutions are not feasible, a microfilament mesh filter can be fitted over the screen or a parabolic lighting fixture (louver) can be installed below a conventional fluorescent fixture to reduce screen glare. Rest breaks can also be used to combat visual fatigue. The National Institute of Occupational Safety and Health (NIOSH) [9] recommends, as a minimum, a break should be taken after 2 hr of continuous VDT work. In order to ensure adequate employee visual capacity to perform VDT work, NIOSH [9] advocates visual testing before beginning VDT work and periodically thereafter.

The second frequent complaint among VDT workers is musculoskeletal discomfort. Early NIOSH studies report a prevalence rate exceeding 75% for the "occasional" experience of back, neck, and shoulder discomfort among VDT users [10,11]. As VDT work is stationary and sedentary, operators most often remain seated in fixed, sometimes awkward, postures for extended periods of time. Consequently, joint forces and static loads can be increased to levels causing discomfort. For example, elevation of the arms to reach the keyboard may aggravate neck and shoulder pain. Proper workstation design is the first step necessary to improve the VDT work environment. As previously stated, adjustability is paramount in good workstation design. This is also true for the VDT workstation. Some of the most important VDT workstation features are [12,13]:

Movable keyboards with adjustable height that allow the operator's arms to be approximately parallel to the floor
Adjustable backrest to support the lower back
Adjustable height and depth of the chair seat
Swivel chair with five-point base and casters
Screen between 1 and 2 ft away; middle of screen slightly below eye level; characters large and sharp enough to read easily; brightness and contrast controls; adjustable terminal height and tilt; glareproof surface; no visible flicker of characters
Indirect general lighting 200-500 l, moderate brightness
Direct, adjustable task lighting
Feet resting firmly on the floor, footrest for shorter operators; thighs approximately parallel to the floor
Adequate work-table space for a document holder approximately the same distance as the screen
Additional ventilation or air-conditioning where required to compensate for the heat generated by many VDTs operating in a work space
VDT cables positioned and secured to prevent tripping.

In addition to proper workstation design, consideration to work methods is also required to discourage the health risks associated with VDT operation. First, operators should be aware of ergonomic principles and then trained to adjust their own workstations. Secondly, even with good workstation design, physical stress can result due to the prolonged postures demanded by many VDT tasks. Frequent rest breaks can alleviate some of this stress. Therefore, VDT operators should periodically change positions or stretch during long assignments. Walking around during breaks or performing simple stretching exercises can also be beneficial. Lee et al. [14] provide an excellent review of physical exercise programs recommended for VDT operators.

2.5 SUMMARY

Careful consideration of workstation design, including work methods, should be given whenever the physical work setting changes and should continually be reevaluated to ensure proper person–environment fit. Attention to workstation design can serve as an effective tool in the prevention of work-related health disorders as well as for increasing employee productivity. Consequently, both of these outcomes will result in higher company profit. By consulting current design guidelines, training employees in equipment use as well as in basic ergonomic principles and encouraging employee feedback, an effective workstation environment can be realized.

REFERENCES

1. EL Greene. Cumulative trauma disorders on the rise. Med Trib July 26: 1990.
2. EB Chapnik, CM Gross. Evaluation, office improvements can reduce VDT operator. Occupat Health Safety 56:7, 1987.
3. V Putz-Anderson. Cumulative Trauma Disorders. A Manual for Musculoskeletal Diseases of the Upper Limbs. London: Taylor & Francis, 1988, pp 85–103.
4. American National Standard for Human Factors Engineering of Visual Display Terminal Workstations,

ANSI/HFS 100-1988. Santa Monica, CA: Human Factors Society, 1988.

5. DB Chaffin, GB Andersson. Occupational Biomechanics. New York: John Wiley Sons, 1991.
6. E Grandjean. Fitting the task to the man. Philadelphia, PA, Taylor & Francis, 1988.
7. R Carson. Ergonomically Designed Chairs Adjust to Individual Demands. Occupat Health Safety. June: 71–75, 1993.
8. SL Sauter, TM Schnorr. Occupational health aspects of work with video display terminals. In: WN Rom, ed. Environmental and Occupational Medicine. Boston, Toronto, London: Little Brown and Company, 1992.
9. BL Johnson, JM Melius. A review of NIOSH's VDT studies and recommendations. Presented at Work with Display Units International Conference, Stockholm, Sweden, May, 1996.
10. SL Sauter, MS Gottlieb, KC Jones, VN Dodson. Job and health implications of VDT use: initial results of the Wisconsin-NIOSH study. Commun Assoc Comput Machinery 26: 1982, pp 284–294.
11. MJ Smith, BGF Cohen, LW Stammerjohn. An investigation of health complaints and stress in video display operations. Hum Factors 23: 1981.
12. M Sullivan. Video display health concerns. AAOHN J 37:7, 1989, pp 254–257.
13. JP Shield. Video display terminals and occupational health. Prof Safety Dec: 1990, pp 17–19.
14. K Lee, N Swanson, S Sauter, R Wickstrom, A Waikar, M Magnum. A review of physical exercises recommended for VDT operators. Appl Ergon 23:6, 1992, pp 387–408.

Chapter 9.3

Physical Strength Assessment in Ergonomics

Sean Gallagher
National Institute for Occupational Safety and Health, Pittsburgh, Pennsylvania

J. Steven Moore
The University of Texas Health Center, Tyler, Texas

Terrence J. Stobbe
West Virginia University, Morgantown, West Virginia

James D. McGlothlin
Purdue University, West Lafayette, Indiana

Amit Bhattacharya
University of Cincinnati, Cincinnati, Ohio

3.1 INTRODUCTION

Humankind's interest in the measurement of human physical strength probably dates to the first humans. At that time, life was truly a struggle in which the fittest survived. To a great extent, fittest meant strongest. It is perhaps ironic that in a modern civilized world, children still emphasize the relative importance of physical size and strength in determining the status hierarchy within a group. It is equally ironic that current interest in human physical strength comes from 1970s–1980s vintage research which demonstrated that persons with adequate physical strength were less likely to be injured on physically demanding jobs. Survival in many modern workplaces may still be a case of survival of the strongest.

There is, however, a flip side to this issue. If persons with limited strength are likely to be injured on "hard" jobs, what we know about physical strength can be applied to job design so that "hard" jobs are changed into jobs the are within the physical strength capability of most people. Thus, since human physical strength is important, it is necessary to find ways to quantify it through testing. This chapter is about human physical strength testing. Its purpose is not to recommend any particular type of testing, but rather to describe the types of testing that are available, and the uses to which strength testing has been put. It is up to individual users of the strength testing to decide which testing technique is the most appropriate for his or her particular application. This chapter discusses four types of strength testing: isometric, isoinertial, psychophysical, and isokinetic.

3.1.1 Human Strength

Before describing the different types of strength measurement, the term strength must be defined and the concept of strength measurement must be explained. Strength is defined as the capacity to produce force or torque with a voluntary muscle contraction. Maximum strength is defined as the capacity to produce force or torque with a maximum voluntary muscle contraction [1,2]. These definitions have some key words which must be explained.

A voluntary muscle contraction is "voluntary." When a person's physical strength is measured, only the voluntary effort the person is willing to put forth at the time is measured. Thus, when we test a person's maximum strength, we do not measure their maximum; we measure some smaller number that represents what they are comfortable expressing at the time with the existing equipment and environmental conditions. It is interesting to note that researchers have experimented with startling persons being tested (for example by setting off a starter's pistol behind them during a test) and have found significant increases in measured strength [3]. It has been hypothesized that the lower strength displayed by persons during normal testing provides a margin of safety against overloading and damaging muscle tissue. It is also true that the test equipment and the tested person's familiarity with the process will influence their "voluntary" strength output. This is particularly true of the interface between the tested person and the test equipment. A poorly designed interface will induce localized tissue pressures which vary from uncomfortable to painful. In this situation, you are measuring voluntary discomfort tolerance—not strength. It is important for strength researchers to keep the "voluntary" nature of their data in mind when they are designing their equipment and protocols.

The definition of strength also speaks of force or torque. Strength researchers and users of strength data must also understand this distinction. We commonly use the terms "muscle force" and "muscle strength" to describe the strength phenomenon. Technically, this is incorrect. For most human movements and force exertions, there is actually a group of individual muscles (a functional muscle group) which work together to produce the observable output. In complicated exertions, there are a number of functional muscle groups working together to produce the measured output. Elbow flexion strength, for example, is the result of the combined efforts of the biceps brachii, brachialis, and the brachioradialis, and a squat lift is the result of the combined efforts of the legs, back, and arms. In elbow flexion, each individual muscle's contribution to the functional muscle group's output depends on the posture of the arm being measured. Thus, when we measure elbow flexion strength, we are measuring the strength of the elbow flexor muscle group, not the strength of any individual muscle.

Furthermore, we are measuring (recording) the force created by the functional muscle group(s) against the interface between the person and the equipment (a set of handles for example). Consider the elbow flexion measurement depicted in Fig. 1. The force generated by the elbow flexor muscle group is shown by F_m. This force acts through lever arm a. In so doing, it creates a torque about the elbow joint equal to $F_m a$. The measured force (Q, R, or S) will depend upon how far (b, c, or d) the interface (force cuff) is from the elbow. Assuming that the exertion is static (nothing moves) in this example, the measured force (on the gage) will equal the elbow flexor torque divided by the distance that the gage's associated force cuff is from the elbowjoint. That is,

$$Q = (F_m a)/b \tag{1}$$

or

$$R = (F_m a)/c \tag{2}$$

or

$$S = (F_m a)/d \tag{3}$$

As we move the interface (force cuff) from the elbow to the hand, the measured force will decrease. This example highlights four points. First, "muscular strength is what is measured by an instrument" [4]. Second, people publishing/using strength data must report/understand in detail how the measurements were done. Third, the differences in published strengths of the various body parts may be due to differences in the measurement methods and locations. Fourth, interface locations selected using anthropometric criteria will result in more consistent results across the population measured [5].

In summary, a record of a person's strength describes what the instrumentation measured when the person voluntarily produced a muscle contraction in a specific set of circumstances with a specific interface and instrumentation.

3.1.2 Purposes of Strength Measurement in Ergonomics

There are a number of reasons people may want to collect human strength data. One of the most common is collecting population strength data which can be used to build an anthropometric database; create

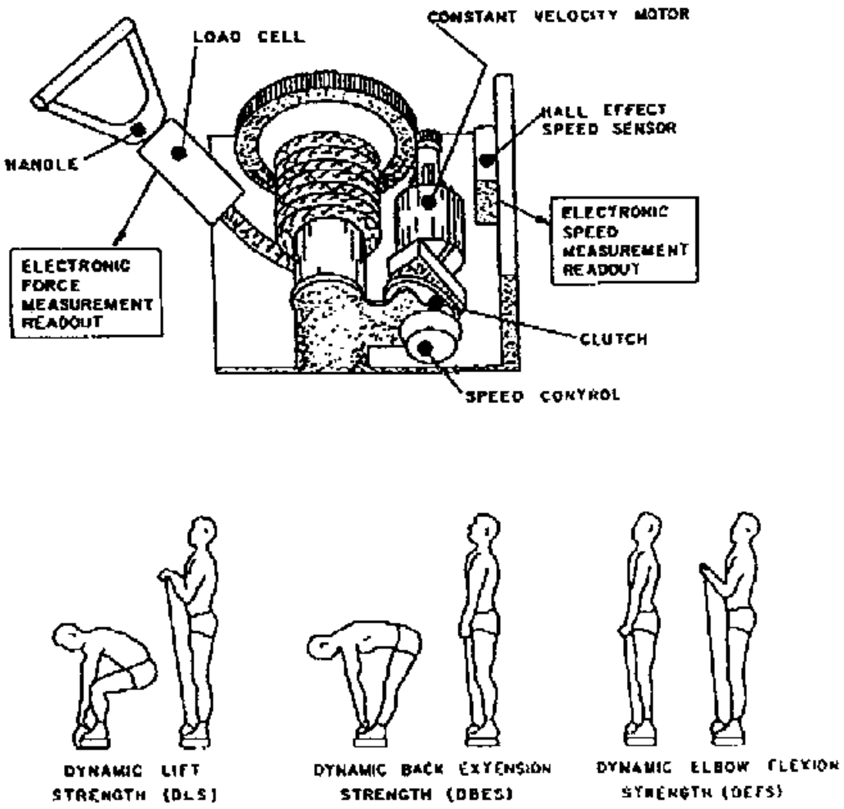

CONSTANT VELOCITY MOTOR
LOAD CELL
HANDLE
HALL EFFECT SPEED SENSOR
ELECTRONIC SPEED MEASUREMENT READOUT
ELECTRONIC FORCE MEASUREMENT READOUT
CLUTCH
SPEED CONTROL
DYNAMIC LIFT STRENGTH (DLS)
DYNAMIC BACK EXTENSION STRENGTH (DBES)
DYNAMIC ELBOW FLEXION STRENGTH (DEFS)

Figure 1 Given a constant muscle force (F_m), forces measured at various distances from the elbow will result in different force readings (F_Q, F_R, or F_S).

design data for products, tasks, equipment, etc.; and for basic research into the strength phenomenon. This chapter will focus on two common uses of physical strength assessment in ergonomics: *worker selection and placement and job design.*

3.1.2.1 Worker Selection and Placement

The purpose of worker selection and placement programs is to ensure that jobs which involve heavy physical demands are not performed by those lacking the necessary strength capabilities [6]. It should be noted that this method is not the preferred strategy of the ergonomist, but is a provisional measure for the control of work-related musculoskeletal disorders (WMSDs) where job design cannot be used to alleviate task demands. Nonetheless, this method can be effective in reducing the harmful physical effects caused by the mismatch of worker and job, *given adherence to two fundamental principles*. These principles are: (1) ensuring that the strength measures are a close simulation of the actual high-strength elements in a job, and (2) that strength assessment is performed only under circumstances where they can predict who may be at risk of WMSD. The following paragraphs describe these issues in more detail.

It has become quite clear over the past several years that strength, in and of itself, is a poor predictor of the risk of future injury to a worker [7–9]. Evaluation of worker strength capacity is only predictive o of those at risk of injury when it is carefully equated with job demands [10]. All too often, emphasis is placed on collecting data on the former attribute, while the latter receives little or no attention. Recent evidence has illustrated that the analysis of job demands cannot be a generalized description of "light" versus "heavy" jobs [11], there needs to be a careful biomechanical evaluation of strenuous tasks as performed by the worker.

The need to analyze strength in relation to specific job demands can be illustrated using the following scenario. An employer has an opening for a physically demanding job and wishes to hire an individual with strength sufficient for the task. This employer decides to base his employment decision on a strength test given to a group of applicants. Naturally, he selects the applicant with the highest strength score to perform the job. The employer may have hired the strongest job applicant; however, what this employer must understand is that he may not have decreased the risk of injury to his employee if the demands of his job still exceed this individual's maximum voluntary strength capacity. This example should make it clear that only through knowing both about the person's capabilities *and* the job demands might worker selection protect workers from WMSDs.

The second issue that must be considered when worker selection is to be implemented is that of the test's *predictive value*. The predictive value of a test is a measure of its ability to determine who is at risk of future WMSD [6]. In the case of job-related strength-testing, the predictive value appears to hold only when testing individuals for jobs *where high risk is known* (that is, for jobs known to possess high strength demands). Strength testing does not appear to predict the risk of injury or disease to an individual when job demands are low or moderate.

It should be clear from the preceding arguments that worker selection procedures are not the preferred method of reducing the risk of WMSDs, and are not to be applied indiscriminately in the workplace. Instead, care must be exercised to ensure that these strength testing procedures are applied only in select circumstances. This procedure appears only to be effective when jobs are known to entail high strength demands, and only when the worker's strength is evaluated in the context of the high strength elements of a job. However, if attention is paid to these limitations, worker selection can be an effective tool to decrease the risk of WMSDs.

3.1.2.2 Job Design

The use of physical strength assessment in ergonomics is not limited to its role in worker selection, it can also be used for the purpose of *job design*. Job design has been a primary focus of the psychophysical method of determining acceptable weights and forces. Rather than determining *individual* worker strength capabilities and comparing these to job demands, the psychophysical method attempts to determine workloads that are "acceptable" (a submaximal strength assessment) for *populations* of workers. Once the acceptable workloads for a population are determined, the job or task is designed to accommodate the vast majority of the population. For example, a lifting task might be designed by selecting a weight that is acceptable to 75% of females and 90% of males. The use of strength assessment for job design has been shown to be an effective method of controlling WMSDs. It has been estimated that proper design of manual tasks using psychophysical strength assessment might reduce the risk of back injuries by up to 33% [12].

3.1.3 Purpose of This Chapter

Muscular strength is a complicated function which can vary greatly depending on the methods of assessment. As a result, there is often a great deal of confusion and misunderstanding of the appropriate uses of strength testing in ergonomics. It is not uncommon to see these techniques misapplied by persons who are not thoroughly familiar with the caveats and limitations inherent with various strength assessment procedures. The purposes of this chapter are: (1) to familiarize the reader with the four most common techniques of strength assessment used in ergonomics (isometric, isoinertial, psychophysical, and isokinetic); and (2) to describe the proper applications of these techniques in the attempt to control WMSDs in the workplace.

This chapter contains four parts, one for each of the four strength measurement techniques listed above. Each part describes the strength measurement technique and reviews the relevant published data. Equipment considerations and testing protocols are described, and the utility of the tests in the context of ergonomics are also evaluated. Finally, each part concludes with a discussion of the measurement technique with regard to the Criteria for Physical Assessment in Worker Selection [6]. In this discussion, each measurement technique is subjected to the following set of questions:

1. Is it safe to administer?
2. Does it give reliable, quantitative values?
3. Is it related to specific job requirements?
4. Is it practical?
5. Does it predict risk of future injury or illness?

It is hoped that this part of the chapter will provide a resource that can be used to better understand and properly apply these strength assessment techniques in the effort to reduce the risk of WMSDs.

3.2 PART I: ISOMETRIC STRENGTH

3.2.1 Introduction and Definition

Isometric strength is defined as the capacity to produce force or torque with a voluntary isometric [muscle(s) maintain(s) a constant length] contraction. The key thing to understand about this type of contraction and strength measurement is that there is no body movement during the measurement period. The tested person's body angles and posture remain the same throughout the test.

Isometric strength has historically been the one most studied and measured. It is probably the easiest to measure and the easiest to understand. Some strength researchers feel that isometric strength data may be difficult to apply to some "real life" situations because in most real circumstances people are moving—they are not static. Other researchers counter that it is equally difficult to determine the speed of movement of a person group of persons doing a job (each moves in their own unique manner and at their own speed across the links and joints of the body). Thus, dynamic strength test data collected on persons moving at a different speed and/or in a different posture from the "real world" condition will be just as hard to apply. In truth, neither is better—they are different measurements and both researchers and users should collect/use data which they understand and which fits their application.

3.2.2 Workplace Assessment

When a worker is called upon to perform a physically demanding lifting task moments (or torques) are produced about various joints of the body by the external load [13]. Often these moments are augmented by the force of gravity acting on the mass of various body segments. For example, in a biceps curl exercise, the moment produced by the forearm flexors must counteract the moment of the weight held in the hands, as well as the moment caused by gravity acting on the center of mass of the forearm. In order to successfully perform the task, the muscles responsible for moving the joint must develop a greater moment than that imposed by the combined moment of the external load and body segment. It should be clear that for each joint of the body, there exists a limit to the strength that can be produced by the muscle to move ever increasing external loads. This concept has formed the basis of isometric muscle strength prediction modelling [13].

The following procedures are generally used in this biomechanical analysis technique. First, workers are observed (and usually photographed or videotaped) during the performance of physically demanding tasks. For each task, the posture of the torso and the extremities are documented at the time of peak exertion. The postures are then recreated using a computerized software package, which calculates the load moments produced at various joints of the body during the performance of the task. The values obtained during this analysis are then compared to population norms for isometric strength obtained from a population of industrial workers. In this manner, the model can estimate the proportion of the population capable of performing the exertion, as well as the predicted compression forces acting on the lumbar disks resulting from the task.

Figure 2 shows an example of the workplace analysis necessary for this type of approach. Direct observations of the worker performing the task provide the necessary data. For example, the load magnitude and direction must be known (in this case a 200 N load acting downward), the size of the worker, the postural angles of the body (obtained from photographs or videotape), and whether the task requires one or two hands. Furthermore, the analysis requires accurate measurement of the load center relative to the ankles and the low back. A computer analysis program

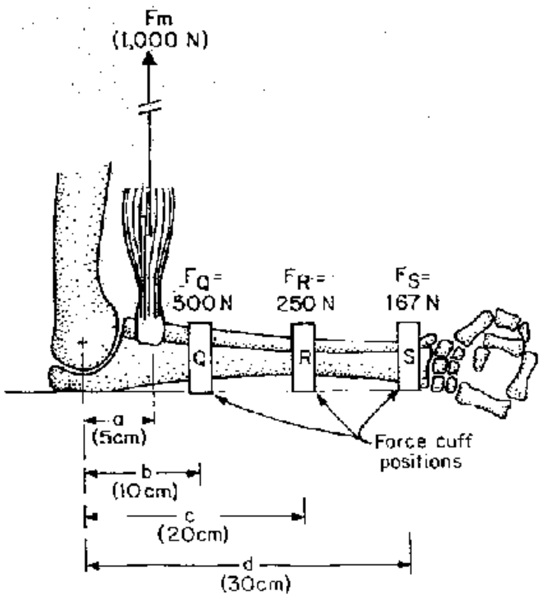


Figure 2 Postural data required for analysis of joint moment strengths using the isometric technique.

can be used to calculate the strength requirements for the task, and the percentage of workers who would be likely to have sufficient strength capabilities to perform it. Results of this particular analysis indicate that the muscles at the hip are most stressed, with 83% of men having the necessary capabilities but on slightly more than half of women would have the necessary strength in this region. These results can then be used as the basis for determining those workers who have adequate strength for the job. However, such results can also the used as ammunition for recommending changes in job design [13].

3.2.3 Isometric Testing Protocol

The basic testing protocol for isometric strength testing was developed by Caldwell et al. [1] and published in an AIHA ergonomics guide by Chaffin [2]. The protocol outlined herein includes additional information determined by researchers since that time. When conducting isometric testing, there are a number of factors that must be considered and controlled (if possible) to avoid biased results. These factors include the equipment used to make the measurements, the instructions given to the person tested, the duration of the measurement period, the person's posture during the test, the length of the rest period between trials, the number of trials a person is given for each test, the tested person's physical state at the time of testing, the type of postural control used during the tests, and the environmental conditions during the test.

3.2.4 Test Duration

The length of an isometric strength test can impact the result in two ways. If it is too long, the subject will fatigue and their strength score will decline. If it is too short, the subject will not reach their maximum force level before the test is terminated. The existing AIHA Guide suggests a 4 sec test with the score being the average strength displayed during seconds 2–4. The appropriate 3 sec period can be determined as follows.

If the measuring equipment has the capability, collect strength data by having the person begin their contraction with the equipment monitoring the force until some preselected threshold is reached (usually 20–30% below the expected maximum force for the person and posture), have the equipment wait 1 sec, and then have the equipment average the displayed force for the next 3 sec. This is easily done with computerized systems.

If the equipment does not have the above capability, then have the person tested begin the test and gradually increase their force over a 1 sec period. The force should be measured and averaged over the next 3 sec. In complex whole body tests where multiple functional muscle groups are involved, it may take persons a few seconds to reach their maximum. Under these conditions, the data collector must adjust the premeasurement time interval accordingly, and they must carefully monitor the progress of the testing to insure that they are fact measuring the maximal force during the 3 sec period.

3.2.5 Instructions

The instructions to the person tested should be factual, include no emotional appeals, and be the same for all persons in a given test group. This is most reliably accomplished with standardized written instruction, since the test administrator's feelings about the testee or the desired outcome may become evident during verbal instruction.

The following additional factors should also be considered. The purpose of the test, the use of the test results, the test procedures, and the test equipment should be thoroughly explained to the persons tested. Generally, the anonymity of the persons tested is maintained, but if names may be released, the tested person's written permission must be obtained. Any risks inherent to the testing procedure should be explained to the persons tested, and an informed consent document should be provided to, and signed by, all participating persons. All test participants should be volunteers.

Rewards, performance goals, encouragement during the test (for example, "pull, pull, pull, you can do it," etc.), spectators, between person competition, and unusual noises will all affect the outcome of the tests and must be avoided. Feedback to the tested person should be positive and qualitative. Feedback should not be provided during the test exertion, but may be provided after a trial or test is complete. No quantitative results should be provided during the testing period because they may change the person's incentive and thus their test result.

To the tested person, a 4 sec maximal exertion seems to take a long time. During the test, feedback in the form of a slow four count or some other tester–testee agreed-upon manner should be provided so the tested person knows how much longer a test will last.

3.2.6 Rest Period Length

Persons undergoing isometric strength testing will generally be performing a series of tests, with a number of trials for each test. Under these conditions, a person could develop localized muscle fatigue, and this must be avoided, since it will result in underestimating strength. Studies by Schanne [14] and Stobbe [5] have shown that a minimum rest period of 2 min between trials of a given test or between tests is adequate to prevent localized muscle fatigue. The data collector must be alert for signs of fatigue, such as a drop in strength scores as a test progresses. The person tested must be encouraged to report any symptoms of fatigue and the rest periods should be adjusted accordingly. Whenever possible, successive tests should not stress the same muscle groups.

3.2.7 Number of Trials for Each Test

The test–retest variability for this type of testing is about 10%. It is higher for people with limited experience with either isometric testing or with forceful physical exertion in general. In addition, these people will often require a series of trials of a test to reach their maximum. The use of a single trial of a test will generally underestimate a person's maximum strength, and may underestimate it by more than 50%. A two-trial protocol results in less of an underestimate, but it may still exceed 30% [15].

For this reason, the preferred approach to determining the number of trials for each test is to make the choice on the basis of performance. Begin by having the subject perform two trials of the test. The two scores are then compared and if they are within 10% of each other the highest of the two values is used as the estimate of the person's maximal strength, and you proceed to the next test. If the two values differ by more than 10%, additional trials of the same test are performed until the two largest values are within 10% of each other. Using this approach, Stobbe and Plummer averaged 2.43 trials per test across 67 subjects performing an average of 30 different strength tests [15]. In any case, a minimum of two trials is needed for each test.

3.2.8 When to Give Tests

A person's measured strength is, for a variety of reasons, somewhat variable. It will not be constant over time, nor over a workday. However, in the absence of specific muscle strength training, it should remain within a relatively narrow range. It is generally higher at the beginning of a workday than at the end. The fatigue-induced strength decrement will vary from person to person and will depend on the nature of the work done during the day. A person who performs repetitive lifting tasks all day can be expected to have a large lifting strength decrement over a workday, whereas a sedentary worker should have little or no decrement. Based on these results, the fairest evaluation of a person's maximum strength can be done at the beginning of, or at least early in, a workday.

3.2.9 Test Posture

Measured strength is highly posture dependent. Even small changes in the body angles of persons being tested and/or changes in the direction of force application can result in large changes in measured strength. When collecting strength data, a researcher should first determine what type of data is sought, and then one or more strength tests which will provide that specific type of data should be designed. If, for example, the test is being done to determine whether people are physically fit for a job, the test posture should emulate, to the extent possible, the posture required on the job.

Once the test posture has been determined, the researcher must ensure that the same posture is used on each trial of the test. The researcher must monitor the test to ensure that the person's posture does not change during the test. If these things are not done, the test results will be erratic, and may seriously overestimate or underestimate the person's actual maximal strength.

3.2.10 Restraint Systems

Restraint systems are generally used either to confine a person to the desired test posture, or to isolate some part of the tested person's body so that a specific muscle group (or groups) can be tested (see Fig. 3). In addition, restraint systems help to assure that all persons participating in a given study will be performing the same test. The type and location of restraint system used can have a major impact on test results. Similarly, the lack of a restraint system can allow the posture to vary or allow the use of the wrong or additional muscle groups, both of which will impact test results.

Any restraint system used should be comfortable, it should be padded in a manner that prevents local tissue stress concentrations during the test; it should be positioned so that the correct muscle group(s) and

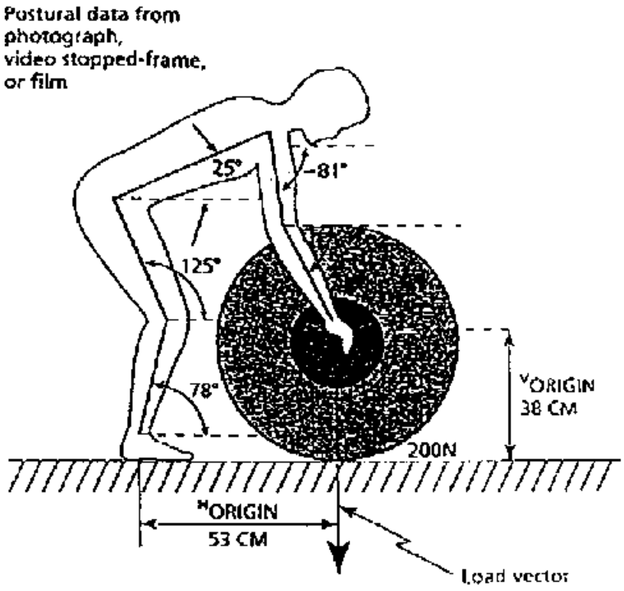


Figure 3 Example of a test fixture designed to restrain various body segments during isometric strength testing.

posture(s) are used and maintained. This latter item often requires some experimentation to correctly achieve.

For many strength tests, the use of a restraint system will be necessary if consistent and meaningful results are to be achieved. Researchers reporting strength testing results should describe the restraints used and their location in detail so that other researchers and persons applying their data will be able to interpret it correctly. The nonuse of restraints should also be reported.

