INDUSTRIAL CONTROL SYSTEMS

Mathematical and Statistical Models and Techniques

Adedeji B. Badiru
Oye Ibidapo-Obe
Babatunde J. Ayeni

CRC Press
Taylor & Francis Group

INDUSTRIAL CONTROL SYSTEMS

Mathematical and Statistical Models and Techniques

Industrial Innovation Series

Series Editor

Adedeji B. Badiru
Department of Systems and Engineering Management
Air Force Institute of Technology (AFIT) – Dayton, Ohio

PUBLISHED TITLES

Computational Economic Analysis for Engineering and Industry
Adedeji B. Badiru & Olufemi A. Omitaomu

Conveyors: Applications, Selection, and Integration
Patrick M. McGuire

Global Engineering: Design, Decision Making, and Communication
Carlos Acosta, V. Jorge Leon, Charles Conrad, and Cesar O. Malave

Handbook of Industrial Engineering Equations, Formulas, and Calculations
Adedeji B. Badiru & Olufemi A. Omitaomu

Handbook of Industrial and Systems Engineering
Adedeji B. Badiru

Handbook of Military Industrial Engineering
Adedeji B.Badiru & Marlin U. Thomas

Industrial Control Systems: Mathematical and Statistical Models and Techniques
Adedeji B. Badiru, Oye Ibidapo-Obe, & Babatunde J. Ayeni

Industrial Project Management: Concepts, Tools, and Techniques
Adedeji B. Badiru, Abidemi Badiru, and Adetokunboh Badiru

Inventory Management: Non-Classical Views
Mohamad Y. Jaber

Kansei Engineering - 2 volume set

- **Innovations of Kansei Engineering,** *Mitsuo Nagamachi & Anitawati Mohd Lokman*
- **Kansei/Affective Engineering,** *Mitsuo Nagamachi*

Knowledge Discovery from Sensor Data
Auroop R. Ganguly, João Gama, Olufemi A. Omitaomu, Mohamed Medhat Gaber, and Ranga Raju Vatsavai

Learning Curves: Theory, Models, and Applications
Mohamad Y. Jaber

Moving from Project Management to Project Leadership: A Practical Guide to Leading Groups
R. Camper Bull

Quality Management in Construction Projects
Abdul Razzak Rumane

Social Responsibility: Failure Mode Effects and Analysis
Holly Alison Duckworth & Rosemond Ann Moore

STEP Project Management: Guide for Science, Technology, and Engineering Projects
Adedeji B. Badiru

Systems Thinking: Coping with 21st Century Problems
John Turner Boardman & Brian J. Sauser

Techonomics: The Theory of Industrial Evolution
H. Lee Martin

Triple C Model of Project Management: Communication, Cooperation, Coordination
Adedeji B. Badiru

FORTHCOMING TITLES

Essentials of Engineering Leadership and Innovation
Pamela McCauley-Bush & Lesia L. Crumpton-Young

Modern Construction: Productive and Lean Practices
Lincoln Harding Forbes

Project Management: Systems, Principles, and Applications
Adedeji B. Badiru

Statistical Techniques for Project Control
Adedeji B. Badiru

Technology Transfer and Commercialization of Environmental Remediation Technology
Mark N. Goltz

INDUSTRIAL CONTROL SYSTEMS

Mathematical and Statistical Models and Techniques

Adedeji B. Badiru
Oye Ibidapo-Obe
Babatunde J. Ayeni

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 2011909

International Standard Book Number-13: 978-1-4200-7559-5 (eBook - PDF)

To Iswat, Sola, and Flora, the rocks behind the writing

Contents

Preface

This book presents the mathematical foundation for building and implementing industrial control systems. It contains mathematically rigorous models and techniques for control systems, in general, with specific orientation toward industrial systems. Industrial control encompasses several types of control systems. Some common elements of industrial control systems include supervisory control and data acquisition systems, distributed control systems, and other generic control system configurations, such as programmable logic controllers, that are often found in industrial operations and engineering infrastructures.

Industrial control systems are not limited to production or manufacturing enterprises, as they are typically used in general industries such as electrical, water, oil and gas, and data acquisition devices. Based on information received from remote sensors, automated commands can be sent to remote control devices, which are referred to as field devices. Field devices are used to control local operations. These may include opening and closing valves, tripping breakers, collecting data from sensors, and monitoring local operating conditions. All of these are governed by some form of mathematical representation. Thus, this book has great importance in linking theory and practice.

Distributed control systems are used to control industrial processes such as electric power generation, oil and gas refineries, water and wastewater treatment, and chemical, food, and automotive production. They are integrated as a control architecture containing a supervisory level of control overseeing multiple, integrated subsystems that are responsible for controlling the details of a localized process. Product and process controls are usually achieved by deploying feed-back or feed-forward control loops, whereby key product and/or process conditions are automatically maintained around a desired set point. To accomplish the desired product and/or process tolerance around a specified set point, specific programmable controllers are used.

Programmable logic controllers provide Boolean logic operations, timers, and continuous control. The proportional, integral, and/or differential gains of the continuous control feature may be tuned to provide

the desired tolerance as well as the rate of self-correction during process disturbances. Distributed control systems are used extensively in process-based industries. Programmable logic controllers are computer-based solid-state devices that control industrial equipment and processes. They are used extensively in almost all industrial processes. That makes this book very important and appealing to a wide range of readers, including students, professors, researchers, practitioners, and industrialists. The book is an amalgamation of theoretical developments, applied formulations, implementation processes, and statistical control. The contents of this book include

- Industrial innovations and systems analysis
- Systems fundamentals
- Technical systems
- Production systems
- Systems filtering theory
- Systems control
- Linear and nonlinear systems
- Switching in systems
- Systems communication
- Transfer systems
- Statistical experimental design models (factorial design and fractional factorial design)
- Response surface models (central composite design and Box–Behnken design)

Adedeji B. Badiru
Oye Ibidapo-Obe
Babatunde J. Ayeni

For MATLAB® and Simulink® product information, please contact:

The MathWorks, Inc.
3 Apple Hill Drive
Natick, MA, 01760-2098 USA
Tel: 508-647-7000
Fax: 508-647-7001
E-mail: info@mathworks.com
Web: www.mathworks.com

Acknowledgments

Developing this book was an arduous undertaking that went through several stages of trials and tribulations stretching over several years. When physical and mental exhaustion dictated that we should quit, we somehow found the strength to continue. Such strength often came from family members, friends, and professional colleagues. For this, we thank everyone. We thank the editorial and production staff members at CRC Press for their steadfast support throughout the production of the book. Several individuals, too numerous to be listed here, played major roles in contributing to the final quality and content of the book. For this, we are very grateful.

Authors

Adedeji B. Badiru is a professor and department head of the Department of Systems and Engineering Management at the Air Force Institute of Technology, Dayton, Ohio. He was previously professor and department head of industrial and information engineering at the University of Tennessee in Knoxville. Prior to that, he was professor of industrial engineering and dean of University College at the University of Oklahoma. He is a registered professional engineer, a certified project management professional, a fellow of the Institute of Industrial Engineers, and a fellow of the Nigerian Academy of Engineering. He received his BS in industrial engineering, MS in mathematics, and MS in industrial engineering from Tennessee Technological University and his PhD in industrial engineering from the University of Central Florida. His research interests include mathematical modeling, project modeling and analysis, economic analysis, systems engineering, and productivity analysis and improvement. He is the author of several books and technical journal articles. He is also the editor of the *Handbook of Industrial and Systems Engineering* and coeditor of the *Handbook of Military Industrial Engineering*.

Professor Badiru is a member of several professional associations, including the Institute of Industrial Engineers, the Institute of Electrical and Electronics Engineers, the Society of Manufacturing Engineers, the Institute for Operations Research and Management Science, the American Society for Engineering Education, the American Society for Engineering Management, the New York Academy of Science, and the Project Management Institute. He has served as a consultant to several organizations around the world and has won many awards for his teaching, research, publications, administration, and professional accomplishments. He is on the editorial and review boards of several technical journals and book publishers. He has served as an industrial development consultant to the United Nations Development Program, and is also an engineering program evaluator for ABET.

Oye Ibidapo-Obe is a professor of systems engineering and former vice-chancellor of the University of Lagos, Nigeria. His major academic contributions are in the areas of control and information systems, including specialized interests in stochastic/optimization problems in engineering and reliability studies and simulation and animation studies (with applications to urban transportation, water resources, biomedics, expert systems, etc.). He pioneered studies on stochastic methods in mechanics, and was also part of the pioneering efforts on robotics developments and applications, especially in the area of sensor and actuator placement. Professor Ibidapo-Obe served at the University of Lagos from 1972 to 1983, starting as a graduate assistant, then as lecturer grade II and finally as a professor. He subsequently became the deputy vice-chancellor in 2000 and took on the role of acting vice-chancellor between September 2000 and April 2002. He was appointed substantive vice-chancellor in May 2002, and successfully served until April 2007. He was the chairman of the Committee of Vice Chancellors of Nigerian Universities. He was awarded the Best Vice Chancellor's Prize for the Nigerian University System (NUS) in 2004 and 2005. In 2004, Professor Ibidapo-Obe received the prestigious Fellowship of the Academy of Science (of which he is the current president (2009–). He is a member of the Nigerian Academy of Engineering, the Nigerian Computer Society, and the Mathematical Association of Nigeria. He also received national honor as Officer of the Federal Republic (OFR) by the Federal Republic of Nigeria.

Babatunde J. Ayeni is currently an advanced application development specialist with 3M Company. He possesses a high level of expertise in industrial statistics and petroleum engineering with significant application experience in major 3M businesses in such areas as quality, R&D, oil and gas, Six sigma, supply chain, manufacturing and laboratory. He is a recognized expert in the field of statistics both within and outside 3M. He has excellent customer focus and has been providing leadership on statistical methods since 1988. He has had substantial success both in the United States and Europe and on many occasions has received commendations from his clients and management. He also has a strong research capability as evidenced by his more than 30 technical publications and a book on quality and process improvement published in 1993.

Dr. Ayeni received his BS and MS in petroleum engineering and his PhD in statistics, all from the University of Louisiana. Before joining 3M Statistical Consulting Group in 1988, he served as a professor and chairman in the Department of Technology at Southern University in Louisiana. He has provided consulting support to most 3M business units in the United States and Europe, with major contributions to the polyester film plant in Caserta, Italy, and the development of a feedback controller design that controlled a printing plate process in Sulmona, Italy.

Dr. Ayeni has received several academic honors: Pi Mu Epsilon (Mathematics Honor), Phi Eta Sigma (Freshman Honor), Tau Beta Pi (Engineering Honor), Pi Epsilon Tau (Petroleum Engineering). He has served on the editorial boards of *Industrial Honor Engineering Journal* and *Journal of Operations and Logistics*. Dr. Ayeni is a member of the Society of Petroleum Engineers and the American Statistical Association. Ayeni was a recepient of Federal Government Nigeria Scholarships leading to BS, MS, and PhD degrees. He has contributed substantially to quality and productivity improvements, cost savings, and product innovations in 3M laboratory R&D, quality, oil and gas, and manufacturing environments. He is a recipient of the 1997 3M Specialty Materials Film Division Quality Achievement Award, the 1995 3M Kenneth D. Kotnour Award for Leadership in Statistical Thinking, the 1994 professional achievement award given by Minnesota Institute for Nigerian Development (MIND), the 1994 3M Information Technology Division Pyramid of Excellence Award, and the 1992 3P Award from 3M in recognition of a cost-savings approach to pollution prevention based on a new multivariable sampling plan.

From 1990–1992, Dr. Ayeni served as a quality examiner on the Minnesota Quality Award Board of Examiners. He is profiled in *Tall Drums: Portraits of Nigerians Who Are Changing America*. In 1993, Dr. Ayeni served as an industrial consultant to the United Nations TOKTEN program.

chapter one

Mathematical modeling for product design[*]

Product design has normally been performed by teams, each with expertise in a specific discipline such as material, structural, electrical, systems. Traditionally, each team would use its members' experience and knowledge to develop the design sequentially. Collaborative design decisions explore the use of optimization methods to solve the design problem, incorporating a number of disciplines simultaneously. It is known that the optimum of the product design is superior to the design found by optimizing each discipline sequentially due to the fact that it is enabled to exploit the interactions between the disciplines. In this chapter, a bi-level decentralized framework based on memetic algorithm (MA) is proposed for a collaborative design decision using a forearm crutch as the case. Two major decisions are considered: weight and strength. In this chapter, we introduce two design agents for each of the decisions (Wu et al., 2011). At the system level, one additional agent termed "facilitator agent" is created. Its main function is to locate the optimal solution for the system objective function that is derived from the Pareto concepts; thus, Pareto optimum for both weight and strength is obtained. It is demonstrated that the proposed model can converge to Pareto solutions.

Introduction

Under a collaborative design paradigm, the first common topic is multidisciplinary design optimization (MDO), which is defined as "an area of research concerned with developing systematic approaches to the design of complex engineering artifacts and systems governed by interacting physical phenomena" (Alexandrov, 2005). Researchers agree that interdisciplinary coupling in the engineering systems presents challenges in formulating and solving MDO problems. The interaction between design analysis and optimization modules and multitudes of users is complicated by departmental and organizational divisions. According to Braun and Kroo (1997), there are numerous design problems where the product is so complex that a coupled analysis driven by a single design optimizer

* T. Wu, S. Soni, M. Hu, F. Li, and A. Badiru, The application of Memetic algorithms for forearm crutch design: A case study, *Mathematical Problems in Engineering*, 3 (4), 1–15, 2011.

is not practical as the method becomes too time-consuming either because of the lead time needed to integrate the analysis or because of the lag introduced by disciplinary sequencing. Some researchers have taken project management as a means to facilitate and coordinate the design among multidisciplines (Badiru and Theodoracatos, 1995; Thal et al., 2007).

Early advances in MDO involve problem formulations that circumvent the organizational challenges, one of which is to protect disciplinary privacy by not sharing full information among the disciplines. It is assumed that a single analyst has complete knowledge of all the disciplines. As indicated by Sobieszczanski-Sobieski and Haftka (1997), most of the work at this phase aims to tackle the problems by a single group of designers within one single enterprise environment where the group of designers shares a common goal and requires less disciplinary optimum. The next phase of MDO gives birth to two major techniques: optimization by linear decomposition (OLD) and collaborative optimization (CO). These techniques involve decomposition along disciplinary lines and global sensitivity methods that undertake overall system optimization with minimal changes to disciplinary design and analysis. However, Alexandrov and Lewis (2000) explore the analytical and computational properties of these techniques and conclude that disciplinary autonomy often causes computational and analytical difficulties that result in severe *convergence* problems.

Parallel to these MDO developments there also evolves the field of decision-based design (Simpson et al., 2001; Hernandez and Seepersad, 2002; Choi et al., 2003; Wassenaar and Chen, 2003; Huang, 2004; Li et al., 2004), which provides a means to model the decisions encountered in design and aims at finding "satisfying" solutions (Wassenaar et al., 2005; Nikolaidis, 2007). Research in decision-based design includes the use of adaptive programming in design optimization (Simon, 1955) and the use of discrete choice analysis for demand modeling (Hernandez and Seepersad, 2002; Choi et al., 2003). In addition, there has been extensive research ranging from single-objective decision-based design (Hernandez and Seepersad, 2002) to multi-objective models (Lewis and Mistree, 1998, 1999). It combines game theory, utility theory, and decision sciences for collaborative design that can be conducted among a group of designers from different enterprises. This technique has several computational difficulties in calculating the "best reply correspondence" and the rational reaction sets, especially when the designs are very complex. Besides, several approximations like using response surfaces within these techniques make them prone to errors (Fernandez et al., 2005).

Note most methods reviewed so far have strict assumptions on the utility functions and/or constraints (e.g., convexity and quasi-linear of the functions), which limits the application to product design. In this research, we explore the use of a heuristic method, memetic algorithm (MA), and a combination of local search (LS) and genetic algorithm (GA) to a forearm crutch

design that has a non-convex objective for one of the decisions. Forearm crutches had been exclusively used by people with permanent disability. Nowadays, it is beginning to serve for some shorter-term purposes as well. The design of the forearm crutch needs to consider multidisciplinary decisions. For example, the structure designer wants to ensure that the design is lightweight. The material engineer wants composite material to have the right number of layers at right angles to make the product durable. The outsourcing engineer wants the supplier to provide low-cost, highly reliable, lightweight parts. Another important factor impacting the design is cost. Here, we introduce the design agent for each disciplinary decision problem and one system agent facilitating the communication among the design agents and guiding the design to convergence. To achieve this, the overall decision space is partitioned into two sets: one for coupled variables (the ones shared by at least two designers) and one for local variables (the ones that can be fully controlled by each designer). Next, an iterative process between design agent decisions on local variables and facilitator agent decisions on the whole design space launches. It is demonstrated that a converged Pareto optimum is achieved after a number of iterations for the forearm crutch design which has nonlinear-form decision functions.

This chapter is organized as follows: the related literature is briefly reviewed in Section 1.2, followed by the detailed explanation on the proposed bi-level decentralized framework in Section 1.3. The forearm crutch case is explained in Section 1.4 with the conclusions being drawn in Section 1.5.

Literature review

CO, introduced by Braun and Kroo (1997), is a bi-level optimization approach where a complex problem is hierarchically decomposed along disciplinary boundaries into a number of subproblems which are brought into multidisciplinary agreement by a system-level coordination process. With the use of local subspace optimizers, each discipline is given complete control over its local design variables subject to its own disciplinary constraints. The system-level problem sets up target values for variables from each discipline. Each discipline sets the objectives to minimize the discrepancy between the disciplinary variable values and the target values. The system-level optimization problem is formulated as minimizing a global objective subject to interdisciplinary consistency constraints. The interdisciplinary consistency constraints are equality constraints that match the system-level variables with the disciplinary variables. In OLD (Sobieszcanaski-Sobieski, 1982, 1988; Sobieszcanaski-Sobieski et al., 1985), the disciplines are given the autonomous task of minimizing disciplinary design infeasibility while maintaining system-level consistency. The system-level problem is to drive design infeasibility to zero. At the

local-level problem the disciplines use their local degrees of freedom to minimize the violation of the disciplinary design constraints, subject to matching the target value for the disciplinary output that is fed into the discipline. Balling and Sobieszcanaski-Sobieski (1994) introduce a combination of CO and OLD where the disciplinary subproblems minimize the discrepancy in the system-level targets as well as the disciplinary design infeasibility given the disciplinary design constraints.

Both CO and OLD depend on a design problem's amenability to hierarchical decomposition with the system objective explicitly defined. On the other hand, concurrent subspace optimization (CSSO) (Sobieszcanski-Sobieski, 1988) is a nonhierarchic system optimization algorithm that optimizes decomposed subspaces concurrently, followed by a coordination procedure for directing system problem convergence and resolving subspace conflicts. In CSSO, each subspace optimization is a system-level problem with respect to the subset of the total system design variables. Within the subspace optimization, the nonlocal variables that are required to evaluate the objective and the constraint functions are approximated using global sensitivity equation (GSE). Interested readers are referred to Sobieszczanski-Sobieski (1988) for detailed description of GSEs.

The bi-level integrated synthesis (BLISS) (Sobieszczanski-Sobieski and Kodiyalam, 1998) method uses a gradient-guided path to reach the improved system design, alternating between the set of modular design spaces (the disciplinary problems) and the system-level design space. BLISS is an all-in-one method in that the complete system analysis is performed to maintain multidisciplinary feasibility at the beginning of each cycle of the path. The overall problem is decomposed such that a set of local optimization problems deal with the detailed local variables which are large in number, and one system-level optimization problem deals with a small number of global variables.

Decision-based design (Simpson et al., 2001; Fernandez et al. 2002; Hernandez and Seepersad 2002; Choi et al., 2003) is a paradigm focusing on distributed and collaborative design efforts. For the cases where continuous variables are used, adaptive linear programming (Lewis and Mistree, 1999) is employed; in case of mixed discrete and continuous variables, foraging-directed adaptive linear programming has been used (Lewis and Mistree, 1999). In a noncooperative environment, game theoretic principles are used to arrive at the best overall design (Lewis and Mistree, 1998, 1999). Recently, design-for-market systems grows out of the decision-based design and emerges as an area focusing on establishing a solid basis in decision theory, by taking microeconomics into account to support engineering design. Kumar et al. (2007) propose a hierarchical choice model based on discrete choice analysis to manage and analyze customer preference data in setting design targets. Azarm's group studies new product designs that are robust from two perspectives—from the

engineering perspective in terms of accounting for uncertain parameters and from the market perspective in terms of accounting for variability in customer preferences measurement (Besharati et al., 2006). They conclude incorporating consumer heterogeneity in considering that the variability in customer preferences may have significant impact on the ultimate design. Research led by Michalek explores the use of game-theoretic approach to finding market equilibrium under various regulation scenarios (Shiau and Michalek, 2007). A metric for agility measurement is introduced by Seiger et al. (2000) to explore the product development for mass customization.

In general, some common criticisms and/or challenges facing collaborative design decisions are the convergence and information-sharing issues:

- Will the decision model converge? If yes, under what condition (assumptions on the function form and design spaces) will it converge? How fast will it converge?
- Most models (CO, OLD, BLISS, etc.) take a top-down approach with the full knowledge of the design space (e.g., the form of utility functions, constraints) being available. For the cases when the design information is partially known, what decision model is appropriate?

To address these challenges, we propose a general decision framework based on MA that allows distributed design teams to arrive at Pareto solutions which is explained in Section 1.3.

Memetic algorithm and its application to collaborative design

MA is one of the emerging areas in evolutionary computation. It integrates GA with LS to improve the efficiency of searching complex spaces. In MA, GA is used for global exploration while LS is employed for local exploitation. The complementary nature of GA and LS makes MA an attractive approach for large-scale, complex problems, for example, collaborative design.