3.2.11 Environmental Conditions

The environmental conditions selected for the testing periods should be appropriate to the purpose of the test. For most testing, the environmental conditions found in a typical office building or medical department will be acceptable. In some cases, the effects of the environment on measured strength or physical performance must be determined, and then appropriate conditions can be established (e.g., worksites requiring exposure to hot or cold temperature extremes).

3.2.12 Equipment

Isometric strength-testing equipment has not been standardized. Any equipment which has the capability to perform the necessary timing and averaging described above under "test duration" is probably acceptable. Today, this varies from dedicated force measurement devices such as the force monitor developed in the 1970s at University of Michigan, to a force transducer coupled to a PC via an analog-to-digital converter and managed by appropriate software, to complex multiple-mode strength-measuring devices manufactured by companies like Cybex, Chattex, Loredan, and Isotechnologies. The associated prices vary from $1000 or so to as high as $50,000–$100,000.

However, equipment price is not the issue; rather, it is equipment function. Researchers should select or build equipment suited to their needs. Researchers must also understand what is happening inside the device (and its associated software) they are using so that the data they collect can be properly interpreted.

The human–equipment interface is another matter which can impact the test results. The interface must be appropriate to the task measured, it should be comfortable (unless discomfort effects are being studied), and it should give the person tested a sense of security about the test. Persons will generally be providing a maximal exertion in a situation where there is no movement. If they fear that the testing system may fail or move unexpectedly, they will not give a maximal performance. Similarly, the equipment must be strong enough to remain intact under the maximum load placed on it. If it fails unexpectedly, someone is going to be injured—perhaps severely.

3.2.13 Subjects

The subjects selected for strength testing will determine the results obtained. This means that when strength data are collected, the selection of subjects must appropriately represent the population it claims to describe (for example, design data for retired persons should be collected on retired persons, and design data for entry-level construction workers should be collected on young healthy adults).

For general research purposes, persons participating in a strength testing project should not have a history of musculoskeletal injuries. There are other medical conditions, including hypertension, which may pose a threat of harm to a participant. Whenever possible, prospective participants should be medically evaluated and approved before participating in a strength-testing project.

The following data should be provided about the subject population when reporting strength-testing results:

1. Gender
2. Age distribution
3. Relevant anthropometry (height, weight, etc.)
4. Sample size
5. Method by which sample was selected and who it is intended to represent
6. Extent of strength training done by participants, and their experience with isometric testing
7. Health status of participants (medical exam and/or health questionnaire recommended.

3.2.14 Strength Data Reporting

The minimum data which should be reported for strength-testing projects are:

1. Mean, median, and mode of data set
2. Standard deviation of data set
3. Skewness of data set (or histogram describing data set)
4. Minimum and maximum values.

3.2.15 Evaluation According to Physical Assessment Criteria

A set of five criteria have been purposed to evaluate the utility of all forms of strength testing. Isometric strength testing is evaluated with respect to these criteria in the following sections.

3.2.15.1 Is It Safe to Administer?

Any form of physical exertion carries with it some risk. The directions for the person undergoing an isometric test specifically state that the person is to slowly increase the force until they reach what they feel is a maximum, and to stop if at any time during the exertion they feel discomfort or pain. The directions also expressly forbid jerking on the equipment. When isometric testing is performed in this manner it is quite safe to administer because the tested person is deciding how much force to apply, over what time interval, and how long to apply it. The only known complaints relating to participation in isometric testing are some residual soreness in the muscles which were active in the test(s), and this is rarely reported.

3.2.15.2 Does the Method Provide Reliable Quantitative Values?

The test-retest variability for isometric testing is 5–10%. In the absence of a specific strength training program, individual isometric strength remains relatively stable over time. When the number of trials is based on the 10% criterion discussed earlier, the recorded strength is near or at the tested person's maximum voluntary strength. Assuming the above factors, and that test postures are properly controlled, isometric strength testing is highly reliable and quantitative.

3.2.15.3 Is Method Practical?

Isometric strength testing has already been used successfully in industry for employee placement, in laboratories for the collection of design data, and in rehabilitation facilities for patient progress assessment.

3.2.15.4 Is the Method Related to Specific Job Requirements (Content Validity)?

Isometric strength testing can be performed in any posture. When it is conducted for employee placement purposes, the test postures should be as similar as possible to the postures that will be used on the job. The force vector applied by the tested person should also be similar to the force vector that will be applied on the job. When these two criteria are met, isometric strength testing is closely related to job requirements. However, it should be noted that results obtained using isometric strength testing loses both content and criterion-related validity as job demands become more dynamic.

3.2.15.5 Does the Method Predict the Risk of Future Injury or Illness?

A number of researchers have demonstrated that isometric strength testing does predict risk of future injury or illness for people on physically stressful job [16,17]. The accuracy of this prediction is dependent on the quality of the job evaluation on which the strength tests are based, and the care with which the tests are administered.

3.3 PART II: MAXIMAL ISOINERTIAL STRENGTH TESTING

3.3.1 Definition of Isoinertial Strength

Kroemer [18–20] and Kroemer et al. [4] define the isoinertial technique of strength assessment as one in which *mass properties of an object are held constant,* as in lifting a given weight over a predetermined distance. Several strength assessment procedures possess the attribute in this definition. Most commonly associated with the term is a specific test developed to provide a relatively quick assessment of a subject's maximal lifting capacity using a modified weight-lifting device [18,21]. The classic psychophysical methodology of assessing maximum acceptable weights of lift is also as an isoinertial technique under this definition [12].

While the definition provided by Kroemer [18] and Kroemer et al. [4] has been most widely accepted in the literature, some have applied the term "isoinertial" to techniques that differ somewhat from the definition given above, such as in a description of the Isotechnologies B-200 strength-testing device [22]. Rather than lifting a constant mass, the B-200 applies a constant force against which the subject performs an exertion. The isoinertial tests described in this chapter apply to situations in which the mass to be moved by a musculoskeletal effort is set to a constant.

3.3.2 Is Isoinertial Testing Psychophysical or Is Psychophysical Testing Isoinertial?

As various types of strength tests have evolved over the pasts few decades, there have been some unfortunate developments in the terminology that have arisen to describe and/or classify different strength assessment procedures. This is particularly evident when one tries to sort out the various tests that have been labelled "isoinertial." One example was cited above. Another problem that has evolved is that the term "isoinertial strength" has developed two different connotations. The first connotation is the conceptual definition—isoinertial strength tests describe any strength test where a constant mass is handled. However, in practice, the term is often used to denote a *specific* strength test where subjects' maximal lifting capacity is determined using a machine where a constant mass is lifted [18,21]. Partially as a result of this dual connotation, the literature contains both references to "isoinertial strength test" as a psychophysical variant [23], and to the psychophysical method as an "isoinertial strength test" [4,24]. In order to lay the framework for the next two parts, the authors would like to briefly discuss some operational definitions of tests of isoinertial and psychophysical strength.

When Ayoub and Mital [23] state that the isoinertial strength test is a variant of the psychophysical method, they refer to the specific strength test developed by Kroemer [18] and McDaniel et al. [21]. Clearly, this isoinertial protocol has many similarities to the psychophysical method: both are dynamic; weight is adjusted in both; both measure the load a subject is willing to endure under specified circumstances, etc. However, while both deal with lifting and adjusting loads, there are significant differences between the psychophysical (isoinertial) technique and the Kroemer–McDaniel (isoinertial) protocol, both procedurally and in use of the data collected in these tests. For purposes of this chapter we will designate the Kroemer–McDaniel protocol maximal isoinertial strength tests (MIST). This part deals with the latter isoinertial technique, which differs from the psychophysical technique on the following counts:

1. In maximal isoinertial strength tests, the amount of weight lifted by the subject is *systematically adjusted by the experimenter, primarily through increasing the load to the subject's maximum*. In contrast, in psychophysical tests, *weight adjustment is freely controlled by the subject, and may be upwards or downwards*.
2. The maximal isoinertial strength tests discussed in this part are designed to quickly establish an individual's *maximal strength* using a *limited number of lifting repetitions*, whereas psychophysical strength assessments are typically performed over a *longer duration of time* (usually at least 20 min), and instructions are that the subject select an *acceptable (submaximal) weight of lift,* not a maximal one. Due to the typically longer duration of psychophysical assessments, greater aerobic and cardiovascular components are usually involved in the acceptable workload chosen.
3. Isoinertial strength tests have traditionally been used as a *worker selection tool* (a method of matching physically capable individuals to demanding tasks). A primary focus of psychophysical methods has been to establish data that can be used for the purpose of *ergonomic job design* [12].

3.3.3 Published Data

There are two primary maximal isoinertial strength test procedures that will be described in this section. One involves the use of a modified weight-lifting machine where the subject lifts a rack of hidden weights to prescribed heights, as depicted in Fig. 4 [21]. Kroemer [18] refers to his technique as LIFTEST, while the Air Force protocol has been named the strength aptitude test (SAT). The other test uses a lifting box, into which weights are placed incrementally at specified times until the lifting limit is reached [25]. The greatest

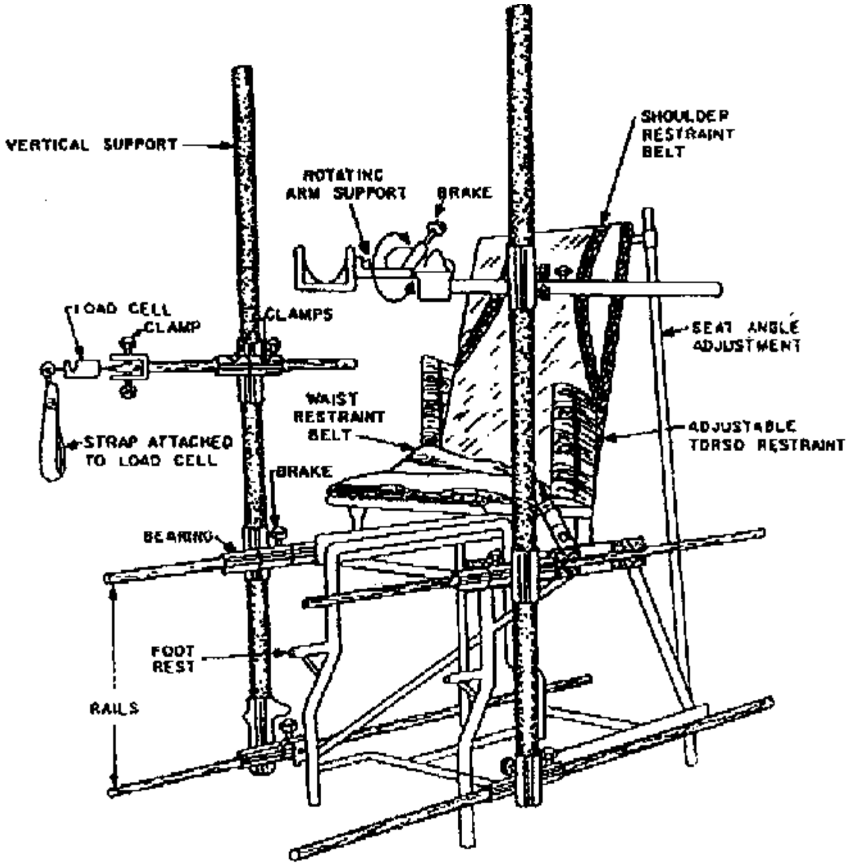


Figure 4 The incremental weight lift machine developed by the Air Force for the strength aptitude test.

bulk of the isoinertial testing literature deals with the former procedure.

3.3.4 The LIFTEST/Strength Aptitude Test Techniques

The LIFTEST and SAT procedures are isoinertial techniques of strength testing that attempt to establish the maximal amount of weight that a person can safely lift [18]. In this technique, a preselected mass, constant in each test, is lifted by the subject (typically from knee height to knuckle height, elbow height, or to overhead reach height). The amount of weight to be lifted is at first relatively light, but the amount of mass is continually increased in succeeding tests until it reaches the maximal amount that the subject voluntarily indicates he or she can handle. This technique has been used extensively by the U.S. Air Force [21], and is applicable to dynamic lifting tasks in industry as well [18,26].

Since a constant mass is lifted in LIFTEST, the acceleration of the load during a test is dependent on the force applied to the load during the test (in accordance with Newton's second law: $F = ma$). The dynamic nature of this procedure, the fact that a constant mass is being lifted, and the subject's freedom to choose the preferred lifting technique, all give the LIFTEST a general similarity to certain types of industrial lifting tasks. A unique aspect of the LIFTEST technique is that it is the only strength measurement procedure discussed in this chapter where results are based on the success or failure to perform a prescribed criterion task. The criterion tasks studied have typically included lifting to shoulder height [20,21,26,27], elbow height [21,26], or knuckle height [20,26]. The USAF also developed a muscular endurance test using an incremental lift machine (ILM) [21].

The LIFTEST shoulder height maximal strength test has demonstrated the highest correlation with manual materials handling activities [26], and has been subjected to a biomechanical analysis by Stevenson et al. [28]. They demonstrated that this criterion task could be separated into three distinct phases: (1) a powerful upward pulling phase, where maximal acceleration, velocity, and power values are observed; (2) a wrist changeover manoeuvre (at approximately elbow height), where momentum is required to compensate for low force and acceleration; and (3) a pushing phase (at or above chest height), characterized by a secondary (lower) maximal force and acceleration profile.

The analysis by Stevenson et al. [28] suggested that successful performance of the criterion shoulder height lift requires a technique quite different from the concept of slow, smooth lifting usually recommended for submaximal lifting tasks. On the contrary, lifting of a maximal load requires a rapid and powerful lifting motion. This is due in large part to the need to develop sufficient momentum to allow successful completion of the wrist changeover portion of the lift. Most lift failures occur during the wrist changeover procedure, probably the result of poor mechanical advantage of the upper limb to apply force to the load at this point in the lift [28]. Stevenson et al. [28] found that certain anatomical landmarks were associated with maximal force, velocity, and power readings (see Fig. 5). Maximal force readings were found to occur at mid-thigh, maximal velocity at chest height, minimum force was recorded at head height, and the second maximal acceleration (pushing phase) was observed at 113% of the subject's stature.

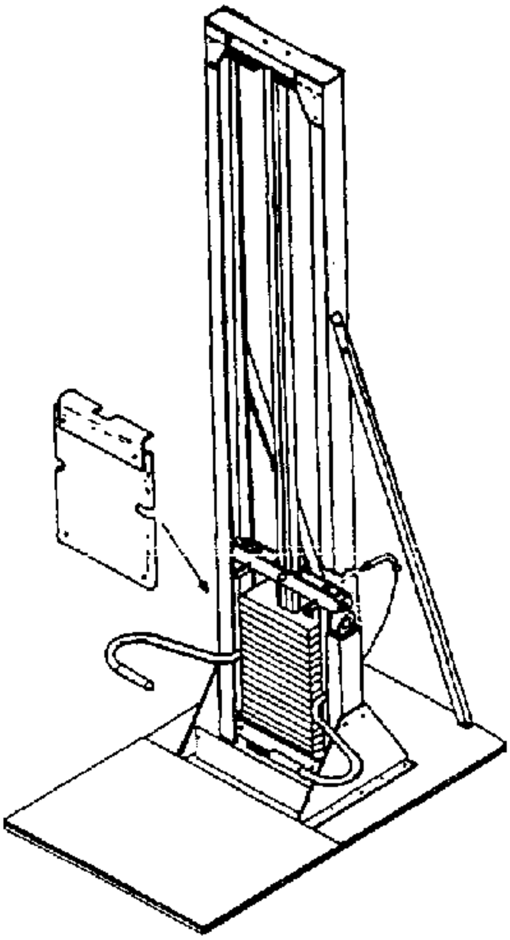


Figure 5 Analysis of the shoulder height strength test indicates three distinct lift phases: (1) a powerful upward pulling phase (where maximal forces are developed), (2) a wrist changeover maneuver (where most failures occur), and (3) a pushing phase (where a secondary, lower, maximal force is observed).

3.3.5 The Strength Aptitude Test

The strength aptitude test (SAT) [21] is a classification tool for matching the physical strength abilities of individuals with the physical strength requirements of jobs in the Air Force (McDaniel, personal communication, 1994). The SAT is given to all Air Force recruits as part of their preinduction examinations. Results of the SAT are used to determine whether the individual tested possesses the minimum strength criterion which is a prerequisite for admission to various Air Force specialties (AFSs). The physical demands of each AFS are objectively computed from an average physical demand weighted by the frequency of performance and the percent of the AFS members performing the task. Objects weighing less than 10 lb are not considered physically demanding and are not considered in the job analysis. Prior to averaging the physical demands of the AFS, the actual weights of objects handled are converted into equivalent performance on the incremental weight lift test using regression equations developed over years of testing. These relationships consider the type of task (lifting, carrying, pushing, etc.), the size and weight of the object handled, as well as the type and height of the lift. Thus, the physical job demands are related to, but are not identical to, the ability to lift an object to a certain height. Job demands for various AFSs are reanalyzed periodically for purposes of updating the SAT.

The first major report describing this classification tool was a study of 1671 basic trainees (1066 males and 605 females) [21]. The incremental weight lift tests started with a 18.1 kg weight which was to be raised to 1.83 m or more above the floor. This initial weight was increased in 4.5 kg increments until subjects were unable to raise the weight to 1.83 m. Maximal weight lift to elbow height was then tested as a continuation of the incremental weight lift test. In the test of lifting the weight to 1.83 m, males averaged 51.8 kg (±10.5), while females averaged 25.8 kg (±5.3). The respective weights lifted to elbow height were 58.6 kg (±11.2) and 30.7 kg (± 6.3). The distributions of weight lifting capabilities for both male and female basic trainees in lifts to 6 ft are provided in Fig. 6. Results of the elbow height lift are presented in Table 1. McDaniel et al. [21] also performed a test of isoinertial endurance. This involved holding a 31.8 kg weight at elbow height for the duration the subject could perform the task. Male basic trainees were able to hold the weight for an average of 53.3 sec (±22.11), while female basic trainees managed to hold the weight an average of 10.3 sec (±10.5 SD).

When developing the SAT, the Air Force examined more than 60 candidate tests in an extensive, four-year research program and found the incremental weight lift to 1.83 m to be the single test of overall dynamic strength capability, which was both safe and reliable (McDaniel, personal communication 1994). This finding was confirmed by an independent study funded by the U.S. Army [29]. This study compared the SAT to a battery of tests developed by the Army (including isometric and dynamic tests) and compared these with representative heavy demand tasks performed within the Army. Results showed the SAT to be superior to all others in predicting performance on the criterion tasks.

Table 1 Weight-Lifting Capabilities of Basic Trainees for Lifts to Elbow Height

	Males		Females	
Percentile	Pounds	Kilograms	Pounds	Kilograms
1	80	36.3	40	18.1
5	93	42.2	48	21.8
10	100	45.4	52	23.6
20	109	49.5	58	26.3
30	116	52.6	61	27.7
40	122	55.4	65	29.5
50	127	57.6	68	30.9
60	133	60.3	71	32.2
70	140	63.5	75	34.0
80	150	68.1	78	35.4
90	160	47.6	85	38.6
95	171	77.6	90	40.8
99	197	89.4	100	45.4
Mean	129.07	58.56	67.66	30.70
SD	24.60	11.16	13.91	6.31
Minimum	50	22.7	< 40	<18.1
Maximum	>200	>90.7	100	49.9
Number	1066		605	

Source: Ref. 21.

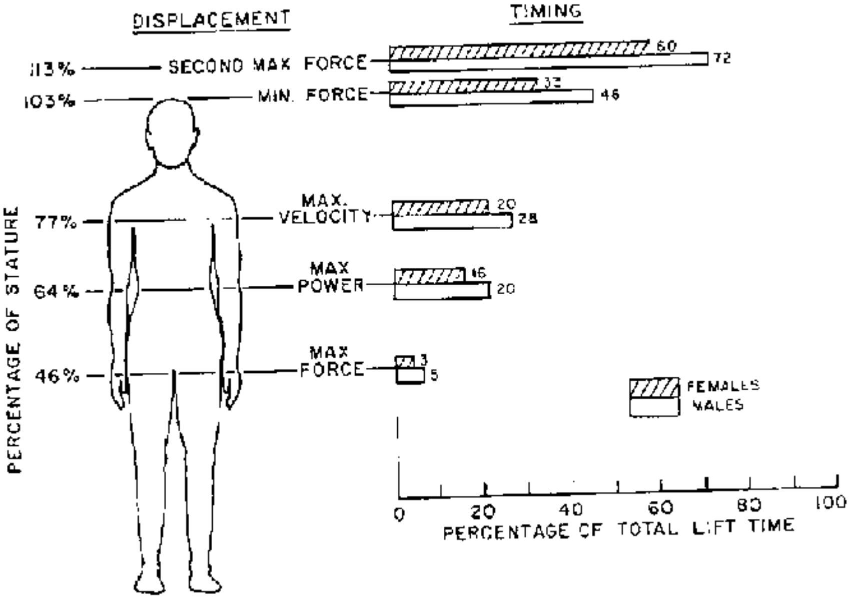


Figure 6 Distribution of weight-lifting capabilities for male and female basic trainees for lifts to 6 ft. (From Ref. 21.)

3.3.6 Virginia Tech Data

Kroemer [18,20] described results of a study using a similar apparatus as the one used by the U.S. Air Force. The sample consisted of 39 subjects (25 male) recruited from a university student population. The procedures were similar to McDaniel et al. [21] with the exception that the minimum starting weight was 11.4 kg, and that maximal lifting limits were established to prevent overexertion. These were 77.1 kg for floor to knuckle height tests, and 45.4 for floor to overhead reach tests. The following procedure was used for establishing the maximal load: if the initial 11.4 kg weight was successfully lifted, the weight was doubled to 22.7 kg. Additional 11.4 kg increments were added until an attempt failed or the maximal lifting limit was reached. If an attempt failed, the load was reduced by 6.8 kg. If this test weight was lifted, 4.5 kg was added; if not, 2.3 kg were subtracted. This scheme allowed quick determination of the maximal load the subject could lift.

In Kroemer's study, six of 25 male subjects exceeded the cutoff load of 100 lb in overhead reach lifts [18,20]. All 14 females stayed below this limit. The 19 remaining male subjects lifted an average of 27 kg. The female subjects lifted an average of 16 kg. In lifts to knuckle height, 17 of the 25 male (but none of the female) subjects exceeded the 77.1 kg cutoff limit. The remaining subjects lifted an average of about 54 kg, with males averaging 62 kg and females 49 kg. The coefficients of variation for all tests were less than 8%. Summary data for this study is given in Table 2.

3.3.7 The Progressive Isoinertial Lifting Evaluation

Another variety of MIST has been described by Mayer et al. [25,30]. Instead of using a weight 3 rack as shown in Fig. 3, the progressive isoinertial, lifting valuation (PILE) is performed using a lifting box with handles and increasing weight in the box as it is lifted and lowered. Subjects perform two isoinertial lifting/lowering tests: one from floor to 30 in. (Lumbar) and one from 30 to 54 in. (Cervical). Unlike the isoinertial procedures described above, there are three possible criteria for termination of the test: (1) voluntary termination due to fatigue, excessive discomfort, or inability to complete the specified lifting task; (2) achievement of a target heart rate (usually 85% of age predicted maximal heart rate); or (3) when the subject lifts a "safe limit" of 55–60% of his or her body weight. Thus, contrary to the tests described above, the PILE test may be terminated due to cardiovascular factors, rather than when an acceptable load limit is reached.

Since the PILE was developed as a means of evaluating the degree of restoration of functional capacity of individuals complaining of chronic low-back pain (LBP), the initial weight lifted by subjects using this procedure is somewhat lower than the tests described above. The initial starting weight is 3.6 kg for women and 5.9 kg for men. Weight is incremented upwards at a rate of 2.3 kg every 20 sec for women, and 4.6 kg every 20 sec for men. During each 20 sec period, four lifting movements (box lift or box lower) are performed. The lifting sequence is repeated until one of

Table 2 Results of Lifts to Shoulder and Knuckle Height for 25 Male and 14 Female Subjects

	All				Male				Female			
	$\bar{X}$	*SD*	*CV*	*N*	$\bar{X}$	*SD*	*CV*	*N*	$\bar{X}$	*SD*	*CV*	*N*
Overhead LIFTEST (kg)	26.95	10.32	3.5%	33	34.72	5.22	3.2%	19	16.34	3.74	3.9%	14
Lift ≥ 45.5 kg	—	—	—	6	—	—	—	6	—	—	—	0
Knuckle LIFTEST (kg)	53.86	13.35	6.9%	22	62.22	7.84	5.2%	8	49.08	13.69	7.8%	14
Lift ≥ 77 kg	—	—	—	17	—	—	—	17	00	00	—	0

Source: Ref. 20.

the three endpoints is reached. The vast majority of subjects are stopped by the "psychophysical" endpoint, indicating the subject has a perception of fatigue or overexertion. The target heart rate endpoint is typically reached in older or large individuals. The "safe limit" endpoint is typically encountered only by very thin or small individuals.

Mayer et al. [25] developed a normative database for the PILE, consisting of 61 males and 31 females. Both total work (TW) and force in pounds (F) were normalized according to age, gender, and a body weight variable. The body weight variable, the adjusted weight (AW), was taken as actual body weight in slim individuals, but was taken as the ideal weight in overweight individuals. This was done to prevent skewing the normalization in overweight individuals. Table 3 presents the normative database for the PILE.

3.3.8 Evaluation According to Criteria for Physical Assessment

3.3.8.1 Is It Safe to Administer?

The MIST procedures described above appear to have been remarkably free of injury. Isoinertial procedures have now been performed many thousands of times without report of verifiable injury. However, reports of transitory muscle soreness have been noted [25]. The temporary muscle soreness associated with isoinertial testing has been similar to that experienced in isokinetic tests, but has been reported less frequently than that experienced with isometric strength tests.

McDaniel et al. [21] present some useful recommendations for design of safe isoinertial weight-lift testing procedures. The following list summarizes the recommendations made by these authors.

1. Weight-lifting equipment should be designed so that the weights and handle move only in a vertical direction.
2. Sturdy shoes should be worn; or the subject may be tested barefoot. Encumbering clothing should not be worn during the test.
3. The initial weight lifted should be low: 20–40 lb. Weights in this range are within the capability of almost everyone. Weight increments should be small.
4. The upper limit should not exceed the largest job related requirement or 160 lb, whichever is less.
5. The starting handle position should be 1–2 ft above the standing surface. If the handle is lower, the knees may cause obstruction. If the handle is too high, the subject will squat to get their shoulders under it prior to lifting. A gap between the handles will allow them to pass outside the subject's knees when lifting, allowing a more erect back and encouraging the use of leg strength.
6. The recommended body orientation prior to lifting should be (a) arms straight at the elbow, (b) knees bent to keep the trunk as erect as possible, and (c) head aligned with the trunk. The lift should be performed smoothly, without jerk.

Table 3 Normative Data

	Males $n = 61$				
	AW	LW/AW	LTW/AW	$CERF/AW$	$CERTW/AW$
Means	161.3	0.50	22.8	0.40	12.3
Standard deviation	19.6	0.10	7.8	0.10	5.1
Standard error of mean	2.51	0.01	1.0	0.01	0.81
	Females ($n = 31$)				
Means	121.6	0.35	17.04	0.25	7.32
Standard deviation	10.65	0.07	7.0	0.04	2.4
Standard error of mean	1.98	0.01	1.3	0.01	0.56

L = Lumbar; CER = Cervical; TW = Total work in lb-ft; AW = Adjusted weight in lbs; F = final force in lbs.
Source: Ref. 25.

7. A medical history of the subject should be obtained. If suspicious physical conditions are identified, a full physical examination should be performed prior to testing. Subjects over 50 years of age or pregnant should always have a physical prior to testing.
8. All sources of overmotivation should be minimized. Testing should be done in private and results kept confidential. Even the test subject should not be informed until the testing is completed.
9. If the subject pauses during a lift, the strength limit has been reached, and the test should be terminated. Multiple attempts at any single weight level should not be allowed.
10. The testing should always be voluntary. The subject should be allowed to stop the test at any time. The subject should not be informed of the criteria prior to or during the test.

It is noteworthy that, as of 1994, over two million subjects have been tested on the SAT without any back injury or overexertion injury (McDaniel, personal communication, 1994).

3.3.8.2 Does It Give Reliable, Quantitative Values?

Kroemer et al. [20] reported LIFTEST coefficients of variation (measures of intraindividual variability in repeated exertions) of 3.5 for all subjects in overhead lifts, and 6.9 in lifts to knuckle height. The same study showed somewhat higher variability in tests of isometric strength (coefficient of variations ranging from 11.6 to 15.4). Test–retest reliability was not reported by McDaniel et al. [21]. Mayer et al. [25] reported correlation coefficients of a reproducibility study of the PILE which demonstrated good test–retest reliability for both floor to 30 in. lifts ($r = 0.87$, $p < 0.001$) and 30–54 in. lifts ($r = 0.93$, $p < 0.001$). Thus, the reliability of isoinertial procedures appears to compare favorably with that demonstrated by other strength assessment techniques.

3.3.8.3 Is It Practical?

Isoinertial techniques generally appear practical in terms of providing a test procedure that requires minimal administration time and minimal time for instruction and learning. Even in a worst case scenario, the isoinertial procedures used by Kroemerz [2] would take only a few minutes to determine the maximal weight lifting capability of the subject for a particular condition. The McDaniel et al. [21] (McDaniel, personal communication, 1994) procedure can be performed in approximately 3–5 min. The PILE test administration time is reported to last on the order of 5 min [25].

Practicality is determined in part by cost of the equipment required, and on this account, the cost of isoinertial techniques is quite modest. In fact, the PILE test requires no more hardware than a lifting box and some sturdy shelves, and some weight. The equipment needed to develop the LIFTEST devices used by McDaniel et al. [21] and Kroemer [18–20] would be slightly more expensive, but would not be prohibitive for most applications. In fact, Kroemer [19] states that the device is easily dismantled and could easily be transported to different sites in a small truck or station wagon, or perhaps in a mobile laboratory vehicle.

3.3.8.4 Is It Related to Specific Job Requirements?

Since industrial lifting tasks are performed dynamically, isoinertial strength tests do appear to provide some useful information related to an individual's ability to cope with the dynamic demands of industrial lifting. McDaniel (personal communication, 1994) has reported that these tests are predictive of performance on a wide range of dynamic tasks, including asymmetrical tasks, carrying, and pushing tasks. Furthermore, Jiang et al. [26] demonstrated that the isoinertial lifting test to 6 ft was more highly correlated with psychophysical tests of lifting capacity than isometric techniques. The PILE test possesses good content validity for industrial lifting tasks, as subjects are able to use a more ''natural'' lifting technique when handling the lifting box.

3.3.8.5 Does It Predict Risk of Future Injury or Illness?

The ability of a strength test to predict risk of future injury or illness is dependent upon performance of prospective epidemiological studies. As of this writing, no such studies have been conducted on the isoinertial techniques described above.

3.4 PART III: PSYCHOPHYSICAL STRENGTH

3.4.1 Theory and Description of the Psychophysical Methodology for Determining Maximum Acceptable Weights and Forces

According to contemporary psychophysical theory, the relationship between the strength of a perceived sensation (S) and the intensity of a physical stimulus (I) is best expressed by a power relationship [31]:

$$S = kI^n \tag{4}$$

This psychophysical principle has been applied to many practical problems, including the development of scales or guidelines for effective temperature, loudness, brightness, and ratings of perceived exertion. Based on the results of a number of experiments using a variety of scaling methods and a number of different muscle groups, the pooled estimate the exponent for muscular effort and force is 1.7 [32].

When applying this principle to work situations, it is assumed that individuals are capable and willing to consistently identify a specified level of perceived sensation (S). For manual materials handling tasks, this specified level is usually the *maximum acceptable weight* or *maximum acceptable force*. The meaning of these phrases are defined by the instructions given to the test subject [33]. "You are to work on an incentive basis, working as hard as you can without straining yourself, or becoming unusually tired, weakened, overheated, or out of breath."

If the task involves *lifting*, the experiment measures the maximum acceptable weight of lift. Similarly, there are maximum acceptable weights for *lowering* and *carrying*. Such tests are isoinertial in nature; however, in contrast to the tests described in Part 2, they are typically used to test submaximal, repetitive handling capabilities. Data are also available for *pushing* and *pulling*. These are reported as maximum acceptable forces and include data for initial as well as sustained pulling or pushing.

3.4.2 Why Use Psychophysical Methods?

Snook identified several advantages and disadvantages to using psychophysical methods for determining maximum acceptable weights [34]. The advantages include:

1. The realistic simulation of industrial work (face validity).
2. The ability to study intermittent tasks (physiological steady state not required).
3. The results are consistent with the industrial. engineering concept of "a fair day's work for a fair day's pay."
4. The results are reproducible.
5. The results appear to be related to low-back pain (content validity).

Disadvantages include:

1. The tests are performed in a laboratory.
2. It is a subjective method that relies on self-reporting by the subject.
3. The results for very high-frequency tasks may exceed recommendations for energy expenditure.
4. The results are insensitive to bending and twisting.

In terms of the application of the data derived from these studies, Liberty Mutual preferred to use it to design a job to fit the worker, since this application represented a more permanent, engineering solution to the problem of low-back pain in industry [12]. This approach not only reduces the worker's exposure to potential low-back pain risk factors, but also reduces liability associated with worker selection [12].

3.4.3 Published Data

3.4.3.1 Liberty Mutual

Snook and Ciriello at the Liberty Mutual Insurance Company have published the most comprehensive tables for this type of strength assessment [35]. The most recent data is summarized in nine tables, organized as follows [35]:

1. Maximum acceptable weight of lift for males
2. Maximum acceptable weight of lift for females
3. Maximum acceptable weight of lower for males
4. Maximum acceptable weight of lower for females
5. Maximum acceptable forces of push for males (initial and sustained)
6. Maximum acceptable forces of push for females (initial and sustained)
7. Maximum acceptable forces of pull for males (initial and sustained)
8. Maximum acceptable forces of pull for females (initial and sustained)
9. Maximum acceptable weight of carry (males and females).

3.4.3.2 Other Sources

Ayoub et al. [36] and Mital [37] have also published tables for maximum acceptable weights of lift. Even though their tables are similar in format and generally in agreement with those from Liberty Mutual, there are some differences. Possible sources for these differences may be differences in test protocol, differences in task variables, and differences in subject populations and their characteristics.

3.4.4 Experimental Procedures and Methods

For the sake of simplicity and convenience, the Liberty Mutual protocol for lifting or lowering and an excerpt from the lifting table will be used as examples for this section. The protocols used by Ayoub et al. [36] and Mital [37] were similar, but not exactly the same. The reader should refer to the original publications for details.

The Liberty Mutual experimental procedures and methods were succinctly reviewed in their most recent revision of the table [35]. The data reported in these revised tables reflect results from 119 second shift workers from local industry (68 males, 51 females). All were prescreened to ensure good health prior to participation. These subjects were employed by Liberty Mutual for the duration of the project (usually 10 weeks). All received 4–5 days of conditioning and training prior to participation in actual test sessions.

Test subjects wore standardized clothing and shoes. The experiments were performed in an environmental chamber maintained at 21°C (dry bulb) and 45% relative humidity. Forty-one anthropometric variables were recorded for each subject, including several isometric strengths and aerobic capacity.

A single test session lasted approximately 4 h and consisted of five different tasks. Each task session lasted 40 min, followed by 10 min rest. Most subjects participated in at least two test sessions per week for 10 weeks. In general, a subject's heart rate and oxygen consumption were monitored during the sessions.

3.4.4.1 Lifting or Lowering Tasks

In a lifting or lowering task session, the subject was given control of one variable, usually the weight of the box. The other task variables would be specified by the experimental protocol. These variables include:

1. *Lifting zone*, which refers to whether the lift occurs between floor level to knuckle height (low), knuckle height to shoulder height (center), or shoulder height to arm reach (high).
2. *Vertical distance of lift,* which refers to the vertical height of the lift within one of these lifting zones. The specified values for distance of lift in the tables are 25 cm (10 in.), 51 cm (20 in.), and 76 cm (30 in.). It is possible to use linear extrapolation for lift distances not exactly equal to one of these values.
3. *Box width,* which refers to the dimension of the box away from the body. The three values of box width are 34 cm (13.4 in.), 49 cm (19.3 in.), and 75 cm (29.5 in.). It is possible to use linear extrapolation between these values.
4. *Frequency of lift*, expressed as one lift per time interval, and include intervals of 5 sec, 9 sec, 14 sec, 1 min, 2 min, 5 min and 8 hr.

These same definitions apply to a lowering task, except the word "lower" is substituted for "lift." The test protocol for lowering was essentially identical to that for lifting, and the results are reported in a similar format. It should be noted, however, that the test protocols for lifting and lowering involved using a special apparatus that returned the box to its original specified location, so that the subject only lifted or lowered, not both.

Per the instructions, the subject was to adjust the weight of the box, according to his or her own perceptions of effort or fatigue, by adding or removing steel shot or welding rods from a box. The box had handles and a false bottom to eliminate visual cues. Each task experiment was broken into two segments so that the initial weight of the box could be randomly varied between high versus low so that the subject approached his or her maximum acceptable weight from above as well as below. If the results met a 15% test–retest criterion, the reported result was the average of these two values. If the results did not meet this criterion, they were discarded and the test repeated at a later time.

In reporting the results, it was assumed that the gender-specific maximum acceptable weights for a particular task were normally distributed. As a consequence, the results were reported as percentages of population, stratified by gender. The Liberty Mutual tables are organized around the following percentages: 90%, 75%, 50%, 25%, and 10% [35]. The 90th percentile refers to a value of weight that 90% of individuals of that gender would consider a maximum acceptable weight (90% "acceptable"), while the 10th percentile refers to a value of weight that only 10% of individuals of that gender would find acceptable (10% "acceptable").

3.4.5 Important Caveats

Snook and Ciriello have identified several important caveats that should be remembered when using the Liberty Mutual tables [35].

1. The data for each experimental situation were assumed to be normally distributed when the maximum acceptable weights and forces acceptable to 10%, 25%, 50%, 75%, and 90% of the industrial population were determined.
2. Not all values in the tables are based on experimental data. Some values were derived by assuming that the variation noted for a particular variable for one type of task would be similar to that observed for another task, e.g., the effects on lowering would be similar to that on lifting.
3. The tables for lifting, lowering, and carrying are based on boxes with handles that were handled close to the body. They recommend that the values in the tables be reduced by approximately 15% when handling boxes without handles. When handling smaller boxes with extended reaches between knee and shoulder heights, they recommend reducing the values by approximately 50%.
4. Some of the reported weights and forces exceed recommended levels of energy expenditure if performed for 8 hr or more per day. These data are italicized in the tables.
5. The data in the tables give results for individual manual materials handling tasks. When a job involves a combination of these tasks, each component should be analyzed separately, and the component with the lowest percent of capable population represents the maximum acceptable weight or force for the combined task. It should be recognized, however, that the energy expenditure for the combined task will be greater than that for the individual components.

Some recent data suggest that persons performing lifting tasks are relatively insensitive to the perception of high disk compression forces on the spine [38]. As a result, there may be some tasks in the tables that exceed recommended levels of disk compression.

3.4.6 Related Research

3.4.6.1 Task and Subject Variables

A variety of researchers have examined the effects of other task and subject variables using the psychophysical protocol. Most of these studies involve a small number (<10) of college students as test subjects. Some experiments used the Liberty Mutual protocol; others used the protocol described by Ayoub et al. [36] and Mital [37]. These "refinements" are summarized in Table 4.

Table 4 Miscellaneous Task Variables Evaluated Using the Psychophysical Methodology

Task variable(s)	Reference(s)
Zone of lift	12, 35–37, 50–52
Distance of lift	12, 35–37, 50–52
Frequency of lift	12, 35–37, 50–52
Box width	12, 35–37, 50–52
Extended work shifts	37
Combinations of lift, carry, and lower	40, 41
Angle of twist	52
Box length	52, 53
Material density	54
Location of center of gravity	54
Center of gravity relative to preferred hand	54
Sleep deprivation	54
Bag versus box	55
Fullness of bag (same weight)	55
Bag ± handles	55
Day 1 to day 5 of work week	48
Asymmetrical loads	57–59
Asymmetrical lifting	52–60
Emergency scenario	61
Handle position	62
Handle Angle	62
Duration of lifting	63, 64
Overreach heights	65
Restricted vs. unrestricted shelf opening clearances	66
Experienced vs. inexperienced workers	67
Nonstandard or restricted postures	49, 68–70

3.4.7 Recommended Applications

3.4.7.1 Job Evaluation

The Liberty Mutual tables were developed for the purpose of evaluating work, not workers [39]. In particular, the tables are intended to help industry in the evaluation and design of manual materials handling tasks that are consistent with worker limitations and abilities [35]. The explicit goal is the control of low-back pain through reductions in initial episodes, length of disability, and recurrences [39].

To apply the tables in the context of job evaluation, it is first necessary to specify the task variables of the job. For a lifting task, this would include the lift zone, distance of lift, box width, frequency of lift, and the presence or absence of box handles. In addition, it would be necessary to measure the weight of the object to be handled, perhaps using a scale or dynamometer. Once these variables are specified, the measured weight can be compared to the data in the table to determine the percent of capable population for males and females. The procedure is similar for pulling or pushing. The required force can be measured with a dynamometer.

Consider the following example. The task is to lift a box 49 cm wide that weighs 20 kg once every minute between floor level to knuckle height for a distance of 51 cm.

From Table 5, an excerpt from the Liberty Mutual tables, it is seen that the weight of the box, 20 kg, is exactly equal to the maximum acceptable weight of lift for 75% of males, i.e., 75% of males would consider this task "acceptable." By contrast, the highest maximum acceptable weight of lift reported for females is 18 kg. As a result, this task is "not acceptable" to over 90% of females.

3.4.7.2 Job Design

To apply the tables in the context of job design, the process is essentially identical. All task-specific parameters must be identified, except the required weight or force (that is what you are determining). You select a desired percent of capable of population, noting gender effects, then identify the maximum acceptable weight or force that corresponds to that desired percent. This is the value recommended for job design.

As an example, suppose you wish to design a lifting task that requires a box 49 cm wide that must be lifted 51 cm once per minute within the floor-to-knuckle zone. You desire to design this job to accommodate 75% of females. According to the data in Table 5, you would recommend that the box weigh no more than 11 kg. This weight would be acceptable to 75% of females and over 90% of males.

Multiple task analysis consisting of a lift, carry, and lower, has also been investigated for the Liberty Mutual data [40]. In this circumstance, it was observed that the maximum acceptable weight for the multiple task was lower than that for only the carrying task when performed separately, but not significantly different from the lifting or lowering maximum acceptable weights when performed separately. For this type of a multiple task, the maximum acceptable weight for the task should be the lowest maximum acceptable weight of the lift or lower as if it were performed separately. One should be careful, however, because the energy expenditure for the multiple task is probably under-

Table 5 Excerpt from the Liberty Mutual Tables for Maximum Acceptable Weight of Lift (kg) for Males and Females

				Floor level to knuckle height, one lift every							
Gender	Box width (cm)	Distance of lift (cm)	Percent capable	5 sec	9 sec	14 sec	1 min	2 min	5 min	30 min	8 hr
Males	49	51	90	7	9	10	14	16	17	18	20
			75	10	13	15	20	23	25	25	30
			50	*14*	17	20	27	30	33	34	40
			25	*18*	*21*	25	34	38	42	43	50
			10	*21*	*25*	*29*	40	45	49	50	59
Females	49	51	90	*6*	7	8	9	10	10	11	15
			75	*7*	9	9	11	12	12	14	18
			50	*9*	10	11	13	15	15	16	22
			25	*10*	*12*	13	16	17	17	19	26
			10	*11*	*14*	15	18	19	20	22	30

estimated when compared to performing the tasks separately. Similar results were reported by Jiang et al. [41].

3.4.8 Validation

3.4.8.1 Content Validity

The concept of content validity, also called face validity, addresses whether the content of the test is identical or highly similar to the content of the job. This is one of the major advantages of the psychophysical methodology, but it is important for the user to realize the limitations of the data, especially the caveats noted earlier.

It is noted that a 40 min test protocol is used to predict an 8 hr maximum acceptable weight or force. The researchers at Liberty Mutual examined this assumption by having subjects select their maximum acceptable weight according to the usual protocol, then having them continue to work, adjusting the weight or force as desired, for a total of 4 hr [39]. There was no statistically significant difference between the values selected after 40 min compared to the values selected after 4 hr. Karwowski and Yates reported similar results [42].

Mital also examined this issue relative to the Ayoub et al. data [43]. Mital found that the test subjects' estimates of their 8 hr maximum acceptable weights of lift were significantly greater than that selected at the end of an actual 8 hr period of work (an average 35% reduction). He "corrected" for this effect in his tables for 8 hr maximum acceptable weights of lift [37].

3.4.8.2 Criterion Related Validity

This type of validity, also called predictive validity, deals with the question of whether the results of the this type of job analysis predicts risk of future injury or illness. This is generally demonstrated by the presence of a statistically significant correlation between a test "score" and a particular outcome in an appropriately conducted epidemiological study.

There are two such studies relevant to the criterion-related validity of the psychophysical methodology.

Liberty Mutual Data. In 1978, Snook et al. published an investigation of three preventive approaches to low-back injuries in industry [44]. They distributed 200 questionnaires to Liberty Mutual Loss Prevention representatives throughout the United States. These representatives were asked to complete the questionnaire for the most recent compensable back injury. If the specific act or movement associated with the injury were some form of manual handling task, a task evaluation was completed to estimate the percent of capable working population that could perform the task without overexertion, i.e., what percent of the population could perform the task without exceeding their maximum acceptable weight or force. The investigators received 192 questionnaires, one with incomplete data.

They observed that 70% of these 191 low-back injuries were associated with manual materials handling tasks. They also compared the observed number of injuries to an expected number of injuries according to whether the percent capable population was greater than or less than 75%. This analysis is summarized as follows:

Capable population	Observed	Expected
≥ 75%	98	145.9
< 75%	93	45.1

The expected values were derived from control data that revealed that 23.6% of jobs involve handling tasks that less than 75% of the population could perform without overexertion.

$$X^2 = 66.6 \qquad p < 0.01$$

Based on these results, the authors concluded:

1. A worker is three times more susceptible to low-back injury if he or she performs a job that less than 75% of the working population can perform without overexertion.
2. At best, the ergonomic approach could reduce low-back injuries associated with manual material handling tasks by 67% by designing the jobs so that percent capable population was ≥75%. The remaining 33% of back injuries will occur regardless of the job demands.
3. Since only 50% of the industrial back injuries are related to manual materials handling tasks where the percent capable population is less than 75%, the overall reduction in low-back injuries would be 33%. This reduction would be higher if the percent capable population were raised to 90%.

Ayoub et al. Data. Ayoub and coworkers proposed the use of a severity index, called the job severity index (JSI), for purposes of validation [45]. The JSI is a ratio of job demands to worker capability. Since a job may consist of multiple tasks, they defined the JSI as a time-

and frequency-weighted average of the maximum weight required by each task divided by the task-specific worker capacity. Their validation studies included 101 jobs, performed by 385 males and 68 females, and involved four steps:

1. Selection of candidate jobs
2. Analysis of candidate jobs in terms of lifting requirements and morbidity data
3. Determination of the JSI for jobs and operators
4. Determination of the relationship between JSI and observed morbidity.

Individual JSIs were calculated for each worker that were subsequently grouped in to four categories: $0.00 \leq \text{JSI} < 0.75$; $0.75 \leq \text{JSI} < 1.5$; $1.5 \leq \text{JSI} < 2.25$; and $\text{JSI} \geq 2.25$.

The morbidity data was classified into five groups: musculoskeletal injuries to the back; musculoskeletal injuries to other parts of the body; surface-tissue injuries due to impact; other surface-tissue injuries; and miscellaneous injuries, and reported as incidence rates per 100 workers per year. Data for severity (days lost) and cost were also collected.

Their results revealed that the incidence of back injuries and the incidence of disabling back injuries increased substantially if the JSI was greater than or equal to 1.5. The relationships were nonlinear. The severity for disabling back injuries was increased if the JSI was greater than 2.25. The authors did not report any statistical analyses.

Another aspect of their validation involved classifying jobs according to the percent of capable population. Each job was categorized according to the percentage of the and population "overstressed," i.e., JSI greater than 1.5. The ranges were: $>75\%$; $<5\%$ and $\leq 75\%$; and $\leq 5\%$. They observed that the incidence of back injuries, incidence of disabling injuries, days lost per injury, and total cost increased as the percent of population "overstressed" increased. The authors did not report any statistical analyses.

Both Sets of Data. Another study that examined the predictive validity of the psychophysical methodology was published by Herrin et al. [46]. These investigators performed detailed biomechanical and psychophysical evaluations on 55 industrial jobs from five major industries. The psychophysical analyses involved determining the minimum percent of capable population from the Liberty Mutual tables for each individual task (PSY.MIN) as well as an average percent of capable population when the job involved multiple tasks (PSY.AVG). Additional comparison variables included the JSI and lifting strength ratio (LSR). These investigators modified the definition of JSI to represent a frequency- and time-weighted ratio of weights lifted compared to the average task-specific lifting strength of males and females, averaged across all tasks. By contrast, the lifting strength ratio represented the worst case scenario in that the LSR was the largest single ratio identified among all the tasks.

After the jobs were characterized as described above, injury and illness data for 6912 incumbent workers were monitored for two years retrospectively and one year prospectively (>12.6 million man-hours). Morbidity was categorized as contact incidents, musculoskeletal disorders (excluding the back), and back incidents, and expressed as incidence rates (number of incidents per 100 workers per year). Severity data was also examined (lost time versus no lost time).

The results revealed a significant negative correlation between the minimum percent capable population (PSY.MIN) and all three incidence rates, i.e., the incidence rates increased as the percent capable population decreased. A similar correlation was noted between PSY.MIN and severity. There was no correlation between the average percent capable population (PSY.AVG) with any incidence rate or severity. The incidence rates for musculoskeletal disorders and back disorders were positively and significantly correlated with the LSR. The LSR was also correlated with severity. The only correlated with severity, not incidence.

The authors offered the following conclusions:

1. Overexertion injuries can be related to physical job stresses.
2. Indices representing the extremes of the job requirements (PSY.MIN and LSR) are generally more predictive of risk than indices representing averages (PSY.AVG and JSI).
3. The percent capable population for the most stressful aspect of the job, either isometric or psychophysical, is the most simple index of this type.

3.4.9 Evaluation According to Criteria for Physical Assessment

3.4.9.1 Is It Safe to Administer?

According to Snook, there has been one compensable injury among the 119 industrial worker test subjects [47]. This single episode involved a chest wall strain associated with a high lift. It was also associated with four days restricted activity, but no permanent disability.

3.4.9.2 Does the Protocol Give Reliable Quantitative Values?

The Liberty Mutual protocol incorporates a criterion for test–retest reliability (maximum difference of 15%). Legg and Myles reported that 34% of their data did not meet this criterion [48]. In contrast, Gallagher reported that only 3% of tests in their study had to be repeated because of violating the 15% test–retest criterion [49]. Clearly, the maximum acceptable weights and forces are quantitative.

3.4.9.3 Is It Practical?

There are two major sources of impracticality associated with this type of strength assessment: (1) it is conducted in a laboratory, and (2) the duration of testing is somewhat prolonged compared to other strength assessment methods. It is possible, however, to have the subjects use objects that are actually handled in the workplace. Equipment is not very costly.

3.4.9.4 Is It Related to Specific Job Requirements (Content Validity)?

The content validity of this method of strength assessment is one of its greatest assets. One potential weakness, however, is its insensitivity to bending and twisting.

3.4.9.5 Does It Predict Risk of Future Injury or Illness (Predictive Validity)?

The results of two epidemiological studies suggest that selected indices derived from the psychophysical data are predictive of risk for contact injury, musculoskeletal disorders (excluding the back), and back disorders [44,45]. These indices are correlated to the severity of these injuries. A third study demonstrated predictive value [46]. It should be noted, however, that at high frequencies, test subjects selected weights and forces that often exceeded consensus criteria for acceptable levels of energy expenditure. In addition, test subjects may also select weights and forces that exceed consensus levels of acceptable disk compression.

3.5 PART IV: ISOKINETIC STRENGTH

3.5.1 Theory and Description of Isokinetic Strength Measurement

The concept of isokinetic measurement of strength was originally related by Hislop and Perrine [71]. Characteristics of an isokinetic exertionare constant velocity throughout a predetermined range of motion. Strictly speaking, a means of speed control, and not a load in the usual sense, is applied in isokinetic exertion [71]. However, load and resistance are definitely present in this technique. In this case, the load is a result of the energy absorption process performed by the device to keep the exertion speed constant. Energy cannot be dissipated through acceleration in isokinetic exercise, because this is prevented by the device. Because the energy is not dissipated in the process, it is converted into a resistive force, which varies in relation to the efficiency of the skeletal muscle.

Since the speed of motion is held constant in isokinetic exercise, the resistance experienced during a contraction is equivalent to the force applied throughout the range of motion. For this reason, the technique of isokinetic exercise has sometimes been referred to as *accommodating resistance exercise*. This type of exercise allows the muscle to contract at its maximum capability at all points throughout the range of motion. At the extremes of the range of motion of a joint, the muscle has the least mechanical advantage, and the resistance offered by the machine is correspondingly lower. Similarly, as the muscle reaches its optimal mechanical advantage, the resistance of the machine increases proportionally. It must be understood, however, that while isokinetic devices control the speed of the exertion, this does not assure a constant speed of muscle contraction.

It should be noted that while the speed of isokinetic contractions is constant during individual exertions, it is also possible to compare muscular performance over a wide range of isokinetic velocities. Increasing the isokinetic speed of contraction will place increasing demands on Type II muscle fibers (fast twitch and fast oxidative glycolytic).

3.5.2 Workplace Assessment

It is clear that isometric strength testing cannot substitute for dynamic strength assessment when examining highly dynamic occupational job demands. As most industrial work tasks contain a significant dynamic component, analysis of isokinetic strength capabilities would appear to offer some advantage to isometric testing in this regard. However, it must be recognized that isokinetic devices are not entirely realistic in comparison with free dynamic lifting, where subjects may use rapid acceleration to gain a weight lifting advantage.

The majority of isokinetic devices available on.the market focus on quantifying strength about isolated joints or body segments, for example, trunk extension and flexion (see Fig. 7). This may be useful for rehabilitation or clinical use, but isolated joint testing is generally not appropriate for evaluating an individual's ability to perform occupational lifting tasks. One should not make the mistake of assuming, for instance, that isolated trunk extension strength is representative of an individual's ability to perform a lift. In fact, lifting strength may be almost entirely unrelated to trunk muscle strength. Strength of the arms or legs (and not the trunk) may be the limiting factor in an individual's lifting strength. For this reason, machines that measure isokinetic strengths of isolated joints or body segments should not be used as a method of evaluating worker capabilities related to job demands in most instances.

Many investigators have used dynamic isokinetic lifting devices specifically designed to measure whole-body lifting strength [72,73] (see Fig. 8). These devices typically have a handle connected by a rope to a winch, which rotates at a specified isokinetic velocity when the handle is pulled. Studies using this type of device have demonstrated good correlations between isokinetic dynamic lift strength (i.e., a lift from floor to chest height) and the maximum weights individuals were willing to lift for infrequent tasks [72]. Thus, under certain circumstances, this device appears to possess some validity for assessment of job related dynamic lifting strength capabilities of individuals. However, many of these isokinetic lifting devices are limited to analysis of relatively simple lifting tasks (i.e., a simple sagittal plane lift). Unfortunately, such rudimentary lifting tasks are rare in industry. Some investigators have attempted to modify this type of instrument by providing means to mount it so that isokinetic strength can be measured in vertical, horizontal, and transverse planes [74]. In spite of efforts to improve the versatility of these devices, however, it remains clear that complex lifting tasks are not well simulated by current isokinetic apparatus.

3.5.3 Evaluation According to Criteria for Physical Assessment

3.5.3.1 Is It Safe to Administer?

Given proper procedures and supervision, isokinetic musculoskeletal testing appears to be a reasonably safe method of evaluating muscular strength and endurance. Certain risks associated with use of free weights, weight machines, and other isotonic methods

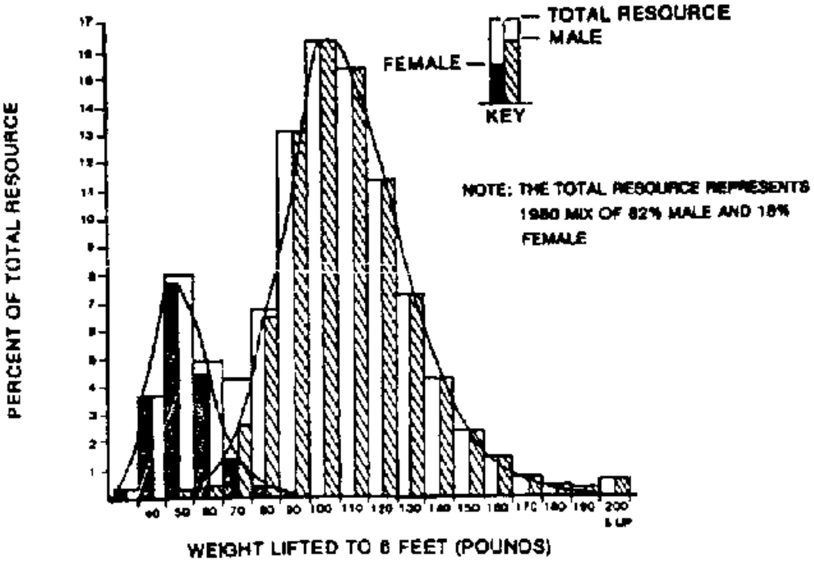


Figure 7 Many isokinetic devices are designed to evaluate isolated joint muscle strengths. Such devices can be of great benefit in a clinical setting, but may not be as conducive to workplace assessment procedures.

Figure 8 An isokinetic whole-body strength measurement system. This device allows the experimenter to assess various muscular strengths (such as those shown) at a constant velocity.

of assessing strength are not present in isokinetic testing. In addition, since the resistance or load experienced by the subject is directly related to the force the subject voluntarily applies, theoretically there would be decreased risk of injury due to overloading of the or musculature, because the subject can control his or her own effort. However, it should be noted that some investigators have reported that lower velocity isokinetic exertions may be painful [75,76].

Certain precautions have been suggested to reduce injury risk in performance of isokinetic musculoskeletal evaluations:

1. Warm-up and stretching of the involved muscle groups
2. Performance of 5–10 submaximal trial repetitions to assess proper alignment, subject comfort, and subject familiarization with the test requirements
3. Postexercise stretching
4. Ice/compression/elevation any time postexercise effusion or swelling occurs.

In addition, subjects should wear tennis or running shoes during isokinetic muscle testing when performing standing exertions.

Guidelines have been established by the American Academy of Orthopaedic Surgeons (AAOS) that should be met when testing dynamic muscle performance [77]. The following summarize the guidelines developed by the AAOS Human Performance Testing Task Force:

1. The equipment must be determined safe for both the subject and the tester.
2. The reliability and validity of the equipment should be documented.
3. The equipment should be designed to ensure freedom of movement with subject comfort, and isolation of the motion should be achieved via proper stabilization techniques.
4. Training and education in correct use of the equipment should be available.