Proposed framework for collaborative design

Let us consider a general collaborative design with z design teams. The problem can be represented as

$$\text{Min } \mathbf{J}(\mathbf{x}, \mathbf{y})$$

$$\text{St. } \mathbf{g}(\mathbf{x}) \leq 0$$

$$\mathbf{h}(\mathbf{x}, \mathbf{y}) \leq 0$$

$$x_i^{LB} \le x_i \le x_i^{UB} \quad (i = 1, \ldots, n_1)$$

$$y_i^{LB} \le y_i \le y_i^{UB} \quad (i = 1, \ldots, n_2)$$

where

$\mathbf{J} = [J_1(\mathbf{x}, \mathbf{y}) \ldots J_Z(\mathbf{x}, \mathbf{y})]^T$
$\mathbf{x} = [x_1 \ldots x_{n1}]^T$
$\mathbf{y} = [y_1 \ldots y_{n2}]^T$
$\mathbf{g} = [g_1(\mathbf{x}) \ldots g_{m1}(\mathbf{x})]^T$
$\mathbf{h} = [h_1(\mathbf{x}, \mathbf{y}) \ldots h_{m2}(\mathbf{x}, \mathbf{y})]^T$
$\mathbf{x}$ is the set of n_1 local variables
$\mathbf{y}$ is the set of n_2 coupled variables
$\mathbf{g}$ is the set of m_1 local constraints
$\mathbf{h}$ is the set of m_2 coupled constraints

Figure 1.1 illustrates the iterative decision process between system facilitator agent and disciplinary design agents. First, the facilitator initializes the global solution space over both local and coupled variables. For any solution, for example, [**x***, **y***], each design agent will execute local optimizer over the sub-design space which consists of **x** only, that is, Min **J**(**x**, **y***). The results fed back to the facilitator are the value of objective function and the gradient of objective function over coupled variables. The facilitator will (1) employ local search for the recent results updated by each designer using the related gradient information for the improved design (2) next, traditional GA operators, crossover, and mutation are applied to introduce new candidates to the solution space.

Pseudo Code

This section shows the layout of the Pseudo Code for the proposed methodology as illustrated in Figure 1.2.

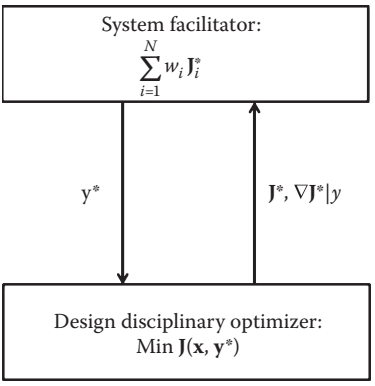


Figure 1.1 Overall decision framework.

(1) //**Initialization**
(2) The set of final Pareto solutions $FP = \emptyset$;
(3) The set of GA population $PS = \emptyset$;
(4) The set of weights combination $WS = \emptyset$;
(5) Given N objective functions, we have $\sum_{i=1}^{N} w_i J_i(x_i, \mathrm{y})$
(6) **Begin** (at facilitator agent level)
(7) //*Enumerate weights combination*
(8) Set $w_1 = w_2 = \cdots = w_{N-1} = 0$;
(9) Given weight step size $1/W$;
(10) Let each w_i ($i = 1, \ldots, N-1$) increases $1/W$, $w_N = 1 - w_1 - \cdots w_{N-1}$, and add ($w_1, w_2, \ldots, w_N$) to WS;
(11) //*Weights loop*
(12) For each weights combination ($w_1, w_2, \ldots, w_N$) in WS, $\sum_{i=1}^{N} w_i J_i(x_i, \mathrm{y})$ is constructed;
(13) //*GA loop*
(14) //*Initialization*
(15) Generate random population of P solutions and add them to PS;
(16) For $n = 1$ to maximum # of generations for GA loop;
(17) //*Crossover and mutation*
(18) Random select two parents p_a and p_b from PS;
(19) Generate two offspring p'_a and p'_b by crossover operator;
(20) if p'_a and/or p'_b are not feasible, generate new feasible offspring
(21) p''_a and/or p''_b using mutation operator;
(22) //*Selection*
(23) Using fitness function $\left(\sum_{i=1}^{N} w_i J_i(x_i, \mathrm{y})\right)$ evaluate the solution, update PS with improved solutions;
(24) //*Local search loop*
(25) For each chromosome p_j in PS;
(26) Call each Design Agent for local optimization on x (note different optimization engines can be employed based on the design disciplines);
(27) Given updates from Design Agent on x, Facilitator agent employs subgradient algorithm [19] as local search algorithm to iteratively locate improved solution p'_j with respect to y;
(28) Next p_j;
(29) //*Pareto filter*:
(30) For each chromosome p_j in the set PS;
(31) If p_j is not dominated by all the solutions in the set FP;
(32) Add P_j to the set FP;
(33) Else If there are solutions in the set FP are dominated by p_j;
(34) ← Remove those solutions in the set FP;
(35) End If;
(36) Next p_i;
(37) Next n;
(38) **End;**

Figure 1.2 Pseudo Code 1.

Parameters

N, no. of disciplinary design agents
w_i, weight for the objective function of ith disciplinary design agent, where $i = 1,\ldots, N$
$1/W$, weight step size
P, population size

Pseudo Code lines

As shown in the Pseudo Code, there exist three loops, from outer to inner in the proposed method: weight enumeration (lines 11–37), GA loop (lines 13–37), and local search loop (lines 24–28). That is, given a weights combination (e.g., $w_1 = 0.3$, $w_2 = 0.7$ for two agents), GA is triggered, which applies crossover and mutation operators and selection mechanism (in this case study, elitism selection is employed) for the population update. In addition, for the updated population, local search is further employed to identify improved solutions within the neighborhood. This is achieved by having sub-gradient information from each designer on the coupled variables fed back to the facilitator. Specifically, given any chromosome from the population, each design agent assumes the coupled variables are set and thus conducts optimization on the local variables only. Each design agent would also study the gradients on the coupled variables. Thus, given the values of the coupled variables, both the optimal design on local variables and the sub-gradient on the coupled variables are sent back to the facilitator. Since the priorities of the objective functions reflected by the weight assignments are enumerated exhaustively, all the possible Pareto solutions are located forming the Pareto frontier. In some cases where the priority is known, the weight loop can be removed. Please note that the Pareto filter operation (lines 29–36) is triggered by the facilitator within each weight combination. That is, it is possible that some Pareto solutions given a specific weight may be dominated by the Pareto solutions obtained with other weights.

One distinguishable feature of this proposed approach from other existing methods is that the information exchanged iteratively between the facilitator and the design agent is values instead of function forms. For example, passing from the facilitator to the design agent (top-down) is the values of the coupled variable; passing from the design agent back to the facilitator (bottom-up) is the values of the objective function and associated gradient values, passing from the facilitator to the design agents (top-down) is the values of the coupled variables. The main advantage of this approach is a "black box" disciplinary optimizer can be easily plugged in. Secondly, since the facilitator explores the solution space based on the knowledge of the solution candidates ($\mathbf{x}^*$, $\mathbf{y}^*$), the candidate performance ($\mathbf{J}^*$) instead of the function formulation, a truly decentralized decision without the full

knowledge of the design space can be implemented. An industry case is explored to demonstrate the applicability of the proposed framework.

Forearm crutch design

Crutches are mobility aids used to counter mobility impairment or an injury that limits walking ability. Forearm crutches are used by slipping the arm into a cuff and holding the grip (Figure 1.3). It has been increasingly used for patients with shorter-term needs. Earlier study conducted by National Medical Expenditure Survey (NMHS) in 1987 indicates that an estimated 9.5 million (4%) noninstitutionalized U.S. civilians experience difficulty in performing basic life activities; some need crutches for leg support for walking. This number increases due to the baby boomer effect.

Typical forearm crutches are made of aluminum and are criticized by customers for being heavy, noisy, and less durable. Customers suggest that a small reduction in the weight of forearm crutches would significantly reduce the fatigue experienced by crutch users. However, the reduction in weight should not be accompanied by a strength reduction. Most crutches on the market are designed for temporary use and wear out quickly.

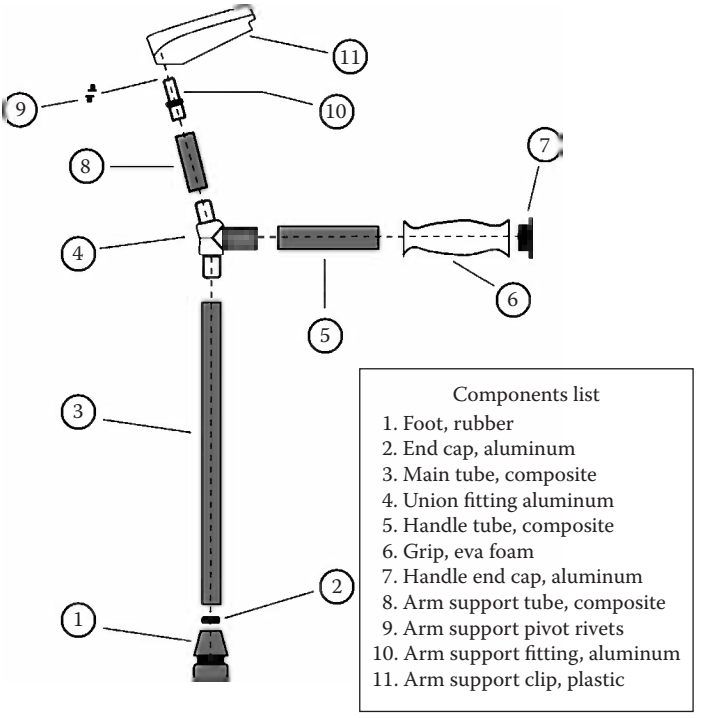


Figure 1.3 Exploded view of a forearm crutch.

Crutch users commonly have to replace their crutches two to three times a year. This drives the need to redesign forearm crutches which are robust, appropriate for a wide range of users from lighter-weight adults to users weighing up to 250 lb with considerable upper body strength and who may use them aggressively on a continuous basis.

One solution is to use composite material for crutch which is *light-weight* with good performance in *strength*. However, it comes with relatively expensive cost. After in-depth marketing survey, the design team decides to outsource the aluminum union fitting (component #4 in Figure 1.3), use appropriate composite tube, and apply adhesive from Hysol to bond the tubes with union fitting.

Aluminum union

The design team first develops a computer model based on finite element method to determine the necessary wall thickness and to calculate the load on the handle necessary to produce yielding. An aluminum union which costs $150 and stands ≥630 lb is used. The use of Hysol adhesive to bond the union with the tube needs to be tested to ensure the strength requirement is satisfied.

Composite tube

A typical composite tube is 39 in. in length. The tube can be cut into smaller pieces for the forearm crutch assembly. Approximately 2½ tubes are needed to make a pair of crutches. Here three smaller tubes are used as: handle (component 5 in Figure 1.3) which is fixed as 4.75 in., arm support tube (component 8 in Figure 1.3) which usually ranges from 6.5 to 7.8 in., main tube (component 3 in Figure 1.3) which ranges from 30.69 to 34.25 in. The inner diameter of the tube is critical to maintain the proper bond-line thickness for each adhesive joint. It ranges from 0.7605 to 0.7735 in. The outer diameter is determined by the number of plies and it ranges from 0.922 to 0.935 in. Usually, the arm support tube is less concerned with strength; the main tube needs to be tested for the strength consideration. Thus, we have two decision problems constructed: weight and strength.

Design problem formulation

Design agent for weight decision

In this research, we focus on the weights of the tubes (arm support and main tubes) and a minimization problem is introduced as

$$\text{Min: } W = W_u + W_L$$

$$\text{St: } W_u = \mathbf{rp}\left[\left(\frac{D_o}{2}\right)^2 - \left(\frac{D_i}{2}\right)^2\right] \times L_u$$

$$W_L = \rho\left[\left(\frac{D_o}{2}\right)^2 - \left(\frac{D_i}{2}\right)^2\right] \times L_L$$

$$30.6875 \le L_L \le 34.25$$

$$6.5 \le L_u \le 7.8$$

$$0.922 \le D_o \le 0.935$$

$$0.7605 \le D_i + 2T/1000 \le 0.7735$$

where

W_u (in.) is the weight of arm support tube
W_L (in.) is the weight of main tube
ρ (lb/in.^{3}) is the density of the composite tube, which is 0.08
L_u (in.) is the length of the arm support tube
L_L (in.) is the length of the main tube
D_o (in.) is the outer diameter
D_i (in.) is the inner diameter
T (mils, 1 mils = 0.001 in.) is the bondline adhesive material thickness

Design agent for strength decision

Since the strength from aluminum fitting is satisfactory from finite element analysis (FEM), the strength model will consider two potential failures: the adhesive applied joint and the strength of the main tube. Thus, the problem is constructed as

$$\text{Max: } S = \text{Min}(S_L, S_A)$$

St:

$$S_L = \frac{\rho E I}{L_L^2}$$

$$I = \rho(D_o^4 - D_i^4)/64$$

$$12 \le E \le 16$$

$$S_A = (-6.0386T^2 + 7.7811T + 4644.5) \times \frac{\rho}{4} \times (D_o^2 - D_i^2)$$

$$0.922 \le D_o \le 0.935$$

$$0.7605 \le D_i + 2T/1000 \le 0.7735$$

$$30.6875 \le L_L \le 34.25$$

$$0 \le T \le 17$$

where

S_L (lb) is the strength of the bottom of the lower tube
E (msi, 1 msi = 10^6 psi) is the modulus of elasticity
I (in.^{4}) is the area moment of inertia
S_A (lb) is the strength of the joint after applying adhesive

Implementation

For the decision problems explained earlier, optimization code written in MATLAB® is executed. Here, we provide a detailed explanation of how the system problem is constructed and how the facilitator agent guides the design agents to converge to the solution using MA.

Step 1 Initialization: Given w_1, w_2, construct system search space as $w_1W^* - w_2S^*$ (W^* and S^* are the values of the objectives from each design agent, $w_1 + w_2 = 1$).

Step 2 Real Code Genetic Algorithm: The chromosome is represented with real numbers, that is, (L_u, L_L, D_o, D_i, T, E). Note L_L, D_o, D_i, T are coupled variables, L_u is the local variable for weight agent and E is the local variable for strength agent.

Step 2.1 (Initial population): For (L_u, L_L, D_o, D_i, T, E), without losing generalization, assume a and b represent the lower bound and upper bound of one of the variables, r, be a random number $r \in [0, 1]$, we get $(b - a)r + a$. Thus, a new chromosome is generated as for the initial population. A pool of 40 chromosomes is created.

Step 2.2 (Selection of parents): To ensure all chromosomes have the chance to be selected, solutions are classified into three groups according to their fitness: high fitness level, medium fitness level, and low fitness level. The fitness is assessed based on $w_1W^* - w_2S^*$, the lower, the better.

Step 2.3 (Crossover): Given two chromosomes $C_1 = (L_u^1, L_L^1, D_o^1, D_i^1, T^1, E^1)$ and $C_2 = (L_u^2, L_L^2, D_o^2, D_i^2, T^2, E^2)$, the offspring are generated as

$$C_1' = \theta C_1 + (1-\theta)C_2$$

$$C_2' = (1-\theta)C_1 + \theta C_2$$

where $\theta \in [0,1]$.

Step 2.4 (Mutation): Mutation is applied by simply generating a new feasible solution to replace the infeasible one.

Step 3 (Local Search): The facilitator agent applies sub-gradient-method-based LS over coupled variables to improve the solutions. First, each design agent evaluates the gradients of the design decision problems

(disciplinary) with respect to the coupled variables. For example, given the coupled variables $L_L = L_L^*, D_o = D_o^*, D_i = D_i^*, T = T^*$, each decision problem is solved independently for W^* and S^*. The gradients are obtained as

$$\mathbf{l}_{W,L_L} = \left.\frac{\partial W}{\partial L_L}\right|_{L_L=L_L^*,D_o=D_o^*,D_i=D_i^*,T=T^*}, \quad \mathbf{l}_{S,L_L} = \left.\frac{\partial S}{\partial L_L}\right|_{L_L=L_L^*,D_o=D_o^*,D_i=D_i^*,T=T^*}$$

$$\mathbf{l}_{W,D_o} = \left.\frac{\partial W}{\partial D_o}\right|_{L_L=L_L^*,D_o=D_o^*,D_i=D_i^*,T=T^*}, \quad \mathbf{l}_{S,D_o} = \left.\frac{\partial S}{\partial D_o}\right|_{L_L=L_L^*,D_o=D_o^*,D_i=D_i^*,T=T^*}$$

$$\mathbf{l}_{W,D_i} = \left.\frac{\partial W}{\partial D_i}\right|_{L_L=L_L^*,D_o=D_o^*,D_i=D_i^*,T=T^*}, \quad \mathbf{l}_{S,D_i} = \left.\frac{\partial S}{\partial D_i}\right|_{L_L=L_L^*,D_o=D_o^*,D_i=D_i^*,T=T^*}$$

$$\mathbf{l}_{W,T} = \left.\frac{\partial W}{\partial T}\right|_{L_L=L_L^*,D_o=D_o^*,D_i=D_i^*,T=T^*}, \quad \mathbf{l}_{S,T} = \left.\frac{\partial S}{\partial T}\right|_{L_L=L_L^*,D_o=D_o^*,D_i=D_i^*,T=T^*}$$

The gradients of the system problem are then calculated as

$$\mathbf{l}_{L_L} = w_1\mathbf{l}_{w,L_L} - w_2\mathbf{l}_{S,L_L}$$

$$\mathbf{l}_{D_o} = w_1\mathbf{l}_{w,D_o} - w_2\mathbf{l}_{S,D_o}$$

$$\mathbf{l}_{D_i} = w_1\mathbf{l}_{w,D_i} - w_2\mathbf{l}_{S,D_i}$$

$$\mathbf{l}_T = w_1\mathbf{l}_{w,T} - w_2\mathbf{l}_{S,T}$$

Based on $\mathbf{l} = [\mathbf{l}_{L_L}, \mathbf{l}_{D_o}, \mathbf{l}_{D_i}, \mathbf{l}_T]$, the facilitator agent non-summable diminishing method to update the coupled variables. That is, at iteration $k + 1$,

$$\begin{bmatrix} L_L \\ D_o \\ D_i \\ T \end{bmatrix}^{(k+1)} = \begin{bmatrix} L_L \\ D_o \\ D_i \\ T \end{bmatrix}^{(k)} - \mathbf{a}_{k+1} \begin{bmatrix} \mathbf{l}_{L_L} \\ \mathbf{l}_{D_o} \\ \mathbf{l}_{D_i} \\ \mathbf{l}_T \end{bmatrix}^{(k)}$$

where step size α_k satisfies

$$\begin{cases} \lim_{k\to\infty} \mathbf{a}_k = 0 \\ \sum_{k=1}^{\infty} \mathbf{a}_k = \infty \end{cases}$$

The coupled variables are updated based on the aforementioned subgradient method until no further improvement of the weighted system problem is required.

Results and analysis

The Pareto frontier obtained by the proposed decentralized framework is shown in Figure 1.4. Note that the problem has min–max structure. Since this project focuses on the composite tube design (main tube and handle tube), the weight for the handle tube (component #5) is computed as

$$\texttt{rp}\left[\left(\frac{D_o}{2}\right)^2-\left(\frac{D_i}{2}\right)^2\right]\times 4.75$$

Other components in Figure 1.3 are outsourced with the weights summarized in Table 1.1.

We choose Pareto solution (A and B) to compare with the composite crutch from Ergonomics and the Invacare crutch which are two

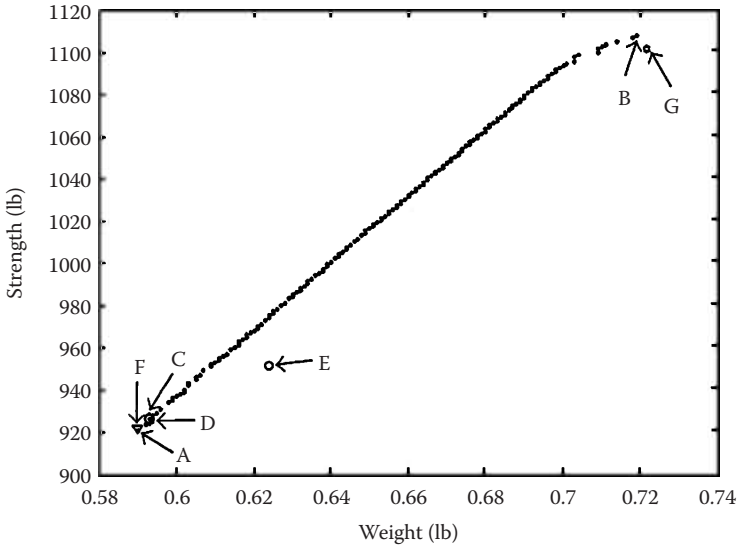


Figure 1.4 Pareto frontier in performance space for the crutch design.

Table 1.1 Weight for Each Component of the Crutch

Components (Figure 1.3)	Weight (lb)
#2	0.006
#4	0.05
#7	0.0074
#10	0.025
Others (#1, #6, #9, #11)	0.2

Table 1.2 Comparison of Crutch Weight and Strength

Crutch design	Weight (lb)	Strength (lb)
Invacare	2.3	630
Ergonomics	1	715
Pareto design (A)	0.9498	921
Pareto design (B)	1.0945	1107
Nash equilibrium (C)	0.9532	926
Weight leader strength follower (D)	0.9532	926
Strength leader weight follower (E)	0.9879	951
Weight complete control (F)	0.9499	922
Strength complete control (G)	1.0952	1100

commercial products in Table 1.2. Apparently, most composite crutches outperform Ergonomics and Invacare for both weight and strength except that Design B outweighs Ergonomics by 0.09 lb. However, Design B is more durable, with strength being 1105 lb compared with 715 lb of Ergonomics.

It is expected that the cost of the composite crutch will be high. In this case, it is around $460 in total (tube and other components shown in Figure 1.3). The price of the crutch produced by Invacare and Ergonomics ranges from $60 to $250. Although the composite crutch is several times more expensive, it lasts much longer. Instead of having replacement two to three times a year, it can be used for a number of years since the lighter composite crutch could sustain greater than 1100 lb load.

Conclusion

Collaborative design decisions involve designers from different disciplines with different specific domain knowledge. The decision process is a sequence of phases or activities where mathematical modeling can be employed. In this chapter, a bi-level distributed framework based on MA is proposed. Since the information communicated is neither the form of the decision function nor the decision space, private information is protected. In addition, in the cases where the information is not complete, the proposed framework can still guarantee the convergence to Pareto solutions. To demonstrate the applicability of the framework, a forearm crutch design is studied in detail. The results confirm converged Pareto set can be obtained for any form of decision function. While promising, the decision problems constructed are deterministic; our next step is to explore the use of this framework for design decisions under uncertainty. Computational-efficient approach in the area of reliability-based design optimization would be explored.

References

Alexandrov, N. M., Editorial—Multidisciplinary design optimization, *Optimization and Engineering*, 6 (1), 5–7, 2005.

Alexandrov, N. M. and Lewis, R. M., An analysis of some bilevel approaches to multidisciplinary design optimization, Technical report, Institute for Computer Applications in Science and Engineering, Mail Stop 132C, NASA Langley Research Center, Hampton, VA 23681-2199, June 2000.

Badiru, A. and Theodoracatos, V., Analytical and integrative expert system model for design project management, *Journal of Design and Manufacturing*, 4, 195–213, 1995.

Balling, R. and Sobieszczanski-Sobieski, J., Optimization of coupled systems: A critical overview, in *AIAA-94-4330, AIAA/NASA/USAF/ISSMO Fifth Symposium on Multidisciplinary Analysis and Optimization*, Panama City Beach, FL, September 1994, publication in *AIAA J.*, 1995.