3.5.3.2 Does It Give Reliable, Quantitative Values?

Several studies have reported on the reliability of values obtained using isokinetic devices. Results have generally indicated high reliability for isokinetic equipment. In a study examining the isokinetic movement of the knee extensors using a CYBEX II dynamometer, Johnson and Siegel [78] found reliability coefficients ranged from 0.93 to 0.99. Furthermore, these authors reported that reliability appeared to be affected more by testing over days than when comparing different trials performed on the same day. Pipes and Wilmore [79] reported test reliability in isokinetic exertions of a similar magnitude ($r = 0.92$–0.99) when testing bench press strength and leg press strength. Moffroid et al. [80] performed a test of reliability for torque measurements at various velocities with a CYBEX device and found that peak torque was reliably measured ($r = 0.999$) at velocities ranging from 4 to 12 rpm. Intratest, intertest, and intertester reliability of isokinetic strength measurements was examined in a study quantifying strength in children using a CYBEX dynamometer [81]. The authors concluded that none of these sources of measurement error constituted a significant source of inaccuracy.

While good reliability for the CYBEX dynamometer has been reported, some authors have expressed concern about a torque "overshoot" artifact that may appear in CYBEX torque measurements [82]. This artifact is evidenced as an initial prominent spike in the torque output curve, which is then followed by a series of progressively diminishing secondary oscillations. The cause of this phenomenon appears to be a result of "overspeeding" of the dynamometer's input lever during a free acceleration period prior to engagement of its resistance mechanism. The authors concluded that the prominent initial spikes represent inertial forces and should not be confused with actual muscle tension development. Proper signal damping procedures may suppress this "overshoot"; however, damping should not be used when absolute torque values are required.

Many other isokinetic devices have been developed since the introduction of the CYBEX in 1980. Most of these devices have demonstrated reliability similar to the CYBEX. Klopfer and Greij [75] analyzed the liability of torque production on the Biodex B-200 at high isokinetic velocities (300°–450° sec) and found that coefficients of ta correlation ranged from 0.95 to 0.97, reflecting a high degree of reliability of the test equipment. Other authors reported reliability of between 0.94 and 0.99 with the same equipment [83]. A study analyzing the reliability of the Kinetic Communicator (KINCOM) device reported intraclass correlation coefficients of 0.94–0.99 [84]. Reliability of the Lido isokinetic system appears somewhat lower than the others reported above, ranging from 0.83–0.94 [85]. The reliability of the Mini-Gym (the isokinetic device best suited to analysis of occupational tasks) does not appear to have been reported in the literature.

The foregoing data suggests that isokinetic strength testing equipment generally exhibits a high degree of reliability. However, it should be noted that results obtained using one system may not be comparable to results collected on other systems. Several studies have attempted to compare results between systems, and all have found significant differences. Torque values may vary as much as 10–15% when comparing different systems [86,87]. These discrepancies indicate that data collected on different devices cannot be compared, and that normative data generated on one system cannot be used on other systems.

3.5.3.3 Is It Practical?

Several issues may impact the practicality of using isokinetic devices to examine an individual's muscular capabilities. Not the least of these is the significant cost of purchasing an isokinetic measurement system. Many of the systems discussed in this section cost tens of thousands of dollars, which may render such systems impractical for many applications. Another important issue related to practicality in terms of job specific strength assessment is the ability of these devices to easily simulate a variety of occupational tasks. While certain isokinetic devices have been specifically designed to mimic lifting tasks [72], many are designed simply for quantification of strength of isolated muscle groups in a clinical setting without regard to accurate simulation of work tasks.

3.5.3.4 Is it related to specific job requirements?

The answer to this question depends upon the type of isokinetic device and how it is used. As discussed previously, isokinetic machines that test isolated muscle groups do not meet this criterion if the job requires use of many muscle groups or body segments. On the other hand, the Mini-Gym can be used to evaluate the dynamic strength necessary to perform many types of occupational tasks, and results of strength tests using this device appear to be related to lifting capacity, at least under certain conditions [72]. However, many industrial tasks are clearly too complicated to be evaluated using current isokinetic technologies. Great care must be taken to ensure that isokinetic strength measurements are appropriate for analysis of strength requirements associated with specific occupational tasks.

3.5.3.5 Does It Predict Risk of Future Injury or Illness?

A recent prospective epidemiological investigation analyzed whether isokinetic lifting strength was able to predict those at risk of occupational low-back pain (LBP) [82]. Subjects were required to perform maximal whole-body lifting exertions using an isokinetic linear lift task device, and were then followed for two years to evaluate whether this measure of strength was predictive of those who would experience LBP. Results of this study indicated that isokinetic lifting strength was a poor predictor of subsequent LBP or injury. However, it should be noted that no attempt was made in this study to compare job strength requirements to individual strength capabilities. Whether isokinetic strength tests can be used to predict future LBP when a careful comparison of job demands and individual strength capacity is made has yet to be determined.

3.6 SUMMARY

In spite of advances in measurement techniques and an explosive increase in the volume of research, our understanding of human strength remains in its introductory stages. It is clear that muscle strength is a highly complex and variable function dependent on a large number of factors. It is not surprising, therefore, that there are not only large differences in strength between individuals, but even within the same individual tested repeatedly on a given piece of equipment. The issue is compounded by the fact that correlations of strength among different muscle groups in the same individual are generally low, and that tests of isometric strength do not necessarily reflect the strength an individual might exhibit in a dynamic test. As a result of these and other influences, it is evident that great care needs to be exercised in the design, evaluation, reporting, and interpretation of muscular strength assessments.

Traditionally, tests of muscular strength were in the domain of the orthopedist, physical therapist, and exercise physiologist. However, such tests are also an important tool for the ergonomist due to the high strength demands required of workers in manual materials handling tasks. In some cases, it has been shown that task demands may approach or even exceed the strength that an individual is voluntarily willing to exert in a test of strength. In such cases, there is evidence to suggest that the likelihood of injury is significantly greater than when the task demands lie well within an individual's strength capacity. Because the relationship between strength capabilities, job demands, and musculoskeletal injury has been established, it becomes apparent that tests of muscular strength may be of benefit to the ergonomist both in the design of jobs, and in ensuring that individuals have sufficient strength to safely perform physically demanding jobs.

Several different strength assessment techniques have been employed for these purposes, each possessing unique characteristics and applicability to job design and/or worker selection procedures. The main purpose of this chapter has been to elucidate these strengths and weaknesses of the various procedures, so that tests of strength may be properly applied in the design of jobs and the selection of workers.

One of the crucial points emphasized in this chapter is that any test of strength used in job design or worker selection *must be directly related to the demands of the job* [89]. For example, if an occupational lifting task has a high dynamic component, a test of isometric strength is not likely to provide the data necessary for proper design of the job. Of course, use of dynamic strength tests to assess a job requiring isometric exertions would also be a misapplication. Another potential pitfall is the use of tests of strength on isolated muscle groups, and assuming that these tests are indicative of whole-body strength. For example, one might mistakenly assume that dynamic trunk extension strength is representative of a person's capability to perform a lifting task. However, an individual's lifting capacity may be entirely unrelated to trunk extension strength. Instead, lifting capacity may be limited by an individual's arm or leg strength, depending upon the task being performed.

It should be clear from the parts discussed in this chapter that tests of muscular strength are a tool that can be used in the prevention of occupational musculoskeletal disease. However, it is imperative that the use of these techniques be applied with a clear understanding of the advantages and limitations associated with each technique. The paragraphs that follow summarize the tests of muscular strength covered in this chapter. *Isometric strength* is defined as the capacity to produce force or torque with a voluntary isometric (muscles maintain a constant length) contraction. A characteristic of this type of strength measurement is the absence of body movement during the measurement period. Isometric strength testing has a long history, and it may be the easiest to measure and understand. The basic procedures for testing isometric strength are well established. Risk of injury appears to be small, and of relatively minor nature. Residual soreness of muscle groups tested is occasionally reported. Tests of isometric strength appear reliable, with test–retest variability on the order of 5–10%. The approach appears quite practical and has been applied in many industrial situations. The major limitation of isometric strength testing is in its inability to accurately model materials handling tasks that have a significant dynamic component. It is therefore recommended that tests of isometric strength be applied when there is little or no dynamic movement involved. In spite of this limitation, it should be duly noted that of all the procedures reviewed in this chapter, tests of isometric strength are the only strength tests that have shown the ability to predict individuals with a high risk of future injury or illness on physically stressful jobs [89]. The accuracy of this prediction appears to be dependent on the quality of the job evaluation on which the strength tests are based, and on the care with which the tests are administered.

Tests of *isoinertial strength* are defined as those in which the mass properties of an object are held constant, as in lifting a given weight (mass) over a predetermined distance. Several strength tests reviewed in this chapter possess the attribute in this definition. However, there are significant philosophical and procedural differences among the different isoinertial procedures in use, and the authors have subdivided isoinertial strength tests into maximal isoinertia strength tests [19,21,25], and psychophysical strength tests [12]. The following distinctions are made between these techniques:

1. In maximal isoinertial strength tests, the amount of weight lifted by the subject is *systematically adjusted by the experimenter*. In contrast, in psychophysical tests, *weight adjustment is freely controlled by the subject*.
2. Maximal isoinertial strength tests are designed to quickly establish an individual's *maximal strength* using a *limited number of lifting repetitions*, whereas psychophysical strength assessments are typically performed over a *longer duration of time* (usually at least 20 min), and instructions are that the subject select an *acceptable (submaximal) weight of lift*, not a maximal one.
3. Maximal isoinertial strength tests have traditionally been used as a *worker selection tool* (a method of matching physically capable individuals to demanding tasks). A primary focus of psychophysical methods has been to establish data that can be used for the purpose of *ergonomic job design* [12].

Two primary maximum isoinertial strength tests have been described. One involves the use of a modified weightlifting machine where a subject lifts a rack of unseen weights to various prescribed heights (often termed the LIFTEST). The other, the progressive isoinertial lifting evaluation (PILE), uses a standard lifting box, into which weights are placed incrementally until the lifting limit is reached. Both procedures appear to be safe to administer and remarkably free of injury. These techniques also appear to compare favorably to other strength tests in terms of test–retest reliability. Both tests are practical in that they require relatively inexpensive hardware, and can be administered quickly with minimal time needed for subject instruction and learning. The dynamic nature of the LIFTEST gives the procedure a similarity to certain industrial lifting tasks, and has correlated well with psychophysical test results [41].

A vast and expanding base of literature is devoted to psychophysical strength assessment. The psychophysical method, as applied to strength, has been used to determine maximum acceptable weights and forces associated with manual materials handling tasks for healthy adult male and female industrial workers [33,35]. The focus of this approach is to establish data that can be used to improve the design of manual materials handling activities. Psychophysical strength tests appear very safe, with isolated reports of muscle strain. Psychophysical results are very reproducible and seem to be related to low back pain [34]. The cost of the procedure is extremely low, except in the time that it takes to administer the tests. Of all the strength techniques reviewed in this chapter, the psychophysical approach is the one best suited to simulating specific industrial work tasks. However, it should be noted that at high lifting frequencies, test subjects may select weights and forces that exceed manual materials handling limits based on metabolic or disk compression criteria. Furthermore, there is some question as to whether psychophysical lifting tests are sensitive to bending and twisting motions, which are often associated with the onset of low-back pain. At this time, the use of psychophysical methods of strength assessment for the prediction of future risk of injury, illness, impairment, or disability for an individual has not been validated.

The characteristics of *isokinetic strength tests* are variable displacement and constant velocity of motion [71]. The majority of isokinetic devices focus on quantifying torques about isolated joints or body segments. Isolated joint testing may be most useful in rehabilitation or in clinical use, but is more limited in terms of evaluating occupational job demands. However, devices that measure isokinetic whole-body lifting strength, consisting of a handle connected by rope to a winch (which rotates at a specified isokinetic velocity) have also been developed. Studies using this type of device have shown good correlations between an isokinetic lift from floor to chest height and psychophysically acceptable weights for infrequent lifting tasks [72,74]. Given proper procedures and supervision, isokinetic strength tests appear to be a reasonably safe method of evaluating muscular strength and endurance. However, some investigators have indicated that low velocity isokinetic exertions may be painful [75]. There are numerous isokinetic devices on the market, and all appear to possess high reliability. The practicality of isokinetic strength testing may well hinge on the considerable cost associated with purchase of the equipment. Another issue in terms of practicality is the ability of isokinetic devices to easily simulate a variety of occupational tasks. Many industrial tasks are clearly too complicated to be evaluated using current isokinetic technologies. Thus far, prospective studies have shown that generic isokinetic lifting tests are poor predictors of future low back disorders [88]. Whether isokinetic tests can be used to predict injury or illness when careful comparisons of job demands and individual strength capabilities are performed has not yet been investigated.

A final point on strength assessment should be made. An individual's strength capability cannot be considered a fixed human attribute. Strength training regimens can increase an individual's strength capability by 30–40%. Whether such changes have a preventive effect when a person performs heavy physical work has yet to be established in epidemiologic studies; however, some anecdotal evidence supports the possibility [89].

REFERENCES

1. LS Caldwel, DB Chaffin, FN Dukes-Dobos, KHE Kroemer, LL Laubach, SH Snook, et al. A proposed standard procedure for static muscle strength testing. Am Ind Hyg Assoc J 35:201–206, 1974.
2. DB Chaffin. Ergonomics guide for the assessment of human static strength. Am Ind Hyg Assoc J 36:505–511, 1975.
3. M Ikai, AH Steinhaus. Some factors modifying the expression of strength. J Appl Physiol 16:157–163, 1991.
4. KHE Kroemer, WS Marras, JD McGlothlin, DR McIntyre, M Nordin. On the measurement of human strength. Int J Indust Ergon, 6:199–210, 1990.
5. TJ Stobbe. The development of a practical strength testing program in industry. Unpublished PhD dissertation, University of Michigan, Ann Arbor, MI, 1982.
6. DB Chaffin, GBJ Andersson. Occupational Biomechanics. 2nd ed. New York: John Wiley and Sons, 464–466, 1991.
7. Troup, JDG, JW Martin, DCEF Lloyd, Back pain in industry. A prospective study. Spine 6:61–69, 1981.
8. MC Battie, SJ Bigos, LD Fisher, TH Hansson, ME Jones, MD Wortley. Isometric lifting strength as a predictor of industrial back pain. Spine 14:851–856, 1989.
9. RA Mostardi, DA Noe, MW Kovacik, JA Porterfield. Isokinetic lifting strength and occupational injury: A prospective study. Spine 17(2):189–193, 1992.
10. DB Chaffin. Ergonomic basis for job-related strength testing. In: Disability Evaluation. SL Demeter, GBJ Anderson, GM Smith, eds. Louis, MO: Mosby, 1996, 159–167.
11. V Mooney, K Kenney, S Leggett, B Holmes. Relationship of Lumbar Strength in Shipyard Workers to Workplace Injury Claims. Spine 21:2001–2005, 1996.
12. SH Snook. The design of manual handling tasks. Ergonomics 21 (12):963–985, 1978.
13. DB Chaffin, GBJ Andersson. Occupational Biomechanics. 2nd ed. New York: John Wiley and Sons, pp. 105–106, 1991.
14. FT Schanne. Three dimensional hand force capability model for a seated person. Unpublished PhD dissertation, University of Michigan, Ann Arbor, MI, 1992.
15. TJ Stobbe, RW Plummer. A test–retest criterion for isometric strength testing. Proceedings of the Human Factors Society 28th Annual Meeting, Oct 22–26, 1984, San Antonio, TX, pp. 455–459, 1984.
16. DB Chaffin, GD Herrin, WM Keyserline. Pre-employment strength testing: an updated position. J Occupat Med 20(6): 403–408, 1978.
17. WM Keyserling, GD Herrin, DB Chaffin. Isometric strength testing as a means of controlling medical incidents on strenuous jobs. J Occupat Med 22(5):332–366, 1980.
18. KHE Kroemer. Development of LIFTEST: A dynamic technique to assess the individual capability to lift material. Final Report, NIOSH Contract 210-79-0041. Blacksburg, VA: Ergonomics Laboratory, IEOR Department, Virginia Polytechnic Institute and State University, 1982.
19. KHE Kroemer. An isoinertial technique to assess individual lifting capability. Hum Factors 25(5):493–506, 1983.
20. KHE Kroemer. Testing individual capability to lift material: repeatability of a dynamic test compared with static testing. J Safety Res 16(1):1–7, 1985.
21. JW McDaniel, RJ Shandis, SW Madole. Weight lifting capabilities of Air Force basic trainees. AFAMRL-TR-83-0001. Wright-Patterson AFBDH, Air Force Aerospace Medical Research Laboratory, 1983.
22. M Parnianpour, M Nordin, N Kahanovitz, V Frankel. The triaxial coupling of torque generation of trunk muscles during isometric exertions and the effect of fatiguing isoinertial movements on the motor output and movement patterns. Spine 13(9):982–992, 1988.
23. MM Ayoub, A Mital. Manual Materials Handling. London: Taylor and Francis, 1989, pp 241–242.
24. DB Chaffin, GBJ Andersson. Occupational Biomechanics. New York: John Wiley and Sons, 1991, pp 152–153.
25. TG Mayer, D Barnes, ND Kishino, G Nichols, RJ Gatchell, H Mayer, V Mooney. Progressive isoinertial lifting evaluation—I. A standardized protocol and normative database. Spine 13(9):993–997, 1988.
26. BC Jiang, JL Smith, MM Ayoub. Psychophysical modelling of manual materials-handling capacities using isoinertial strength variables. Hum Factors 28(6):691–702, 1986.
27. LT Ostrom, JL Smith, MM Ayoub. The effects of training on the results of the isoinertial 6-foot incremental lift strength test. Int J Indust Ergon 6:225–229, 1990.
28. JM Stevenson, JT Bryant, SL French, DR Greenhorn, GM Andrew, JM Thomson. Dynamic analysis of isoinertial lifting technique. Ergonomics 33(2): 161–172, 1990.
29. DO Myers, DL Gebhardt, CE Crump, EA Fleishman. Validation of the Military Entrance Physical Strength Capacity Test (MEPSCAT). U.S. Army Research

Institute Technical Report 610, NTIS No. AD-A142 169, 1984.

30. TG Mayer, D Barnes, G Nichols, ND Kishino, K Coval, B Piel, D Hoshino, RJ Gatchell. Progressive isoinertial lifting evaluation—II. A comparison with isokinetic lifting in a chronic low-back pain industrial population. Spine 13(8):998–1002, 1988.
31. SS Stevens. On the psychophysical law. Psychol Rev 64:153–181, 1957.
32. LA Jones. Perception of force and weight: Theory and research. Psychol Bull 100(1):29–42, 1986.
33. SH Snook. Psychophysical acceptability as a constraint in manual workbility of the psychophysical approach to manual materials handling activities. Ergonomics 29:237–248, 1986.
34. SH Snook. Psychophysical considerations in permissible loads. Ergonomics 28(1):327–330, 1985.
35. SH Snook, VM Ciriello. The design of manual handling tasks: revised tables of maximum acceptable weights and forces. Ergonomics 34(9):1197–1213, 1991.
36. MM Ayoub, NJ Bethea, S. Devanayagam, SS Asfour, GM Bakken, D Liles, A Mital, M Sherif. Determination and modeling of lifting capacity, final report. HEW (NIOSH) Grant No. 5-RO1-OH-00545-02.
37. A Mital. Comprehensive maximum acceptable weight of lift database for regular 8 h shifts. Ergonomics 27:1127–1138, 1978.
38. DD Thompson, DB Chaffin. Can biomechanically determined stress be perceived? Human Factors and Ergonomics Society, Proceedings of the 37th Annual Meeting, Seattle WA. 1993, pp 789–792.
39. SH Snook. Approaches to the control of back pain in industry: Job design, job placement, and education/training. Spine: State Art Rev 2:45–59, 1987.
40. VM Ciriello, SH Snook, AC Blick, PL Wilkinson. The effects of task duration on psychophysically determined maximum acceptable weights and forces. Ergonomics 33:187–200, 1990.
41. BC Jiang, JL Smith, MM Ayoub. Psychophysical modelling for combined manual materials-handling activities. Ergonomics 29(10):1173–1190, 1986.
42. W Karwowski, JW Yates. Reliability of the psychophysical approach to manual materials handling activities. Ergonomics 29:237–248, 1986.
43. A Mital. The psychophysical approach-in-manual lifting—a verification study. Hum Factors 25(5):485–491, 1983.
44. SH Snook, RA Campanelli, JW Hart. A study of three preventive approaches to low back injury. J Occup Med 20(7):478–481, 1978.
45. MM Ayoub, JL Selan, DH Liles. An ergonomics approach for the design of manual materials-handling tasks. Hum Factors 25(5):507–515, 1983.
46. GD Herrin, M Jaraiedi, CK Anderson. Prediction of overexertion injuries using biomechanical and psychophysical models. Am Ind Hyg Assoc J 47(6):322–330, 1986.
47. SH Snook. Assessment of human strength: Psychophysical methods. Roundtable presentation at the American Industrial Hygiene Conference and Exposition, Boston, 1992.
48. SJ Legg, WS Myles. Metabolic and cardiovascular cost, and perceived effort over an 8 hour day when lifting loads selected by the psychophysical method. Ergonomics 28(1):337–343, 1985.
49. S Gallagher. Acceptable weights and psychophysical costs of performing combined manual handling tasks in restricted postures. Ergonomics 34(7):939–952, 1991.
50. VM Ciriello, SH Snook. A study of size, distance, height, and frequency effects on manual handling tasks. Hum Factors 25(5):473–483, 1983.
51. A Mital, MM Ayoub. Effect of task variables and their interactions in lifting and lowering loads. Am Ind Hyg Assoc J 42:134–142, 1981.
52. SS Asfour, MM Ayoub, AM Genaidy. A psychophysical study of the effect of task variables on lifting and lowering tasks. J Hum Ergol 13:3–14, 1984.
53. A Garg, A Mital, SS Asfour. A comparison of isometric and dynamic lifting capability. Ergonomics 23(1):13–27, 1980.
54. A Mital, I Manivasagan. Maximum acceptable weight of lift as a function of material density, center of gravity location, hand preference, and frequency. Hum Factors 25(1):33–42, 1983.
55. SJ Legg, DR Haslam. Effect of sleep deprivation on self selected workload. Ergonomics 27(4):389–396, 1984.
56. JL Smith, BC Jiang. A manual materials handling study of bag lifting. Am Ind Hyg Assoc J 45(8):505–508, 1984.
57. A Mital, HF Fard. Psychophysical and physiological responses to lifting symmetrical and asymmetrical loads symmetrically and asymmetrically. Ergonomics 29(10):1263–1272, 1986.
58. A Mital. Maximum weights of asymmetrical loads acceptable to industrial workers for symmetrical lifting. Am Ind Hyg Assoc J 48(6):539–544, 1987.
59. A Mital. Psychophysical capacity of industrial workers for lifting symmetrical loads and asymmetrical loads symmetrically and asymmetrically for 8 hour work shifts. Ergonomics 35(718):745–754, 1992.
60. CG Drury, JM Deeb, B Hartman, S Wooley, CE Drury, S Gallagher. Symmetric and asymmetric manual materials handling. Part 1. Physiology and psychophysics. Ergonomics 32(5):467–489, 1989.
61. SL Legg, CM Pateman. Human capabilities in repetitive lifting. Ergonomics 28(1):309–321, 1985.
62. CG Drury, JM Deeb. Handle positions and angles in a dynamic lifting task. Part 2. Psychophysical measures and heart rate. Ergonomics 29(6):769–777, 1986.
63. A Mital. Maximum acceptable weights of lift acceptable to male and female industrial workers for extended work shifts. Ergonomics 27(11):1115–1126, 1984.

64. JE Fernandez, MM Ayoub, JL Smith. Psychophysical lifting capacity over extended periods. Ergonomics 34(1):23–32, 1991.

65. A Mital, F Aghazadeh. Psychophysical lifting capabilities for overreach heights. Ergonomics 30(6):901–909, 1987.

66. A Mital, L-W Wang. Effects on load handling of restricted and unrestricted shelf opening clearances. Ergonomics 32(1):39–49, 1989.

67. A Mital. Patterns of differences between the maximum weights of lift acceptable to experienced and inexperienced materials handlers. Ergonomics 30(8):1137–1147, 1987.

68. JL Smith, MM Ayoub, JW McDaniel. Manual materials handling capabilities in non-standard postures. Ergonomics 35(7/8):807–831, 1992.

69. S Gallagher, WS Marras, TG Bobick. Lifting in stooped and kneeling postures: Effects on lifting capacity, metabolic costs, and electromyography of eight trunk muscles. Int J Ind Ergon 3:65–76, 1988.

70. S Gallagher, CA Hamrick. Acceptable workloads for three common mining materials. Ergonomics 35(9):1013–1031, 1992.

71. H Hislop, JJ Perrine. The isokinetic concept of exercise. Phys Therapy 47:114–117, 1967.

72. JL Pytel, E Kamon. Dynamic strength test as a predictor for maximal and acceptable lift. Ergonomics 24(9):663–672, 1981.

73. ND Kishino, TG Mayer, RJ Gatchel, MM Parish, C Anderson, L Gustin, V Mooney. Quantification of lumbar function: Part 4: isometric and isokinetic lifting simulation in normal subjects and low-back dysfunction patients, Spine 10(10):921–927, 1985.

74. A Mital, R Vinayagormoothy. Three-dimensional dynamic strength measuring device: a prototype. Am Ind Hyg Assoc J 45:B9–B12, 1984.

75. DA Klopfer, SD Greij. Examining quadriceps/hamstrings performance at high velocity isokinetics in untrained subjects. J Orthop Sports Phys Therapy 10:18–22, 1988.

76. RC Elsner, LR Pedegana, J Lang. Protocol for strength testing and rehabilitation of the upper extremity. J Orthop Sports Phys Therapy 4:229, 1983.

77. American Academy of Orthopaedic Surgeons (AAOS). Human Performance Testing Task Force, October 1988.

78. J Johnson, D Siegel. Reliability of an isokinetic movement of the knee extensors. Res Q 49:88–90.

79. TV Pipes, JH Wilmore. Isokinetic vs. Isotonic strength training in adult men. Med Sci Sports Exercise 7:262–271, 1975.

80. M Moffroid, R Whipple, J Hofkosh, et al. A study of isokinetic exercise. Phys Therapy 49:735, 1969.

81. GE Molnar, J Alexander, N Gutfield. Reliability of quantitative strength measurements in children. Arch Phys Med Rehab 60:218, 1979.

82. AA Sapega, JA Nicholas, D Sokolow, D., A Saraniti. The nature of torque “overshoot” in CYBEX isokinetic dynamometry. Med Sci Sports Exercise 14(5):368–375, 1982.

83. KE Wilk, RE Johnson. The reliability of the Biodex B-200 (abstract). Phys Therapy 68:792, 1988.

84. M Farrell, JG Richards. Analysis of the reliability and validity of the , kinetic communicator exercise device. Med Sci Sports Exercise 18:44–49.

85. J Lord, S Aitkins, M McCrory, M., et al. Reliability of the Lido Digital Isokinetic system for the measurement of muscular strength (abstract). Phys Therapy 67:757, 1987.

86. KE Wilk, RE Johnson, et al. A comparison of peak torque values of knee extensor and flexor muscle groups using Biodex, Cybex, and Kin-Com isokinetic dynamometers. Phys Therapy 67:789, 1987.

87. KE Timm. Comparison of knee extensor and flexor group performance using the Cybex 340 and the Merac isokinetic dynamometers. Phys Therapy 69:389, 1989.

88. RA Mostardi, DA Noe, MW Kovacik, JA Porterfield. Isokinetic Lifting Strength and Occupational Injury: A prospective study. Spine 17(2): 189–193.

Chapter 10.1

Engineering Economy

Thomas R. Huston
University of Cincinnati, Cincinnati, Ohio

1.1 INTRODUCTION

Engineering is the profession that is devoted to the application of scientific knowledge for practical purposes. Through the application of scientific knowledge, engineers are continually developing products, processes, and services for the benefit of society. Engineers have been instrumental in many of the advances of society. For example, the state of modern transportation can be linked to the efforts of engineers.

While undertaking such pursuits, the engineer is typically faced with a variety of alternatives. These alternatives may include material selections, the degree of computer automation, the selection of an applicable safety system, and the means of manufacturing. Each alternative will have inherent technical advantages and disadvantages that the engineer must evaluate. The evaluation of any alternative will also have to consider the constraints of the particular problem or project.

The engineer will typically be well informed about the technical aspects of various alternatives. However, the engineer must also have a sound understanding of the economic feasibility of the various alternatives. Indeed, money is a scarce resource that must be allocated in a prudent fashion.

This chapter provides a foundation in the basic principles of engineering economics. Through the application of these basic principles, the engineer will be able to address economic issues. One such issue is the economic feasibility of alternatives. Engineering economics offers a means to assess any receipts and disbursements associated with an alternative. Such an assessment will consider the magnitude and timing of the receipts and disbursements. Inflation and taxes may also be factors that enter into the economic evaluation of an alternative. The basic principles of engineering economics also provide methods for the comparison of alternatives and the subsequent selection of an optimal alternative. For example, an engineer may be confronted with the selection of machinery from a variety of sources. As another example, the engineer may face the economic decision of manufacturing a part versus purchasing a part.

It should also be recognized that there are limitations to engineering economics. Certain problems may not have the potential to be evaluated properly in economic terms. Some problems may be highly complex wherein economics is a minor consideration. Still other problems may not be of sufficient importance to warrant engineering economic analysis.

1.2 ELEMENTARY CONCEPTS OF ENGINEERING ECONOMICS

There are several fundamental concepts that form a foundation for the application of the methods of engineering economics. One fundamental concept is the recognition that money has a time value. The value of a given amount of money will depend upon when it is received or disbursed. Money possessed in the present will have a greater value than the same amount of money at some point in the future.

It would be preferable to receive \$1000 in the present rather than receiving \$1000 five years hence. Due to the earning power of money, the economic value of \$1000 received at the present will exceed the value of \$1000 received five years in the future. The \$1000 received today could be deposited into an interest bearing savings account. During the intervening period of five years, the \$1000 would earn additional money from the interest payments and its accumulated amount would exceed \$1000.