Besharati, B., Luo, L., Azarm, S., and Kannan, P. K., Multi-objective single product robust optimization: An integrated design and marketing approach, *Journal of Mechanical Design*, 128 (4), 884–892, 2006.

Braun, R. D. and Kroo, I. M., Development and application of the collaborative optimization architecture in a multidisciplinary design environment. N. Alexandrov and M. Y. Hussaini (eds.), in *Multidisciplinary Design Optimization: State of the Art. Proceedings of the ICASE/NASA Langley Workshop on Multidisciplinary Design Optimization*, Hampton, VA, March 13–16, 1997, pp. 98–116.

Choi, H., Panchal, J. H., Allen, K. J., Rosen, W. D., and Mistree, F., Towards a standardized engineering framework for distributed collaborative product realization, in *Proceedings of Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Chicago, IL, 2003.

Fernandez, M. G., Panchal, J. H., Allen, J. K., and Mistree, F., An interactions protocol for collaborative decision making—Concise interactions and effective management of shared design spaces, in *ASME Design Engineering Technical Conferences and Computer and Information in Engineering Conference*, Long Beach, CA, Paper No. DETC2005-85381, 2005.

Fernandez, G. M., Rosen, W. D., Allen, K. J., and Mistree, F., A decision support framework for distributed collaborative design and manufacture, in *Ninth AIAA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, Atlanta, GA, September 4–6, 2002.

Hernandez, G. and Seepersad, C. C., Design for maintenance: A game theoretic approach, *Engineering Optimization*, 34 (6), 561–577, 2002.

Huang, C. C., A multi-agent approach to collaborative design of modular products, *Concurrent Engineering: Research and Applications*, 12 (2), 39–47, 2004.

Kumar, D., Hoyle, C., Chen, W., Wang, N., Gomez-levi, G., and Koppelman, F., Incorporating customer preferences and market trends in vehicle packaging design, in *Proceedings of the DETC 2007, ASME 2007 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Las Vegas, NV, September 2007.

Lewis, K. and Mistree, F., Collaborative, sequential and isolated decisions in design, *ASME Journal of Mechanical Design*, 120 (4), 643–652, 1998.

Lewis, K. and Mistree, F., FALP: Foraging directed adaptive linear programming. A hybrid algorithm for discrete/continuous problems, *Engineering Optimization*, 32 (2), 191–218, 1999.

Li, W., Ong, W. D., Fuh, S. K., Wong, J. Y. H., Lu, Y. S., and Nee, A. Y. C., Feature-based design in a distributed and collaborative environment, *Computer-Aided Design*, 36, 775–797, 2004.

Nikolaidis, E., Decision-based approach for reliability design, *ASME Journal of Mechanical Design*, 129 (5), 466–475, 2007.

Shiau, C. S. and Michalek, J. J., A game-theoretic approach to finding market equilibria for automotive design under environmental regulation, in *Proceedings of the ASME International Design Engineering Technical Conferences*, Las Vegas, NV, 2007.

Sieger, D., Badiru, A., and Milatovic, M., A metric for agility measurement in product development, *IIE Transactions*, 32, 637–645, 2000.

Simon, H., A behavioral model of rational choice, *Quarterly Journal of Economics*, 6, 99–118, 1955.

Simpson, T. W., Seepersad, C. C., and Mistree, F., Balancing commonality and performance within the concurrent design of multiple products in a product family, *Concurrent Engineering Research and Applications*, 9 (3), 177–190, 2001.

Sobieszczanski-Sobieski, J., A linear decomposition method for large optimization problems blueprint for development, National Aeronautics and Space Administration, NASA/TM-83248-1982, 1982.

Sobieszczanski-Sobieski, J., Optimization by decomposition: A step from hierarchic to nonhierarchic systems, Technical Report TM 101494, NASA, Hampton, VA, September 1988.

Sobieszczanski-Sobieski, J. and Haftka, R. T., Multidisciplinary aerospace design optimization: Survey of recent developments, *Structural Optimization*, 14 (1), 1–23, August 1997.

Sobieszczanski-Sobieski, J., James, B., and Dovi, A., Structural optimization by multilevel decomposition, *AIAA Journal*, 23, 1775–1782, 1985.

Sobieszczanski-Sobieski, J. and Kodiyalam, S., BLISS/S: A new method for two-level structural optimization, *Structural Multidisciplinary Optimization*, 21, 1–13, 1998.

Thal., A. E., Badiru, A., and Sawhney, R., Distributed project management for new product development, *International Journal of Electronic Business Management*, 5 (2), 93–104, 2007.

Wassenaar, H. and Chen, W., An approach to decision-based design with discrete choice analysis for demand modeling, *ASME Journal of Mechanical Design*, 125 (3), 480–497, 2003.

Wassenaar, H., Chen, W., Cheng, J., and Sudjianto, A., Enhancing discrete choice demand modeling for decision-based design, *ASME Journal of Mechanical Design*, 127 (4), 514–523, 2005.

Wu, T., Soni, S., Hu, M., Li, F., and Badiru, A., The application of Memetic algorithms for forearm crutch design: A case study, *Mathematical Problems in Engineering*, 3 (4), 1–15, 2011.

chapter two

Dynamic fuzzy systems modeling

Introduction: Decision support systems and uncertainties

This chapter discusses scientific uncertainty (Beer, 2006; Benjamin and Cornell, 1970), fuzziness, and application of stochastic–fuzzy models in urban transit, water resources, energy planning, and education (universities' admission process and other higher educational institutes [HEIs] in developing economies). It enunciates the prime place of decision support systems (DSS) models in providing a robust platform for enabled action on developmental issues. Scientists now recognize the importance of studying scientific phenomenon having complex interactions among their components. These components include not only electrical or mechanical parts but also "soft science" (human behavior, etc.) and how information is used in models. Most real-world data for studying models are uncertain. Uncertainty exists when facts, state, or outcome of an event cannot be determined with probability of 1 (in a scale of 0–1). If uncertainty is not accounted for in model synthesis and analysis, deductions from such models become at best uncertain. The "lacuna" in understanding the concept of uncertainty and developmental policy formulation/implementation is not only due to the non-acceptability of its existence in policy foci, but also the radically different expectations and modes of operation that scientists and policymakers use. It is therefore necessary to understand these differences and provide better methods to incorporate uncertainty into policymaking and developmental strategies (Figure 2.1) (Ibidapo-Obe, 1996; Ibidapo-Obe and Asaolu, 2006; Ibidapo-Obe and Ogunwolu, 2004).

Scientists treat uncertainty as a given, a characteristic of all data and information (as processed data). Over the years, however, sophisticated methods to measure and communicate uncertainty, arising from various causes, have been developed. In general, more uncertainty has been uncovered rather than **absolute precision** (Zadeh, 1965, 1973). Scientific inquiry can only set boundaries on the limits of knowledge. It can define the **edges of the envelope of known possibilities**, but often the envelope is very large and the probabilities of the content (the known possibilities) occurring can be a complete mystery. For instance, scientists can describe the range of uncertainty about global warming and toxic chemicals and perhaps about the relative probabilities of different outcomes, but in

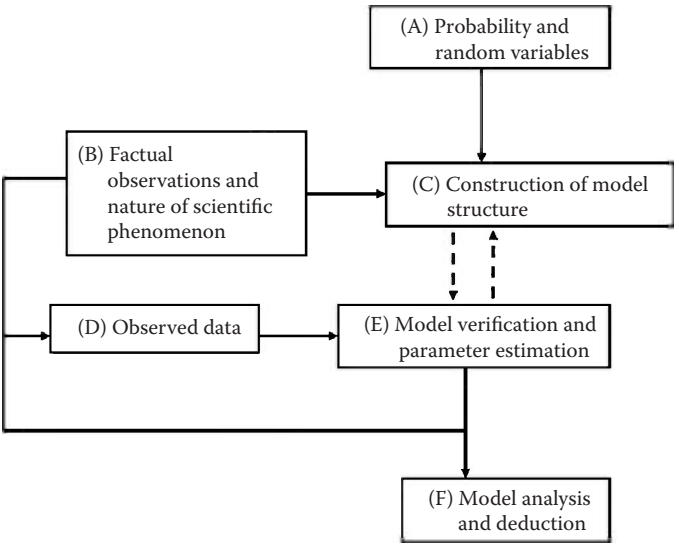


Figure 2.1 Basic cycle of probabilistic modeling and analysis.

most important cases, they cannot say which of the possible outcomes will occur at any particular time with any degree of accuracy. Current approaches to policymaking, however, try to avoid uncertainty and gravitate to the edges of the scientific envelope. The reasons for this bias are clear. Policymakers want to make unambiguous, defensible decisions that are to be codified into laws and regulations (Mamdani and Assilian, 1975).

Although legislative language is often open to interpretation, regulations are much easier to write and enforce if they are stated in absolutely certain terms. **Science defines the envelope while the policy process gravitates to an edge**—usually the edge that best advances the policymaker's political agenda! But to use science rationally to make policy, the whole envelope and all of its contents must be considered.

Decision support systems

A decision is a judgment, choice, or resolution on a subject matter made after due consideration of alternatives or options. It involves setting the basic objectives, optimizing the resources, determining the main line of strategy, planning and coordinating the means to achieve them, managing relationships, and keeping things relevant in the operating environment. A competitive situation exists when conflicting interests must be resolved. A good decision must be **auditable**; that is, it must be the best decision at the time it was taken and must be based on facts or assumptions as well as other technical, socio-political, and cultural considerations (Rommelfanger, 1988).

DSS are computer-based tools for problem formulation through solution, simulation implementation, archiving, and reporting (Ogunwolu, 2005). These tools include operations research software tools, statistical and simulation programs, expert systems, and project management utilities. DSS scheme is as follows:

- Problem statement (modeling—dry testing)
- Data information requirements (collection and evaluation procedure)
- Performance measure (alternatives—perceived worth/utility)
- Decision model (logical framework for guiding project decision)
- Decision making (real-world situation, sensitivity analysis)
- Decision implementation (schedule and control)

Uncertainty

In science, information can be, for example, objective, subjective, dubious, incomplete, fragmentary, imprecise, fluctuating, linguistic, data based, or expert specified. In each particular case this information must be analyzed and classified to be eligible for quantification. The choice of an appropriate uncertainty model primarily depends on the characteristics of the available information. In other words, the underlying reality with the sources of the uncertainty dictates the model (Rayward-Smith, 1995).

A form of uncertainty stems from the *variability* of the data. In equal and/or comparable situations each datum in question may not show identical states. This kind of uncertainty is typically found in data taken from plants and animals and reflects the rich variability of nature. Another kind of uncertainty is the impossibility of observing or measuring to a certain level of precision. This kind of precision depends not only on the power of the sensors applied but also on the environment, including the observer. This type of uncertainty can also be termed as uncertainty due to *partial ignorance* or *imprecision*. Finally, uncertainty is introduced by using a natural or professional language to describe the observation as a datum. This *vagueness* is a peculiar property of humans and uses the special structure of human thinking. Vagueness becomes more transparent in a case where we deal with grades, shades, or nuances, expressed verbally and represented by marks or some natural numbers. Typical phrases used in such vague descriptions include "in many cases," "frequently," "small," "high," "possibly," "probably," etc. All these kinds of uncertainty (uncertainty due to variability, imprecision, and vagueness) can also occur in combinations.

One approach is to investigate uncertainty by the use of *sensitivity analysis* whereby the given data are subjected to small variations to see how these variations will influence the conclusions drawn from the data. The problem with sensitivity analysis, however, has to do with its "point-oriented" approach as to where and in what dimensions the variations

are to be fixed. Another approach is the use of *interval mathematics*, in which each datum is replaced by a set of "possible" surrounding data on the real line. The problem with interval mathematics is the difficulty of specifying sharp boundaries for the data sets, for example, the ends of the intervals. A third approach for tackling uncertainty is the *stochastic* approach. This involves realization of each datum as a random variable in time, that is, the datum is assumed to be chosen from a hypothetical population according to some fixed probability law. This approach works well with modeling of variability and small imprecision. Finally, uncertainty can be taken into account using notions and tools from *fuzzy set theory*. In this approach each datum is represented by a fuzzy set over a suitable universe. *The main idea of fuzzy set is the allowance of membership, to a grade, for every element of a specified set*. With this notion, uncertainty can be modeled mathematically more adequately and subtly using only the common notion of membership of an element to a set. Fuzzy set models both imprecision and vagueness.

However, in typical real-world systems and decision-making processes, virtually all the three types of uncertainty (variability, imprecision, and vagueness) manifest. Since, stochasticity captures variability and small imprecision well and fuzzy set captures imprecision and vagueness in data description well, then for a comprehensive treatment of uncertainty in data, it is advisable to exploit a combined effect of stochastic and fuzzy types of uncertainty. Simply put, randomness caters for objective data while fuzziness caters for subjective data.

Fuzziness

Fuzzy logic (and reasoning) is a scientific methodology for handling uncertainty and imprecision. Unlike in conventional (crisp) sets, the members of fuzzy sets are permitted varying degrees of membership. An element can belong to different fuzzy sets with varying membership grades in each set. The main advantage of fuzzy sets is that it allows classification and gradation to be expressed in a more natural language; this modeling concept is a useful technique when reasoning in uncertain circumstances or with inexact information which is typical of human situations. Fuzzy models are constructed based on expert knowledge rather than on pure mathematical knowledge; **therefore, they are both quantitative and qualitative, but are considered to be more qualitative than quantitative**. Therefore, a fuzzy expert system is a computer-based decision tool that manipulates imprecise inputs based on the knowledge of an expert in that domain (Zimmermann, 1992).

A fuzzy logic controller (FLC) makes control decisions by its well-known fuzzy IF–THEN rules. In the antecedence of the fuzzy rules (i.e., *the IF part*), the control space is partitioned into small regions with

respect to different input conditions. Membership function (MF) is used to fuzzify each of the input variables. For continuity of the fuzzy space, the regions are usually overlapped by their neighbors. By manipulating all the input values in the fuzzy rule base, an output will be given in the consequent (i.e., *the THEN part*). FLCs can be classified into two major categories: the Mamdani (M) type FLC that uses fuzzy numbers to make decisions and the Takagi–Sugeno (TS) type FLC that generates control actions by linear functions of the input variables.

Fuzzy set specifications

In classical set theory, the membership of elements in relation to a set is assessed in binary terms according to a crisp condition. An element either belongs or does not belong to the set; the boundary of the set is crisp. As a further development of classical set theory, fuzzy set theory permits the gradual assessment of the membership of elements in relation to a set; this is described with the aid of a membership function.

If **X** represents a fundamental set and x are the elements of this fundamental set, to be assessed according to an (lexical or informal) uncertain proposition and assigned to a subset *A* of **X**, the set

$$\tilde{A} = \left\{(x, \mathbb{m}_A(x)) | x \in \mathbf{X}\right\} \tag{2.1}$$

is referred to as the uncertain set or fuzzy set on **X** (Figure 2.2).

$\mu_A(x)$ is the membership function of the fuzzy set and may be continuous or discrete with elements assessed by membership values.

The uncertainty model fuzziness lends itself to describing imprecise, subjective, linguistic, and expert-specified information. It is capable of representing dubious, incomplete, and fragmentary information

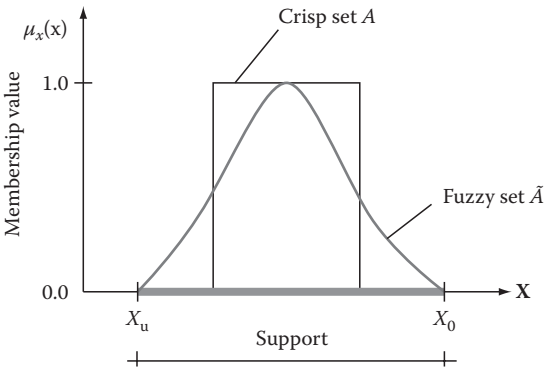


Figure 2.2 Fuzzy set.

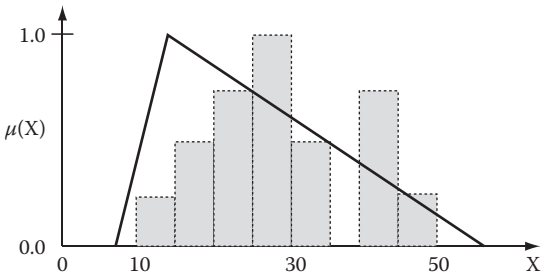


Figure 2.3 Fuzzification of information from a very small sample.

and can additionally incorporate objective, fluctuating, and data-based information in the fuzziness description. Requirements regarding special properties of the information generally do not exist. With respect to the regularity of information within the uncertainty, the uncertainty model fuzziness is less rigorous in comparison with probabilistic models. It specifies lower information content and thus possesses the advantage of requiring less information for adequate uncertainty quantification.

Primarily, fuzzification (Figure 2.3) is a subjective assessment, which depends on the available information. In this context, four types of information are distinguished to formulate guidelines for fuzzification. If the information consists of various types, different fuzzification methods may be combined.

Information type I: Sample of very small size

The membership function is specified on the basis of existing data comprising elements of a sample. The assessment criterion for the elements $\underline{x}$ is directly related to numerical values derived from $\underline{\mathbf{X}}$. An initial draft for a membership function may be generated with the aid of simple interpolation algorithms applied to the objective information, for example, represented by a histogram. This is subsequently adapted, corrected, or modified by means of subjective aspects.

Information type II: Linguistic assessment

The assessment criterion for the elements $\underline{x}$ of $\underline{\mathbf{X}}$ may be expressed using linguistic variables and associated terms, such as "low" or "high" as shown in Figure 2.4. As numerical values are required for a fuzzy analysis, it is necessary to transform the linguistic variables to a numerical scale. By combining the terms of a linguistic variable with modifiers, such as "very" or "reasonably," a wide spectrum is available for the purpose of assessment.

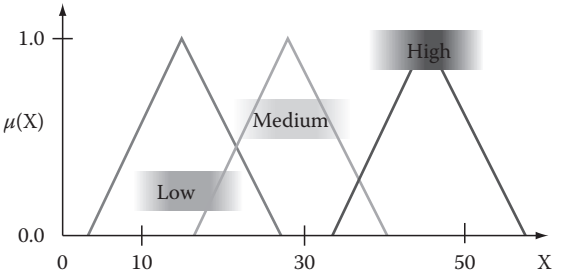


Figure 2.4 Fuzzification of information from a linguistic assessment.

Information type III: Single uncertain measured value

If only a single numerical value from $\underline{\mathbf{X}}$ is available as an uncertain result of measurement $\tilde{\underline{X}}_m$, the assessment criterion for the elements $\underline{x}$ may be derived from the uncertainty of the measurement, which is quantified on the assigned numerical scale. The uncertainty of the measurement is obtained as a "grey zone" comprising more or less trustworthy values. This can be induced, for example, by the imprecision of a measurement device or by a not clearly specifiable measuring point.

The experimenter evaluates the uncertain observation for different membership levels. For the level $\mu_A(\underline{x}) = 1$, a single measurement or a measurement interval is specified in such a way that the observation may be considered to be "as crisp as possible." For the level of the support, $\mu_A(\underline{x}) = 0$, a measurement interval is determined that contains all possible measurements within the scope of the observation. An assessment of the uncertain measurements for intermediate levels is left to the experimenter. The membership function is generated by interpolation or by connecting the determined points $(\underline{x}, \mu_A(\underline{x}))$. Figure 2.5 shows an example.

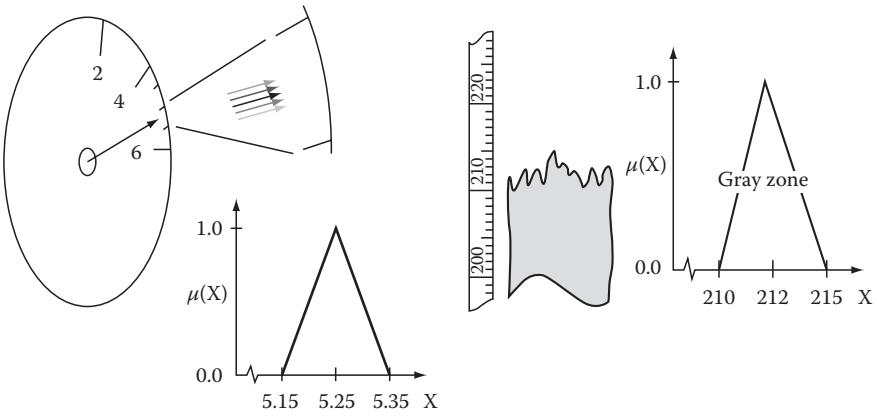


Figure 2.5 Fuzzification of a single uncertain measurement due to imprecision of the measuring device or imprecise measuring point.

Information type IV: Knowledge based on experience

The specification of a membership function generally requires the consideration of opinions of experts or expert groups, of experience gained from comparable problems, and of additional information where necessary. Also, possible errors in measurement, and other inaccuracies attached to the fuzzification process may be accounted for. These subjective aspects generally supplement the initial draft of a membership function. If neither reliable data nor linguistic assessments are available, fuzzification depends entirely on estimates by experts.

As an example, consider a single measurement carried out under dubious conditions, which only yields some plausible value range. In those cases, a crisp set may initially be specified as a kernel set of the fuzzy set. The boundary regions of this crisp kernel set are finally "smeared" by assigned membership values $\mu_A(\underline{x}) < 1$ to elements close to the boundary and leading the branches of $\mu_A(\underline{x})$ beyond the boundaries of the crisp kernel set monotonically to $\mu_A(\underline{x}) = 0$. By this means, elements that do not belong to the crisp kernel set, but are located "in the neighborhood" of the latter, are also assessed with membership values of $\mu_A(\underline{x}) > 0$. This approach may be extended by selecting several crisp kernel sets for different membership levels (α-level sets) and by specifying the $\mu_A(\underline{x})$ in level increments.

Stochastic–fuzzy models

Fuzzy randomness simultaneously describes objective and subjective information as a fuzzy set of possible probabilistic models over some range of imprecision. This generalized uncertainty model contains fuzziness and randomness as special cases (Möller and Beer, 2004).

Objective uncertainty in the form of observed/measured data is modeled as **randomness**, whereas subjective uncertainty (see Figure 2.6), for example, due to a lack of trustworthiness or imprecision of measurement results, of distribution parameters, of environmental conditions, or of the data sources, is described as **fuzziness**. The fuzzy-random model then combines but does not mix objectivity and subjectivity; these are separately visible at any time. It may be understood as an imprecise probabilistic model, which allows for simultaneously considering all possible probability models that are relevant to describing the problem.