The time value of money is also related to the purchasing power of money. The amount of goods and services that a quantity of money will purchase is usually not static. Inflation corresponds to a loss in the purchasing power of money over time. Under the pressures of inflation, the cost of a good or service will increase. As an example, during the period of 1967 to 1997 the cost of a U.S. first-class postage stamp rose to 32 cents from 5 cents. Deflation is the opposite condition of inflation. Historically, inflationary periods have been far more common than periods of deflation.

A fundamental concept that is related to the time value of money is interest. Money is a valuable commodity, so businesses and individuals will pay a fee for the use of money over a period of time. Interest is defined as the rental fee paid for the use of such a sum of money. Interest is usually quantified by the interest rate where the interest rate represents a percentage of the original sum of money that is periodically applied. For instance, a financial institution may charge 1% per month for a borrowed sum of money. This means that at the end of a month, a fee of 1% of the amount borrowed would have to be paid to the financial institution.

The periodic payment of interest on a loan represents a cash transaction. During such a transaction, a borrower would view the associated interest as a disbursement while the interest would be a receipt for the lender. In engineering economics analysis, a point of view must be selected for reference. All analysis should proceed from a sole viewpoint.

Engineering economic analysis should also only consider and assess feasible alternatives. Alternatives that ordinarily would be feasible may be infeasible due to the particular constraints of a problem or project. A frequently overlooked alternative is the do-nothing alternative. Under the do-nothing alternative, the option of doing nothing is preferable to any of the other feasible alternatives.

It is the inherent differences between alternatives that must be evaluated. Indeed it is the differences in alternatives that will lead to the selection of an optimal alternative. Such an evaluation will utilize money as a common unit of measurement to discern the differences between alternatives. The evaluation of alternatives should also utilize a uniform time horizon to reveal the differences in alternatives.

It is essential to recognize that any decisions about alternatives will only affect the present and the future. Therefore, past decisions and any associated costs should be ignored in engineering economic analysis. The associated costs from past decisions are known as sunk costs. Sunk costs are irrelevant in engineering economic analysis.

1.3 ECONOMIC EQUIVALENCE AND CASH FLOW FORMULAS

1.3.1 Economic Equivalence

In engineering, two conditions are said to be equivalent when each condition produces the same effect or impact. The concept of equivalence also pertains to engineering economics. Two separate alternatives will have economic equivalence whenever each alternative possesses the same economic value. Any prospective economic equivalence between two alternatives will be dependent upon several factors. One factor is the respective magnitudes of the cash flow for each alternative. Another factor is the timing of the receipts and disbursements for each alternative. A third factor is the interest rate that accounts for the time value of money.

Through a combination of these factors, two cash flows that differ in magnitude may possess the same inherent economic value. The concept of economic equivalence is revealed through the cash flows associated with a routine loan. Suppose an individual borrowed \$10,000 at 6% compounded annually to be repaid in annual instalments of \$2374 over five years. One cash flow would be the sum of \$10,000 at the present. The other cash flow would entail five annual payments of \$2374 that totaled \$11,870. Although each cash flow occurs at distinct points in time and has a different magnitude, both cash flows would be equivalent at the interest rate of 6% compounded annually.

1.3.2 Simple and Compound Interest

There are different ways in determining the amount of interest that a sum of money will produce. One way is simple interest. Under simple interest, the amount of interest accrued, I, on a given sum of money, P, is calculated by

$$I = Pni \tag{1}$$

where P is the principal amount, n the number of interest periods, and i the interest rate. Hence with simple interest, a sum of money would increase to

$$F = P + I = P + Pni \tag{2}$$

With simple interest, any interest earned during an interest period does not earn additional interest in forthcoming interest periods.

In contrast, with compound interest, the interest is. determined by the principal sum of money and on any interest that has accumulated to date. So any previous interest will earn interest in the future. For example, if a sum of money, P, is deposited into an interest-bearing account at an interest rate, i, after one period the amount of money available, F, would be determined by

$$F = P(1 + i) \tag{3}$$

If the sum of money were deposited for two periods, the amount of money available, F, would be determined by

$$F = (P(1 + i))(1 + i) = P(1 + i)^2 \tag{4}$$

In general, the amount of money, F, that would accumulate with n additional periods would be

$$F = P(1 + i)^n \tag{5}$$

Compound interest is more prevalent in financial transactions than simple interest, although simple interest is often encountered in bonds.

1.3.3 Cash Flow Diagrams and End-of-Period Convention

In engineering, diagrams are frequently drawn to help the individual understand a particular engineering issue. A cash flow diagram is often used to depict the magnitude and the timing of cash flows in an engineering economics issue. A cash flow diagram presumes a particular point of view. A horizontal line is used to represent the time horizon, while vertical lines from the horizontal line depict cash flows. An upward arrow indicates a receipt of money, while a downward arrow is a disbursement (see Fig. 1).

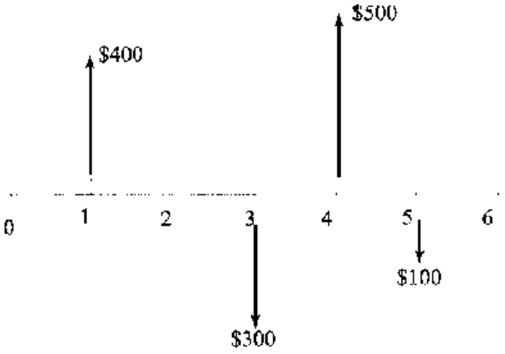


Figure 1 Cash flow diagram.

In this chapter, there is an assumption that cash flows will be discrete and will occur at the end of a period. Continuous flows of cash over a period will not be considered. An extensive discussion of continuous cash flows is offered in the references.

1.3.4 Cash Flow Patterns

In financial transactions, a cash flow may undertake a variety of patterns. The simplest pattern is the single cash flow. Under this cash flow pattern, a single present amount is transformed into a single future amount (see Fig. 2).

The uniform series is another cash flow pattern. With this pattern, all of the cash flows are of the same magnitude and the cash flows occur at equally spaced time intervals (see Fig. 3).

A cash flow that increases or decreases by the same amount in each succeeding period would be a uniform gradient cash flow pattern (see Fig. 4). Whereas, a cash flow that increases or decreases by the same percentage in each succeeding period would be a geometrical gradient cash flow pattern. (see Fig. 5).

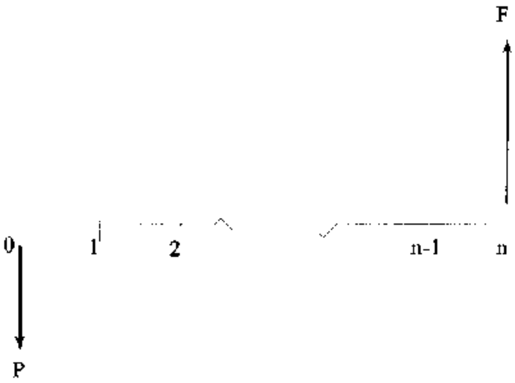


Figure 2 Present amount and future amount.

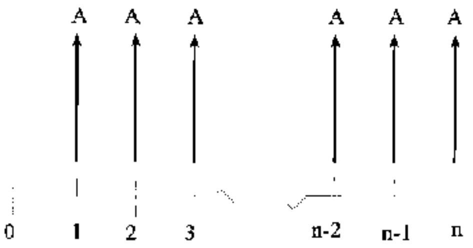


Figure 3 Uniform series.

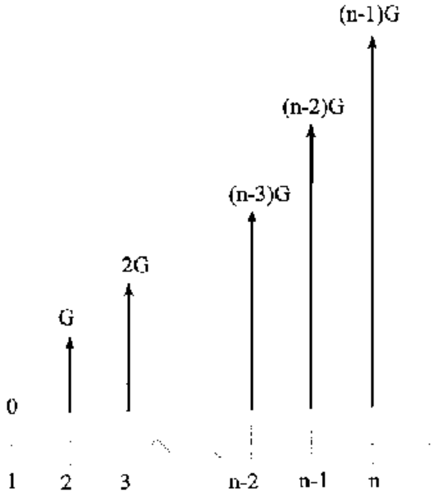


Figure 4 Uniform gradient.

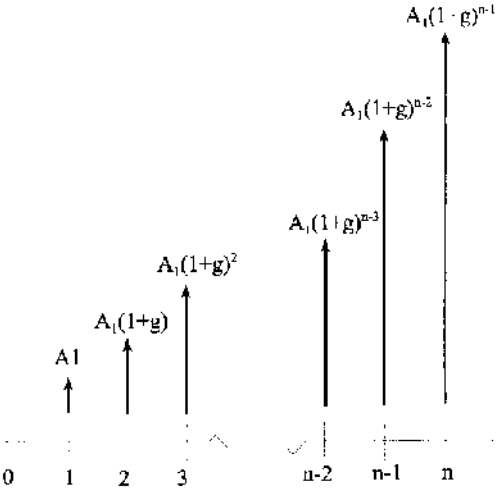


Figure 5 Geometrical gradient series.

An irregular cash flow pattern would occur whenever the cash flow did not maintain one of the aforementioned regular patterns. Occasionally, a portion of an irregular cash flow pattern may exhibit a regular pattern (see Fig. 6). In Fig. 6, the overall cash flow pattern would be classified as irregular but in the final three years there is a uniform series pattern.

Equivalent relationships between the various cash flow patterns may be developed mathematically. Due to the time value of money, such relationships will be dependent upon the prevailing interest rates and the duration of the associated cash flows.

1.3.5 Single-Payment Compound Amount Factor

Due to the time value of money, a single cash flow, P, will increase over time to an equivalent future value, F. The future value, F, will depend upon the length of time, the prevailing interest rate, and the type of interest. If the single cash flow, P, is invested at a constant compound interest rate, i, for a given number of interest periods, n, then the future value, F, will be determined by Eq. (5). Eq. (5) may be rewritten to introduce the following notation:

$$F = P(1+i)^n = P(F|P, i, n) \qquad (6)$$

The conversion factor, $(F|P, i, n)$, is referred to as the single-payment compound amount factor. It is interpreted as "to find the equivalent future amount, F, given the present amount, P, at the interest rate, i, for n periods." The single-payment compound amount factor, $(F|P, i, n)$, is simply the quantity $(1+i)^n$. The evaluation of the single-payment compound amount factor is an easy calculation. Tabulated values of the single-payment compound amount factor for interest rates of 1%, 8%, and 10% may be found in Tables 1 to

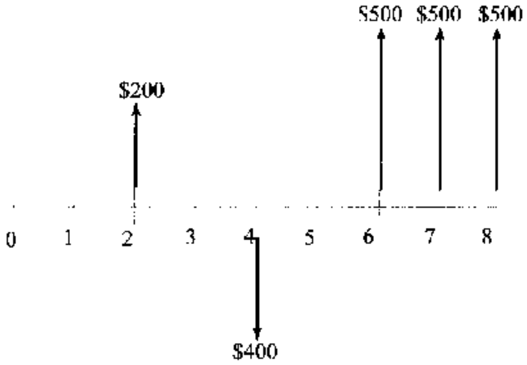


Figure 6 Irregular cash flow.

3. Note, other economic equivalence factors can also be found in Tables 1 to 3.

Example 1. *A sum of $5000 is deposited into an account that pays 10% interest compounded annually. To determine the future value of the sum of money 20 years hence, utilize Eq. (6):*

$$F = \$5000(1 + 0.10)^{20} = \$33{,}637$$

1.3.6 Single Payment Present-Worth Amount Factor

Through simple algebra, Eq. (6) can be solved for P, wherein the resulting factor, $(P|F, i, n)$, is designated as the single-payment present worth factor:

$$P = F(1+i)^{-n} = F(P|F, i, n) \tag{7}$$

Example 2. *In the settlement of a litigation action, a boy is to receive a lump sum of $250,000 10 years in the future. What is the present worth of such a payment presuming an annual compound interest rate of 8%?*

$$P = \$250{,}000(1 + 0.08)^{-10} = \$115{,}798$$

Example 3. *What annual rate of interest was earned if an investment of $11,000 produced a value of $21,000 after 10 years?*

$$F = P(1+i)^n$$

$$\$21{,}000 = \$11{,}000(1+i)^{10}$$

$$1.909 = (1+i)^{10}$$

$$i = 1.909^{1/10} - 1 = 0.066 = 6.6\%$$

1.3.7 Compound Amount Factor

Formulas can be derived that relate a single future cash flow pattern, F, to a uniform series of cash flow patterns, A. The equivalent future amount, F, that a uniform cash flow pattern, A, will produce is

$$F = A\left[\frac{(1+i)^n - 1}{i}\right]$$

$$= A(F|A, i, n) \tag{8}$$

Example 4. *A design engineer expects to collect $5000 per year on patent royalties. The patent remains in effect for the next 10 years. What is the future value of this series of patent royalties if it is deposited into a fund that earns 10% compounded annually?*

$$F = A(F|A, i, n) = \$5{,}000(F|A, 10\%, 10)$$

$$= \$5000(15.937)$$

$$= \$79{,}685$$

1.3.8 Sinking Fund Factor

Similarly, an equivalent uniform series cash flow pattern, A, can be obtained from a single future cash flow pattern, F:

$$A = F\left[\frac{i}{(1+i)^n - 1}\right]$$

$$= F(A|F, i, n) \tag{9}$$

1.3.9 Present-Worth Factor

It is often desirable to be able to relate a present amount, P, to a uniform series cash flow pattern, A. The present worth factor converts a uniform series cash flow pattern, A, to a single present amount, P. The formula for the present-worth factor is

$$P = A\left[\frac{(1+i)^n - 1}{i(1+i)^n}\right]$$

$$A = A(P|A, i, n) \tag{10}$$

1.3.10 Capital Recovery Factor

The capital recovery factor is the reciprocal of the present-worth factor. This conversion factor transforms a single present amount, P, to a uniform series of cash flows:

$$A = P\left[\frac{i(1+i)^n}{(1+i)^n - 1}\right]$$

$$= P(A|P, i, n) \tag{11}$$

Example 5. *To purchase a new machine, a manufacturer secures a $50,000 loan to be paid off in annual payments over the next five years. If the interest rate of the loan is 8% compounded annually, what is the periodic payment that the manufacturer must pay?*

$$A = P(A|P, I, n) = \$50{,}000(A|P, 8\%, 5)$$

$$= \$50{,}000(0.2505)$$

$$= \$12{,}525$$

Table 1 One Percent: Compound Interest Factors

N	$F\|P$	$P\|F$	$F\|A$	$A\|F$	$P\|A$	$A\|P$	$A\|G$
1	1.0100	0.99010	1.0000	1.00000	0.99010	1.01000	0.00000
2	1.0201	0.98030	2.0100	1.49751	1.97040	0.50751	0.49751
3	1.0303	0.97059	3.0301	0.33002	2.94099	0.34002	0.99337
4	1.0406	0.96098	4.0604	0.24628	3.90197	0.25628	1.48756
5	1.0510	0.95147	5.1010	0.19604	4.85343	0.20604	1.98010
6	1.0615	0.94205	6.1520	0.16255	5.79548	0.17255	2.47098
7	1.0721	0.93272	7.2135	0.13863	6.72819	0.14863	2.96020
8	1.0829	0.92348	8.2857	0.12069	7.65168	0.13069	3.44777
9	1.0937	0.91434	9.3685	0.10674	8.56602	0.11674	3.93367
10	1.1046	0.90529	10.4622	0.09558	9.47130	0.10558	4.41792
11	1.1157	0.89632	11.5668	0.08645	10.36763	0.09645	4.90052
12	1.1268	0.88745	12.6825	0.07885	11.25508	0.08885	5.38145
13	1.1381	0.87866	13.8093	0.07241	12.13374	0.08241	5.86073
14	1.1495	0.86996	14.9474	0.06690	13.00370	0.07690	6.33836
15	1.1610	0.86135	16.0969	0.06212	13.86505	0.07212	6.81433
16	1.1726	0.85282	17.2579	0.05794	14.71787	0.06794	7.28865
17	1.1843	0.84438	18.4304	0.05426	15.56225	0.06426	7.76131
18	1.1961	0.83602	19.6147	0.05098	16.39827	0.06098	8.23231
19	1.2081	0.82774	20.8109	0.04805	17.22601	0.05805	8.70167
20	1.2202	0.81954	22.0190	0.04542	18.04555	0.05542	9.16937
21	1.2324	0.81143	23.2392	0.04303	18.85698	0.05303	9.63542
22	1.2447	0.80340	24.4716	0.04086	19.66038	0.05086	10.09982
23	1.2572	0.79544	25.7163	0.03889	20.45582	0.04889	10.56257
24	1.2697	0.78757	26.9735	0.03707	21.24339	0.04707	11.02367
25	1.2824	0.77977	28.2432	0.03541	22.02316	0.04541	11.48312
26	1.2953	0.77205	29.5256	0.03387	22.79520	0.04387	11.94092
27	1.3082	0.76440	30.8209	0.03245	23.55961	0.04245	12.39707
28	1.3213	0.75684	32.1291	0.03112	24.31644	0.04112	12.85158
29	1.3345	0.74934	33.4504	0.02990	25.06579	0.03990	13.30444
30	1.3478	0.74192	34.7849	0.02875	25.80771	0.03875	13.75566
31	1.3613	0.73458	36.1327	0.02768	26.54229	0.03768	14.20523
32	1.3749	0.72730	37.4941	0.02667	27.26959	0.03667	14.65317
33	1.3887	0.72010	38.8690	0.02573	27.98969	0.03537	15.09946
34	1.4026	0.71297	40.2577	0.02484	28.70267	0.03484	15.54410
35	1.4166	0.70591	41.6603	0.02400	29.40858	0.03400	15.98711
36	1.4308	0.69892	43.0769	0.02321	30.10751	0.03321	16.42848
37	1.4451	0.69200	44.5076	0.02247	30.79951	0.03247	16.86822
38	1.4595	0.68515	45.9527	0.02176	31.48466	0.03176	17.30632
39	1.4741	0.67837	47.4123	0.02109	32.16303	0.03109	17.74278
40	1.4889	0.67165	48.8864	0.02046	32.83469	0.03046	18.17761
41	1.5038	0.66500	50.3752	0.01985	33.49969	0.02985	18.61080
42	1.5188	0.65842	51.8790	0.01928	34.15811	0.02928	19.04237
43	1.5340	0.65190	53.3978	0.01873	34.81001	0.02873	19.47231
44	1.5493	0.64545	54.9318	0.01820	35.45545	0.02820	19.90061
45	1.5648	0.63905	56.4811	0.01771	36.09451	0.02771	12.32730
46	1.5805	0.63273	58.0459	0.01723	36.72724	0.02723	20.75235
47	1.5963	0.62646	59.6263	0.01677	37.35370	0.02677	21.17578
48	1.6122	0.62026	61.2226	0.01633	37.97396	0.02633	21.59759
49	1.6283	0.61412	62.8348	0.01591	38.58808	0.02591	22.10778
50	1.6446	0.60804	64.4632	0.01551	39.19612	0.02551	22.43635
55	1.7285	0.57853	72.8525	0.01373	42.14719	0.02373	24.50495
60	1.8167	0.55045	81.6697	0.01224	44.95504	0.02224	26.53331
65	1.9094	0.52373	90.9366	0.01100	47.62661	0.02100	28.52167
70	2.0068	0.49831	100.6763	0.00993	50.16851	0.01993	30.47026
75	2.1091	0.47413	110.9128	0.00902	52.58705	0.01902	32.37934
80	2.2167	0.45112	121.6715	0.00822	54.88821	0.01822	34.24920
85	2.3298	0.42922	132.9790	0.00752	57.07768	0.01752	36.08013
90	2.4486	0.40839	144.8633	0.00690	59.16088	0.01690	37.87245
95	2.5735	0.38857	157.3538	0.00636	61.14298	0.01638	39.62648

Table 2 Eight Percent: Compound Interest Factors

N	*F*|*P*	*P*|*F*	*F*|*A*	*A*|*F*	*P*|*A*	*A*|*P*	*A*|*G*
1	1.0800	0.92593	1.0000	1.00000	0.92593	1.08000	0.00000
2	1.1664	0.85734	2.0800	0.48077	1.78326	0.56077	0.48077
3	1.2597	0.79383	3.2464	0.30803	2.57710	0.38803	0.94874
4	1.3605	0.73503	4.5061	0.22192	3.31213	0.30192	1.40396
5	1.4693	0.68058	5.8666	0.17046	3.99271	0.25046	1.84647
6	1.5869	0.63017	7.3359	0.13632	4.62288	0.21632	2.27635
7	1.7138	0.58349	8.9228	0.11207	5.20637	0.19207	2.69366
8	1.8509	0.54027	10.6366	0.09401	5.74664	0.17401	3.09852
9	1.9990	0.50025	12.4876	0.08008	6.24689	0.16008	3.49103
10	2.1589	0.46319	14.4866	0.06903	6.71008	0.14903	3.87131
11	2.3316	0.42888	16.6455	0.06008	7.13896	0.14008	4.23950
12	2.5182	0.39711	18.9771	0.05270	7.53608	0.13270	4.59575
13	2.7196	0.36770	21.4953	0.04652	7.90378	0.12652	4.94021
14	2.9372	0.34046	24.2149	0.04130	8.24424	0.12130	5.27305
15	3.1722	0.31524	27.1521	0.03683	8.55948	0.11683	5.59446
16	3.4259	0.29189	30.3243	0.03289	8.85137	0.11298	5.90463
17	3.7000	0.27027	33.7502	0.02963	9.12164	0.10963	6.20375
18	3.9960	0.25025	37.4502	0.02670	9.37189	0.10670	6.49203
19	4.3157	0.23171	41.4463	0.02413	9.60360	0.10413	6.76969
20	4.6610	0.21455	45.7620	0.02185	9.81815	0.10185	7.03695
21	5.0338	0.19866	50.4229	0.01983	10.01680	0.09983	7.29403
22	5.4365	0.18394	55.4568	0.01803	10.20074	0.09803	7.54118
23	5.8715	0.17032	60.8933	0.01624	10.37106	0.09642	7.77863
24	6.3412	0.15770	66.7648	0.01498	10.52876	0.09498	8.00661
25	6.8485	0.14602	73.1059	0.01368	10.67478	0.09368	8.22538
26	7.3964	0.13520	79.9544	0.01251	10.80998	0.09251	8.43518
27	7.9881	0.12519	87.3508	0.01145	10.93516	0.09145	8.63627
28	8.6271	0.11591	95.3388	0.01049	11.05108	0.09049	8.82888
29	9.3173	0.10733	103.9659	0.00962	11.15841	0.08962	9.01328
30	10.0627	0.09938	113.2832	0.00883	11.25778	0.08883	9.18971
31	10.8677	0.09202	123.3459	0.00811	11.34980	0.08811	9.35843
32	11.7371	0.08520	134.2135	0.00745	11.43500	0.08745	9.51967
33	12.6760	0.07889	145.9506	0.00685	11.51389	0.08685	9.67370
34	13.6901	0.07305	158.6267	0.00630	11.58693	0.08630	9.82075
35	14.7583	0.06763	172.3168	0.00580	11.65457	0.08580	9.96107
36	15.9682	0.06262	187.1021	0.00534	11.71719	0.08534	10.09490
37	17.2456	0.05799	203.0703	0.00492	11.77518	0.08492	10.22246
38	18.6253	0.05369	220.3159	0.00454	11.82887	0.08454	10.34401
39	20.1153	0.04971	238.9412	0.00419	11.87858	0.08419	10.45975
40	21.7245	0.04603	259.0565	0.00386	11.92461	0.08386	10.56992
41	23.4625	0.04262	280.7810	0.00356	11.96723	0.08356	10.67473
42	25.3395	0.03946	304.2453	0.00329	12.00760	0.08329	10.77441
43	27.3666	0.03654	329.5830	0.00303	12.04324	0.08303	10.86915
44	29.5560	0.03383	356.9496	0.00280	12.07707	0.08280	10.95917
45	31.9204	0.03133	386.5056	0.00259	12.10840	0.08259	11.04465
46	34.4741	0.02901	418.4261	0.00239	12.13741	0.10823	11.12580
47	37.2320	0.02686	452.9002	0.100221	12.16427	0.08221	11.20280
48	40.2106	0.02487	490.1322	0.00204	12.18914	0.08204	11.27584
49	43.4274	0.02303	530.3427	0.00189	12.21216	0.08189	11.34509
50	46.9016	0.02132	573.7702	0.00174	12.23348	0.08174	11.41071
55	68.9139	0.01451	848.9232	0.00118	12.31861	0.08118	11.69015
60	101.2571	0.00988	1,253.2133	0.00080	12.37655	0.08080	11.90154
65	148.7798	0.00672	1,847.2481	0.00054	12.41598	0.08054	12.06016
70	218.6064	0.00457	2,720.0801	0.00037	12.44282	0.08037	12.17832
75	321.2045	0.00311	4,002.5566	0.00025	12.46108	0.08025	12.26577
80	471.9548	0.00212	5,886.9354	0.00017	12.47351	0.08017	12.33013
85	693.4565	0.00144	8,655.7061	0.00012	12.48197	0.08012	12.37725
90	1,018.9151	0.00098	12,723.9386	0.00008	12.48773	0.08008	12.41158

Table 3 Ten Percent: Compound Interest Factors

N	$F\|P$	$P\|F$	$F\|A$	$A\|F$	$P\|A$	$A\|P$	$A\|G$
1	1.1000	0.90909	1.0000	1.00000	0.90909	1.10000	0.00000
2	1.2100	0.82645	2.1000	0.47619	1.73554	0.57619	0.47619
3	1.1300	0.75131	3.3100	0.30211	2.48685	0.40211	0.93656
4	1.4641	0.68301	4.6410	0.21547	3.16987	0.31547	1.38117
5	1.6105	0.62092	6.1051	0.16380	3.79079	0.26380	1.81013
6	1.7716	0.56447	7.7156	0.12961	4.35528	0.22961	2.22356
7	1.9487	0.51316	9.4872	0.10541	4.86842	0.20541	2.62162
8	2.1436	0.46651	11.4359	0.08744	5.33493	0.18744	3.00448
9	2.3579	0.42410	13.5796	0.07364	5.75902	0.17364	3.37235
10	2.5937	0.38554	15.9374	0.06275	6.14457	0.16276	3.72546
11	2.8531	0.35049	18.5312	0.05396	6.49508	0.15396	4.06405
12	3.1384	0.31863	21.3843	0.04676	6.81369	0.14676	4.38840
13	3.4523	0.28966	24.5227	0.04078	7.10336	0.14078	4.69879
14	3.7975	0.26333	27.9750	0.03575	7.36669	0.13575	4.99553
15	4.1772	0.23939	31.7725	0.03147	7.60608	0.13147	5.27893
16	4.5960	0.21763	35.9497	0.02782	7.82371	0.12782	5.54934
17	5.0545	0.19784	40.5447	0.02466	8.02155	0.12466	5.80710
18	5.5599	0.17986	45.5992	0.02193	8.20141	0.12193	6.05256
19	6.1159	0.16351	51.1591	0.01955	8.36492	0.11955	6.28610
20	6.7275	0.14864	57.2750	0.01746	8.51356	0.11746	6.50808
21	7.4002	0.13513	64.0025	0.01562	8.64869	0.11562	6.71888
22	8.1403	0.12285	71.4027	0.01401	8.77154	0.11401	6.91889
23	8.9543	0.11168	79.5430	0.01257	8.88322	0.11257	7.10848
24	9.8497	0.10153	88.4973	0.01130	8.98474	0.11130	7.28805
25	10.8347	0.09230	98.3471	0.01017	9.07704	0.11017	7.45798
26	11.9182	0.08391	109.1818	0.00916	9.16095	0.10916	7.61865
27	13.1100	0.06728	121.0999	0.00826	9.23722	0.10826	7.77044
28	14.4210	0.06934	134.2099	0.00745	9.30657	0.10745	7.91372
29	15.8631	0.06304	148.6309	0.00673	9.36961	0.10673	8.04886
30	17.4494	0.05731	164.4940	0.00608	9.42691	0.10608	8.17623
31	19.1943	0.05210	181.9434	0.00550	9.47901	0.10550	8.29617
32	21.1138	0.04736	201.1378	0.00497	9.52638	0.10497	8.40905
33	23.2252	0.04306	222.2515	0.00450	9.56943	0.10450	8.51520
34	25.5477	0.03914	245.4767	0.00407	9.60857	0.10407	8.61494
35	28.1024	0.03558	271.0244	0.00369	9.64416	0.10369	8.70860
36	30.9127	0.03235	299.1268	0.00334	9.67651	0.10334	8.79650
37	34.0039	0.02941	330.0395	0.00303	9.70592	0.10303	8.87892
38	37.4043	0.02673	364.0434	0.00275	9.72265	0.10275	8.95617
39	41.1448	0.02430	401.4478	0.00249	9.75696	0.10249	9.02852
40	45.2593	0.02209	442.5926	0.00226	9.77905	0.10226	9.09623
41	49.7852	0.02009	487.8518	0.00205	9.79914	0.10205	9.15958
42	54.7637	0.01826	537.6370	0.00186	9.81740	0.10186	9.21880
43	60.2401	0.01660	592.4007	0.00169	9.83400	0.10169	9.27414
44	66.2641	0.01509	652.6408	0.00153	9.84909	0.10153	9.32582
45	72.8905	0.01372	718.9048	0.00139	9.86281	0.10139	9.37405
46	80.1795	0.01247	791.7953	0.00126	9.87528	0.10126	9.41904
47	88.1975	0.01134	871.8749	0.00115	9.88662	0.10115	9.46099
48	96.0172	0.01031	960.1723	0.00104	9.89693	0.10104	9.50009
49	106.7190	0.00937	1,057.1896	0.00095	9.90630	0.10095	9.53651
50	117.3909	0.00852	1,163.9085	0.00086	9.91481	0.10086	9.57041
55	189.0591	0.00529	1,880.5914	0.00053	9.94711	0.10053	9.70754
60	304.4816	0.00328	3,034.8164	0.00033	9.96716	0.10033	9.80229
65	490.3707	0.00204	4,893.7073	0.00020	9.97961	0.10020	9.86718
70	789.7470	0.00127	7,887.4696	0.00013	9.98734	0.10013	9.91125
75	1,271.8954	0.00079	12,708.9537	0.00008	9.99214	0.10008	9.94099
80	2,048.4002	0.00049	20,474.0021	0.00005	9.99512	0.10005	9.96093
85	3,298.9690	0.00030	32,979.6903	0.00003	9.99697	0.10003	9.97423
90	5,313.0226	0.00019	53,120.2261	0.00002	9.99812	0.10002	9.98306
95	8,556.6760	0.00012	85,556.7605	0.00001	9.99883	0.10001	9.98890
100	13,780.6123	0.00007	137,796.1234	0.00001	9.99927	0.10001	9.99274

1.3.11 Uniform Gradient Series Factor

As previously discussed, a cash flow series is not always uniform. Gradient series are frequently encountered in engineering economics. Formulas for conversion factors of gradient series have likewise been developed. Specifically, a uniform gradient series can be expressed as a uniform series of cash flows by

$$A = G\left[\frac{(1+i)^n - in - 1}{i(1+i)^n - i}\right] = A(A|G, i, n) \qquad (12)$$

1.3.12 Geometrical Gradient Present-Worth Factor

Cash flow series that increase or decrease by a constant percentage, g, with each succeeding period can be converted to a present amount by the geometric gradient present worth factor.

$$P = \frac{A_1}{1+g}\left[\frac{(1+g^*)^n - 1}{g^*(1+g)^n}\right] = \frac{A_1}{1+g}(P|A, g^*, n) \qquad (13)$$

where

$$g^* = \left[\frac{(1+i)}{(1+g)} - 1\right]$$

Example 6. *A manufacturer has established a new production line at an existing facility. It has been estimated that the additional energy costs for the new production line are \$5,000 for the first year and will increase 3% for each subsequent year. The production line is expected to have a life span of 10 years. Given an annual compound interest rate of 5% what is the present worth of the energy costs for the new production line?*

First calculate the value of g given that g = 3% and i = 5%:*

$$g^* = \left[\frac{(1+i)}{(1+g)} - 1\right] = \left(\frac{1.05}{1.03}\right) - 1 = 0.0194175$$

then

$$P = \frac{A_1}{1+g}\left[\frac{(1+g^*)^n - 1}{g^*(1+g^*)^n}\right]$$

$$= \frac{\$5000}{1.03}\left[\frac{(1.10194175)^{10} - 1}{0.0194175(1.0194175)^{10}}\right]$$

$$P = \$43{,}738$$

1.3.13 Frequency of Compounding and Its Impact on Equivalence

Compounding periods may assume a variety of durations. Interest can be compounded annually, semiannually, quarterly, monthly, daily, or continuously. Perhaps, the most common compounding period is annual. Similarly, the flow of funds also can occur over a variety of periods. For example, the periodic payments on a loan may be monthly.