The uncertainty model fuzzy randomness is particularly suitable for adequately quantifying uncertainty that comprises only some (incomplete, fragmentary) objective, data-based, randomly fluctuating information, which can simultaneously be dubious or imprecise and may additionally be amended by subjective, linguistic, expert-specified evaluations.

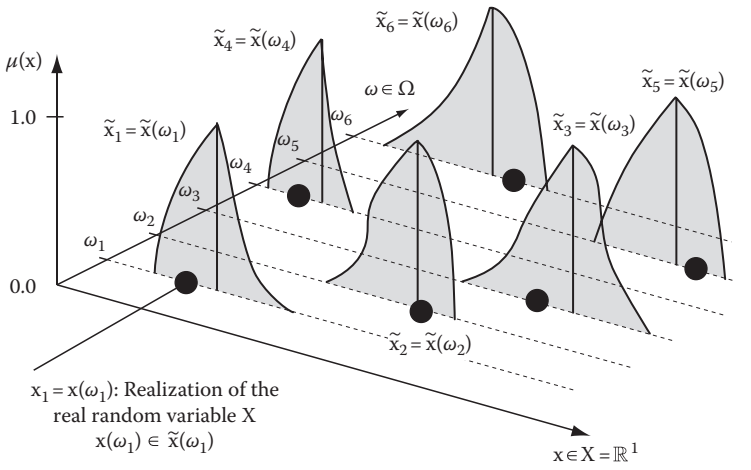


Figure 2.6 Model of a fuzzy–random variable.

This generalized uncertainty model is capable of describing the whole range of uncertain information reaching from the special case of **fuzziness** to the special case of **randomness**. That is, it represents a viable model if the available information is too rich in content to be quantified as fuzziness without a loss in information but, on the other hand, cannot be quantified as randomness due to imprecision, subjectivity, and non-satisfied requirements. **This is probably the most common case in applied science**.

Uncertainty quantification with fuzzy randomness represents an imprecise probabilistic modeling, which incorporates imprecise data as well as uncertain or imprecise subjective assessments in terms of probability. The quantification procedure is a *combination of established methods from mathematical statistics for specifying the random part and of fuzzification methods for describing the fuzzy part* of the uncertainty.

Applications

Development model

Development is the vital summation of all efforts made by man to increase the quality of life, while sustainability is the continued successful upholding and enhancement of this quality of life by getting replenished the necessary ingredient/resources such as human labor and ecological resources. Development occurs when the intrinsic aspect is applied through technology to generate the physical aspect. Technology confirms the existence of the intrinsic aspects and creates the physical aspect to manifest development (Olunloyo, 2005).

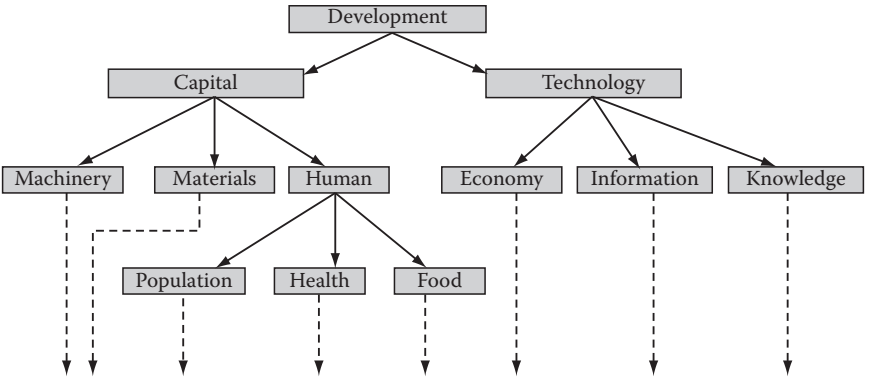


Figure 2.7 Hierarchical representation of macro- and micro-development variables.

Considering development as a hierarchical (Figure 2.7) multivariate nonlinear function

$$D = (C, T, I, K, F, H) \tag{2.2}$$

where

C is capital
T is technology
I is information technology
K is knowledge
F is food
H is health
clearly, C, T, I, K, F, and H are not independent variables

A compact analysis of the behavior of D can be facilitated by the following **dimensionality reduction**.

Capital can be defined as an inventory of infrastructures machinery, m, that is, plant, equipment, etc., materials, m_2, and human resources, h.

$$C = C(m_1, m_2, h) \tag{2.3}$$

Since human resources, h, itself can be viewed as function of population, health, and food, the functional representation for h alone becomes

$$h = h(P, H, F) \tag{2.4}$$

Thus, capital C can be represented as a compound functional representation as

$$C = C(m_1, m_2, h(P, H, F)) \tag{2.5}$$

Following similar argument, since the technology variable T depends on the average performance measure of all productive processes in the economy: E, the level of development and utilization of information technology: I, and the average experience and intelligence in the society, that is, knowledge, K, T becomes

$$T = T(E, I, K) \tag{2.6}$$

Thus, the development nonlinear equation can be written as

$$D = D(C, T) = D(C(m_1, m_2, h(P, H, F)), T(E, I, K)) \tag{2.7}$$

This is obviously functionalizing development at a macro-level. At the micro-level are myriads of variables on which each of the aforementioned variables depends. The decision-making organelle in optimizing for development definitely consists of optimization at the macro- and the micro-level functions.

Uncertainty of different kinds influences decision making and hence development. In optimizing functions of development, the effects of uncertainty cannot be ignored or rationalized else the results of such decision will at best be out of relevance to effect economic growth. At the micro-level, where the numerous variables that influence decision making are influenced by the macro level, cognizance must be taken of uncertainty in quantifying development variables. Inherent in the quantification are elements of uncertainty in terms of variability, imprecision, vagueness of description, randomness. As amply explained in earlier sections, such quantities are better realized as combinations of fuzzy and random variables (Takagi and Sugeno, 1985).

For example, food security (see Table 2.1) is influenced by a myriad of variables including level of mechanization of agriculture, population, incentives, soil conditions, climatic conditions, etc. These variables may not be measured precisely. Some may be fuzzy, some stochastic, and others manifesting a combination of stochasticity and fuzziness in their quantification. In effect, quantification of **food** (F) in the development model is a **fuzzy–stochastic** variable.

The other variables in the aforementioned development can also be viewed as fuzzy–stochastic variables; thus, the **developmental model** can be viewed as a **fuzzy–stochastic developmental model**.

$$D = D\left(\tilde{\tilde{C}}, \tilde{\tilde{T}}\right) = D\left(C\left(\tilde{\tilde{m}}_1, \tilde{\tilde{m}}_2, h\left(\tilde{\tilde{P}}, \tilde{\tilde{H}}, \tilde{\tilde{F}}\right)\right) T\left(\tilde{\tilde{E}}, \tilde{\tilde{I}}, \tilde{\tilde{K}}\right)\right) \tag{2.8}$$

where the symbols _ and ~ represent fuzzification and randomization, respectively, of the various variables.

Table 2.1 A Typical Uncertain Description of Variables Influencing Food Production

Food-related variable	Description	Class of uncertainty
Mechanization	Linguistic description, e.g., "high" level of mechanization	Fuzzy
Population	A range of numbers, random over a time space	Fuzzy–stochastic
Incentives	A range of numbers	Fuzzy
Soil conditions	A random variable subject to variation over time	Stochastic
Climatic conditions	Quantitative, and variable over time space	Fuzzy–stochastic
Yield	Quantitative description	Fuzzy–stochastic

The optimization of the fuzzy–stochastic development model

The realization of the relational effects of the variables of development is in hierarchies, even at the macro-level. A good strategy for optimizing the function is to use the concept of multilevel optimization model.

Multilevel-optimizing models are employed to solve decentralized planning decision problems in which decisions are made at different hierarchical decision levels in a top-to-down fashion. Essentially, the features of such multilevel planning organizations are as follows:

- Interactive decision-making units exist within a predominantly hierarchical structure.
- Execution of decisions is sequential, from top to lower level.
- Each unit independently maximizes its own net benefit but is affected by the actions of other units through externalities.
- The external effect of a decision maker's (DM's) problem can be reflected in both the objective function and the feasible decision space.

The mode of execution of such decision problem is as follows:

- The upper-level DM sets his goal and accepts the independent decisions at the lower levels of the organization.
- The upper-level DM modifies it within the framework of the overall benefit of the organization.
- The upper-level DM's action further constrains the lower-level decision space and may or may not be acceptable at that level. If it is not acceptable, the upper-level DM can still make a consensus that the constraints are relaxed further.
- This process is carried out until a satisfactory solution to all levels and units of decision making is arrived at.

In solving the **fuzzy–stochastic development model**, each level of the macro-level model is taken as a decision level optimizing variables of its concern.

The complete **fuzzy–stochastic multilevel** formulation of the **fuzzy–stochastic development model** is, therefore,

$$\underset{\bar{C},\bar{T}}{Max}\ \tilde{\tilde{D}} \tag{2.9}$$

where, $\bar{C},\bar{T}$ is obtainable from

$$\underset{V_3}{Max}\ f(V_2) \tag{2.10}$$

with $V_2 = \left\{\tilde{\tilde{m}}_1, \tilde{\tilde{m}}_2, \tilde{\tilde{h}}, \tilde{\tilde{E}}, \tilde{\tilde{I}}, \tilde{\tilde{K}}\right\}$, which again is obtainable from

$$\underset{V_3}{Max}\ f(V_2) \tag{2.11}$$

where $V_2 = \left\{\tilde{\tilde{X}}, \tilde{\tilde{P}}, \tilde{\tilde{H}}, \tilde{\tilde{F}}\right\}$ and $\tilde{\tilde{X}}$ = relevant variables at the micro-level of the development model.

Subject to

$$\tilde{\tilde{C}}, \tilde{\tilde{T}}, v_2, v_3, \tilde{\tilde{X}} \in S$$

where S is the constrained space of development variables (constrained by the limitations in realizing the various variables).

Realistically, this type of model may be difficult to solve for large spaces such as the national development pursuit but may be solved with smaller decision spaces. The worst solution scenario will be those that are not amenable to analytical solutions for which there are many heuristics to be coupled with simulation-optimization techniques.

Urban transit systems under uncertainty

A relevant work (Ibidapo-Obe and Ogunwolu, 2004) on DSS under uncertainty is the investigation and characterization of combinations of effects of fuzzy and stochastic forms of uncertainty in urban transit time-scheduling. The results of the study vindicate the necessity for taking both comprehensive combinations of fuzzy and stochastic uncertainties

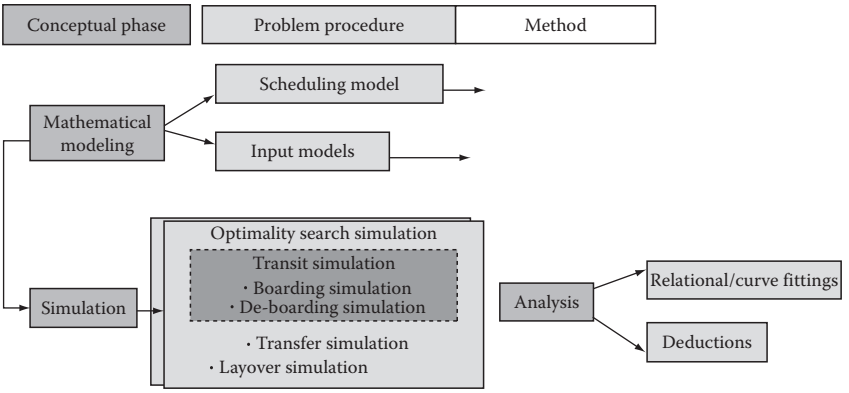


Figure 2.8 Conceptual, procedural, and methodic frame of the transit time-scheduling research study.

as well as multi-stakeholders' objective interests into account in urban transit time-scheduling. It shows that transit time-scheduling performance objectives are better enhanced under fuzzy–stochastic and stochastic–fuzzy uncertainties and with the multi-stakeholders' objective formulations than with lopsided single stakeholder's interests under uncertainty. Figure 2.8 shows the modeling framework, which covers the following:

- Mathematical modeling
- Algebraic, fuzzy, and stochastic methods
- Multi-objective genetic algorithm
- Algorithmic computer simulation technique
- Mixed integer models
- **M**ax–min techniques

Water resources management under uncertainty

An ongoing research on water resources management also suggests the need for incorporating comprehensive scientific uncertainty into account in developmental models (Figure 2.9). This involves using a three-level hierarchical system model for the purpose of optimal timing and sequencing of project development for water supply, the optimal allocation of land, water, and funds for crop growth, and optimal timing for irrigation. Major functions and inputs in the hierarchical fuzzy–stochastic dynamic programming developmental model are to be realized as fuzzy, stochastic, and/or fuzzy–stochastic inputs of the model. Particularly, the demand and supply of water over a time horizon are taken as fuzzy–stochastic.

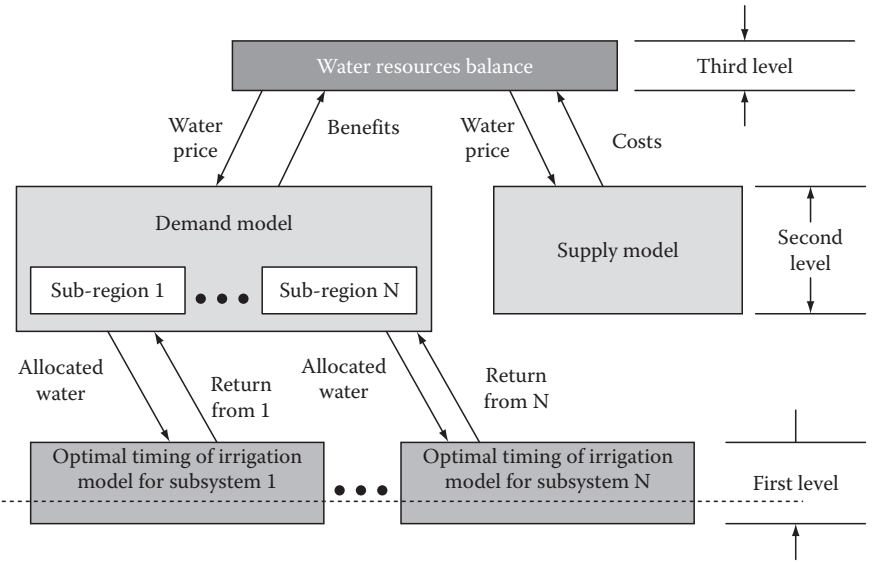


Figure 2.9 Multilevel dynamic programming structure for planning and management of agricultural/water system.

Energy planning and management under uncertainty

One of the immediate future thrusts is on ***energy planning and management modeling under uncertainty***. Energy information systems are organized bodies of energy data often indicating energy sources and uses with information on historical as well as current supply–demand patterns, which are naturally bugged with uncertainties (see Figure 2.10). These information systems draw upon energy surveys of various kinds as well as upon other sources of information such as the national census (another potential source of uncertainty in evaluation over time and space); information on energy resources and conversion technologies as well as consumption patterns (which are also better realized considering inherent uncertainties). A typical structure of national energy planning system is shown in Figure 2.10.

University admissions process in Nigeria: The post-UME test selection saga

There are many issues related to the Nigerian education system (universal basic education UBE, the 3–3 component of the 6–3–3–4 system, funding, higher education institution admissions, consistency of policies, etc.).

In spite of the opening up of the HEIs space to more states, private, open, and transnational institutions, the ratio of available space to the

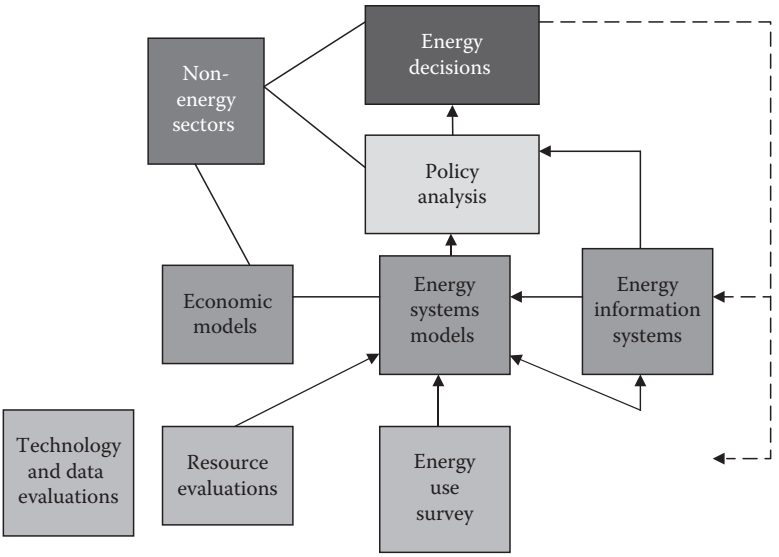


Figure 2.10 A structure for national energy planning.

number of eligible candidates is 87:1000 (University of Lagos, 2006). It is, therefore, imperative that the selection mechanism be open and rational. The realization problem for rational access is further compounded with the introduction of post-UME test. With two assessment scores per candidate given the department/faculty/university of choice quota, the question is how best can we optimize a candidate's opportunity to be admitted? That is,

- How do we establish that a particular candidate actually sat for the JAMB examination?
- Is the JAMB examination score enough to measure competence?
- Is it sufficient for the various universities to select based on their own examinations?
- To what extent will the university post-UME test help judge candidates' suitability?
- Could there be a rational middle-of-the-road approach in determining a candidate's admission to the university based on the two regimes of scores he or she holds?
- To what confidence level can a university assert that it has selected rationally?
- Answers to these and many more questions and issues on this subject matter are subjects of wide variability, imprecision, and vagueness. The objective is to formulate a fuzzy–stochastic decision support model for the resolution of this matter.

- The just-concluded University of Lagos admission process was reflective of the underlying uncertainty in rationally giving candidates to its programs based on the two regimes of scores:
  - *Stochastic aspect*: A candidate for admission must reach a particular scholastic level at the JAMB UME examination. UME is a highly stochastic scheme with randomness in performance over time and space. The various performance levels of all candidates for a particular course/faculty/university may be assumed as normally distributed.
  - *Fuzzy aspect*: Candidates for admission need another scholastic assessment based on the score of the post-UME screening exercise conducted by the university. This is fuzzy based on a reference course faculty/university mean score(s). The post-UME screening score may be assumed as having membership grade distribution which is equivalent to the fuzzy set (represented by the course/faculty/university mean score (x_i) to 100/400).

In other words, candidates must satisfy post-UME screening (fuzzy) criterion before being considered for admission based on the UME scores (stochastic).

Mathematically expressed,

$$Max\,A = A(Pu_s, U_s/q_0) \tag{2.12}$$

subject to

$$q_0 = q_0(\text{merit}, \text{catchment}, \text{ELDS})$$

where

A is the probability of admission at the university of first choice and first course
U_s is the UME score
q_0 is the course admission quota
Pu_s is the post-UME screening score
"merit" is the absolute performance (independent of state of origin of candidate)
"catchment" is the contingent states of location of institution
"ELDS" is the educationally less disadvantaged states in the federation

Conclusions

This chapter outlined scientific uncertainty in relation to optimal DSS management for development. The two principal inherent forms of uncertainty (in data acquisition, realization, and processing), fuzziness and

stochasticity, as well as their combinations are introduced. Possible effects on developmental issues and the necessity for DSS which incorporate combined forms of this uncertainty are presented.

- In particular, the major focus of this submission is the proposal to deal with scientific uncertainty inherent in developmental concerns.
- Uncertainty should be accepted as a basic component of developmental decision making at all levels and thus the quest is to correctly quantify uncertainty at both macro- and micro-levels of development.
- Secondly, expert and scientific DSS which enhance correct evaluation, analysis, and synthesis of uncertainty inherent in data management should be utilized at each level of developmental planning and execution.

References

Beer, M., Uncertain structural design based on non-linear fuzzy analysis, *Special Issue of ZAMM* 84 (10–11), 740–753, 2004.

Benjamin, J. R. and Cornell, C. A., *Probability, Statistics and Decision for Civil Engineers*. McGraw-Hill, New York, 1970.

Ibidapo-Obe, O., *Understanding Change Dynamics in a Stochastic Environment*. University of Lagos Press, Lagos, Nigeria, Inaugural Lecture Series, 1996.

Ibidapo-Obe, O., Modeling, identification/estimation in stochastic systems. In A. B. Badiru (ed.), *Handbook of Industrial and Systems Engineering*, CRC Press, Taylor & Francis Group, New York, Chapter 14, 2006, pp. 1–16.

Ibidapo-Obe, O. and Asaolu, S., Optimization problems in applied sciences: From classical through stochastic to intelligent metaheuristic approaches. In A. B. Badiru (ed.), *Handbook of Industrial and Systems Engineering*, CRC Press, Taylor & Francis Group, New York, Chapter 22, 2006, pp. 1–18.

Ibidapo-Obe, O. and Ogunwolu, L., Simulation and analysis of urban transit system under uncertainty—A fuzzy-stochastic approach, in *Proceedings of the Practice and Theory of Automated Timetabling PATAT*, Pittsburg, PA, 2004.

Mamdani, E. H. and Assilian, S., An experiment in linguistic synthesis with a fuzzy logic controller, *International Journal of Man-Machine Studies*, 7, 1–13, 1975.

Möller, B. and Beer, M., *Fuzzy Randomness—Uncertainty in Civil Engineering and Computational Mechanics*. Springer, Berlin/Heidelberg, Germany, 2004.

Ogunwolu, F. O., Time-scheduling in urban transit systems: A multi-objective approach under fuzzy and stochastic conditions, PhD thesis, University of Lagos, Lagos, Nigeria, 2005.

Olunloyo, V. O. S., Project Nigeria and Technology. Nigerian National Merit Award Winner's Lecture. (2005).

Rayward-Smith, V. J., *Applications of Modern Heuristic Methods*. Alfred Walter Limited Publishers in association with UNICOM, London, England, pp. 145–156, 1995.

Rommelfanger, H., *Fuzzy Decision Support-System*. Springer, Berlin, Germany, 1988.

Takagi, T. and Sugeno, M., Fuzzy identification of systems and its applications to modeling and control, *IEEE Transactions on Systems, Man, and Cybernetics*, 15, 116–132, 1985.

University of Lagos/Joint Admissions and Matriculations Board-Undergraduate Admissions, Lagos, Nigeria, 2006.

Zadeh, L. A., Fuzzy sets, *Information and Control*, 8, 338–353, 1965.

Zadeh, L. A., Outline of a new approach to the analysis of complex systems and decision processes, *IEEE Transactions on Systems, Man, and Cybernetics*, 3, 28–44, 1973.

Zimmermann, H., *Fuzzy Set Theory and Its Applications*. Kluwer Academic Publishers, Boston, MA, 1992.

chapter three

Stochastic systems modeling

Introduction to model types

Proper modeling of a system is the first step to the formulation of an optimization strategy for the system. There are different types of models.

First, a **system** is a collection of **interrelated elements** brought together **to achieve a specified objective**. Normally, a system processes an input to yield an output or response.

Next, models are needed to represent the system.

A model is an abstraction of a real-world system. It is a simplified representation of reality which employs descriptive (linguistic, physical, or mathematical) concepts and/or symbols; that is, a model mirrors or approximates only some aspects of reality and evaluates others. Models provide a safe and cost-effective means for the study of entities and phenomena and their interactions. Two major types of models are *material/iconic* (including robotic/expert) models as well as *mathematical* models. Proper modeling of a system is the first step toward formulating an optimization strategy for the system. The criterion for objective/cost selection of a model would be to minimize the errors between the model and the actual system. The "goodness-of-fit" criterion can be evaluated when the model and the system are forced by sample inputs.

Material/iconic models

The material/iconic models simulate the actual system as a prototype in the physical space. It could be a scaled model of an empirical system or a direct physical analogue. The study of its behavior under various conditions possible is undertaken, for example, wind-tunnel laboratories, test piloting, moot trials, etc.

Robotic/expert models

This is a feedback control system—a device that can measure its own state and take actions based on it.

Mathematical models

Mathematical modeling involves the application of mathematics/empirical knowledge to real-life problems. Stimulators to this approach include the advent of high-speed electronic computers (for large environmental problems—parallel computers), developments in information and communication technology, progress in applied mathematics (functional analysis and numerical methods), and progress in empirical knowledge (engineering). A mathematical model consists of a set of mathematical formulae giving the validity of certain fundamental "natural laws" and various hypotheses relating to physical processes.

The direct engineering problem is to find the output of a system, given the input (see Figure 3.1), and the inverse problems are of three main divisions—the design/synthesis, control/instrumentation, and modeling/identification (Figure 3.2) (Liebelt, 1967; Sage and Melsa, 1971).

1. *Design/synthesis*: Given an input and output, find a system description which fits such a physically realizable relationship optimally.
2. *Control/instrumentation*: Given a system description and a response, find the input which is responsible for the response (output).
3. *Modeling/identification*: Given a set of inputs and corresponding outputs from a system, find a mathematical description (model) of the system.

The criterion for objective/cost function selection would be to minimize the errors between the model and the actual system. The "goodness of fit" criterion can be evaluated when both the model and the system are forced by sample inputs (see Figure 3.3).

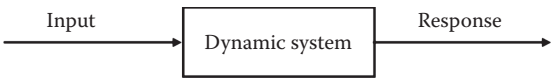


Figure 3.1 Input–response relationship in system modeling.

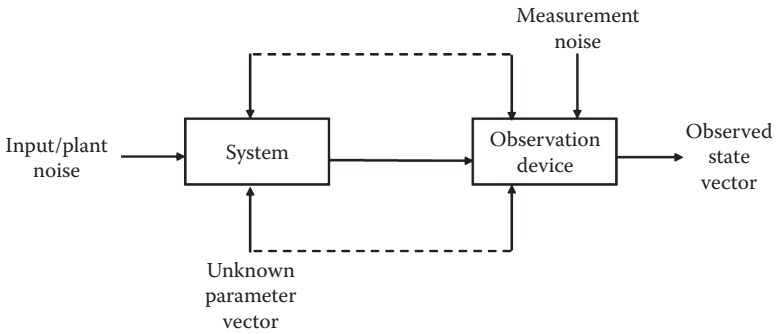


Figure 3.2 General system configuration.

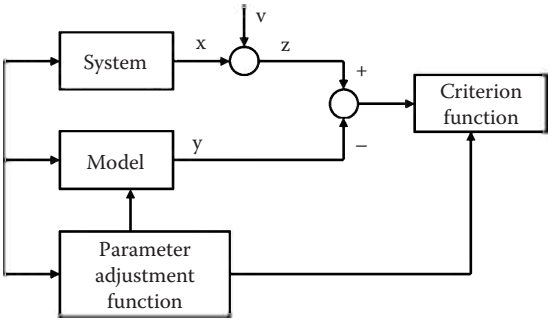


Figure 3.3 Parameter models.

General problem formulation

Let

$$\frac{dx(t)}{dt} = f(x(t), u(t), w(t), p(t), t) \tag{3.1}$$

be the system equation where

$x(t)$ is the system state vector
$u(t)$ is the input signal/control
$w(t)$ is the input disturbance/noise
$p(t)$ is the unknown parameters

Assume that the observation is of the form

$$z(t) = h(x(t), u(t), w(t), p(t), v(t)t) \tag{3.2}$$

where v(t) is the observation noise.

The identification/estimation problem is to determine $p(t)$ (and perhaps $x(t)$ as well as the mean and variance coefficients of system noise $w(t)$ and observation noise $v(t)$).

System:

$$\frac{dx(t)}{dt} = f(x(t), u(t), w(t), p(t), t) \tag{3.3}$$

Observation:

$$z(t) = Dy + Eu, \quad \text{D, E are matrices} \tag{3.4}$$

Model:

$$\frac{dy(t)}{dt} = g(y(t), u(t), w(t), p', t) \tag{3.5}$$

Criterion function:

$$J(T, p') = \int_0^T \left\| x(t) - y(t) \right\|_w dt \tag{3.6}$$

W is an appropriate weighing matrix (Bekey, 1970).

Problem

Seek an optimum set of parameters p* which minimizes J. That is,

$$\mathrm{J(T, p^*) = \min J(T, p')} \tag{3.7}$$

Analytical expressions for p* is possible, namely,

$$\frac{dJ}{dp'} = 0 \quad \text{provided } \frac{d^2J}{dp'^2} > 0 \text{ in special cases.} \tag{3.8}$$

Search techniques are useful when the number of parameters is small. The technique consists of

1. Random selection of preselected grid pattern for parameters p'_1, p'_2, … and corresponding J_1, J_2, …
2. Simple comparison test for the determination of minimum J

Gradient methods are based on finding the values of p′ for which the gradient vector equals zero, namely,

$$\nabla_0 J = \left[\frac{\partial J}{\partial p_1}, \frac{\partial J}{\partial p_2}, \ldots \frac{\partial J}{\partial p_k} \right] = 0 \tag{3.9}$$

and

$$p^{(i+1)} = p^{(i)} - K \nabla_0 J(p^{(i)}) \tag{3.10}$$

where for

1. Steepest descent, $\mathrm{K = kJ}$, k = constant (3.11)

2. Newton–Raphson, $K = \dfrac{J(p)}{\left\| \nabla J(p) \right\|^2}$ (3.12)

3. Newton, $\mathrm{K} = H^{-1} = \left[\dfrac{\partial^2 J}{\partial p_j \partial p_k} \right]^{-1}$ (3.13)

4. Gauss–Newton, $\mathrm{K} = \mathrm{d}^{-1} = \left[\int_0^T 2 \nabla y \nabla y' dt \right]^{-1}$ (3.14)

It is desirable to have online or recursive identification so as to make optimum adaptation to the system goal possible in the face of environmental uncertainty and changing environmental conditions.

Systems filtering and estimation

Identification

Identification problems can be categorized into two broad areas, namely, the total ignorance/"black box" identification and the grey box identification. In the grey box identification, the system equations may be known or deductible from the basic physics or chemistry of the process up to the coefficients or parameters of the equation. The methods of solution consist of classical (deconvolution, correlation, etc.) and modern techniques (Gelb, 1974; Albert and Gardner, 1967) (Figure 3.4).

Given u(t) and y(t) for $0 \le t \le T$, determine h(t).

Observe input and output at N periodical sampled time intervals, say Δ seconds apart in [0,T] such that NΔ = T.
It is known that

$$y(t) = \int_0^t h(t-\tau)u(\tau)d\tau \tag{3.15}$$

called the convolution integral.
Assume that

$$\mathrm{u(t)} = \mathrm{u}(E\Delta) \quad \text{or} \quad u(t) \approx \frac{1}{2}\{u(n\Delta) + \mathrm{u(n+1)}\Delta\} \tag{3.16}$$

for nΔ < t < (n + 1)Δ

$$h(t) \approx h\left(\frac{2n+1}{2}\Delta\right), \quad \mathrm{n}\Delta <= \mathrm{t} < (\mathrm{n}+1)\Delta, \tag{3.17}$$

If

$$\mathrm{y}(n\Delta) = \Delta\sum_{i=0}^{n-1} h\left(\frac{2n-1}{2}\Delta - i\Delta\right)u(i\Delta) \tag{3.18}$$

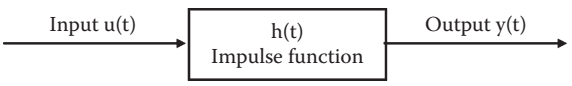


Figure 3.4 Single input–output function.

If

$$y(T) = \begin{bmatrix} y(\Delta) \\ y(2\Delta) \\ . \\ y(N\Delta) \end{bmatrix}, \quad h(T) = \begin{bmatrix} h\left(\frac{\Delta}{2}\right) \\ h\left(\frac{3\Delta}{2}\right) \\ . \\ h\left(\frac{(2N-1)\Delta}{2}\right) \end{bmatrix} \tag{3.19}$$

then

$$y(T) = \Delta U h(T) \tag{3.20}$$

where

$$U = \begin{bmatrix} u(0) & 0 & 0 & 0 & 0 \\ u(\Delta) & u(0) & 0 & 0 & 0 \\ u(2\Delta) & u(\Delta) & u(0) & . & 0 \\ . & . & . & . & . \\ u((N-1)\Delta) & u((n-2)\Delta) & . & u(0) & (0) \end{bmatrix} \tag{3.21}$$

From Equation 3.20,
$h(T) = U^{-1}\Delta^{-1}y(T)$, so that

$$h_n \cong h\left(\frac{2n-1}{2}\Delta\right), \quad h_1 = \frac{y(\Delta)}{\Delta u(0)}$$
$$= \frac{1}{u(0)}\left\{\frac{y(n\Delta)}{\Delta} - \sum_{i=1}^{n-1} h_{n-i}u(i\Delta)\right\}$$

Advantages (Stewart, 1973):

1. Simple.
2. Quite effective for many identification problems.
3. FFT (fast Fourier transform) may be used to reduce the computational requirements.
4. Any input may be used (no need for special test inputs).

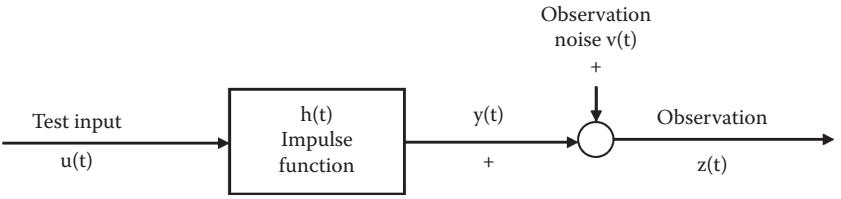


Figure 3.5 Configuration.

Disadvantages:

1. Sequential/online use of algorithm is impossible unless the time interval of interest is short.
2. Numerical round-off errors make the technique inaccurate for $m \to \infty$ (Figure 3.5).

Correlation techniques

Correlation techniques use white noise test signal, $u(t)$; hence, it is necessary to have wide bandwidth to detect high-frequency components of h(t). For zero error, u(t) must be proper "white" (infinite bandwidth) (Figure 3.6).

1. It is assumed that steady state is reached.
2. Noise u(t) and v(t) ergodic and Gaussian distribution with zero mean

$$x_0(t) = \frac{1}{t}\int_0^t x(1)d1 \tag{3.22}$$

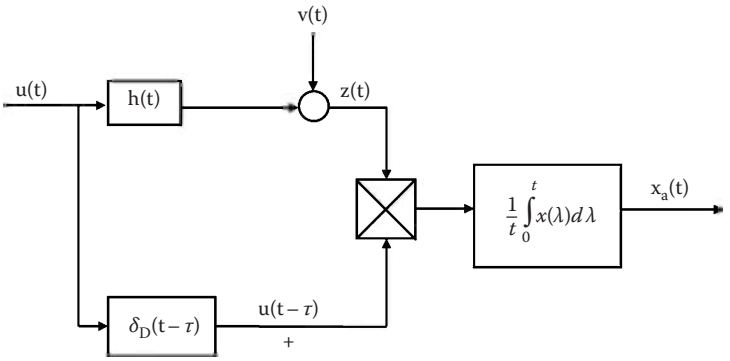


Figure 3.6 Identification correlator.

$$x(t) = z(t)u(t-\tau) \tag{3.23}$$

$$z(t) = y(t) + v(t) \tag{3.24}$$

$$y(t) = \int_0^t h(\eta)u(t-\eta)d\eta \tag{3.25}$$

Now,

$$\begin{aligned} E\{x_a(t)\} &= \frac{1}{t}\int_0^t E\{x(\lambda)\}d\lambda \\ &= E\{x\} \\ &= E\{z(t)u(t-\tau)\} \\ &= R_{uz}(\tau) \\ R_u(\tau) &= E\{u(t)z(t+\tau)\} \end{aligned} \tag{3.26}$$

From Equations 3.24 and 3.25, $E\{x_a(t)\} = R_{uz}(\tau) = \int_0^\infty h(\eta)R_u(t-\eta)d\eta$

Taking Fourier transforms:

$$R_{uz}(s) = h(s)R_u(s)$$

If the assumption on bandwidth holds,

$$R_{uz}(s) = kh(s), \quad R_{uz}(\tau) = kh(\tau)$$

And if u(t) is white, δ_D is a Dirac-delta:

$$R_u(\tau) = R_u\delta(I), \quad R_u(s) = R_w$$

And then, for $R_u = 1$,

$$R_{uz}(\tau) = E\{x_a(t)\} = h(\tau)$$

Complete system identification is subsequently obtained by using N correlators in parallel, such that the quantities

$$R_{uz}(\tau_i) = h(\tau_i), \quad i = 1, 2, ...N \text{ are measured.}$$

Advantages

1. Not critically dependent on normal operating record.
2. By correlating over a sufficiently long period of time, the amplitude of the test signal can be set very low such that the plant is essentially undisturbed by the white noise test signal.
3. No a priori knowledge of the system to be tested is required.

System estimation

A new formulation of the Wiener (classical) theory expressing the results of the estimation in the time domain rather than in the frequency domain was initiated in 1960 (Sage and Melsa, 1971). The modern theory is more fundamental, requires minimum mathematical background, is ideal for digital computation, and provides a general estimator, whereas the classical method can only deal with restricted dimension, is rigorous, and has limited applicability to nonlinear systems.

Problem formulation

Given

$$\Theta = \Theta(x) \tag{3.27}$$

where

Θ is a vector of m observations
x is a vector whose variables are to be estimated

The estimation problem is continuous if Θ is a continuous function of time, otherwise it is a discrete estimation problem. *Estimating the past is known as smoothing, estimating the present as filtering while estimating the future is prediction/forecasting.*

Nomenclature

1. An estimate $\hat{x}$ of x is unbiased if $E(\hat{x}) = x$.
2. Let $e = \hat{x} - x$ and $C_e = E[(\hat{x} - x)(\hat{x} - x)']$.

Maximum likelihood

Let

$$\Theta = Bx + v \tag{3.28}$$

where v is the noise. The maximum likelihood method takes $\hat{x}$ as the value which maximizes the probability of measurements that actually occurred taking into account known statistical properties of v. The conditional probability density function for Θ given x, is the density of

v centered around Bx. If v is zero mean Gaussian distributed with covariant matrix C_v then

$$p(\Theta|x) = \frac{1}{(2\mathrm{p})^{1/2}|C_v|^{1/2}} e^{\left[\frac{-1}{2}(\Theta - Bx)C_v^{-1}(\Theta - Bx)\right]} \tag{3.29}$$

So that

$$Maximum\ p(\Theta|x) = Max\left[\frac{-1}{2}(\Theta - Bx)C_v^{-1}(\Theta - Bx)\right]$$

Hence

$$x = (B'C_v^{-1}B)^{-1}B'C_v^{-1}\Theta \tag{3.30}$$

Least squares/weighted least squares

The least squares choose $\hat{x}$ as that value which minimizes the sum of squares of the deviations $\theta_i - \hat{\theta}_i$, that is, minimizes

$$J = (\Theta - Bx)'(\Theta - Bx) \tag{3.31}$$

thus setting $\partial J/\partial x = 0$ yields

$$\hat{x} = (B'B)^{-1}B'\Theta \tag{3.32}$$

For weighted least squares, minimize

$$J = (\Theta - Bx)'W^{-1}(\Theta - Bx) \quad \text{or} \quad J = \|\Theta - Bx\|_{W^{-1}} \tag{3.33}$$

yielding

$$\hat{x} = (B'W^{-1}B)^{-1}B'W^{-1}\Theta \tag{3.34}$$

Bayes estimators

For Bayes estimators, statistical models for both x and Θ are assumed available. The a posteriori conditional density function is $p(x|\Theta)$, since it contains all the statistical information of interest.

$$p(x|\Theta) = \frac{p(\Theta|x)p(x)}{p(\Theta)} \quad \text{(Bayes' rule)} \tag{3.35}$$

$\hat{x}$ is computed from $p(x|\Theta)$.

Minimum variance

Minimize

$$J = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(\hat{x}-x)'S(\hat{x}-x)\,p(x\,|\,\Theta)dx_1\ldots dx_n \tag{3.36}$$

where S is an arbitrary, positive, semi-definite matrix.

$$\text{Set } \partial J/\partial x = 0 \text{ to yield}$$

$$\hat{x} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} x\,p(x\,|\,\Theta)dx_1\ldots dx_n \tag{3.37}$$

$$\therefore \hat{x} = E[x\,|\,\Theta] \tag{3.38}$$

For linear minimum variance unbiased (Gauss–Markov)

Let $\hat{x} = A\theta$, where A is an unknown parameter and $C_x = E(xx')$, $C_{x\theta} = E(x\theta')$ and $C_\theta = E(\theta\theta')$. Now,

$$\begin{aligned}
C_e &= E(ee') = E[(\hat{x}-x)(\hat{x}-x)'] \\
&= E[(A\mathrm{q}-x)(A\mathrm{q}-x)'] \\
&= E[(A\mathrm{q}\mathrm{q}'A' - x\mathrm{q}'A' - A\mathrm{q}x' + xx')] \\
&= E[A\mathrm{q}\mathrm{q}'A'] - E[x\mathrm{q}'A'] - E[A\mathrm{q}x'] + E[xx'] \\
&= AC_{\mathrm{q}}A' - C_{x\mathrm{q}}A' - AC_{\mathrm{q}}x + C_x \\
&= (A - C_{x\mathrm{q}}C_{\mathrm{q}}^{-1})C_{\mathrm{q}}(A - C_{x\mathrm{q}}C_{\mathrm{q}}^{-1}) - C_{x\mathrm{q}}C_{\mathrm{q}}^{-1}C'_{x\mathrm{q}} + C_x
\end{aligned}$$

Minimum is obtained when $A - C_{x\mathrm{q}}C_{\mathrm{q}}^{-1} = 0$.

That is

$$A = C_{x\mathrm{q}}C_{\mathrm{q}}^{-1} \tag{3.39}$$

$$\therefore x = C_{x\mathrm{q}}C_{\mathrm{q}}^{-1}\mathrm{q}, \quad C_e = C_x - C_{x\mathrm{q}}C_{\mathrm{q}}^{-1}C'_{x\mathrm{q}} \tag{3.40}$$

(i) If $\theta = Bx + v$, then

$$\begin{aligned} C_{xq} &= E[xq'] \\ &= E[x(Bx+v)'] \\ &= E[xx'B'] + E[xv'] \end{aligned}$$

$$C_{xq} = C_x B' + C_{xv}$$

$$\begin{aligned} C_q &= E[qq'] = E[(Bx+v)(Bx+v)'] \\ &= BC_xB' + C'_{xv}B' + BC_{xv} + C_v \end{aligned} \tag{3.41}$$

$$\hat{x} = [C_xB' + C_{xv}][BC_xB' + (BC_{xv})' + BC_{xv} + C_v]^{-1}q \tag{3.42}$$

$$C_e = C_x - (C_xB' + C_{xv})(BC_xB' + (BC_{xv})' + BC_{xv} + C_v)^{-1}(C_xB' + C_{xv})' \tag{3.43}$$

(ii) If $\theta = Bx + v$, $C_{xv} = 0$, then

$$\hat{x} = (C_xB')[BC_xB' + C_v]^{-1}q$$

$$C_e = C_x - (C_xB')(BC_xB' + C_v)^{-1}(C_xB')' \tag{3.44}$$

or

$$\hat{x} = (B'C_v^{-1}B)^{-1}B'C_v^{-1}q$$

$$\hat{x} = (C_x^{-1} + B'C_vB)^{-1}B'C_v^{-1}q \tag{3.45}$$

$$C_e = (C_x^{-1} + B'C_v^{-1}B)^{-1} \tag{3.46}$$

(iii) If in (ii), $C_x \to \infty$ (no information on state), then

$$\hat{x} = (B'C_v^{-1}B)^{-1}B'C_v^{-1}q \tag{3.47}$$

$$C_e = (B'C_v^{-1}B)^{-1} \tag{3.48}$$

Partitioned data sets

Let θ^r be a set of measurements of dimension r and $\hat{x}^r$ the estimate obtained using θ^r. Let $\theta^r = B^r x + v^r$ such that $(C_v^r)^{-1}$ exists, $C_{xv}^r = C_{vx}^r = 0$; $C_x^{-1} \to 0$, then

$$\hat{x}^r = (B'^r C_v{}^r B^r)^{-1} B'^r (C_v{}^r)^{-1} \mathbf{q}^r \tag{3.49}$$

$$C_e = (B'^r (C_v{}^r)^{-1} B^r)^{-1} \tag{3.50}$$

Suppose now that an additional set of data θ^s is taken:

$$\mathbf{q}^s = B^s x + v^s \tag{3.51}$$

provided $C_{xv}^s = 0$, $C_{vv}^{rs} = 0$, so that

$$\begin{bmatrix} \mathbf{q}^r \\ \mathbf{q}^s \end{bmatrix} = \begin{bmatrix} B^r \\ B^s \end{bmatrix} x + \begin{bmatrix} v^r \\ v^s \end{bmatrix}$$