Hence, there are three conditions that can occur, concerning the frequency of the compounding periods and the frequency of the periods for the cash flow. First, the frequency of the compounding periods and that of the cash flow are synchronized. Secondly, the compounding periods are shorter than the periods for the cash flow. Third, the compounding periods are longer than the corresponding periods of the cash flow.

If the periods of the compounding and the flow of funds are synchronized, the aforementioned conversion factors can be utilized to determine any equivalent cash flow. When the compounding periods and the periods of the cash lows are not synchronized, then intermediate steps to synchronize the periods must be undertaken prior to utilizing the aforementioned conversion factors.

Example 7. *What is the present value of a series of annual payments of \$90,000 over 10 years at the rate of 12% compounded monthly?*

Convert i = 1%/month to an effective annual interest rate:

$$i_e = (1+0.01)^{12} - 1 = 0.126825$$

$$P = A(P|A, 0.126825, 10)$$

$$= \$90{,}000\left(\frac{(1.126825)^{10} - 1}{0.126825(1.126825)^{10}}\right)$$

$$= \$494{,}622$$

For the condition where the compounding periods are less frequent than the cash flows, it should be noted that interest is not earned on funds that are not on deposit for the entire interest period. To synchronize the timing of the flows of funds with the compounding periods, any cash receipt or disbursement is moved to the end of its respective time period. With the movement of cash receipts and disbursements to the end of the time periods, economic equivalence can be determined with the use of the aforementioned conversion factors.

1.3.14 Amortized Loans

The capital needed to finance engineering projects will not always be available through retained earnings. Indeed, money will often have to be borrowed. There are many types of loans that exist, but this chapter will focus upon the standard amortized loan. With an amortized loan, the loan is repaid through installments over time.

The most prevalent amortized loan has monthly installments with interest that is compounded monthly. Also, the monthly installments are fixed. Each installment consists of a portion that pays the interest on the loan and a portion that repays the outstanding balance. With each succeeding installment, the interest portion will diminish, while the portion devoted to the repayment of the outstanding balance will increase.

The magnitude of an installment payment is determined through the use of the capital recovery conversion factor. In short, the payment, A, is found by

$$A = P(A|P, i, n) \tag{14}$$

Noting that each installment payment consists of an interest portion and a remaining balance portion, the following notation is introduced:

I_j = interest payment in period j

Pr_j = principal payment in period j

B_j = outstanding balance at end of period j

The interest portion of any installment payment is simply the product of the outstanding balance times the prevailing interest rate:

$$I_j = (B_{j-1})i \tag{15}$$

The portion of the installment that may be applied to the outstanding balance:

$$\mathrm{Pr}_j = A - I_j \tag{16}$$

Example 8. *A consulting firm obtains a \$10,000 loan to purchase a computer workstation. The terms of the loan are 12 months at a nominal rate of 12% compounded monthly. What is the monthly installment payment? How does the interest portion of the installment payment vary monthly?*

The installment payment is calculated by merely applying the capital recovery conversion factor, $(A|P, i, n)$:

$$A = \$10{,}000(A|P, 1\%, 12) = \$10{,}000(0.08885) \qquad \text{from Eq. (14)}$$
$$= \$888.50$$

The interest portion of the first installment would be

$$I_1 = (B_{1-1})i = (\$10{,}000)(0.01) \qquad \text{from Eq. (15)}$$
$$= \$100.00$$

Hence, the portion of the first installment applied to the principle would be the difference between A and I_1:

$$\mathrm{Pr}_1 = A - I_1 = \$888.50 - 100.00 \qquad \text{from (16)}$$
$$= \$788.50$$

The new outstanding balance would be

$$B_1 = \$10,000 - 788.50$$
$$= \$9211.50$$

Through an iterative process, the values for I_j and B_j can be found for the remaining 11 months. Obviously, the iterative nature of this problem is ideal for a computer application.

Installment no.	Payment (\$)	Principal (\$)	Interest (\$)	Balance (\$)
1	888.50	788.50	100.00	9211.50
2	888.50	796.39	92.11	8415.11
3	888.50	804.35	84.15	7610.76
4	888.50	812.40	76.10	6798.36
5	888.50	820.52	67.98	5977.34
6	888.50	828.73	59.77	5148.61
7	888.50	837.02	51.48	4311.59
8	888.50	845.39	43.11	3466.20
9	888.50	853.84	34.66	2612.36
10	888.50	862.38	26.12	1749.98
11	888.50	871.01	17.49	878.97
12	888.50	870.19	8.78	0.00

There are also formulas that enable one to determine the interest portion of any installment payment without having to engage an iterative solution:

$$I_j = (B_{j-1})i + A(P|A, i, n - j + 1)i \tag{17}$$

The corresponding remaining balance after $n - j$ payments may also be found via the following formula:

$$B_j = A(P|A, i, n - j) \tag{18}$$

Likewise, the principal payment for a particular installment would be obtained by subtracting the interest paid for a particular installment from the periodic payment:

$$\mathrm{Pr}_j = A - I_j \tag{19}$$

Returning to Example 8, the interest portion of the sixth payment may be found as follows:

$$\begin{aligned} I_6 &= A(P|A, i, n-j+1)i \\ &= \$888.50(P|A, 1\%, 12-6+1)0.01 \\ &= \$888.50(P|A, 1\%, 7)0.01 \qquad \text{from Eq. (17)} \\ &= \$888.50(6.7282)0.01 \\ &= \$59.77 \end{aligned}$$

1.4 ECONOMIC EVALUATION OF ALTERNATIVES

Often an engineer will be faced with the responsibility of determining the economic feasibility of various projects and propositions known as alternatives. In short, the engineer will have to make a decision on whether to proceed with an alternative. With a thorough understanding of cash flow patterns and the compounding of interest, one may apply a variety of techniques to evaluate various alternatives.

The application of these techniques requires that one be able to classify alternatives. Alternatives are classified as independent whenever the decision to proceed with the alternative or to reject the alternative has no bearing on other prospective alternatives. For example, the decision for a consulting firm to purchase a new computer system would ordinarily be unrelated to the firm's decision as to whether the firm should utilize a particular long-distance carrier for its telecommunication system.

Alternatives may also be classified as mutually exclusive. Such a condition exists when there are a series of alternatives from which only one alternative may be selected. If an engineer had to select a machine for a workstation from three distinct machines each having unique first costs, maintenance costs, and salvages, then this would be a condition where the alternatives were mutually exclusive. With mutually exclusive alternatives, the selection of an alternative prevents the selection of another alternative.

Often when identifying alternatives, an individual must include the do-nothing alternative. The do-nothing alternative simply represents the opportunity to maintain the existing conditions. In many instances, after the careful evaluation of a series of alternatives, the optimal decision will be to do-nothing or to maintain the existing conditions. The selection of the do-nothing alternative will preserve the scarce resource of capital.

The comparison of various alternatives involves the estimation of cash flows for each alternative. These estimated cash flows also extend over several time periods. A decision will have to made as to the duration of the planning horizon. The planning horizon represents the time period over which the alternatives will be evaluated. The selection of the planning horizon is important. If the planning horizon is too short, one runs the risk of rejecting alternatives that are initially expensive but generate large returns in the future. Conversely, a planning horizon that is too long can result in an entity collapsing before it reaps any benefits from accepted alternatives.

Further, the basic concept of time value of money must be incorporated into the evaluation of alternatives. This is accomplished through the selection of an interest rate that will be used to adjust the various cash flows in the panning horizon. This interest rate has been identified with a variety of names: minimum attractive rate of return (MARR), discount rate, return on capital, and cost of money. In this chapter, the term MARR will be used.

The determination of the value of the MARR is important. The value of the MARR should not be arbitrarily assigned. The MARR should recognize the cost of capital and should compensate for the risk associated in adopting an alternative. If the MARR is set unnecessarily high, an entity may needlessly reject worthwhile projects. Similarly, if the MARR is set too low, an entity can be exposed to the potential of investing in projects that are expensive and wasteful.

1.4.1 Present-Worth Method

This technique for evaluating economic alternatives entails the conversion of any pertinent estimated cash flows to the present. The cash flows are converted by the methods previously discussed in this chapter. In short, all cash flows are converted to an equivalent P pattern that is referred to as the present worth (PW). The conversions utilize a chosen MARR and a specified planning horizon. Each alternative must be evaluated over the same planning horizon. If the economic alternative is an independent alternative, then the alternative is accepted by entity whenever the present worth has a value greater than zero.

Example 9. *A consulting company is considering undertaking a project. The initial cash outlay for the 10 year project would be \$50,000. The project is estimated to yield \$8000 per year for 10 years. If the MARR is 10%, should the project be undertaken?*

$$\begin{aligned}\mathrm{PW} &= -\$50{,}000 + \$8000(P|A, 10\%, 10)\\ &= -\$50{,}000 + \$8000(6.1446)\\ &= -\$843.20\end{aligned}$$

Due to the negative present worth, this project should be rejected.

When faced with mutually exclusive alternatives, the optimal alternative is the alternative with the highest present worth. Indeed, the present worth of each alternative can be used to rank the alternatives. It should also be noted that on occasion each of the mutually exclusive alternatives may have a negative present worth. In such a situation, one would select the alternative that was the least costly by choosing the alternative that had the highest present worth.

Example 10. *Two different machines are being considered by a manufacturing company. Due to constraints, the manufacturing company must select one of the two machines. Machine A has an initial cost of \$75,000 and an estimated salvage of \$25,000 after five years. The annual operating costs for Machine A are assessed at \$7500. Machine B has an initial cost of \$50,000 and its salvage is negligible after five years. Its operating costs are \$9000 per year. Given a MARR of 10%, which machine should the company select?*

Machine A:

$$\begin{aligned}\mathrm{PW} &= -\$75{,}000 + \$25{,}000(P|F, 10\%, 5)\\ &\qquad - \$7500(P|A, 10\%, 5)\\ \mathrm{PW} &= -\$75{,}000 + \$25{,}000(0.6209) - \$7500(3.7908)\\ \mathrm{PW} &= -\$87{,}909\end{aligned}$$

Machine B:

$$\begin{aligned}\mathrm{PW} &= -\$50{,}000 - \$9000(P|A, 10\%, 5)\\ \mathrm{PW} &= -\$50{,}000 - \$9000(3.7908)\\ \mathrm{PW} &= -\$84{,}117\end{aligned}$$

Therefore, Machine B is preferred over Machine A.

1.4.2 Annual-Worth Method

This technique is similar to the present-worth technique. However, this technique involves the conversion of the estimated cash flows into a uniform annual cash flow that is known as the annual worth (AW). The conversion of the cash flows is based on an identified MARR and a specified time horizon. Each alternative must be evaluated over the same planning horizon. An independent alternative will be accepted, if its annual worth exceeds zero.

For mutually exclusive alternatives, the annual worth of each alternative provides a ranking. The alternative with the greatest annual worth is the optimal alternative. It is also possible, that each of the alternatives may have a negative annual worth. The best alternative still would be the alternative that had the greatest annual worth. This would be the least costly alternative.

Example 11. *Two different machines are being considered by a manufacturing company. Due to constraints, the manufacturing company must select one of the two machines. Machine A has an initial cost of \$75,000 and an estimated salvage of \$25,000 after five years. The annual operating costs for Machine A are assessed at \$7500. Machine B has an initial cost of \$50,000 and its salvage is negligible after five years. Its operating costs are \$9000 per year. Given a MARR of 10%, which machine should the company select? Utilize the annual worth approach.*

Machine A:

$$\begin{aligned}\mathrm{AW} &= -\$75{,}000(A|P, 10\%, 5)\\ &\qquad + \$25{,}000(A|F, 10\%, 5) - \$7500\\ \mathrm{AW} &= -\$75{,}000(0.2638) + \$25{,}000(0.1638)\\ &\qquad - \$7500\\ \mathrm{AW} &= -\$23{,}190\end{aligned}$$

Machine B:

$$\begin{aligned}\mathrm{AW} &= -\$50{,}000(A|P, 10\%, 5) - \$9000\\ \mathrm{AW} &= -\$50{,}000(0.2638) - \$9000\\ \mathrm{AW} &= -\$22{,}190\end{aligned}$$

Hence, Machine B is preferred to Machine A. Also note the consistency between the annual worth and present-worth methods. See Example 10.

1.4.3 Rate of Return Method

For independent alternatives, this technique relies on the concept of determining the interest rate where an alternative's receipts will be equivalent to its disbursements. The interest rate is known as the rate of return (ROR). The rate of return is then compared to the MARR. If the rate of return exceeds the MARR, then the alternative is viewed favorably and funds are expended for it.

Example 12. *A consulting company is considering undertaking a project. The initial cash outlay for the project would be $50,000. The project is estimated to yield $8000 per year for 10 years. If the MARR is 10%, should the project be undertaken? Solve using the rate-of-return approach.*

The ROR is the interest rate where

$$\mathrm{PW} = 0 = -\$50{,}000 + \$8000(P|A, i, 10)$$

Via trial and error:

$$\mathrm{PW}(i = 9\%) = -\$50{,}000 + \$8000(6.4177) = \$1341$$

$$\mathrm{PW}(i = 10\%) = -\$50{,}000 + \$8000(6.1446) = -\$843$$

Thus ROR is between 9 and 10%. Using interpolation the ROR is found to be 9.6%. The project is rejected because the ROR (9.6%) is less than the MARR (10%). Note the consistency between methods of evaluating alternatives. See Example 9.

For mutually exclusive alternatives, the optimal alternative is not decided from the individual rates of return of each alternative. Rather, an incremental analysis is employed. The incremental analysis compares pairs of alternatives. First, the alternatives are ranked according to the initial investments. Then for the alternative that has the smallest initial investment, its rate of return is calculated. Provided that its rate of return is greater than the MARR, then it is accepted as viable alternative. The viable alternative is then compared to next most expensive alternative. The comparison is based on the incremental additional investment and the incremental additional cash flows. If the incremental investment yields a rate of return greater than the MARR, then the more expensive alternative is selected. Conversely, if the incremental investment does not have a rate of return greater than the MARR, then the more expensive alternative is rejected. This pairwise comparison continues until all of the alternatives have been examined.

Example 13. *Consider four mutually exclusive alternatives, each of which has an eight-year useful life:*

	First cost ($)	Annual income ($)	Salvage value ($)
A	500	61	375
B	400	60	250
C	300	50	250
D	250	61	0

If the MARR is 8%, and utilizing the incremental rate of return approach, which alternative if any should be selected?

Alternative D vs. Do-nothing:

$$\begin{aligned}\mathrm{PW} &= -\$250 + \$61(P|A, 8\%, 8)\\ &= -\$250 + \$61(55.7466) = \$100.54\end{aligned}$$

Conclude that the incremental ROR > MARR*; hence accept Alternative D and reject do-nothing.*

Alternative C vs. Alternative D:

$$\begin{aligned}\mathrm{PW} &= (-\$300 - \$250) + (\$50 - \$61)(P|A, 8\%, 8)\\ &= -\$50 + (-\$11)(5.7466) = \$21.86\end{aligned}$$

Conclude that the incremental ROR > MARR*; hence accept Alternative C and reject Alternative D.*

Alternative B vs. Alternative C:

$$\begin{aligned}\mathrm{PW} &= (-\$400 - \$300)\\ &\quad + (\$60 - \$50)(P|A, 8\%, 8)\\ &= (-\$100) + \$10(5.7466) = -\$42.53\end{aligned}$$

Conclude that the incremental ROR < MARR*; hence reject Alternative B and keep Alternative C.*

Alternative A vs. Alternative C:

$$\begin{aligned}\mathrm{PW} &= (-\$500 - \$300) + (\$61 - \$50)(P|A, 8\%, 8)\\ &= (-\$200) + \$11(5.7466) = \$136.78\end{aligned}$$

Conclude that the incremental ROR < MARR*; hence reject Alternative A and keep Alternative C.*

Through the pairwise comparison of the incremental ROR, Alternative C is accepted as the optimal alternative.

1.4.4 Benefit–Cost Ratio

This technique operates on the simple concept that in order for an alternative to be deemed worthwhile it benefits must outweigh its costs. To make such a comparison requires that the benefits and costs be presented in equivalent economic terms. Ordinarily, the benefits and costs are expressed as either equivalent P patterns or equivalent A patterns. These equivalent P patterns or A patterns are determined using a given MARR and a stated planning horizon.

For independent alternatives, the benefits are then compared to the costs by means of a ratio. If the ratio of benefits to costs exceeds unity, then the alternative should be accepted.

Example 14. *A local government is evaluating a construction proposal for a roadway. The initial cost of the roadway is $1,650,000. It is estimated that the annual maintenance costs on the roadway will be $75,000. The estimated annual benefits to be derived from the roadway are $310,000. The useful life of the roadway is 20 years without a salvage. Using the benefit–cost approach, should the road be constructed with a 10% MARR?*

Annual projected benefits: $310,000
Annual project costs:

$$\$1{,}650{,}000(A|P, 10\%, 20) + \$75{,}000 =$$

$$\$1{,}650{,}000(0.1175) + \$75{,}000 = \$269{,}075$$

Benefit–cost ratio:

$$\frac{B}{C} = \frac{\$310{,}000}{\$269{,}075} = 1.15$$

Based on the benefit–cost ratio being greater than 1, the roadway should be constructed.

Mutually exclusive alternatives require an incremental analysis. One cannot select the optimal alternative by merely examining individual benefit–cost ratios. Initially, the alternatives are ranked in ascending order of equivalent first costs. The first viable alternative is then found by selecting the alternative with the smallest initial costs that has an individual benefit–cost ratio greater than 1. Once a viable alternative is found, then any remaining alternatives are evaluated on a pairwise basis to analyze whether additional costs are justified by the additional benefits. Throughout this procedure, once a viable alternative is found, it remains the alternative of choice unless the incremental pairwise analysis yields a superior alternative. Then the superior alternative becomes the alternative of choice. This incremental pairwise comparison continues until all alternatives have been examined.

1.4.5 Payback Method

This is a technique that is often used due to its simplicity and its ease of application. In short, the payback method determines the length of time for an alternative to pay for itself. Under the most common form of the payback method, any relevant cash flows are not adjusted for their inherent time value. For mutually exclusive alternatives, the optimal alternative would be the one with the shortest payback.

There are inherent disadvantages to this common form of the payback method. It ignores the time value of money. Also, the common form of the payback method ignores the duration of the alternatives.

Example 15. *Examine the following four alternatives. Note that each alternative has a payback period of two years. However, the alternatives are obviously not equivalent.*

Year end	Alt. I ($)	Alt. II ($)	Alt. III ($)	Alt. IV ($)
0	−1500	1500	−1500	−1500
1	750	500	1000	0
2	750	1000	500	1500
3	750	1500	0	1500
4	750	2000	0	1500
5	750	2500	0	1500

Another form of the payback method is known as the discounted payback method. Here, the cash flows are converted by means of a MARR. The alternatives are then evaluated by the length of time that it takes for the alternative to pay for itself. However, the discounted payback method still fails to account for the duration of the alternatives.

Both payback methods provide estimates that may be useful in explaining economic alternatives. However, due to the inherent flaws with these methods, it is recommended that the payback methods only be used as an ancillary technique.

1.5 AFTER-TAX ANALYSIS

The consideration of taxes must often be included in the evaluation of economic alternatives. In such an evaluation, taxes are simply another expenditure. Indeed, an alternative that may initially appear to be viable may lose its viability with the inclusion of taxes.

The magnitude of this taxation depends upon the prevailing federal, state, and local tax laws. These tax laws have been passed by legislatures so that these governments can operate. Taxation occurs in many different forms. A partial listing of various forms of taxation includes federal income tax, state income tax, local income tax, local property tax, state and local sales tax, federal excise tax, and federal and state gasoline tax.

The topic of taxes is complex. Tax laws are continually changing due to political forces and underlying economic conditions. In this chapter, the concentration will be upon federal income taxes. The techniques introduced will be applicable notwithstanding the inconstant nature of taxes.

1.5.1 Depreciation

The term depreciation has several meanings. In one sense, depreciation refers to the deterioration of an asset. For example, as a machine ages, its downtime will often increase and its overall productivity will diminish. Similarly, depreciation can be equated with obsolescence. A desktop computer from the mid-1980s is obsolete in the late 1990s.

However, in engineering economics, the concept of depreciation that is utilized by accountants is adopted. Depreciation is simply the accounting procedure that amortizes the cost of an asset over the estimated life of the asset. In short, the cost of an asset is not expensed at the time of purchase but rather is rationally spread throughout the useful life of the asset. Such a concept is adopted because this concept of depreciation is utilized in the calculation of federal income taxes.

It should be noted that depreciation does not represent a cash flow. Rather it is an accounting procedure. Heretofore, all of the economic analysis has concentrated upon cash flows. Depreciation must be included in any after-tax economic analysis because depreciation will affect the amount of taxes owed.

There are some basic terms associated with depreciation. Book value, BV_j, denotes the undepreciated value of an asset. The cost basis of an asset is usually the acquisition cost. Hence, the book value is the difference between the cost basis and the accumulated depreciation costs. The book value is given in the following formula:

$$\mathrm{BV}_j = \mathrm{CB} - (D_1 + D_2 + \cdots + D_t) \qquad (20)$$

where BV_j is the book value, CB is the cost basis, and D_t is the depreciation charge for year t.

The salvage value of an asset is the estimated value of an asset at the end of its estimated life.

Over the years, a variety of methods have been used to calculate depreciation charges. These methods are prescribed by the Internal Revenue Service (IRS). Prior to 1981, the permissible depreciation methods were straight-line, declining balance, and sum-of years digits. The Economic Recovery Tax Act of 1981 introduced the accelerated cost recovery system (ACRS). In 1986, the Tax Reform Act again modified allowable depreciation methods with the introduction of the modified accelerated cost recovery system (MACRS). This chapter will examine the MACRS method of depreciation. The MACRS applies to assets placed in service after December 31, 1986. The references offer rigorous examinations of the other depreciation methods for assets placed in service prior to December 31, 1986.

The MACRS categorizes assets into eight classifications known as the recovery period: 3-year, 5-year, 7-year, 10-year, 15-year, 20-year, 27.5-year, and 39-year. The IRS has guidelines that determine into which classification an asset should be placed. These guidelines are found in the IRS Publication 946 *How to Depreciate Property* [1]. Table 4 gives examples of some common assets and their pertinent recovery periods.

For each MACRS classification, the IRS has specific depreciation rates. The depreciation rates are the recovery allowance percentages. The MACRS method also uses a half year convention so that all property is treated as if it were placed into service at the midyear. Hence, depreciation charges exist for an additional tax year beyond the class designation. For instance, 3-year property will be allocated over four tax years. Table 5 sets forth the recovery allowance percentages for the various classifications.

The depreciation charges then for any given year depend upon the acquisition cost and the appropriate recovery allowance percentage. The depreciation charge is then simply the product of the acquisition cost and the appropriate recovery allowance percentage.

Example 16. *A computer system with an initial cost of $20,000 is purchased in 1997 by an engineering consulting company. Compute the allowable annual depreciation charges and the corresponding book values.*

Computers are classified as having a 5-year recovery period.

Table 4 MACRS Classifications of Depreciable Property

Classification	Property
3-year	Fabricated metal products; special handling devices for food and beverage manufacture; tractor units for over-the-road use; certain livestock
5-year	Automobiles; light and heavy trucks; computers and copiers; equipment used in research and experimentation; equipment used in oil wells
7-year	All other property not assigned to another classification; office furniture and equipment; single-purpose agricultural structures; railroad track; telephone station equipment
10-year	Assets used in petroleum refining; assets used in manufacture of castings, forgings, tobacco, and certain food products; vessels and water transportation equipment
15-year	Waste-water plants; telephone distribution equipment; industrial steam and electrical generation equipment; railroad wharves and docks; storage tanks
20-year	Municipal sewers; barges and tugs; electrical power plant
27.5-year	Residential rental property
39-year	Nonresidential rental property

Table 5 MACRS Recovery Allowance Percentages

Recovery year	3-year class	5-year class	7-year class	10-year class	15-year class	20-year class
1	33.33%	20.00%	14.29%	10.00%	5.00%	3.750%
2	44.45	32.00	24.49	18.00	9.50	7.219
3	14.81	19.20	7.49	14.40	8.55	6.677
4	7.41	11.52	12.49	11.52	7.70	6.177
5		11.52	8.93	9.22	6.93	5.713
6		5.76	8.92	7.37	6.23	5.285
7			8.93	6.55	5.90	4.888
8			4.46	6.55	5.90	4.522
9				6.56	5.91	4.462
10				6.55	5.90	4.461
11				3.28	5.91	4.462
12					5.90	4.461
13					5.91	4.462
14					5.90	4.461
15					5.91	4.462
16					2.95	4.461
17						4.462
18						4.461
19						4.462
20						4.461
21						2.231

$$D_1 = \$20{,}000(0.20) = \$4000$$

$$\mathrm{BV}_1 = \$20{,}000 - 4000 = \$16{,}000$$

$$D_2 = \$20{,}000(0.32) = \$6400$$

$$\mathrm{BV}_2 = \$16{,}000 - 6400 = \$9600$$

$$D_3 = \$20{,}000(0.192) = \$3840$$

$$\mathrm{BV}_3 = \$9600 - 3840 = \$5760$$

$$D_4 = \$20{,}000(0.1152) = \$2304$$

$$\mathrm{BV}_4 = \$5760 - 2304 = \$3456$$

$$D_5 = \$20{,}000(0.1152) = \$2304$$

$$\mathrm{BV}_5 = \$3456 - 2304 = \$1152$$

$$D_6 = \$20{,}000(0.0576) = \$1152$$

$$\mathrm{BV}_6 = \$1152 - 1152 = \$0$$

1.5.2 Income Tax Rates

Federal income tax rates for both corporations and individuals have varied over the years. Note, the top federal income tax rate for an individual in 1970 was 70%, while in 1995 it was 39.6%. The income tax rates are also graduated so that the rate depends upon the taxable income. In 1997, a corporation with a taxable income of \$40,000 was taxed at the rate of 15%, whereas the corporation with a taxable income of \$1,000,000 would be taxed at the 34% level for all taxable income over \$335,000.

In engineering economic analysis an effective income tax rate is usually used. The effective income-tax rate is simply a percentage. The product of the effective income tax rate and the taxable income then yields the tax owed. The concept of an effective income tax rate often combines the federal, state, and local income tax rates.

1.5.3 Factors Affecting Taxable Income

The taxable income reflects the quantity from which income taxes are determined. Therefore, the taxable income includes before-tax cash flows such as income and expenses. Also, included in the taxable income are any applicable depreciation charges. Recall, these depreciation charges are not cash flows. Depreciation charges will further reduce the taxable income and in turn reduce the tax liability.

Loans are commonly used to finance business operations. The interest paid on such loans is ordinarily viewed as an expense by the federal government. Hence, any interest paid on a loan by a corporation would be deductible from the taxable income. Note, only the interest payments on a loan and not the principal portion is deductible. In essence, this reduces the effective cost of borrowing through the alleviation of tax liability.

1.5.4 After-Tax Analysis

In order to proceed with an after-tax analysis on an alternative there are several preliminary considerations. The MARR must be established. The MARR used for after-tax analysis should not be the same MARR used for before-tax analysis. The effective income tax rate must be identified. Remember that the effective income tax rate is often based upon the prevailing federal, state, and local income tax rates. If necessary, the appropriate depreciation method and associated depreciation charges must be calculated. Similarly, any relevant interest on loans must be determined. Also, the length of the time horizon needs to be set.

The underlying concept is to try to calculate an after-tax cash flow for each period within the time horizon. After securing these after-tax cash flows, then one can proceed to utilize any of the previously mentioned means of evaluating alternatives. For example, an after-tax present-worth analysis is simply where the present-worth technique is applied to the after-tax cash flows. Similarly, an after-tax rate of return utilizes the rate-of return technique on after-tax cash flows. The following example illustrates the procedures for completing an after-tax cash flow analysis.

Example 17. *With the purchase of a \$100,000 computer system, a consulting firm estimated that it could receive an additional \$40,000 in before-tax income. The firm is in the 30% income tax bracket and expects an after-tax MARR of 10%. If the funds for the computer are borrowed on a 4-year direct 8% reduction loan with equal annual payments, what is the present worth of the after-tax cash flow?*

First find the annual loan payment:

$$A = P(A|P, 8\%, 4) = \$100{,}000(0.3019) = \$30{,}190$$

Then determine the interest paid each year on the loan:

$$I_j = A(P|A, j, n - j + 1)i \qquad \text{from Eq. (17)}$$

$$I_1 = \$30{,}190(P|A, 8\%, 4)(0.08)$$

$$= \$30{,}190(3.3121)(0.08) = \$7999$$

$$I_2 = \$30{,}190(P|A, 8\%, 3)(0.08)$$

$$= \$30{,}190(2.5771)(0.08) = \$6224$$

$$I_3 = \$30{,}190(P|A, 8\%, 2)(0.08)$$

$$= \$30{,}190(1.7833)(0.08) = \$4307$$

$$I_4 = \$30{,}190(P|A, 8\%, 1)(0.08)$$

$$= \$30{,}190(0.9259)(0.08) = \$2236$$

Note that computers are classified as having a 5-year recovery period. Hence, the annual depreciation expenses are:

$$D_1 = \$100{,}000(0.20) = \$20{,}000$$

$$D_2 = \$100{,}000(0.32) = \$32{,}\ 000$$

$$D_3 = \$100{,}000(0.192) = \$19{,}200$$

$$D_4 = \$100{,}000(0.1152) = \$11{,}520$$

$$D_5 = \$100{,}000(0.1152) = \$11{,}520$$

$$D_6 = \$100{,}000(0.0576) = \$5760$$

Construct a table to calculate the after-tax cash flow (amounts in dollars):

Year (1)	Before tax (2)	Prin. paid (3)	Int. paid (4)	Deprec. expense (5)	Taxable income (6)	Taxes paid (7)	After tax (8)
1	40,000	−22,191	−7,999	−20,000	12,001	−3,600	6,210
2	40,000	−23,966	−6,224	−32,000	1,776	−533	9,277
3	40,000	−25,883	−4,307	−19,200	16,493	−4,948	4,862
4	40,000	−27,954	−2,236	−11,520	26,244	−7,873	1,937
5	40,000	0	0	−11,520	28,480	−8,544	31,456
6	40,000	0	0	−5,760	34,240	−10,272	29,728
7	40,000	0	0	0	40,000	−12,000	28,000
8	40,000	0	0	0	40,000	−12,000	28,000
9	40,000	0	0	0	40,000	−12,000	28,000
10	40,000	0	0	0	40,000	−12,000	28,000

The taxable income is calculated by summing the before-tax amount, the interest paid, and the depreciation expense [add columns (2), (4), and (5)]. The taxes paid are simply the product of the taxable income and the tax rate. The after-tax cash flow is then the sum of the before-tax, the principal paid, the interest paid, and taxes paid [add columns (2), (3), (4), and (7)]. The after-tax column then is analyzed for its present worth:

$$\text{PW} = \$6210(P|F, 10\%; 1) + \$9277(P|F, 10\%, 2)$$

$$+ \cdots + \$28{,}000(P|F, 10\%, 10)$$

$$\text{PW} = \$6{,}210(0.9091) + \$9{,}277(0.8265)$$

$$+ \cdots + \$28{,}000(0.3856)$$

$$\text{PW} = \$104{,}\ 700$$

Therefore the project is sound.

1.6 INFLATION

The purchasing power of money is not static over time. Prices for goods and services are rarely constant from one year to the next. With inflationary pressures, the cost of goods and services increases with time. Whereas, decreasing prices would signify the condition of deflation.

Recall that the time value of money is based upon the earning power of money and also the purchasing power of the money. When evaluating alternatives or computing economic equivalence, it is often desirable to separate the earning power of money from its purchasing power. In short, an "inflation-free" analysis is frequently preferred.

1.6.1 Measures of Inflation

Indexes of inflation are frequently used to monitor changes in prices. The Consumer Price Index (CPI) is perhaps the most widely referenced index of inflation. Indexes of inflation are merely weighted averages of a series of goods and services. The index then tracks how the prices of the goods and services vary from period to period.

Care should be undertaken in the selection of an inflation index. One should verify that a particular inflation index is tracking the factors that are needed for a particular analysis. For example, rises in the cost of groceries may not be a significant factor to an equipment manufacturer.

An inflation index will have a base year or time period. Subsequent changes in price are measured against the base year or period. For example, the CPI has a base year of 1967 with a value of 100.00. In 1990, the CPI index had a value of 391.4. This indicates that comparable goods and services that cost \$100 in 1967 would have cost \$391.40 in 1990.

With the selection of an appropriate inflation index, it is possible to analyze economic alternatives on an inflation-free basis. Such an approach requires that one convert all of the cash flows to a particular year or time period based on the inflation index. Once the conversion has been made, then an alternative can be evaluated using any of the previously mentioned techniques for the evaluation of alternatives. However, one must still include a means to account for the time value of money based upon the earning power of money. This interest rate is generally called the inflation-free interest rate and is denoted as i'.

Example 18. *In 1985, a manufacturing company invested in a new process that cost \$4,500,000. In the subsequent four years, the net profit after taxes made by the facility, along with the price index was:*

Year	Net profit (actual \$)	Price index (1967 = 100)
1985	—	322.2
1986	2,200,000	328.3
1987	1,700,000	340.4
1988	1,900,000	354.4
1989	1,500,000	371.4

If the inflation-free rate of return, i', was 3%, determine the present worth of the investment in 1985 dollars. Was the investment a sound one?

First express net profit in terms of 1985 dollars:

$$1986: \quad \$2{,}200{,}000\left(\frac{322.2}{328.3}\right) = 2{,}159{,}122$$

$$1987: \quad \$1{,}700{,}000\left(\frac{3.22}{340.4}\right) = 1{,}609{,}107$$

$$1988: \quad \$1{,}900{,}000\left(\frac{322.2}{354.4}\right) = 1{,}727{,}370$$

$$1989: \quad \$1{,}500{,}000\left(\frac{322.2}{371.4}\right) = 1{,}301{,}292$$

Then find present worth:

$$\begin{aligned}
\text{PW} &= -\$4{,}500{,}000 + \$2{,}159{,}122(P|F', 3\%, 1)\\
&\quad + \$1{,}609{,}107(P|F', 3\%, 2)\\
&\quad + \$1{,}727{,}370(P|F', 3\%, 3)\\
&\quad + \$1{,}301{,}292(P|F', 3\%, 4)\\
&= \$4{,}500{,}000 + \$2{,}159{,}122(0.9709)\\
&\quad + \$1{,}609{,}107(0.9426)\\
&\quad + \$1{,}727{,}370(0.9152) + \$1{,}301{,}292(0.8885)\\
&= \$1{,}850{,}123
\end{aligned}$$

Thus, the investment was sound.

1.6.2 Average Inflation Rate

A difficulty associated with an inflation index is that the index tracks past inflationary patterns. It will not necessarily give reliable estimates of future inflationary trends. Also, from the examination of an inflation index, it is obvious that inflation is rarely constant.

Hence, an average inflation rate is often used to account for the variation in the inflation rates over a number of years. An average inflation rate can be calculated from an inflation index by the following equation:

$$(\text{Index})_t(1 + \bar{f})^n = (\text{Index})_{t+n} \tag{21}$$

1.6.3 Actual and Constant Dollars

In any analysis where inflation is taken into account, there are a few fundamental terms and relationships that must be understood. Constant dollars represent money where the money has been adjusted for inflationary effects. Cash flow patterns may be expressed in constant dollars. A notation with a prime superscript often denotes a constant dollar cash flow pattern. For instance, an F' would denote a constant dollar future cash flow.

Actual or current dollars represent a monetary value that incorporates both inflation and the earning power of money. Estimates in actual dollars represent the true sums of money that one could anticipate to receive or disburse.

The inflation-free interest rate, i', is an estimate of the earning power of money without inflation, whereas the market interest rate, i, combines the earning power of money and the effects of inflation. The market interest rate is what one will encounter in common everyday experiences. The interest rate on a standard mortgage is an example of a market interest rate.

The inflation-free interest rate, the market interest rate, and the average inflation rate are related by the following equation:

$$i = i' + f + i'f \tag{22}$$

Therefore, a series of cash flows can then be expressed either in constant dollars or in actual dollars. The conversion from actual dollars to constant dollars in any given period would be accomplished by multiplying by the following factor:

$$(\text{Constant dollar})_n = (\text{Actual dollar})_n(1 + f)^{-n} \tag{23}$$

Similarly, the conversion from constant dollars to actual dollar utilizes this factor:

$$(\text{Actual dollar})_n = (\text{Constant dollar})_n(1 + f)^n \tag{24}$$

For after-tax analysis where the average inflation rate is estimated, it is recommended that the subject cash flows be converted to actual dollars. Such a conversion will enable one to readily assess the pertinent tax liabilities.

There are two approaches for a before-tax analysis with inflation. One approach calls for all of the cash flows to be expressed in terms of actual dollars with the subsequent analysis to use the market interest rate, i. Under the second approach, all of the cash flows are expressed in terms of constant dollars with the subsequent analysis utilizing the inflation-free interest rate, i'.

Example 19. *A manufacturing corporation is constructing a new production line. The associated production costs for the new line are estimated at \$2.5 million. Over the ensuing years, the production costs are expected to increase \$100,000 per year in actual dollars. The yearly inflation rate is presumed to be 4% and the market interest rate is 8%. Given a life span of 10 years, find the annual worth of the production costs in terms of constant dollars.*

Recall,

$$i = i' + f + i'f \qquad \text{from Eq. (22)}$$

$$0.08 = i' + 0.04 = i' + i'(0.04)$$

Via algebra,

$$i' = 0.03846$$

Next find the annual worth of production costs in actual dollars:

$$A = A_1 + G(A|G, 8\%, 10)$$

$$A = \$2{,}500{,}000 + 100{,}000(3.8713)$$

$$A = \$2{,}887{,}130$$

Convert to present worth:

$$P = \$2{,}887{,}130(P|A, 8\%, 10)$$

$$P = \$2{,}887{,}130(6.7101)$$

$$P = \$19{,}372{,}931$$

Convert to constant annual worth using $i' = 3.846\%$*:*

$$A' = \$19{,}372{,}931(A'|P, 3.846\%, 10)$$

$$A' = \$19{,}372{,}931(0.12234) \qquad \text{from Eq. (11)}$$

$$A' = \$2{,}370{,}250/\text{year}$$

REFERENCES

EL Grant, WG Ireson, RS Leavenworth. Principles of Engineering Economy. New York: John Wiley & Sons, 1990.

IRS Publication 946. How to Depreciate Property. Washington DC: United States Government Printing Office, 1997.

DG Newnan, B Johnson. Engineering Economic Analysis. San Jose, CA: Engineering Press, 1995.

GJ Thuesen, WJ Fabrycky. Engineering Economy. Englewood Cliffs, NJ: Prentice Hall.

CS Park. Contemporary Engineering Economics. Menlo Park, CA: Addison-Wesley, 1997.

Chapter 10.2

Manufacturing-Cost Recovery and Estimating Systems

Eric M. Malstrom[†] **and Terry R. Collins**
University of Arkansas, Fayetteville, Arkansas

2.1 INTRODUCTION

This chapter overviews cost recovery and estimating systems typically used by manufacturing organizations. The chapter begins by overviewing conventional manufacturing-cost estimating systems. Cost centers are described, as are types of costs and use of performance standards in making cost estimates. Process design and its effect on manufacturing costs is addressed, as is the integration of learning curves into estimating procedures. Contingency allowances are addressed, as is the concept of making cost reviews or re-estimates based on the progress of a manufacturing project.

Conventional cost recovery systems are next described. In these systems direct labor is used as a recovery basis variable. Concepts of capital budgeting are introduced as are subsequent adjustments of obtained labor/overhead rates.

Quick-response estimating is next described with an emphasis on cost variable identification and construction of estimating relationships. The relationship to group technology and production mix/volume scenarios is addressed as well. Cost estimating software development concepts are introduced. The merits of purchasing commercially available software versus in-house software development are described.

Activity based costing is described in some detail. Topics include mechanics of the recovery procedure, and the identification and selection of cost drivers. The chapter concludes by discussing how high levels of manufacturing automation impact the processes of manufacturing cost estimating and recovery.

2.2 CONVENTIONAL COST ESTIMATING PROCEDURES

Manufacturing-cost estimating is a topic that has not enjoyed high visibility in university curricula. In 1981, only three books existed that addressed this subject in sufficient depth to permit them to be used is textbooks for university courses on this subject [1–3]. In 1981, almost no university faculty specialized in the field of cost estimating. This continues to be true at present.

The result has been limited availability of suitable texts on this subject. To be an effective cost estimator requires prior industrial experience. Comparatively few university faculty members have significant work experience outside academia. Consequently, scant academic research on this subject has, or is being accomplished.

2.2.1 Cost Estimating Defined

Cost estimating may be described as the process by which a forecast of costs required to manufacture a product or complete a specified task can be made. The estimate consists of the costs of people, materials, methods, and management. The accuracy of a cost

[†] Deceased.

estimate is a function of the degree of design or project definition available at the time the estimate is made. The more finalized and firm the design and definition, the more accurate the estimate is likely to be. Estimate accuracy is also a function of the time and resources that the estimator has available to compile a bid. Accuracy may also be affected by the quantity of units that are to be fabricated or produced.

2.2.2 Role of Engineers in Cost Estimating

Engineers have had historically a limited role in the cost estimating process. Few engineering curricula have formal courses on this subject. Good estimating skills require detailed knowledge of an organization's product line, production facilities, and manufacturing processes. In many cases nondegreed personnel who have formal shop floor experience have better backgrounds with which to perform cost estimating tasks.

Engineers have a professional need to be knowledgeable about estimating procedures [11]. They need this knowledge to specify cost-effective designs. Often engineers assume managerial positions which encompass or oversee the estimating function.

2.2.3 Basic Steps in the Estimating Process

The steps in compiling a cost estimating include determining whether each part in the bill of materials of an end item should be made in-house or purchased. This is followed by preliminary process sequence planning and the subsequent tallying of labor and material costs. Dependent costs must also be determined and tallied. These include the costs of indirect labor and overhead, manufacturing engineering, and inspection/quality control. Finally, an appropriate contingency allowance must be determined and included.

2.2.4 Types of Manufacturing Costs

Two of the most basic types of manufacturing costs are direct labor and direct material [1, 4]. *Direct labor* is the cost of all ''hands-on'' effort to manufacture a product. Typical direct labor activities include machining, assembly, inspection, testing, and troubleshooting. *Direct material* is the cost of all components and raw materials included in the end product that is be produced. The sum of direct labor and direct material is often referred to as *prime cost*.

Factory expenses may be defined as the total costs for rent, heat, electricity, water, expendable factory supplies, and indirect labor. *Factory cost* is often defined as the sum of prime cost plus factory expenses. *General expenses* are the costs of design engineering, purchasing, office staff salaries, and depreciation. *Manufacturing cost* is the sum of general expenses plus factory cost.

Sales expenses are all costs incurred in selling and delivering the end product. These include the cost of advertising, sales commission, and shipping costs. *Total costs* may be defined as the sum of sales expense plus manufacturing cost.

Finally, the *selling price* of the end product is the sum of the total costs plus the organization's desired profit margin.

2.2.5 Performance Standards

Performance standards are the best prior estimate of the length of time a labor task is likely to require. Such standards can therefore be applied to determine the labor content of a manufacturing cost estimate. Principal data sources for work standards are from time study analyses and predetermined time systems. The role of and use of performance standards in making cost estimates is described in more detail in Refs. 1 and 4.

2.2.6 Cost Centers and Shop Orders

Cost estimating requires the use of both cost centers and shop orders. Often organizational divisions and departments are defined by numerical codes. For example, if a three-digit code is used, the hundreds digit might be used to indicate a department. The tens digit can be used to designate a division within an department. Finally, the units digit may denote a branch within a division within a department. Some sample organization codes indicating both departments and divisions are illustrated in Table 1.

Cost estimating requires establishing an audit trail for charge tracability. An *account number* is used to determine where in the organization a labor charge has occurred. The *cost center* is often a numerical subset of the account number and reflects all or part of the organization code of the department, division, or branch in which the labor charge has occurred.

A *job order* is basically a project number reflecting which manufacturing project has been or should be ''billed'' for labor or material charges. A *shop order* is an authorization to perform work on a given cost center. A shop order code is usually alphanumeric in format. The alpha prefix of the shop order code reflects the type of manufacturing effort on a given cost center

Table 1 Sample Organization Codes

Organizational name	Assigned code
Manufacturing Department	200
Manufacturing Engineering Division	210
Assembly Division	220
Assembly Division	230
Comptroller Department	300
Quality Control Department	400
Quality Engineering Division	410
Inspection Division	420
Standards and Calibration Division	430
Industrial Relations Department	500
Marketing Department	600
Engineering Department	700
Product Design Engineering Division	710
Research and Development Engineering Division	720

that is being performed. A labor charge on a manufacturing project is thus made in conjunction with a cost center, a job order, and a shop order. The cost center specifies where in the organization the work was performed. The job order specifies which manufacturing project was or should be billed for the charge. Finally, the shop order indicates what type of manufacturing effort is being performed on the project.

Example cost centers and shop orders are illustrated in Table 2. Readers desiring more detailed information on cost centers, job orders, and shop orders should consult Refs. 1 and 4.

2.2.7 Making the Initial Cost Estimate

Initial cost estimates are those made prior to the start of production. They are important as their accuracy sets the profit/loss position of the firm. The process begins by reviewing the bill of materials or part explosion structure of the end item to be manufactured. A determination must initially be made on an item-by-item basis as to whether each individual component should be purchased or fabricated in-house. Usually these determinations are made on the basis of which alternative is less expensive. The cost estimator must anticipate the process sequence for each part in the bill of materials which is to be fabricated. Prior shop experience of the estimator is vital in accurately anticipating the process sequence that will be used to actually make each part.

The preliminary sequence is specified by generating a sketch process routing for each "make" part. This routing contains:

Table 2 Example Cost Centers and Shop Orders

Cost center	Activity	Shop order code
21 Manufacturing Engineering	Process	M-XXXXX
	Tool design	D-XXXXX
	Generation of NC tapes	N-XXXXX
	Packaging	K-XXXXX
22 Machining	Chip turning	C-XXXXX
	Sheet metal	S-XXXXX
	Tool fabrication and maintenance	V-XXXXX
	Painting	P-XXXXX
	Plating	L-XXXXX
	Heat treating	H-XXXXX
	Rework	R-XXXXX
23 Assembly	Assembly	A-XXXXX
	Testing	T-XXXXX
	Wire cutting	W-XXXXX
	Troubleshooting	G-XXXXX
	Rework	R-XXXXX
	Encapsulation	J-XXXXX
42 Inspection	Inspection of purchased parts	B-XXXXX
	Inspection of fabricated parts	F-XXXXX
43 Standards and Calibration	Mechanical and electronic calibration	O-XXXXX
70 Engineering	Engineering support to manufacturing	E-XXXXX

The anticipated sequence of manufacturing operations

The departmental location for each operation

The required machines and equipment for each routing operation

A brief description of each production operation

The applicable shop order and cost center

Estimated setup and ‘‘per piece’’ run times for each operation.

The estimated times are obtained from performance standards or time study analyses. Alternately, some machine cycle times are often estimated from formulae.

2.2.8 The Contingency Allowance

The *contingency allowance* is an estimate supplement used to account for work content and materials that are expected to occur, but cannot be accurately anticipated or accounted for at the time the initial cost estimate is made. The contingency allowance varies with both the degree of initial design definition and the time available to make the estimate. Vernon [3] has defined seven separate estimate classes as a function of the type and quantity of information available at the time the estimate is made. These classes are illustrated in Table 3. Malstrom [1, 4] has described typical contingency allowance percentages as a function of estimate confidence and design definition. These percentages are summarized in Table 4.

2.2.9 Aggregating Labor and Material Costs

Malstrom [1, 4] has discussed how labor and material costs are aggregated for compilation of initial estimate totals. Labor hours are tallied by shop order within cost centers. Each cost center has a separate and often different labor overhead rate determined by the capital budgeting process (described later). The *labor/overhead* (LOH) *rate* for each cost center converts labor hours into labor dollars and overhead dollars. *Material costs* are aggregated and totaled by individual cost centers as well.

The labor, overhead, and material cost dollars are totaled for each cost center. The sum of these cost totals over all cost centers is the *estimated production cost* (EPC). The *contingency allowance* is expressed as a fixed percentage of the EPC. The EPC plus the dollar magnitude of the contingency allowance is the *estimated total cost* (ETC). Adding profit to the ETC results in the *total bid cost* that can be submitted to a prospective customer. Malstrom has illustrated how to summarize these costs on an estimate grid. This grid is illustrated in Fig. 1.

2.3 COST REVIEWS

A *cost review* is a follow-on cost estimate of a manufacturing task that has already begun and is in process at the time the cost review is made. The time required to complete a cost review is significant and approxi-

Table 3 Information Defining Numerical Estimate Classes

	Estimate class						
Description of data	1	2	3	4	5	6	7
General design specification, quantity, and production rate	X	X	X	X	X	X	X
Assembly drawings	X	X	X	X	X	X	
Proposed subassemblies	X	X	X	X	X		
Detailed drawings and bill of materials	X	X	X	X			
Test and inspection procedures/equipment	X	X	X	X			
Machine tool and equipment requirements	X	X	X	X			
Packaging/transportation requirements	X	X	X	X			
Manufacturing routings	X	X	X				
Detailed tool, machine, gage and equipment lists	X	X					
Operation analysis and workplace studies	X	X					
Standard time data	X	X					
Material release data	X	X					
Subcontractor cost and delivery date	X	X					
Area and building requirements	X						

Source: Ref. 3.

Table 4 Estimate Classes

Class	Definition
A	Excellent confidence—repeat job, excellent definition, no major changes anticipated.
B	Good confidence—new design, first build, good definition, some design changes anticipated. Contingency 10–20%.
C	Average confidence—preliminary or partial design, verbal information, changes anticipated. Contingency 20–30%.
D	"Ball park" estimates—definition very sketchy and preliminary, many unknowns several design changes anticipated. Contingency 30–40%.
F	Budgeting—an estimate prepared only for the purpose of budgeting funds. Contingency allowance levels vary depending on design definition.
X	Directed estimate—a modification of any previous cost estimate to conform to budget cuts and restrictions which are not based on scope decisions. Adjustments may be increases or reductions in the allowance or in any cost element as directed by top management decisions.

mates the level of effort required to compile an initial cost estimate. Consequently, cost reviews are generally performed only on those jobs where fiscal deficits or surplus are expected to occur.

The procedure begins by selecting a review date. This date is a "snapshot" of the project's fiscal status at a given point in time. Prior to the review date are actual expenditures on the project which have occurred. After the review date are those expenditures expected to occur up until the project being reviewed is expected to be completed.

The mechanics of the cost review procedure are illustrated in the cost review grid illustrated in Fig. 2. This grid lists all cost centers on which direct labor expenditures are expected to occur. The procedure begins by recording all labor hour charges that have occurred, by shop order, prior to the review date. Each cost center has four distinct rows. The *Estimated* row contains estimated labor hours and costs from the most recent prior cost estimate or review for each of the cost centers. The *Expended* row contains dollar expenditures for labor and material by cost centers. The labor dollar expenditures correspond to the labor hours expended by shop order for each cost center. These hourly entries are entered, by shop order in the *Used* column for each cost center.

The cost analyst next estimates the required hours to complete the project for each cost center by shop order. These hour entries are placed in the *To Comp.* column for each cost center and shop order. Material dollar expenditures required for project completion are estimated for each cost center as well. The material dollar expenditures are entered in the *To Complete* row for each cost center in the *Material* column of the grid.

Next, labor and overhead rates are entered in the *To Complete* row for each cost center. These rates may be higher than those used in the most recent prior cost estimate or review. The *To Comp.* hours are then totaled by shop order and entered in the *Hrs.* column of the *To Complete* row for each cost center. The totals in the *Hrs.* column of the *To Complete* row are multiplied by the labor and overhead rates for each cost center. These dollar totals are entered *Labor* and *Overhead* columns of the *To Complete* row for each cost center.

The next step is to add the entries of the *Expended* and *To Complete* rows. The resultant sums are placed in the *Total* row of each cost center as illustrated in Fig. 2. The information is next transferred to the cost estimate grid illustrated in Fig. 3. The hour entries from the *Used* column of Fig. 2 are transcribed to the *Hrs. Exp. to Date* column of Fig. 3 by both shop order and cost center. The hour entries from the *To Comp.* column of Fig. 2 are transferred to the *Hrs. to Compl.* column of Fig. 3 by shop order and cost center.

Hour and dollar entries from the *Total* row of Fig. 2 are next transcribed to the *Direct Labor Hours, Labor, Overhead, Material*, and *Total Cost* columns of Fig. 3 by cost center. The entries in the *Total Cost* column of Fig. 3 are summed. The contingency allowance and profit margin are adjusted as necessary depending on whether a cost deficit or surplus is projected to occur. Readers desiring a more detailed description of the cost review process are urged to consult Refs. 1 and 4.

2.4 LEARNING CURVES

Learning or product improvement curves reflect decreasing labor costs as the production quantity completed increases. These decreases reflect the effects of both human learning and process improvements associated with the startup of production.

2.4.1 Learning Curves Defined

Learning curves may be described by

Direct Labor-Hrs.		Hrs. Exp. To Date	Hrs. To Compl.	Wage Rate	O.H. Rate	Direct Labor-Hrs.	Labor	Overhead	Material	Total Cost
MFG. ENG. CC 21	Process Engineering (M-)		10	13.00	12.00	20	$260	$240	$80	$580
	Tool Design (D-)									
	NC Tapes (N-)									
	Packaging (K-)		10							
MACH. DIV. CC 22	Chip Turning (C-)		201	10.00	8.00	215	$2150	$1720	$2062	$5932
	Sheet Metal (S-)									
	Tool Fab. & Maint. (U-)									
	Painting (P-)									
	Plating (L-)									
	Heating Treating (H-)									
	Rework (R-)		14							
ASSY. DIV. CC 23	Assembly (A-)		4	9.00	7.00	4	$36	$28		$64
	Testing (T-)									
	Wire Cutting (W-)									
	Trouble Shooting (G-)									
	Encapsulation (J-)									
	Rework (R-)									
QUAL. ASSUR. CC 40	Pur. Part Inspection (B-)		15	12.00	11.00	33	$396	$363		$759
	Fab. Part Inspection (F-)		15							
	Calibration (O-)		3							
CC 70	Engr. Support (E-)									
Subtotal (Est. Production Cost)			272							$7335
Allowance		10%								$733
Estimated Total Cost										$8068
Profit										$403
Estimate Total										$8471

Figure 1 Cost estimate grid.

$$Y = KX^n \qquad (1)$$

where

$K =$ Time in hours required to produce the first unit.

$X =$ Total units manufactured

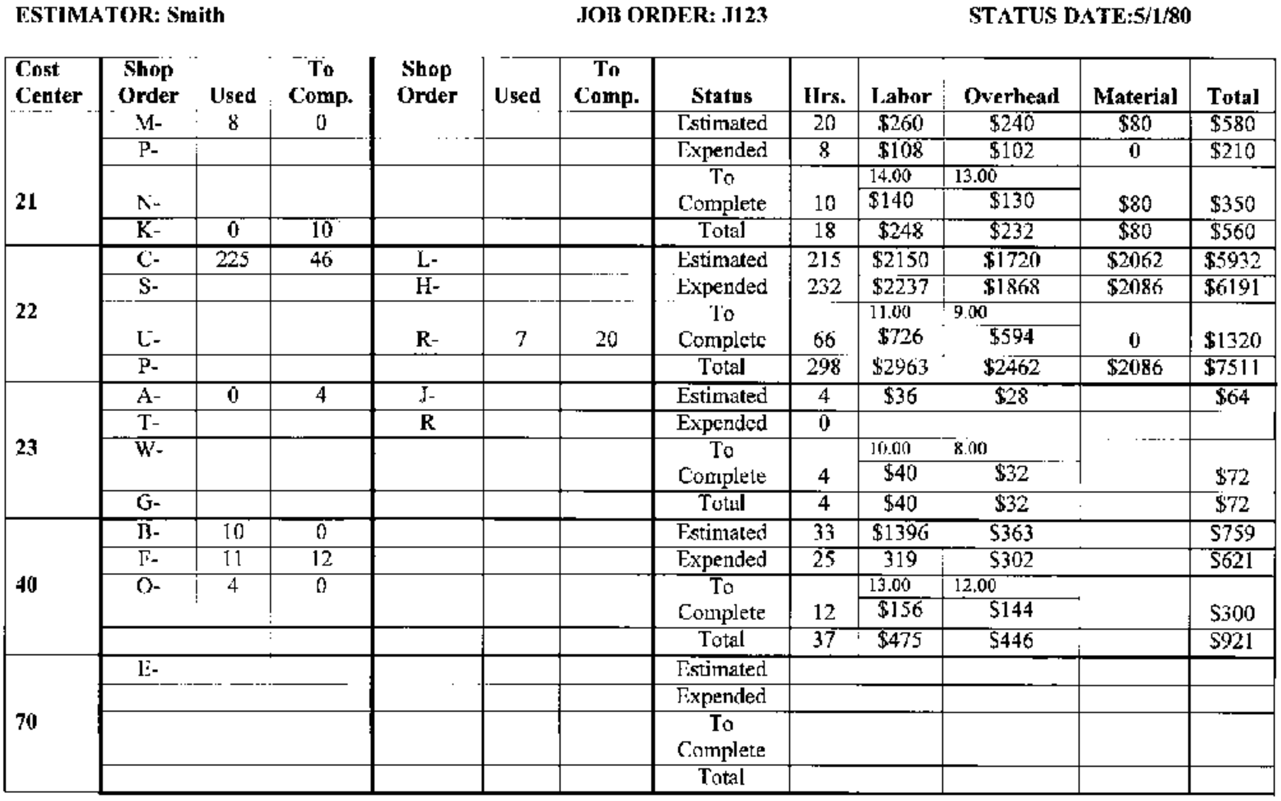
$n =$ A negative exponent which determines the percent by which Y decreases each time X is doubled

$Y =$ The cumulative average time per unit to build a quantity of X units

The definitions above are for cumulative average learning curves. Unit learning curves also exist and may be used for cost analysis purposes. With unit curves, Y is defined as the time in hours required to build the Xth unit. The remaining variables described above are the same for both types of curves. Readers desiring more detailed descriptions of learning curves are urged to consult Refs. 1, 4, 5, 6, and 13.