Let r + s = m so that $\theta^m = (\theta^r, \theta^s)'$, $B^m = (B^r, B^s)'$ and $\mathrm{v}^m = (v^r, v^s)'$

$$\begin{aligned} C_v^m &= E(v^m, v^{m\prime}) \\ &= E\left[\begin{bmatrix} v^r \\ v^s \end{bmatrix}\begin{bmatrix} v^{r\prime} & v^{s\prime} \end{bmatrix}\right] \\ &= E\begin{bmatrix} v^r v^{r\prime} & v^r v^{s\prime} \\ v^s v^{r\prime} & v^s v^{s\prime} \end{bmatrix} \\ &= \begin{bmatrix} C_v^r & C_{vv}^{rs} \\ C_{vv}^{sr} & C_v^s \end{bmatrix} \\ &= \begin{bmatrix} C_v^r & 0 \\ 0 & C_v^s \end{bmatrix} \end{aligned}$$

$$\hat{x}^m = (B'^m (C_v{}^m)^{-1} B^m)^{-1} B'^m (C_v{}^m)^{-1} \mathbf{q}^m \tag{3.52}$$

$$C_e{}^m = (B'^m (C_v{}^m)^{-1} B^m)^{-1} \tag{3.53}$$

$$\hat{x}^m = \left[\begin{pmatrix} B^{r\prime} & B^{s\prime} \end{pmatrix} \begin{bmatrix} C_v^r & 0 \\ 0 & C_v^s \end{bmatrix}^{-1} \begin{pmatrix} B^r \\ B^s \end{pmatrix} \right]^{-1} \begin{pmatrix} B^{r\prime} & B^{s\prime} \end{pmatrix} \begin{bmatrix} C_v^r & 0 \\ 0 & C_v^s \end{bmatrix}$$

thus yielding

$$\hat{x}^m = (B'^r (C_v^r)^{-1} B^r + B^{s\prime} (C_v^s)^{-1} B^s)^{-1} (B^{r\prime} (C_v{}^r)^{-1} \mathbf{q}^r + B^{s\prime} C_v^s) \tag{3.54}$$

$$C_e{}^m = [(C_e{}^r)^{-1} + B^{s\prime} (C_v{}^s)^{-1} B^s]^{-1} \tag{3.55}$$

Kalman form

Given an estimate $\hat{x}^r$, old error matrix $C_e{}^r$, new data $\theta^s = B^s x + v^s$, the new estimate $\hat{x}^m$ based on all the data is found by the sequence

$$k = C_e{}^r B^{s\prime} (C_v^s + B^s C_e{}^r B^{s\prime})^{-1} \tag{3.56}$$

$$\hat{x}^m = \hat{x}^r + k(\mathbf{q}^s - B^s \hat{x}^r) \tag{3.57}$$

$$C_e{}^m = C_e{}^r - k B^s C_e{}^r \tag{3.58}$$

Discrete dynamic linear system estimation

$$\begin{aligned} & x_{i+1} = s(i+1,i)x_i + w_i, \quad s(i+1,i) \text{ is transformation matrix} \\ & E(w_i) = 0, \quad \forall \mathrm{i} \\ & E(w_i w_j{}^T) = 0, \quad \forall \mathrm{i} \neq \mathrm{j} \\ & E(w_i w_j{}^T) = 0, \quad \forall \mathrm{i} \leq \mathrm{j} \\ & W_i = E(w_i w_j{}^T) \end{aligned} \tag{3.59}$$

Observation vector

$$\begin{aligned} & \mathbf{q} = A_i x_i + q_i, \quad A_i \text{ is transformation matrix} \\ & E(q_i q_j{}^T) = 0, \quad \forall \mathrm{i} \neq \mathrm{j} \\ & E(q_i w_j{}^T) = 0, \quad \forall \mathrm{i}, \mathrm{j} \\ & E(q_i x_j{}^T) = 0, \quad \forall \mathrm{i}, \mathrm{j} \\ & Q_i = E(q_i q_j{}^T) \end{aligned} \tag{3.60}$$

Prediction

$$\hat{X}_p^m = S(p,n)\hat{X}_n^m, \quad p \geq n,\ S(.,.) \text{ is the transformation matrix} \tag{3.61}$$

$$C_p^m = S(p,n)C_n^m S^T(p,n) + \sum_{k=n}^{p-1} S(p,k+1)W_k S^T(p,k+1), \quad p > n \geq m \tag{3.62}$$

Filtering

$$\hat{X}_{m+1}^m = S(m+1,m)\hat{X}_m^m \tag{3.63}$$

$$C_{m+1}^m = S(m+1,m)C_m^m S^T S(m+1,m) + W_{m+1} \tag{3.64}$$

$$K = C_{m+1}^m \Delta_{m+1}^T (Q_{m+1} + \Delta_{m+1} C_{m+1}^m \Delta_{m+1}^T)^{-1} \tag{3.65}$$

$$\hat{X}_{m+1}^{m+1} = \hat{X}_m^m + K(\mathrm{q}_{m+1} - \Delta_{m+1}\hat{X}_{m+1}^m) \tag{3.66}$$

$$C_{m+1}^{m+1} = C_{m+1}^m - K\Delta_{m+1} C_{m+1}^m \tag{3.67}$$

Smoothing

$$\hat{X}_r^m = \hat{X}_r^r + J[\hat{X}_{r+1}^m - S(r+1,r)\hat{X}_r^r] \tag{3.68}$$

$$C_r^m = C_r^r + J[C_{r+1}^m - C_{r+1}^r]J^T \tag{3.69}$$

$$J = C_r^r S^T(r+1,r)(C_{r+1}^r)^{-1} \tag{3.70}$$

$$C_{r+1}^r = S(r+1,r)C_r^r S^T(r+1,r) + W_r \tag{3.71}$$

Continuous dynamic linear system

Let

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{3.72}$$

be the system equation and

$$y(t) = Cx(t) + Dv(t) \tag{3.73}$$

be the observation equation. The estimate $\hat{x}(t)$ is restricted as a linear function of $y(\tau)$, $0 \le \tau \le t$; thus,

$$\hat{x}(t) = \int_0^t \mathtt{a}(\mathtt{t}) y(\mathtt{t}) d\mathtt{t} \tag{3.74}$$

The solution to Equation 3.72 is

$$x(t) = \Phi(t)x(0) + \int_0^t \Phi(t)\Phi^{-1}(s)Bu(s)ds \tag{3.75}$$

where $\Phi(.)$ is the transition matrix. From Equation 3.74

$$\hat{x}(t+\mathtt{d}) = \int_0^{t+\mathtt{d}} \mathtt{a}(\mathtt{t}) y(\mathtt{t})\, d\mathtt{t}$$

And from Equation (3.75),

$$x(t+\mathtt{d}) = \Phi(t+\mathtt{d})\left\{ x(0) + \int_0^{t+\mathtt{d}} \Phi^{-1}(\mathtt{s})B(\mathtt{s})u(\mathtt{s})d\mathtt{s} \right\}$$

$$= \Phi(t+\mathtt{d})\Phi^{-1}(t)\Phi(t)\left\{ x(0) + \int_0^{t} \Phi^{-1}(\mathtt{s})B(\mathtt{s})u(\mathtt{s})d\mathtt{s} + \int_t^{t+\mathtt{d}} \Phi^{-1}(\mathtt{s})B(\mathtt{s})u(\mathtt{s})d\mathtt{s} \right\}$$

$$x(t+\mathtt{d}) = \Phi(t+\mathtt{d})\Phi^{-1}(t)\left\{ x(t) + \int_t^{t+\mathtt{d}} \Phi(t)\Phi^{-1}(\mathtt{s})B(\mathtt{s})u(\mathtt{s})d\mathtt{s} \right\}$$

Using the orthogonality principle,

$$E\left\{[x(t+\mathtt{d}) - \hat{x}(t+\mathtt{d})y'(\mathtt{t})]\right\} = 0 \quad \text{for } 0 \le \mathtt{t} \le t$$

and recalling that

$$E\left\{u(\mathtt{s})y'(\mathtt{t})\right\} = 0, \quad \mathtt{s} > t$$

Hence,

$$E\left\{[\Phi(t+\mathtt{d})\Phi^{-1}(t)x(t) - \hat{x}(t+\mathtt{d})]y'(\mathtt{t})\right\} = 0, \quad 0 \le \mathtt{t} \le t$$

Thus,

$$\hat{x}(t+\mathrm{d}) = \Phi(t+\mathrm{d})\Phi^{-1}(t)\hat{x}(t),$$

if

$$u(t) \approx N(0, Q(t)),$$

and

$$v(t) \approx N(0, R(t))$$

given that

$$E\{x(0)\} = \hat{x}_0$$

$$E\{[x(0) - \hat{x}_0][x(0) - \hat{x}_0]'\} = P_0$$

and $R^{-1}(t)$ exist.

The Kalman filter consists of

Estimate:

$$\dot{\hat{x}}(t) = A\hat{x}(t) + K(t)[y(t) - C\hat{x}(t)], \quad \hat{x}(0) = \hat{x}_0 \tag{3.76}$$

Error covariance:

$$\hat{P}(t) = AP(t) + P(t)A' + BQB' - KRK' \tag{3.77}$$

Propagation:

$$P(0) = P_0. \quad \text{For steady state } \hat{P}(t) = 0$$

Kalman gain matrix:

$$\dot{K}(t) = P(T)C'R^{-1}(t) \quad \text{when } E[u(t)'v(\mathrm{t})] = 0 \tag{3.78}$$

and

$$\dot{K}(t) = [P(T)G' + BG]R^{-1}(t) \quad \text{when } E[u(t)v'(\mathrm{t})] = G(t)\mathrm{d}(t - \mathrm{t}) \tag{3.79}$$

The fixed time smoothing algorithm $\hat{x}_{t|T}$ is as follows:

$$P(t \mid T) = (A + BQB'P^{-1})' - BQB'$$

with $\hat{x}(T \mid T) = \hat{x}(t = T)$ and $P(T \mid T) = P(\mathrm{t} = T)$ as initial conditions.

Continuous nonlinear estimation

The analysis of stochastic dynamic systems often leads to differential equations of the form

$$\dot{x}(t) = f(t,x) + G(t,x)u(t), \quad x_{t_0} = c, \quad 0 \le t \le T \le \infty \tag{3.80}$$

or in integral form

$$x(t) = c + \int_0^t f(s,x)ds + \int_0^t G(s,x)dw(s), \quad 0 \le t \le T \le \infty \tag{3.81}$$

where $dw(t)/dt = u(t)$ and w(t) is the Wiener process; x(t) and f(t, s) are n-dimensional, while G(t, x) is (n × m) matrix function and u(t) is m-dimensional (Figure 3.7).

ItÔ rule:

$$\int_0^T G(t,x)dw(t) = \lim_{\Delta \to 0} \sum_{i=0}^{N} G(t_i, x(t_i))(w(t_{i+1}) - w(t_i)) \tag{3.82}$$

For the partition $t_0 < t_1 < t_i \cdots < t_i < t_{i+1} \cdots t_N = T$ and $\Delta = \max_i(t_{i+1} - t_i)$

Stratonovich rule:

$$\int_0^T G(t,x)dw(t) = \lim_{\Delta \to 0} \sum_{i=0}^{N-1} G\left(t_i, \frac{x(t_{i+1}) - x(t_i)}{2}\right)(w(t_{i+1}) - w(t_i)) \tag{3.83}$$

Let the observation be of the form $y(t) = z(t) + v(t)$ where $z(t) = \varphi(x(s), s \le t)$ and $v(t)$ are p-dimensional vectors.

It is further assumed that $E[z(t)z'(t)] < \infty$ and $E[z(t)v'(t)] = 0$ for all t.

Doob (1958) obtained the estimator

$$\hat{x}(t \mid T) = E[x(t) \mid y(s), t_0 \le s \le \mathrm{t}] \tag{3.84}$$

where $t = \tau$ (filtering), $t < \tau$ (smoothing), and $t > \tau$ (prediction).

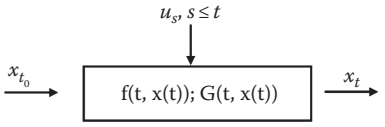


Figure 3.7 Functional configuration.

$$\hat{x}(t) = E[x(t) \mid y(s), t_0 \le s \le t]$$

$$= \int_{-\infty}^{\infty} x(t) P_r(x(t) \mid y(t)) dx(t) \tag{3.85}$$

Let

$$P_t = E[(x(t) - \hat{x}(t))(x(t) - \hat{x}(t))']$$

$$= \int_{-\infty}^{\infty} (x(t) - \hat{x}(t))(x(t) - \hat{x}(t))' P_r(x(t) \mid y(t)) dx(t) \tag{3.86}$$

Assume that

$$E\{u(t)\} = E\{v(t)\} = E\{u(t)v'(t)\} = 0$$

$$E\{u(t)u'(t)\} = Q_t, \quad E\{v(t)v(t)'\} = R_t$$

The Folker–Plank stochastic differential equations for the probability density function P_r:

$$\frac{\partial P_r}{\partial t} = -trace\left(\frac{\partial}{\partial x}\{f(t,x)P_r\}\right) + \frac{1}{2} trace\left(\frac{\partial}{\partial x}\left(\left[\frac{\partial}{\partial x}\right]'\{G(t,x)QG'(t,x)P_r\}\right)\right)$$

$$+ P_r(y - \Phi(t,x))R_t^{-1}(\Phi(t,x) - \Phi(t,\hat{x})) \tag{3.87}$$

Using Equation 3.87 in Equations 3.85 and 3.86:

$$d\hat{x}(t) = \hat{f}(t,x)dt + E\{(x - \hat{x})\Phi'(t,x) \mid y(t)\}R_t^{-1}(y(t) - \Phi(t,x))dt \tag{3.88}$$

$$dP_t + d\hat{x}d\hat{x} = E\{f(t,x)(x - \hat{x}) \mid y(t)\}dt$$

$$+ E\{(x - \hat{x})f'(t,x) \mid y(t)\}dt$$

$$+ E\{G(t,x)Q_tG'(t,x) \mid y(t)\}dt$$

$$+ E\{(x - \hat{x})(x - \hat{x})'[\Phi(t,x) - \Phi(t,\hat{x})]R_t^{-1}(y(t) - \hat{\Phi}(t,x)) \mid y(t)\}dt$$

Extended Kalman filter

The extended Kalman filter results from application of the linear Kalman–Bucy filter to a linearized nonlinear system, where the nonlinear system is relinearized after each observation.

Let

$$f(t,x) = f(t,\hat{x}) - \frac{\partial f}{\partial \hat{x}}(x-\hat{x}) \tag{3.89}$$

$$\Phi(t,x) = \Phi(t,\hat{x}) - \frac{\partial \Phi}{\partial \hat{x}}(x-\hat{x}) \tag{3.90}$$

and

$$G(t,x)Q_tG'(t,x) = G(t,\hat{x})Q_tG'(t,\hat{x}) + (x-\hat{x})\left\{\left(\frac{\partial}{\partial \hat{x}}\right)' G(t,\hat{x})Q_tG'(t,\hat{x})\right\} \tag{3.91}$$

Substituting the above Equation 3.91 into the previous Equation 3.88 we obtain

$$\frac{d\hat{x}}{dt} = f(t,\hat{x}) + P_t \frac{\partial \Phi}{\partial \hat{x}} R_t^{-1}(y(t) - \hat{\Phi}(t,x)) \tag{3.92}$$

$$\frac{dP_t}{dt} = \frac{\partial f}{\partial \hat{x}} P_t + P_t \frac{\partial f'}{\partial \hat{x}} G(t,\hat{x})Q_tG'(t,\hat{x}) - P_t \frac{\partial \Phi'}{\partial \hat{x}} R_t^{-1} \frac{\partial \Phi'}{\partial \hat{x}} P_t \tag{3.93}$$

which can now be solved with appropriate initial conditions

$$\hat{x}\,|_{t_0} = \hat{x}(t_0) \quad \text{and} \quad \mathrm{P}_t\,|_{t_0} = P(t_0)$$

*This is the **extended** Kalman Filter for **nonlinear** systems.*

Partitional estimation

Lainiotis (1974) proposed the partition theorem, general continuous-data Bayes' rule for the posterior probability density.

Let

$$P_r(x(\mathtt{t})\,|\,t,t_0) = \frac{\Lambda(t,t_0\,|\,x(\mathtt{t}))P_r(x(\mathtt{t}))}{\int \Lambda(t,t_0\,|\,x(\mathtt{t}))P_r(x(\mathtt{t}))d\mathtt{t}} \tag{3.94}$$

where

1. $$\Lambda(t, t_0 \mid x(\mathtt{t})) = \exp\left\{ \int_{t_0}^{t} \hat{h}'(\mathtt{s} \mid \mathtt{s}, t_0; x(\mathtt{t})) R_{\mathtt{s}}^{-1} y(\mathtt{s}) d\mathtt{s} - \frac{1}{2} \int_{t_0}^{t} \left\| \hat{h}(\mathtt{s} \mid \mathtt{s}, t_0; x(\mathtt{t})) \right\|^2 R_{\mathtt{s}}^{-1} d\mathtt{s} \right\} \quad (3.95)$$
2. $\hat{h}(\sigma \mid \sigma, t_0; x(\tau)) = E\{h(\sigma, x(\sigma)) \mid y(\sigma); x(\tau)\}$
3. $P_r(x(\tau))$ is the a priori density of $x(\tau)$

The partitioned algorithm for filtering is given by

$$\hat{x}(t) = \int \hat{x}(\mathtt{t}) P_r(x(\mathtt{t}) \mid t, t_0) dx(\mathtt{t}) \quad (3.96)$$

and

$$P_t = \int \left\{ P_{\mathtt{t}} + [\hat{x}(\mathtt{t}) - \hat{x}(t)][\hat{x}(\mathtt{t}) - \hat{x}(t)]' P_r(x(\mathtt{t}) \mid t, t_0) dx(\mathtt{t}) \right\} \quad (3.97)$$

where $P_r(x(\tau) \mid t, t_0)$ is given previously and both $\hat{x}(\tau)$ and P_τ are the "anchored" or conditional mean-square error estimate and error-covariance matrices, respectively.

The partitioned algorithm takes its name from the fact that if the observation interval is partitioned into several small subintervals, repeated use of the filtering equations for each subinterval leads to effective and computationally efficient algorithm for the general estimation problem.

Invariant imbedding

The invariant imbedding approach provides a sequential estimation scheme which does not depend on a priori noise statistical assumptions. The concept in invariant imbedding is to find the estimate $\hat{x}(\tau)$ of x(t) such that the cost function

$$J = \frac{1}{2} \int_0^T \left\{ \left\| y(t) - \Phi(t, \hat{x}(t)) \right\|_{W_1}^2 + \left\| \hat{x}(t) - f(t, \hat{x}(t)) \right\|_{W_2}^2 \right\} dt \quad (3.98)$$

is minimized, where W_1 and W_2 are weighing matrices that afford the opportunity to place more emphasis on the most reliable measurements. The Hamiltonian H is therefore

$$H = \frac{1}{2} W_1 (y(t) - \Phi(t, \hat{x}(t)))^2 + \frac{1}{2} W_2 G^2(t, x(t)) u^2(t) + \mathbb{1}^2(t)(f(t, \hat{x}(t)) + G(t, \hat{x}(t))u(t)) \tag{3.99}$$

For which the necessary conditions for a minimum are

$$\begin{aligned} \dot{\hat{x}}(t) &= \frac{\partial H}{\partial \mathbb{1}} \\ \dot{\mathbb{1}}(t) &= -\frac{\partial H}{\partial \hat{x}} \\ \frac{\partial H}{\partial u} &= 0 \end{aligned} \tag{3.100}$$

which yield the filtering equations

$$\frac{d\hat{x}}{dT} = f(T, \hat{x}(T)) + P \frac{\partial \Phi'}{\partial \hat{x}} (y(T) - \Phi(T, \hat{x}(T))) \tag{3.101}$$

$$\frac{dP}{dT} = \frac{\partial f}{\partial \hat{x}} P + P \frac{\partial f'}{\partial \hat{x}} + P \left(\left[\frac{\partial^2 \Phi'}{\partial \hat{x}^2} \right] (y(T) - \Phi(T, \hat{x}(T))) - \frac{\partial \Phi'}{\partial \hat{x}} \frac{\partial \Phi}{\partial \hat{x}} \right) P + \frac{1}{W} \tag{3.102}$$

Stochastic approximations/innovations concept

Stochastic approximation is a scheme for successive approximation of a sought quantity when the observation and the system dynamics involve random errors. It is applicable to the statistical problem of (Barrett and Lampard, 1955)

1. Finding the value of a parameter which causes an unknown noisy regression function to take on some preassigned value
2. Finding the value of a parameter which minimizes an unknown noisy regression function

Stochastic approximation has wide applications to system modeling, data filtering, and data prediction. It is known that a procedure which is optimal in decision theoretic sense can be nonoptimal. Sometimes the algorithm is too complex to implement, for example, in situations where the nonlinear effects cannot be accurately approximated by linearization or the

noise process is strictly non-Gaussian. A theoretical solution is obtained by using the concepts of innovations and martingales. Subsequently, a numerically feasible solution is achieved through stochastic approximation. The innovations approach separates the task of obtaining a more tractable expression for the equation

$$\hat{x}(t \mid T) = E\{x(t) \mid y(s), 0 \le s \le \mathfrak{t}\} \tag{3.103}$$

into two parts:

1. The data process $\{y(t), 0 \le t \le T\}$ is transformed through a causal and causally invertible filter $v(t) = y(t) - \Phi(\hat{x}(s), s \le t)$ called the innovations process with the same intensity as the observation process.
2. The optimal estimator is determined as a functional of the innovations process.

The following algorithm has been used for several problems:

(i) Pick an $\mathbf{a}_t^i$ gain matrix function, such that for each element $(\mathbf{a}_t^i)_{\mathrm{kl}}$,

$$\int_0^\infty (\mathbf{a}_t^i)_{\mathrm{kl}} dt = \infty, \quad \mathrm{i} = 1 \quad \text{and} \quad \int_0^\infty (\mathbf{a}_t^i)_{\mathrm{kl}}^2 dt < \infty \tag{3.104}$$

(ii) Solve the *suboptimal* problem

$$\frac{d\hat{x}}{dt} = f(t, \hat{x}) + \mathbf{a}_l^i G(t, \hat{x})(y(t) - \Phi(t, \hat{x})) \tag{3.105}$$

where it is assumed without any loss of generality with entries $(\mathbf{a}_1^i, \mathbf{a}_2^i, \ldots \mathbf{a}_n^i)$. The lth component of equation (i) is

$$\frac{d\hat{x}_l}{dt} = f_l(t, \hat{x}) + \sum_{k=1}^{m} \mathbf{a}_l^i g_{\mathrm{lk}}(t, \hat{x})(y_k(t) - \Phi_k(t, \hat{x}))$$

(iii) Compute the innovations process

$$v^i(t) = y(t) - \Phi(t, \hat{x}^i) \tag{3.106}$$

and check for its whiteness (within a prescribed tolerance level) by computing the autocorrelation function as well as the power spectrum.