2.4.2 Learning-Curve Considerations in Estimating Procedures

Cost estimates are dependent on labor hours derived from work or labor standards. Without exception, work or time standards are derived from constant "time per unit" values. The effects of process improvement and human learning are usually not directly considered in labor standard development. To effectively integrate the effects of learning into estimating procedures, it is necessary to determine at what quantity level on the learning curve the standard time is reached. Construction of actual learning curves requires the determination of both K and n in Eq.



ESTIMATOR: Smith **JOB ORDER: J123** **STATUS DATE:5/1/80**

Cost Center	Shop Order	Used	To Comp.	Shop Order	Used	To Comp.	Status	Hrs.	Labor	Overhead	Material	Total
21	M-	8	0				Estimated	20	$260	$240	$80	$580
	P-						Expended	8	$108	$102	0	$210
	N-						To Complete	10	14.00 $140	13.00 $130	$80	$350
	K-	0	10				Total	18	$248	$232	$80	$560
22	C-	225	46	L-			Estimated	215	$2150	$1720	$2062	$5932
	S-			H-			Expended	232	$2237	$1868	$2086	$6191
	U-			R-	7	20	To Complete	66	11.00 $726	9.00 $594	0	$1320
	P-						Total	298	$2963	$2462	$2086	$7511
23	A-	0	4	J-			Estimated	4	$36	$28		$64
	T-			R			Expended	0				
	W-						To Complete	4	10.00 $40	8.00 $32		$72
	G-						Total	4	$40	$32		$72
40	B-	10	0				Estimated	33	$1396	S363		S759
	F-	11	12				Expended	25	319	S302		S621
	O-	4	0				To Complete	12	13.00 $156	12.00 S144		S300
							Total	37	$475	S446		S921
70	E-						Estimated					
							Expended					
							To Complete					
							Total					

Figure 2 Cost review grid.

(1). The nature of labor standard development makes this determination difficult in practice.

An alternate approach is to compile historical ratios of actual to standard hours after manufacturing tasks are complete. These ratios can be compiled and aggregated by both shop order type, production quantity, and product type or family. Multivariate linear regression can be used to determine mathematical functions which specify predicted ratios of actual to standard hours by shop order as functions of both production quantity and product type or family. These ratios may be used as multipliers for hourly totals by shop order compiled by the estimating methods previously described. The effects of learning curves will be embedded in these multipliers.

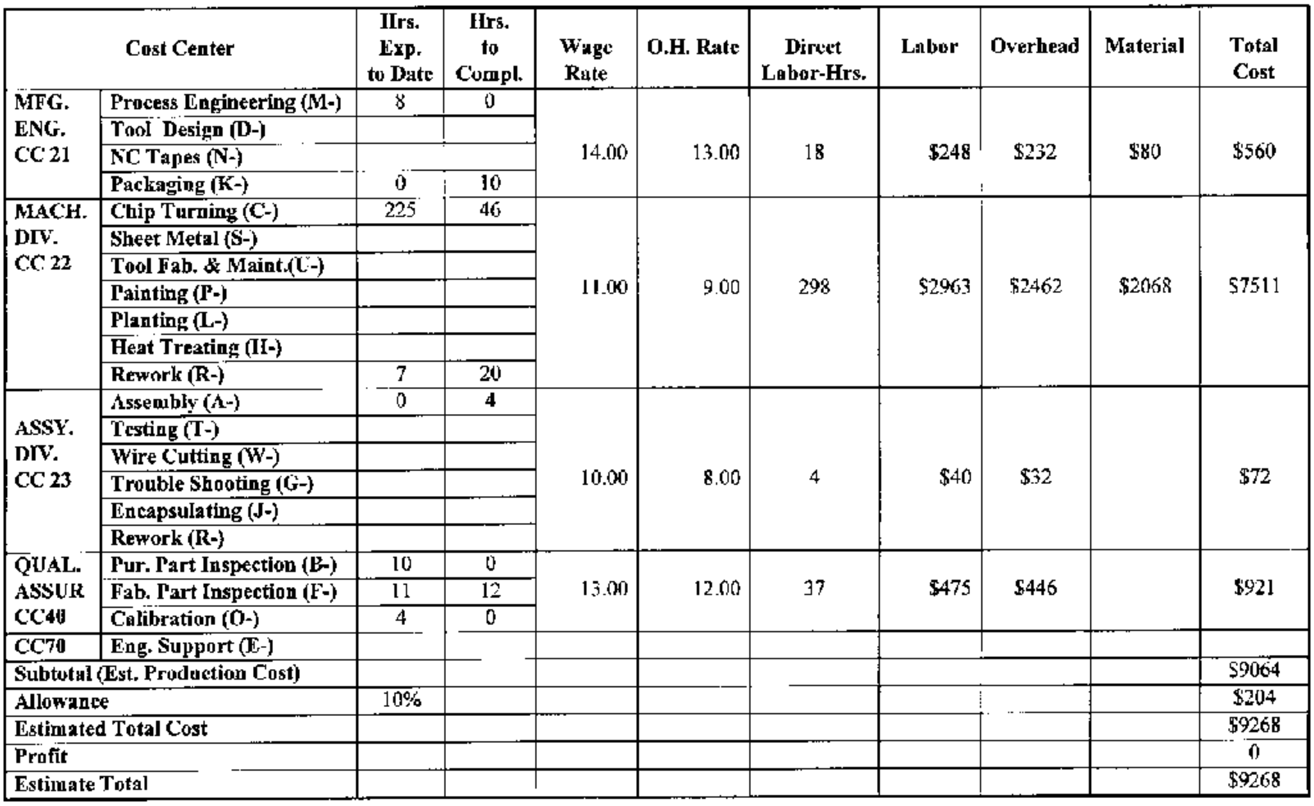
2.5 CAPITAL BUDGETING

Capital budgeting may be defined as the way in which individual labor/overhead rates are determined for each cost center. Most capital budgeting procedures utilize direct labor as a recovery basis variable. The procedure is to estimate the required overhead that must be charged in addition to the direct labor for each cost center, to recover the cost of both indirect labor and burden associated with the operation of a manufacturing facility. The capital budgeting procedure has been described in some detail by Malstrom [5] and is reproduced in some detail in the sections that follow.

2.5.1 Components of LOH Rates

The labor/overhead rate for any cost center has four distinct components [12]. The first of these is the *direct labor rate*. The direct labor rate is the composite average of all direct labor wages (including benefits) on the cost center being analyzed. The second component is the *expense rate*. The expense rate is the sum of all indirect labor dollars estimated to be expended in a budgetary quarter divided by the total number of direct labor hours estimated to be expended during that same quarter.

Burden is the cost, in dollars, of rent, utilities, building/equipment depreciation, and expendable supplies



Cost Center		Hrs. Exp. to Date	Hrs. to Compl.	Wage Rate	O.H. Rate	Direct Labor-Hrs.	Labor	Overhead	Material	Total Cost
MFG. ENG. CC 21	Process Engineering (M-)	8	0	14.00	13.00	18	$248	$232	$80	$560
	Tool Design (D-)									
	NC Tapes (N-)									
	Packaging (K-)	0	10							
MACH. DIV. CC 22	Chip Turning (C-)	225	46	11.00	9.00	298	$2963	$2462	$2068	$7511
	Sheet Metal (S-)									
	Tool Fab. & Maint.(U-)									
	Painting (P-)									
	Planting (L-)									
	Heat Treating (H-)									
	Rework (R-)	7	20							
ASSY. DIV. CC 23	Assembly (A-)	0	4	10.00	8.00	4	$40	$32		$72
	Testing (T-)									
	Wire Cutting (W-)									
	Trouble Shooting (G-)									
	Encapsulating (J-)									
	Rework (R-)									
QUAL. ASSUR CC40	Pur. Part Inspection (B-)	10	0	13.00	12.00	37	$475	$446		$921
	Fab. Part Inspection (F-)	11	12							
	Calibration (O-)	4	0							
CC70	Eng. Support (E-)									
Subtotal (Est. Production Cost)										$9064
Allowance		10%								$204
Estimated Total Cost										$9268
Profit										0
Estimate Total										$9268

Figure 3 Revised cost estimate grid.

for any budgetary quarter. Finally, *general and administrative costs* are the cost of top executives' salaries and centralized plant computing facilities. The labor/overhead rate for any cost center may be described by

$$\text{LOH}_{\text{CC}} = L_D + E_R + B + \text{G\&A} \tag{2}$$

where

$\text{LOH}_{\text{CC}} =$ The labor/overhead rate for a specific cost center in dollars/hr

$L_D =$ The direct labor rate, dollars/hr

$E_R =$ The expense rate, dollars/hr

$B =$ Burden in dollars/hr

$\text{G\&A} =$ General and administrative cost rate in dollars/hr

The dollar amounts for burden and general administrative costs expected to be expended during any quarter are divided by the total number of direct labor hours to be expended to determine the burden and G&A cost rates in Eq. (2).

The mechanics of the capital budgeting process are best illustrated with an example. This example is the subject of the following section and has been adapted from Ref. 5.

2.5.2 A Capital Budgeting Example

Consider a manufacturing plant with a total of 1000 employees. Suppose it is desired to determine the labor/overhead rate for the machining cost center 22. For example purposes we will assume that cost center 22 has a total of 200 employees. Of this total we will further assume that 150 are involved with direct labor activities.

To begin our analysis, we need to know the rest of the cost centers that exist and the respective number of employees in the plant that are associated with them. These staffing levels are illustrated in Table 5. In Table 5, there are a total of four cost centers on which direct labor is performed. These include cost centers 21, 22, 23, and 42 which are Manufacturing Engineering, Machining, Assembly, and Inspection respectively.

Table 5 Employee Staffing Levels by Cost Center

Cost center	Description	Number of employees
21	Manufacturing Engineering	100
22	Machining	200
23	Assembly	200
30	Comptroller	50
41	Quality Engineering	50
42	Inspection	50
43	Standards and Calibration	50
50	Personnel	50
60	Marketing	100
71	Design Engineering	100
72	Research and Development	50
		Total 1000

Each of these four cost centers contain some indirect labor employees as well. Some examples of this indirect labor would be supervisory and secretarial personnel. The remaining cost centers in Table 5 support directly labor activities and contain only indirect labor employees.

We wish to determine the labor/overhead rate associated with cost center 22. Let us assume that 40 hours exist in each work week. Assume further that there are exactly four weeks in each month and that a budgetary quarter consists of three months. Then the number of direct labor hours worked each quarter on cost center 22 is

$$40\,\text{hr/week} \times 4\ \text{weeks/month} \times 3\ \text{months}$$

$$= 4780\,\text{hr}$$

$$\text{Total number of work hours per quarter}$$

$$= 488\,\text{hr/employee} \times 150\ \text{direct labor employees}$$

$$= 72{,}000\,\text{hr}$$

2.5.3 Direct Labor Determination

Our first step is to determine the direct labor rate, L_D, in Eq. (2). This term is merely a composite average of all of the direct labor hourly wage rates, including benefits, on this cost center. For example purposes we will assume that this average is \$10.00 per hour.

2.5.4 Expense Rate Determination

The expense rate in Eq. (2) recovers the cost of indirect labor employees. There are two types of indirect labor costs that need to be recovered. The first is the cost of indirect labor on cost center 22 itself. The second is the cost of indirect labor on pure indirect labor costs centers that support the manufacturing activities of cost center 22.

The average salary levels of indirect labor employees, by cost center, are summarized in Table 6. The average indirect salary on cost center 22 is \$28,000 per year. We need to recover one-fourth of this amount for the next quarter for the 50 indirect employees who work in cost center 22. This amount is

$$\$7000/\text{employee} \times 50\ \text{employees} = \$350{,}000$$

The indirect labor cost centers in Table 5 are cost centers 30, 41, 43, 50, 60, 71, and 72. Direct labor cost centers 21, 23, and 42 also have indirect costs. However, these costs are recovered through the labor/overhead rates that will be determined and associated with these cost centers.

The indirect labor cost centers support all four of the direct labor cost centers. A common way to amortize these costs over the direct labor cost centers is on the basis of the total number of employees (direct and indirect) on each of the direct labor cost centers. From Table 5, cost centers 21, 22, 23, and 42 have employee totals of 100, 200, 200, and 50, respectively. According to the proration logic, pure indirect labor cost centers support the direct labor cost centers proportionately on the basis of people. Therefore, the percentage for cost center 22 would be determined as

$$200/(100 + 200 + 200 + 50) = 200/550$$

The total number of employees on each of the four cost centers are used since the other pure indirect labor cost centers support all of cost center 22, not just the direct

Table 6 Average Salary Levels of Indirect Employees

Cost center	Description	Average salary level (\$)
21	Manufacturing Engineering	32,000
22	Machining	28,000
23	Assembly	28,000
30	Controller	32,000
41	Quality Engineering	36,000
42	Inspection	28,000
43	Standards and Calibration	32,000
50	Personnel	36,000
60	Marketing	40,000
71	Design Engineering	44,000
72	Research and development	48,000

labor activities. The total indirect costs to be recovered are summarized in Table 7.

Column 6 in Table 7 includes no entries for cost centers 21, 23, and 42 because these indirect costs will be captured in their entirety when the budgeting process is repeated for these cost centers. There is no entry in the same column for cost center 22 because the procedure assumes that cost center 22 must pay for all of its own indirect costs. This is applicable for all cost centers that include direct labor activities.

The expense rate, E_R, for the cost center may now be determined. The total of the entries in column 6 of Table 7 is \$1,599,916. The indirect costs on cost center 22 (\$350,000) are added to this total. This sum is divided by the total number of direct labor hours expected to be worked in cost center 22 during the quarter (72,000 hr. The result is

$$E_R = \frac{\$350{,}000 + \$1{,}599{,}916}{72{,}000\,\text{hr}} = \$27.08/\text{hr}$$

2.5.5 Burden Determination

There are two types of burden rates to be considered. The first is the burden for cost center 22 itself. The second portion is for all of the pure indirect labor cost centers. As with the expense rate, the burden on cost center 22 must be recovered in its entirety during the quarter. The burden for the rest of the indirect labor cost centers is charged proportionally to cost center 22 on the basis of the total number of employees on each of the direct labor cost centers.

We assume that the burden on cost center 22 is \$500,000 and that the total burden for the quarter on all of the indirect labor cost centers is \$1,000,000. The burden rate then is calculated as

$$\begin{aligned} \text{Total Burden} &= \$500{,}000 + \$1{,}000{,}000 \times 200/550 \\ &= \$863,\,636 \\ \text{Burden Rate} = B &= \$863{,}636/72{,}000\,\text{hr} \\ &= \$11.99/\text{hr} \end{aligned}$$

2.5.6 G&A Determination

We assume that the total dollar value of G&A to be recovered for the budgetary any quarter is \$400,000. This amount is prorated on the basis of total direct employees and is charged to cost center 22 as

$$\$400{,}000 \times 200/550 = \$145{,}440$$

Therefore, the G&A rate per hour is

$$\text{G\&A} = \$145{,}440/72{,}000\,\text{hr} = \$2.02\,\text{hr}$$

2.5.7 Hourly Rate Determination and Adjustment

The four components of the labor/overhead rate for cost center 22 have now been determined. These components may now be summed as specified by Eq. (2)

Table 7 Proration of Indirect Costs for Cost Center 22

1 Cost center	2 Average indirect salary (\$)	3 Column 2 × 1/4 (\$)	4 Number of indirect employees	5 Column 3 × column 4 (\$)	6 Column 5 × 200/500 (\$)
21	32,000	8,000	50[a]	400,000	
22	28,000	7,000	50[a]	350,000[b]	
23	28,000	7,000	50[a]	350,000	
30	32,000	8,000	50	400,000	145,440
41	36,000	9,000	50	450,000	163,620
42	28,000	7,000	20[a]	140,000	
43	32,000	8,000	50	400,000	145,440
50	36,000	9,000	50	450,000	163,620
60	40,000	10,000	100	1,000,000	363,636
71	44,000	11,000	100	1,100,000	400,000
72	48,000	12,000	50	600,000	218,160
					Total 1,599,916

[a] Number of indirect employees out of total on cost center.
[b] Total must be recovered in its entirety by cost center 22.

to determine the labor/overhead (LOH) rate for cost center 22:

$$\mathrm{LOH}_{22} = \$10.00/\mathrm{hr} + 27.08/\mathrm{hr} + 11.99/\mathrm{hr} + \$2.02/\mathrm{hr} = 51.09/\mathrm{hr}$$

This procedure is repeated for the remaining direct labor cost centers, 21, 23, and 42. Each LOH rate obtained is an estimate or forecast. This is because the dollar values of indirect labor, burden, and G&A, are estimated values. Likewise, it was estimated that 72,000 hr would be worked during the quarter on cost center 22. As the budgetary quarter is completed, actual values of indirect labor, burden, G&A, and the number of direct labor hours worked will differ from the estimated values used in the LOH rate computation. This is also true of the average direct labor wage rate used in Eq. (2).

Cost engineering personnel will collect actual values for each of these parameters at the end of the quarter. This will permit an actual LOH rate to be determined. This actual rate will be either greater or less than the budgeted rate prior to the quarter that was calculated above. If the budgeted rate is greater than the actual rate, excess costs will be recovered. If the budgeted rate is less than the actual rate, insufficient costs will be recovered.

A common practice is to compare these budgeted and actual rates for all direct labor cost centers over several quarters. Subsequent estimates for the parameters in Eq. (2) may be adjusted upward or downward, as appropriate, in subsequent budgetary periods to make the actual and budgeted LOH values more closely approximate one another.

2.6 ACTIVITY BASED COSTING

Activity-based costing (ABC) has gained increased popularity in recent years as a more accurate alternative to conventional careful budgeting methods for some manufacturing activities [14]. Some cost analysts have questioned the accuracy and validity of the use of direct labor hours as a proration basis variable in cost recovery.

Some organizations have found that conventional capital budgeting cost recovery methods are not vernier enough to accurately estimate overhead costs and indirect labor amortization. Implementers of ABC find that their organizations have often been grossly undercharging their customers for small-volume production end items. High-volume items often have inflated production cost values with conventional cost recovery methods.

2.6.1 Cost Center Definition

With ABC, the number of cost centers is usually expanded to encompass small work cells, or even individual machines. Consider a work cell that contains one new and highly automated five-axis machining center. Suppose this cell contains five other manual, conventional machine tools.

With this work cell definition, parts processed by the work cell might not be accurately charged for the manufacturing procedures they require during the fabrication process. The overhead rate for the cell will seek to recover the costs of all of its machines. Parts entering the cell that require processing only by the cheaper, conventional machines, might be unfairly charged overhead for the automated machining center, even though they did not require processing by this piece of equipment.

A better cell definition might be to define the automated machining center as its own work cell. The older conventional machines might be grouped together in a second work cell. This philosophy can result in more than one hundred separate cost centers, many of which may consist of only one machine.

This type of cost center definition enables the cost of manufactured end items to be determined as a function of those manufacturing procedures arid services they consume or require during their manufacturing process sequence. This is true for both direct and indirect labor and materials. The cost recovery methodology for ABC is the subject of the subsection that follows.

2.6.2 ABC Cost Recovery Methodology

Activity-based costing defines a cost rate per unit time that is associated with each defined work center or center. This rate has been described by Ramachandran et al. [7, 8] and is illustrated by

$$R_{D,i} = \frac{(C_{DL} + C_D + C_U + C_{FS} + C_T + C_I + C_{IL} + C_O)}{H_B} \tag{3}$$

where

$R_{D,I}$ = Hourly operation cost of direct labor work center i in dollars per hour
C_{DL} = Cost of direct labor in the work center over the budgetary period in dollars
C_D = Cost of depreciation for equipment in the

work center during the budgetary period in dollars

C_U = Cost of utilities attributable to the work center during the budgetary period in dollars

C_{FS} = Cost of building floor space attributable to the work center during the budgetary period in dollars

C_T = Cost of taxes attributable to the work center during the budgetary *i* period in dollars

C_I = Cost of insuring the work center's equipment and floor space during the budgetary period in dollars

C_{IL} = Cost of indirect labor required to support the work center during the budgetary period in dollars

C_O = Cost of fringe benefits, other overhead, and any supplementary indirect labor wages required by the work center during the budgetary period in dollars

H_B = Estimated or capable number of hours during the budgetary period that the work center is expected to operate

Some of the parameters that make up Eq. (3) require interpretation. Suppose a total of *m* direct labor employees work in a given work center, *i*, during a budgetary period. The direct labor cost can then be described by

$$C_{DL} = \sum_{j=1}^{m} L_j N_{j,i} \tag{4}$$

where:

L_j = Hourly pay rate including benefits for direct labor employee *j* in dollars per hour

$N_{j,i}$ = Number of hours worked by employee *j* in work center *i* during the budgetary period

2.6.3 ABC Operation on a Quarterly Basis

The process of assessing the accuracy of work center rates using Eq. (3) is similar to the capital budgeting process previously described. Prior to a given budgetary quarter, estimates of all parameters is Eq. (3) must be compiled to,determine a budgetary or forecast cost rate for all direct labor work centers. At the end of each quarter, estimated parameters in Eq. (3) are compared with actual values at the end of the quarterly budgeting period. This enables an actual work center cost rate to be determined.

If the actual rate is less than the estimated or budgeted rate, the plant has overcharged for the services of the work center. If the reverse is true, the plant has undercharged products for the work center's use. Comparing actual and projected work center rates on a quarter by quarter basis gives the cost analysis some basis for adjusting follow-on estimates for each of the parameters in Eq. (3). This facilitates the compilation of more accurate budgeted work center rates in future budgetary quarters.

2.7 QUICK-RESPONSE ESTIMATING AND ESTIMATING SOFTWARE

Manufacturers often find it necessary to supply cost estimates with extremely short lead times. Quick-response estimating systems rely on parametric estimating techniques. Parametric estimating entails the development of mathematical relationships that relate the product's manufacturing cost to salient, identifiable features of the product being manufactured. Other factors considered in the estimating equation are production quantity and the number of times the end product has been previously manufactured.

2.7.1 Development of Estimating Equations

Multivariate regression analysis is a tool that can be used to develop parametric estimating relationships. Both linear and nonlinear equations can be fitted to historical cost data using this approach. Many spreadsheet packages currently permit multivariate regression analysis to readily be completed on a personal computer.

It is important to assess the quality of obtained curve fits that relate product cost to features of a the end item being produced. A coefficient of determination (r^2) of at least 80% is recommended. It is also important that developed estimating relationships be validated before being used. This may be accomplished by dividing historical cost data bases into two parts. The parametric equations can be developed using the first half of the historical data. The developed relationships can then be "tested" on the second half, comparing projected cost from the obtained equations, with actual historical costs that were incurred in practice.

2.7.2 Definition of Part Families

Estimating relationships should be developed by part family. For example, cost data used to make an FM radio should obviously not be used to estimate the cost of an automobile drive shaft. *Group technology* (GT) defines manufacturing production cells that are dedicated to producing parts with geometrically similar attributes. Coding and classification methods are used to define families of parts that are manufactured by GT cells. Often, parts within defined GT families are candidates for the use of the same parametric cost relationship.

Group technology lends itself most readily to metalworking operations. For electronics manufacturing, part family definition must be more intuitive. Example part families for electronics manufacturing might be as defined below.

Point-to-point wired cables
Printed circuit board assemblies
Wired chassis assemblies to interface printed circuit boards
Interface wiring between separate chassis assemblies
Metalworking operations to punch and machine the chassis prior to interfacing wiring.

2.7.3 Production Mix

Parametric estimating systems are much easier to implement in high volume, low mix types of production environments. Organizations should expect to make an investment on the order of person-months to person-years in terms of statistician and cost engineering time to develop and validate such relationships prior to their being used.

2.7.4 Development and Use of Estimating Software

In many cases, software is developed to assist in expediting the completion of cost estimates and order quotations to customers. In most cases, it is more expeditious to generate such software "in-house" as opposed to procuring commercially available software packages.

Such software must have the organization's product line "embedded" within its architecture. This requirement often makes it necessary to modify commercial software packages. Required modification costs can be extensive enough to justify "company-generated" software. This justification is enhanced by the fact that many commercially available estimating software packages are extremely expensive. Commercially available software is also usually difficult to interface with the user organization's historical cost records.

Parametric estimating relationships, if previously developed by the organization, are easily incorporated into estimating software that is written "in-house." In-house software is not necessarily less expensive, but is almost always more useful to the organization in the long run.

2.8 IMPACT OF AUTOMATION ON ESTIMATING PROCEDURES

Manufacturing automation has, and continues to have a dramatic effect on conventional estimating procedures. High levels of automation reduce the amount of human direct labor associated with the manufacture of parts. Overhead costs are increased, usually due to the procurement of expensive automated machining centers, industrial robots, and assembly equipment. Indirect labor costs tend to increase due to programming and maintenance costs.

Product quality is generally enhanced as a result of automation. This has a positive effect in reducing warranty and return costs. Cost variation between parts and groups of parts is greatly reduced. Machine cycle times are more constant, and except for unscheduled maintenance, are functions of programs for the machines and transfer mechanisms. This makes the incorporation of learning curves (previously described), of less importance than for manufacturing procedures that incorporate low automation levels.

Production, inspection, and part transfer times that are machine dependent need to be incorporated in cost estimates that will use these manufacturing facilities and machines. The large capital investments tied up in automated manufacturing facilities increasingly mandate the use of activity based costing. The individual work center definition and cost rates provided by ABC more accurately associate automation costs with end products that use or consume automated facilities and resources during the manufacturing process.

Automated equipment is almost always more expensive than its conventional counterpart. Manufacturing automation results in recurring costs associated with more expensive maintenance, programming costs (generation, debugging, and storage). When a plant invests heavily in automation, there is extreme pressure to have high equipment utilization levels. This may prompt the use of such equipment for some parts that might have otherwise been manufactured by cheaper, more conventional methods.

The cost increases and savings associated with manufacturing automation have been described in detail by

Butler et al. [7, 9, 10]. These authors have developed software that enables prospective users to assess the net cost increase or savings associated with manufacturing automation alternatives. Readers desiring more information on this subject are urged to consult these references.

2.9 SUMMARY

This chapter has overviewed manufacturing cost estimating and recovery techniques and methods. There have been many books that have been devoted exclusively to the extensive coverage of these topics. Readers desiring more information on any of the subjects of this chapter are urged to consult the list of references that follows.

REFERENCES

1. M Malstrom. What Every Engineer Should Know About Manufacturing Cost Estimating. New York: Marcel Dekker, 1981, pp 1–48.
2. PF Oswald. Cost Estimating for Engineering and Management. Englewood Cliffs, NJ: Prentice-Hall, p 1–4.
3. R Vernon, ed. Realistic Cost Estimating for Manufacturing. Dearborn, MI: Society of Manufacturing Engineers, 1968, p 1.
4. EM Malstrom, ed. Manufacturing Cost Engineering Handbook. New York: Marcel Dekker, 1984.
5. EM Malstrom. Cost estimating and control. In: R Veilleux, ed. Tool and Manufacturing Engineers Handbook, vol. 5. Dearborn, MI: Society of Manufacturing Engineers, 1988, pp 4.1–12.
6. EM Malstrom, RL Shell. A review of product improvement curves. Manuf Eng 82(5): 1979, pp 70–76.
7. DP Butler, EM Malstrom, K Ramachandran. A computerized ABC model for a job shop environment. AACE Trans, June: 1995, p 9.1–3.
8. K Ramachandran, DP Butler, EM Malstrom. A computer assisted model for activity based costing. Proceedings of the 22nd International Conference on Computers and Industrial Engineering. Cairo, Egypt, December 1997.
9. DP Butler, EM Malstrom, SC Parker. A tutorial model for the economic evaluation of automated manufacturing systems. Cost Eng 38(6): 1996, pp 25–32.
10. DP Butler, EM Malstrom, SC Parker. Assessing the true cost savings associated with the procurement of automated systems. Technical paper MM93-384. Dearborn, MI: Society of Manufacturing Engineers, 1993.
11. CM Creese, M Adithan, BS Pabla. Estimating and Costing for the Metal Manufacturing Industries. New York: Marcel Dekker, 1992, p 12.
12. EM Malstrom, W Beatty. A cost recovery methodology for automated manufacturing workcells. Cost Eng 34(5): 1992, pp 15–20.
13. PE Ostwald, ed. Manufacturing Cost Estimating. Society of Manufacturing Engineers, 1980, p 9.
14. ER Sims, Jr. Precision Manufacturing Costing. New York: Marcel Dekker, 1995.