(iv) If the result of the test conducted in step (iii) is positive, STOP. ELSE, iterate on $\mathbf{a}_i^i$; thus,

$$\mathbf{a}^{i+1}(t) = \mathbf{a}^i(t) + \mathbf{g}^i(t)\Psi(v^i(t)) \tag{3.107}$$

where

$$\mathbf{a}^1(t) = \mathbf{g}^i(t)$$
$$= \left\{ -\frac{a}{t} \quad \text{or} \quad -\frac{a}{b+t} \quad \text{or} \quad -\frac{a+t}{b+t^2} \right\} \tag{3.108}$$

and

$$\Psi(v^i(t)) = v^i(t) - E\{v^i(t)\} \tag{3.109}$$

(v) Go to step (ii)

The optimal trajectories constitute a martingale process and the convergence of the approximate algorithm depends on the assumption that the innovation of the observations is a martingale process, thus the iterations will converge.

Model control—Model reduction, model analysis

Introduction

One of the challenges in systems modeling and subsystem estimation is the need to reduce large-scale systems to lower dimensions in order to effectively control the models (Golub and Reinsch, 1970).

Consider an nth-order linear system S_1 defined by

$$S_1 : \dot{x} = Ax + Bu \tag{3.110}$$

where

x is the n-dimensional state vector
A is an $(n \times n)$ system matrix
u is a p-dimensional input vector

Let z be an m-vector $(m < n)$ related to x by

$$z = Cx \tag{3.111}$$

In model reduction, it is desirable to find an mth-order system S_2 described by

$$S_2 : \dot{z} = Fz + Gu \tag{3.112}$$

The ($m \times n$) matrix C in Equation 3.111 is the aggregation matrix and S_2 is the aggregated system or the reduced model. It is easy to show that $G = CD$ and that F must satisfy the matrix equation

$$FC = CA \tag{3.113}$$

Equation 3.113 defines an over-specified system of equation for the unknown matrix F, and hence F must be approximated. A multivariate linear regression scheme is used to yield a "best" approximation for F in the form

$$\hat{F} = CAC^T(CC^T)^{-1} \tag{3.114}$$

where T and –1 denote matrix transpose and matrix inverse, respectively. The rank of C is assumed to be m. The result given by Equation 3.114 is interpreted as a linear, unbiased, minimum-variance estimate of F and its form agrees with that given by Aoki (1968), following an ad hoc procedure.

In addition, the covariance of $\hat{F}$ is found to be

$$con\left(vec\,\hat{F}\right) = \mathbf{s}^2\left[(CC^T)^{-1} \otimes I_m\right] \tag{3.115}$$

and it is shown that this covariance matrix can be used for model reduction error assessment. In Equation 3.115, the Kronecker product $\otimes$ and the "vec" operator are defined as

$$P \otimes Q = [P_{ij} \quad Q] \tag{3.116}$$

$$vec(P) = \left[P_1, P_2, ...\right]^T \tag{3.117}$$

where

P and Q are matrices of arbitrary dimensions

P_k is the kth column of matrix P

From the computational point of view, it is desirable to circumvent the use of matrix inverses in Equations 3.114 and 3.115, particularly for systems that are large aggregation matrices. In what follows, this is accomplished through the use of matrix singular value decomposition (SVD) which has found useful application in several linear least squares problems.

The SVD concept (Soong, 1977; Aoki, 1968) gives the Moore–Penrose pseudo-inverse of *C* as

$$C^{+} = VAU^{T} \tag{3.118}$$

where U and V are unitary matrices whose columns are the eigenvectors of matrices DD^T, and D^TD, respectively, and

$$A = \begin{bmatrix} \sigma_1^{-1} & & & 0 & \\ & \sigma_2^{-1} & & & \\ & & \ddots & & \\ & 0 & & \sigma_m^{-1} & \\ & & & & 0 \\ & & & & & \ddots \\ & & & & & & 0 \end{bmatrix}_{n \times n} \tag{3.119}$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m \geq 0$, called singular values, are the nonnegative square roots of the eigenvalues of D^TD. A discussion of this decomposition and its properties can be found in Stewart (1973).

Now, Equation 3.113 gives

$$\hat{F} = CAC^{+} \tag{3.120}$$

And, using SVD, we can write

$$\hat{F} = CA\,(VAU^{T}) \tag{3.121}$$

The matrix *C* can also be written in the form

$$D = U\Sigma V^{T} \tag{3.122}$$

where

$$\sum = \begin{bmatrix} \mathbf{s}_1 & & & & \\ & \mathbf{s}_2 & 0 & & \\ & & \ddots & & \\ & 0 & & \mathbf{s}_m & \\ & & & & 0 \\ & & & & & \ddots \\ & & & & & & 0 \end{bmatrix}_{n \times n} \tag{3.123}$$

and we have

$$\hat{F} = U \sum V^T A V A U^T \tag{3.124}$$

Compared with Equation 3.114, either Equation 3.121 or Equation 3.124 provides a more efficient method of computation for $\hat{F}$ due to elimination of the matrix inverse.

Similarly, advantages are realized in the calculation of $con(vec\,\hat{F})$. Following the SVD scheme,

$$\begin{aligned} (DD^T)^{-1} &= (D^T)^+ D^+ \\ &= (UAV^T)(VAU^T) \\ &= (UA^2U^T) \end{aligned} \tag{3.125}$$

Equation (3.115) now takes the form

$$con(vec\,\hat{F}) = \mathbf{s}^2[UA^2U^T \otimes I_m] \tag{3.126}$$

which is clearly of a simpler structure than Equation 3.115.

Modal approach for estimation in distributed parameter systems

We can expand this concept to those describable by distributed parameter systems: based on a scheme, on eigenmode expansion, for the estimation

of system responses in distributed parameter systems with nonlinear sensors. The associated joint conditional probability distribution is realized by using the partition theorem and Wiener functional expansions. The system state and parameters are subsequently computed by obtaining the innovation processes from the observations leading to structural estimation.

A technique for optimal estimation of system parameters in a stochastic environment with nonlinear sensors is hereby presented. It is based on the eigenmode representation of generalized displacements and orthogonal expansion of the conditional joint probability distribution. The algorithm proposed in this chapter for realization of the innovations and system state is the starting point in the search of nonlinear distributed-parameter estimation techniques. A linear dynamical system is used to illustrate the correspondence between the innovations process, the observations process, and the system response.

The proposed algorithm may be applied to several civil engineering and space structures, best described by the distributed parameter equation

$$m(x)u_{ii}(x,t) + Bu_i(x,t) + Cu(x,t) = F_0(x,t) \tag{3.127}$$

with the initial conditions

$$u(x,t_0) = u^0 \quad \text{and} \quad u_i(x,t_0) = u_i{}^0$$

where

$m(x)$ is the mass per unit length (assumed unity)
$u(x, t)$ the vector of generalized displacements
B and C are in general spatial differential operators representing the damping and restoring forces, respectively
$F_0(x, t)$ is the external impressed forces (wind loads, earthquake excitations, etc.)

Modal canonical representation

Let the displacement be expressed in terms of structural dominant mode shapes $\{\phi_i(x)\}_1^N$ with associated frequencies $\{w_i\}_1^N$

$$u(x,t) = \sum_1^N u_i(t)\text{£}_i(x); \quad F_0(x,t) = \sum_1^N F_1(t)\text{£}_i(t)$$

so that

$$\left.\begin{aligned} u(x,t) &= U^T(t)\Phi(x) \\ &\text{and} \\ F_0(x,t) &= F_0{}^T(t)\Phi(x) \end{aligned}\right\} \tag{3.128}$$

where

$$U = \left(u_1, u_2, \ldots, u_N\right)^T, \quad F_0 = \left(F_1{}^0, F_2{}^0, \ldots, F_N{}^0\right)^T$$

and

$$\Phi = \left(\mathtt{f}_1, \mathtt{f}_2, \ldots, \mathtt{f}_N\right)^T$$

Substituting Equation 3.128 to Equation 3.127 gives

$$U^T(t)\ \Phi(x) + BU^T(t)\Phi(x) + CU^T(t)\Phi(x) = F_0{}^T(t)\Phi(x)$$

thus obtaining the *i*th mode amplitude equation

$$\ddot{u}_i(t) + 2\mathtt{x}_i\mathtt{w}_i\,\dot{u}_i(t) + \mathtt{w}_i{}^2 u_i(t) = F_i^0(t) \tag{3.129}$$

where $C\mathtt{f}_i = \mathtt{w}_i{}^2\mathtt{f}_i$ and $B\mathtt{f}_i = 2\mathtt{x}_i\mathtt{w}_i\mathtt{f}_i$ with appropriate initial conditions.

Equation 3.129 can now be put in the state space form:

$$\dot{q}(t) = Aq(t) + F \tag{3.130}$$

such that

$$q = (q_1,\ q_2)^T$$

with

$$q_1 = u_i, \quad q_2 = \dot{u}_i$$

$$A = \begin{pmatrix} 0 & 1 \\ -\mathtt{w}_i{}^2 & -2\mathtt{x}_i\mathtt{w}_i \end{pmatrix}$$

and

$$F = \left(0, F_i^0\right)^T$$

The sensors are assumed to have the form

$$y(t) = z(t) + v(t); \quad z(t) = h[q(s); s \le t] \tag{3.131}$$

where

$$E\{v(t)\} = 0; \quad E\left\{v(t)v^T(s)\right\} = R(t)\mathrm{d}(t-s);$$
$$E\{z(t)\} = 0; \quad [z(t)] \le M < \infty \quad \text{for all } t;$$

$$\int_0^T E\left\{z(t)z^T(t)\right\}dt < \infty; \quad E\left\{v(t)F^T\right\} = 0 \quad \text{and } h \text{ is a functional of } q(s).$$

It is further assumed that the forcing function F is additively (preferably Gaussian) random.

The least squares estimate of $q(t)$ is

$$\hat{q}(t|\mathtt{t}) = E\left\{q(\mathtt{t})|y(s), 0 \le s \le \mathtt{t}\right\} \tag{3.132}$$

Assuming that the property of "causal equivalence" (Clark, 1969) holds, the sensor equations may be transformed into a white Gaussian process $v(t)$, called the strict sense innovations process with a simpler structure so that Equation 3.132 can now be put in the form

$$\hat{q}(t|\mathtt{t}) = \int_0^{\mathtt{t}} E\left\{q(t)v^T(s)|v(\mathtt{s}), 0 \le \mathtt{s} < s\right\}v(s)ds \tag{3.133}$$

where $v(t) = y(t) - \hat{z}(t|t)$, $0 \le t < \mathtt{t}$ and $\int$ is the *Itô* (1951) integral. A differential structure for Equation (3.133) is obtained as

$$\dot{\hat{q}}(t|t) + A\hat{q}(t|t) + K(t)\,v(t) \tag{3.134}$$

where

$$K(t) = E\left\{\hat{q}(t)v^T(t)|v(\mathtt{s}), 0 \le \mathtt{s} < t\right\} \tag{3.135}$$

Conditional joint probability distribution functions are required to obtain explicit solution to the stochastic differential Equation 3.134.

A method for the evaluation of the probability densities is obtained in the form of a partition theorem (Lainiotis, 1971; Clark, 1969):

$$p(q|f^t) = A(t|q)\,p(q_0)\Big|\int P(q_0)A(t|q)dq \tag{3.136}$$

with $f^t = \{y(\tau)\}_0^t$ and

$$A(t|q) = \exp\left\{\int_0^t \hat{h}(s,q)R^{-1}(s)\,y(s)ds - \frac{1}{2}\int_0^t \|\hat{h}(s,q)\|^2_{R^{-1}(s)}\,ds\right\} \tag{3.137}$$

where

$$\hat{h}(s,q) = E\left\{h(s,q)\Big|f^s\right\}$$

and $p(q_0)$ is the a priori density function.

Equation 3.136 can be approximated by expanding the distributions $A(t|q)$ in a Volterra (Barrett and Lampard, 1955; Biglieri, 1973) power series with functional terms

$$A(t|q) = \sum_{i=0}^{\infty} A_i(t|q) \tag{3.138}$$

where

$$A_i(t|q) = \frac{1}{i!}\int_0^\infty d_{\tau_1}\ldots\int_0^\infty d_{\tau_i} g_i(\tau_1,\tau_2,\ldots,\tau_i)\prod_{k-1}^{i} v(t-\tau_k) \tag{3.139}$$

such that $g_i(\ldots)$, $i = 1,\ldots, n$ are the integral kernels describing the system and $v(\cdot)$ is the innovations process. It has been assumed that in order to satisfy the requirements of physical realizability, the kernels are zero for any argument less than zero. The kernels can be identified in the following manner:

1. First-order kernel: $g_1(t)$
   Let $v(t) = A\delta_0(t)$ be an impulse of strength A so that

$$\Lambda(t|q) = Ag_1(t) + A^2 g_2(t,t) + A^3 g_3(t.t,t) + \cdots \tag{3.140}$$

and

$$g_1(t_1) = \left.\frac{\mathrm{d}\Lambda(t_1|q)}{\mathrm{d}A}\right|_{A=0} \tag{3.141}$$

2. Second-order kernel: $g_2(t, t)$
Let $P(t) = A\delta_0(t) + B\delta_0(t + \tau)$ be two impulses at t and $t + \tau$ of strength A and B, respectively, where A and B are specified constants as the preceding ones; then

$$\begin{aligned} A(t|\mathtt{q}) = {} & Ag_1(t) + Bg_1(t+\mathtt{t}) + A^2 g_2(t,t) + 2ABg_2(t,t+\mathtt{t}) \\ & + B^2 g_2(t+\mathtt{t},t+\mathtt{t}) + \cdots \end{aligned}$$

and the second-order kernel

$$g_2(t,\ t+\mathtt{t}) = \frac{1}{2}\left.\frac{\mathrm{d}^2\Lambda(t_1|q)}{\mathrm{d}A\mathrm{d}B}\right|_{\substack{A=0\\B=0}} \tag{3.142}$$

Higher-order kernels can be computed in a similar manner.

The system state estimates can now be obtained as

$$\hat{q}(t|t) = J(t|t_0)\,\hat{q}(t|t_0) + \int_{t_o}^{i} J(t|\mathtt{t})\ k(\mathtt{t})v(\mathtt{t})d\mathtt{t} \tag{3.143}$$

where

$$J(t|t_0) = \exp\{A(t-t_0)\} \tag{3.144}$$

is the transition/fundamental matrix; $k(\tau)$ having been approximated using the Equations 3.136 and 3.138.

As an example, take a hypothetical linear dynamical system, solved in the following, using the Kalman filtering equations, to illustrate the close correlation (in a non-statistical sense) between the observations process, the innovations process, and the systems state. It has been assumed that the distributed parameter system is reducible to a linear ordinary differential equation system as follows:

$$\text{System dynamics: } \dot{x}(t) = 0.5x(t) + u(t), \quad x(0) = 0$$

$$\text{Observation process: } y(t) = x(t) + v(t)$$

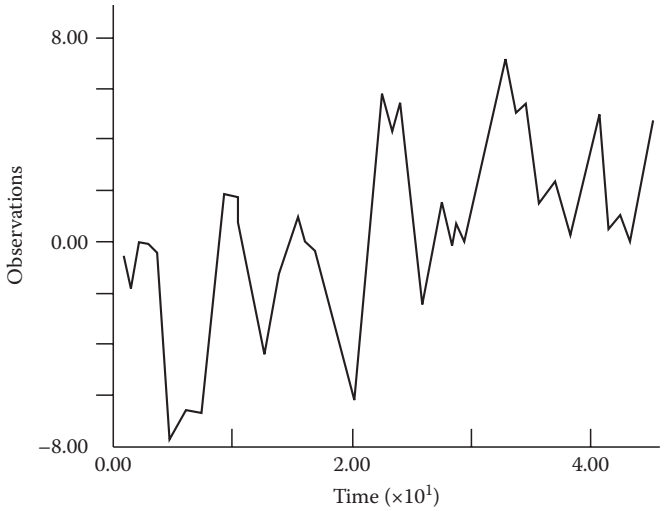


Figure 3.8 Observation process.

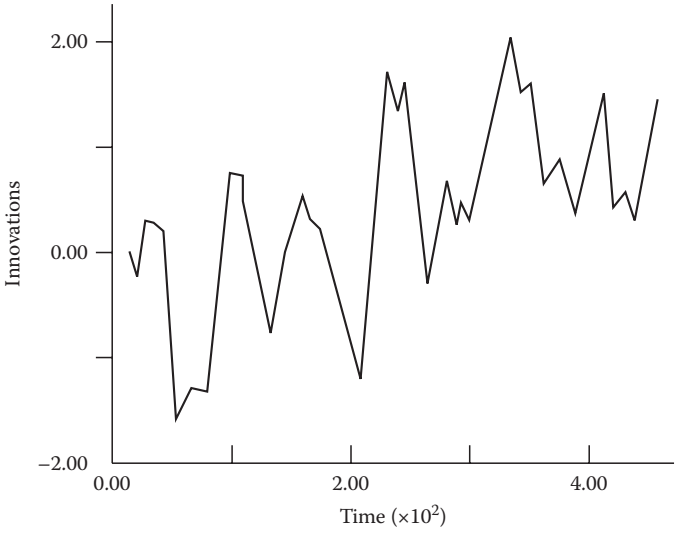


Figure 3.9 Innovations process.

Figure 3.8 represents the observation process, which is simply taken here as the system response plus a white Gaussian noise. Figure 3.9 is the innovations representation. Figure 3.10 is the innovations autocorrelation, which indicates that the process is a white noise. Finally, Figure 3.11 shows the system state $x(t)$.

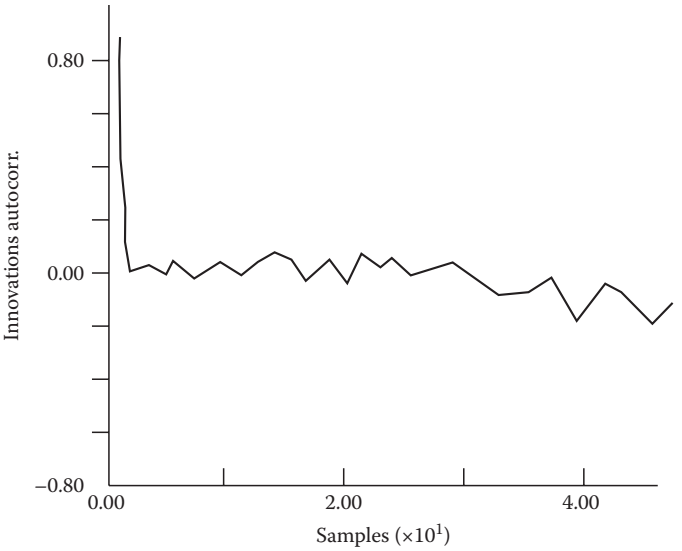


Figure 3.10 Autocorrelation for innovations process.

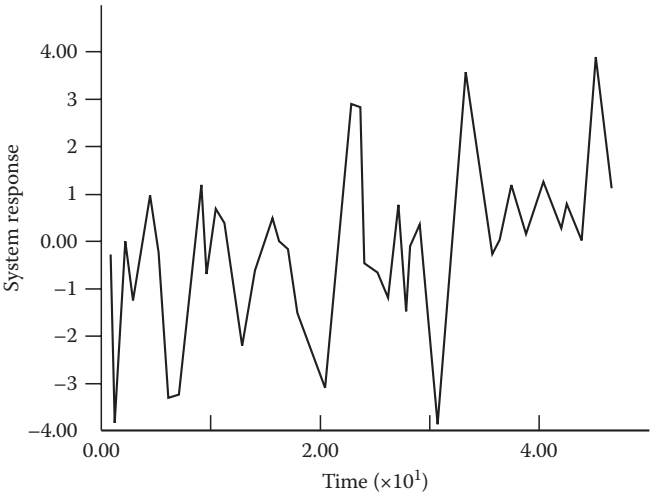


Figure 3.11 System response.

An approximation technique becomes imperative in the nonlinear case since the innovations sequence generated is not a true innovations process and hence a scheme for adaptive improvement has to be employed.

The modal approach for the decomposition of response estimates and conditional probability distributions provides an innovative method for the optimal estimation in some nonlinear distributed parameter systems.

The generation and determination of the innovations for the general (non-Gaussian) case is obtained through a stochastic approximation technique as is illustrated by an example given in this chapter. The method is expected to provide a viable alternative to similar Markov diffusion problems since it is easier to understand and computationally more efficient.

References

Albert, A. E. and Gardner, L. A. Jr., *Stochastic Approximation and Nonlinear Regression*. Cambridge, MA, 1967.

Aoki, M., Control of large scale dynamic systems by aggregation, *IEEE Transactions on Automatic Control*, AC-13, 246–253, 1968.

Barrett, J. F. and Lampard, D. G., *IRE Transactions on Information Theory*, 1, 10, 1955.

Bekey, G. A., System identification—An introduction and a survey, *Simulation*, 1970.

Biglieri, E., *Proceedings of the Institute of the Electrical and Electronics Engineers*, 61, 251, 1973.

Clark, J. M. C., Technical Report, Department of Computing and Control, Imperial College, London, U.K., 1969.

Gelb, A., *Applied Optimal Estimation*. MIT Press, Cambridge, MA, 1974.

Golub, G. H. and Reinsch, C., Singular value decomposition and least squares solutions, *Numerical Mathematics*, 14, 403–420, 1970.

Golub, G. H. and Van Loan, C., Total least squares. T. Gasser and M. Rosenblatt (eds.), In *Smoothing Techniques for Curve Estimation*, Springer Verlag, New York, 1979, pp. 62–76.

Ito, K., *Memoirs of the American Mathematical Society*, 4, 1951.

Lainiotis, D. G., *International Journal of Control*, 14, 1137, 1971.

Lee, R. C. K., *Optimal Estimation, Identification and Control*. MIT Press, Cambridge, MA, 1964.

Liebelt, P. B., *An Introduction to Optimal Estimation*. Addison-Wesley, Reading, MA, 1967.

Rashwan, A. H. and Ahmed, A. S. 1982, Proc. I. F. A. C., India.

Sage, A. P. and Melsa, J. L., *System Identification*. Academic Press, New York, 1971.

Soong, T. T., On model reduction of large scale systems, *Journal of Mathematical Analysis and Applications*, 60, 477–482, 1977.

Stewart, G. W., *Introduction to Matrix Computations*. Academic Press, New York, 1973.

chapter four

Systems optimization techniques

Optimality conditions

Optimization encompasses such areas as the theory of ordinary maxima and minima, calculus of variations, linear and nonlinear programming, dynamic programming, maximum principles, discrete and continuous games, and differential games of varying degrees of complexity.

In this chapter, the methodical development of the optimization problem is examined in relation to applications in applied sciences from classical techniques through stochastic approaches to contemporary and recent methods of intelligent MetaHeuristic. Indeed, computational techniques have grown from classical techniques, which afford closed-form solutions to approximation and search techniques and more recently intelligent search techniques. Such intelligent heuristics include tabu search, simulated annealing (SA), fuzzy systems, neural networks, and genetic algorithms (GAs), among a myriad of evolving modern computational intelligent techniques. Furthermore, important applications of these heuristics to the evolution of self-organizing adaptive systems such as modern economic models, transportation, and mobile robots do exist.

Central to all problems in optimization theory are the concepts of payoff, controllers or players, system, and information sets. In order to define a solution to an optimization problem, the concept of payoff must be defined and the controllers must be identified. If there is only one person on whose decision the outcome of some particular process depends and the outcome can be described by a single quantity, then the meaning of payoff (and hence solution to the optimization problem) and controller or player is clear.

The simplest, of course, is the problem of parameter optimization, which includes the classical theory of maxima and minima, linear and nonlinear programming. In parameter optimization, there is one (deterministic or probabilistic) criterion, one controller, one complete information set and the system state described by static equations and/or inequalities in the form of linear or nonlinear algebraic or difference equations.

On the next rung of complexity are optimization problems of dynamic systems where the state is defined by ordinary or partial differential equations. These can be thought of as limiting cases of multistage (static)

parameter optimization problems where the time increment between steps tends to zero. In this class, developed extensively as **optimal control**, we encounter the classical calculus of variation problems and their extension through various maximum and optimality principles, that is, Pontryagin's minimum principle and the dynamic programming principle. We are still concerned with one criterion, controller, and information set but have added dimension in that the problem is dynamic and might be deterministic or stochastic.

The next level would introduce two controllers (players) with a single conflicting criterion. Here we encounter elementary or finite matrix game theory where the controls and payoff are continuous functions. Each player at this level has complete information regarding the payoff for each strategy but may or may not have knowledge of his opponent's strategy. Such games are known as zero-sum games since the sum of the payoffs to each player for each move is zero, what one player gains the other loses. If the order in which the players act does not matter, that is, the minimum and maximum of the payoff are equal (this minimax is called the value of the game and is unique), the optimal strategies of the players are unaffected by knowledge or lack of knowledge of each other's strategy. A solution to this game involves the value and at least one optimal strategy for each player.

Next, we can consider extensions to dynamical systems where the state is governed by differential equations; we have one conflicting criterion and two players, that is, zero-sum, two-player differential games. The information available to each player might be complete or incomplete. In cases with complete information, the finite game concept of a solution is directly applicable. For the incomplete information case, it is reasonable to expect mixed strategies to form the solution but not much is known of solution methods or whether a solution always exists in the finite game theoretic sense.

In the last group or uppermost rung of the hierarchy we identify a class of optimization problems where the concept of a solution is far from clear. To this class belong multiple criteria, n-person games with complete or incomplete information, nonzero sum (either or both players may lose or gain).

The goal of an optimization problem can be formulated as follows: find the combination of parameters (independent variables) which optimize a given quantity, possibly subject to some restrictions on the allowed parameter ranges. The quantity to be optimized (maximized or minimized) is termed the *objective function*, the parameters which may be changed in the quest for the optimum are called *control or decision variables*, and the restrictions on allowed parameter values are known as *constraints*. A maximum of a function f is a minimum of –f. The general optimization problem may be stated mathematically as

$$\text{Minimize } f(X),\ X = (x_1, x_2, \ldots x_n)^T \tag{4.1}$$

$$\text{subject to } C_i(X) = 0, \quad i = 1, 2, \ldots m'$$

$$C_i(X) > 0, \quad i = m' + 1, m' + 2, \ldots m$$

where

f(X) is the objective function
X is the column vector of the n independent variables
$C_i(X)$ is the set of constraints

Constraint equations of the form $C_i(X) = 0$ are termed equality constraints and those of the form $C_i(X) > 0$ are inequality constraints. Taken together, f(X) and $C_i(X)$ are known as the problem functions (Tables 4.1 and 4.2).

The strict definition of the global optimum X^* of f(X) is that

$$f(X^*) < f(Y) \forall Y \in V(X), \quad Y \neq X^* \tag{4.2}$$

where V(X) is the set of feasible values of the control variables X. Obviously for an unconstrained problem V(X) is infinitely large.

Table 4.1 Optimization Problem Classifications

Characteristics	Property	Classification
No. of decision variables	One	Univariate
	More than one	Multivariate
Types of decision variables	Continuous real numbers	Continuous
	Integers	Discrete
	Both continuous real numbers and integers	Mixed integer
	Integers in permutation	Combinatorial
Objective functions	Linear functions of decision variables	Linear
	Quadratic functions of decision variables	Quadratic
	Other nonlinear functions of decision variables	Nonlinear
Problem formulation	Subject to constraints	Constrained
	Not subject to constraints	Unconstrained
Decision variable realization within the optimization model	Exact	Deterministic
	Subject to random variation	Stochastic
	Subject to fuzzy uncertainty	Fuzzy
	Subject to both random variation and fuzzy uncertainty	Fuzzy–stochastic

Table 4.2 Typical Applications

Field	Problem	Classification
Nuclear engineering	In-core nuclear fuel management	Nonlinear Constrained Multivariate Combinatorial
Computational chemistry	Energy minimization for 3D structure prediction	Nonlinear Unconstrained Multivariate Continuous
Computational chemistry and biology	Distance geometry	Nonlinear Constrained Multivariate Continuous

A point Y* is a strong local minimum of f(X) if

$$f(Y^*) < f(Y) \forall Y \in N(Y^*, \mathrm{h}) Y \neq Y^* \tag{4.3}$$

where (Y^*, η) is defined as the set of feasible points contained in the neighborhood of Y, that is, within some arbitrary small distance of Y. For Y* to be a weak local minimum, only an equality needs to be satisfied:

$$f(Y^*) \leq f(Y) \forall Y \in N(Y^*, \mathrm{h}) Y \neq Y^* \tag{4.4}$$

If f(X) is a smooth function with continuous first and second derivatives for all feasible X, then a point X* is a stationary point of f(X) if

$$\mathrm{g}(\mathrm{X}^*) = 0 \tag{4.5}$$

where g(X) is the gradient of $\hat{f}(X)$. This first derivative vector $\hat{f}(X)$ has components given by

$$g_i(X) = \frac{\partial f(X)}{\partial x_i} \tag{4.6}$$

The point X is also a strong local minimum of f(X) if the Hessian matrix H(X), the symmetric matrix of second derivatives with components

$$H_{ij}(X) = \frac{\partial^2 f(X)}{\partial x_i \partial x_j} \tag{4.7}$$

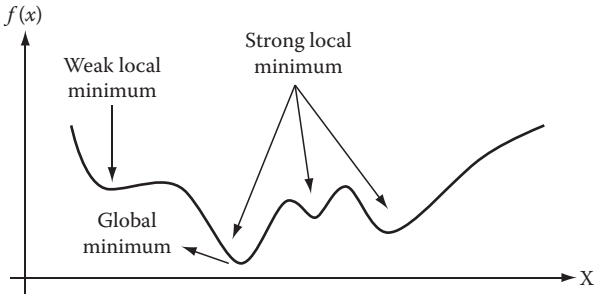


Figure 4.1 Types of minima for unconstrained optimization problems.

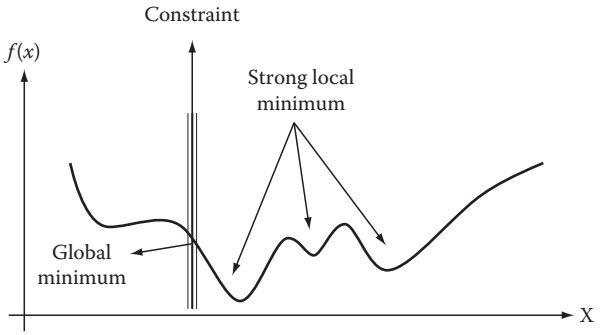


Figure 4.2 Types of minima for constrained optimization problems.

is positive-definite at X*, that is, if for a vector **u**,

$$\mathbf{u}^T H(X^*)u > 0, \quad \forall u \neq 0 \tag{4.8}$$

This condition is a generalization of convexity, or positive curvature to higher dimensions. Figures 4.1 and 4.2 illustrate the different types of stationary points for unconstrained and constrained univariate functions.

As shown in Figure 4.3, the situation is slightly more complex for constrained optimization problems. The presence of a constraint boundary, in Figure 4.4, in the form of a simple bound on the permitted values of the control variable can cause the global minimum to be an extreme value, an *extremum* (i.e., an end point), rather than a true stationary point. Some methods of treating constraints transform the optimization problem into an equivalent unconstrained one, with a different objective function.

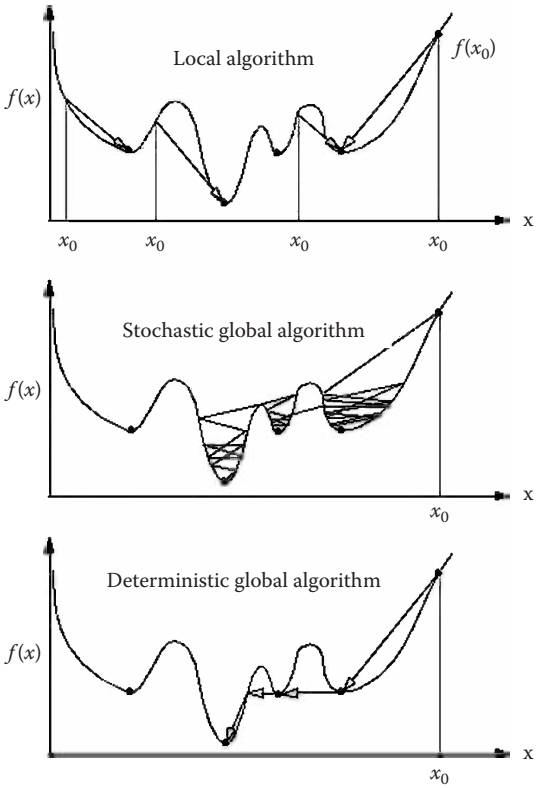


Figure 4.3 Types of structure of local and global minimization algorithms.

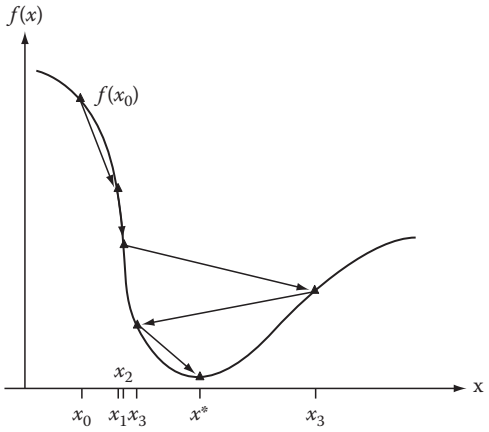


Figure 4.4 The descent structure of local minimization algorithms.

Basic structure of local methods

A starting point is chosen; a direction of movement is prescribed according to some algorithm, and a line search or trust region is performed to determine an appropriate next step. The process is repeated at the new point and the algorithm continues until a local minimum is found. Schematically, a model local minimizer method can be described as follows:

Algorithm 4.1 Basic Local Optimizer
Supply an initial guess x_0

For k = 0, 1, 2, … until convergence,

1. Test x_k for convergence
2. Calculate a search direction p_k
3. Determine an approximate step length λ_k (or modified step s_k)
4. Set x_{k+1} to $x_k + \lambda_k p_k$ (or $x_k + s_k$)

Descent directions

It is reasonable to choose a search vector p that will be a descent direction; that is, a direction leading to function value reduction. A descent direction P is defined as one along which the directional derivative is negative:

$$g(X)^T p < 0 \tag{4.9}$$

when we write the approximation

$$f(X + \mathtt{l}p) \approx f(X) + \mathtt{l} g(X)^T p \tag{4.10}$$

we see that the negativity of the right-hand side guarantees that a lower function value can be found along p for sufficiently small λ.

Steepest descent

Steepest descent (SD) is one of the oldest and simplest methods. At each iteration of SD, the search direction is taken as $-g_k$, the negative gradient of the objective function at x_k. Recall that a descent direction p_k satisfies $g_k{}^T p_k < 0$. The simplest way to guarantee the negativity of this inner product is to choose $p_k = -g_k$. This choice also minimizes the inner product $-g_k{}^T p$ for unit-length vectors and, thus, gives rise to the name *steepest descent*. SD is simple to implement and requires modest storage, O(*n*). However, progress toward a minimum may be very slow, especially near a solution.

Conjugate gradient

The first iteration in conjugate gradient (CG) is the same as in SD, but successive directions are constructed so that they form a set of mutually conjugate vectors with respect to the (positive-definite) Hessian **A** of a general convex quadratic function $qA(X)$.

Algorithm 4.2
CG method to solve AX = –b

1. Set $r_0 = -(Ax_0 + b)$, $d_0 = r_0$
2. For k = 0, 1, 2, … until r is sufficiently small, compute

$$\mathtt{l}_k = r_k{}^T r_k / d_k{}^T A d_k$$

$$x_{k+1} = x_k + \mathtt{l}_k d_k$$

$$r_{k+1} = r_k - \mathtt{l}_k d_k$$

$$\mathtt{b}_k = r_{k+1}{}^T r_{k+1} / r_k{}^T r_k$$

$$d_{k+1} = r_{k+1} + \mathtt{b}_k d_k$$

Newton methods

All Newton methods are based on approximating the objective function locally by a quadratic model and then minimizing that function approximately. The quadratic model of the objective function f at x_k along p is given by the expansion

$$f(x_k + p) \approx f(x_k) + g_k{}^T p + \frac{1}{2} p^T H_k p \tag{4.11}$$

The minimization of the right-hand side is achieved when p is the minimum of the quadratic function:

$$qH_k(p) = g_k{}^T p + \frac{1}{2} p^T H_k p \tag{4.12}$$

Alternatively, such a *Newton direction* P satisfies the linear system of n simultaneous equations, known as the *Newton equation*:

$$H_k p = -g_k \tag{4.13}$$

In the "classic" Newton method, the Newton direction is used to update each previous iterate by the formula $x_{k+1} = x_k + p_k$, until convergence. For the 1D version of Newton's method for solving a nonlinear equation f(X) = 0:

$$x_{k+1} = \frac{x_k - f(x_k)}{f'(x_k)} \tag{4.14}$$

The analogous iteration process for minimizing f(X) is

$$x_{k+1} = \frac{x_k - f'(x_k)}{f''(x_k)} \tag{4.15}$$

Newton variants are constructed by combining various strategies for the aforementioned individual components.

Algorithm 4.3 Modified Newton
For k = 0, 1, 2,…, until convergence, given x_0

1. Test x_k for convergence
2. Compute a descent direction p_k so that $\|H_k p_k + g_k\| \leq \eta_k \|g_k\|$, where η_k controls the accuracy of the solution and some symmetric matrix $\bar{H}_k$ may represent H_k.
3. Compute a step length λ so that for $x_{k+1} = x_k + \lambda p_k$,

$$f(x_{k+1}) \leq f(x_k) + \mathtt{a}\mathtt{l} g_k{}^T p$$

$$\left| g_{k+1}{}^T p_k \right| \leq \mathtt{b} \left| g_k{}^T p_k \right|,$$

   with $0 < \alpha < \beta < 1$.
4. Set $x_{k+1} = x_k + \lambda p_k$.

Newton variants are constructed by combining various strategies for the individual components.

Stochastic central problems

An object that changes randomly both in time and space is said to be ***stochastic***. A basic characteristic of applied optimization problems treated in engineering is the fact that the data for these problems, for example,

parameters of the material (yield stress, allowable stresses, moment capacities, specific gravity, etc.), external loadings, manufacturing errors, cost factors, etc., are not known at the planning stage but have to be considered to be random variables with a certain probability distribution; in addition, there is always some uncertainty in the mathematical modeling of practical problems. Typical problems of this type are

- The limit (collapse) load analysis and the elastic or plastic design of mechanical structures represented mathematically by means of
  a. The equilibrium equation
  b. Hooke's law
  c. The member displacement equation
- Optimal trajectory planning of robots by off-line programming such that the control strategy based on the optimal open-loop control, causes only low online correction expenses. Here, the underlying mechanical system is described by the kinematics and the dynamic equation, and the optimal velocity profile and configuration variables are determined for fixed model parameters by a certain variation problem.

Since the (online) correction of a decision, for example, the decision on the design of a mechanical structure or on the selection of a velocity profile, after the realization/observation of the random data might be highly expensive and time-consuming, the already known prior and statistical information about the underlying probability mechanism generating the random data should be taken into account already in the planning phase. Hence, taking into account stochastic parameter variations already in the planning phase, for off-line programming, that is, applying stochastic programming instead of ordinary mathematical programming methods, the original optimization problem with random data is replaced using appropriate decision criteria by a deterministic substitute problem, for example,

- Using a chance constrained programming approach, the objective function is replaced by its mean value, and the random constraints are replaced by chance constraints.
- Evaluating the violation of the random constraints by means of penalty functions, a weighted sum of the expectation of the primary objective function and the total expected penalty costs are minimized subject to the remaining deterministic constraints, for example, box constraints or the mean value of the objective function is minimized subject to constraints involving upper bounds for the expected penalty cost arising from violations of the original constraints.

A main problem in the solution of these problems is the numerical computation and differentiation of risk functions.

Stochastic approximation

While some nonclassical optimization techniques are able to optimize on discontinuous objective functions, they are unable to do so when complexity of the data becomes very large. In this case, the complexity of the system requires that the objective function be estimated. Furthermore, the models that are used to estimate the objective function may be stochastic due to the dynamic and random nature of the system and processes.

The basic idea behind the stochastic approximation method is the gradient descent method. Here the decision variable is varied in small increments and the impact of this variation (measured by the gradient) is used to determine the direction of the next step. The magnitude of the step is controlled to have larger steps when the perturbations in the system are small and vice versa. Stochastic approximation algorithms based on various techniques have been developed recently. They have been applied to both continuous and discrete objective functions.

General stochastic control problem

The control of a random, dynamic system in some optimal fashion using imperfect measurement data is the general problem. It also constitutes a problem about which it is very difficult to obtain any meaningful insights. Although feedback is used in order to compensate for unmodeled errors and inputs, most controllers are designed and analyzed in a deterministic context. The control inputs for the system generally must be based on imperfect observations of some of the variables which describe the system. The control policy that is utilized must be based on a priori knowledge of the system characteristics, on the time history of the input variables.

The mathematical model of the system is described by a nonlinear difference equation

$$x_{k+1} = f(x_k, H_k) + w_k, \quad k = 0, 1, \ldots, N \tag{4.16}$$

The noise w has been assumed to be additive primarily for reasons of convenience. The state x is n-dimensional and the input u is p-dimensional. In general, a probabilistic model for the initial state x_o and for the plant W_k is assumed to be known except for some unknown parameters. With rare

exception these variables are regarded as having a Gaussian distribution such that

$$E[x_0] = U_0, \quad E[w_k] = 0, \quad \forall \mathrm{k} \tag{4.17}$$

$$E\left[(x_0 - H_0)(x_0 - U_0)\right] = M_0, \quad E\left[x_0 \mathbf{w}_k{}^T\right] = 0, \quad \forall \mathrm{k} \tag{4.18}$$

$$E[w_k \mathbf{w}_k{}^T] = \mathrm{O}_k \mathrm{d}_k \tag{4.19}$$

Thus, the plant noise sequence is white and independent of the initial state.

The measurement system is described by a nonlinear algebraic relation to the state. The m-dimensional measurement vector is given by

$$z_k = h_k(x_k) + v_k, \quad k = 0, 1, \ldots, N \tag{4.20}$$

The noise v is considered to be additive for reasons of convenience. It is assumed to be a zero mean, white Gaussian sequence which is independent of the initial state and the plant noise sequence.

$$E[v_k] = 0, \quad \forall \mathrm{k} \tag{4.21}$$

$$E[v_k v_i{}^T] = R_k \mathrm{d}_{ki}, \quad \forall \mathrm{k} \tag{4.22}$$

$$E[v_k x_0{}^T] = 0, \quad \forall \mathrm{k} \tag{4.23}$$

$$E[v_k w_j{}^T] = 0, \quad \forall \mathrm{k}, \mathrm{j} \tag{4.24}$$

The aforementioned equations provide the mathematical description of the system. It is this part of the complete system that represents the physical system that must be controlled. The structure of the controller, of course, depends on the exact form of the system model equations f(·, ·) and h(·).

The behavior of the system is controlled through the input signal u_k, which is introduced at each sampling time t_k. The manner in which the controls are generated can be accomplished in a limitless number of ways. Certainly, the controls are constrained by the objectives that are defined for the control action and by the restrictions on the control and state variables themselves. Generally, there will be more than one control policy that satisfies the system constraints and achieves the prescribed objectives. Then it is reasonable to attempt to select the control policy from among all these admissible policies that is "best" according to some

well-defined performance measure. Optimal stochastic control theory is concerned with the determination of the best admissible control policy for the given system.

The following performance index is assumed:

$$J_0 = E\sum_{i=0}^{N-1} w_i(x_{i+1}, u_i) \tag{4.25}$$

Notice that the summation $E\sum_{i=0}^{N-1} w_i(x_{i+1}, u_i)$ is a random variable. Consequently, it is appropriate to consider its minimization; instead it is mapped into a deterministic quantity by considering its expected value.

Intelligent heuristic models

There had been continuing vast advances in optimal solution techniques for intelligent systems in the last two decades. The heuristic methods offer a very viable approach; however, the design and implementation of problem-specific heuristic can be a long and expensive process and the result is often domain dependent and not flexible enough to deal with changes, which may occur over time. Hence, considerable interest is focused on general heuristic techniques that can be applied to a variety of different combinatorial problems. This has yielded some new generation of intelligent heuristic techniques such as tabu search, simulated annealing, evolutionary algorithms such as genetic algorithms (GAs) and neural networks, etc.

Heuristics

Heuristics are the knowledge used to make good judgments, or strategies, tricks or "rules of thumb" used to simplify the solution of problems. They include "trial and error" (experience-based) knowledge and intelligent guesses/procedures for domain-specific problem solving. They are particularly suitable for ill-defined or poorly posed problems and, poor models such as when there are incomplete data. Heuristics play an important role in such strategies because of the exponential nature of most problems. They help to reduce the number of alternatives from an exponential number to polynomial number and, thereby obtain a solution in tolerable amount of time.

Intelligent systems

Intelligence is the ability to acquire, understand, and apply knowledge or the ability to exercise thought or reasons. It also embodies knowledge and feats both conscious and unconscious, which animate beings have acquired through study and experience. (Artificial) Intelligent systems are thus machines and coded programs aimed at mimicking such feat

and knowledge. Systems have been designed to perform many types of intelligent task. These can be physical systems like robots or mathematical computational systems such as scheduling systems which solve diverse tasks, systems used in planning complex strategies for military and for business, in medical diseases diagnosis and control, and so on.

Algorithm 4.4 General Search Paradigm
The general search algorithm (4.5) is of the following form:

General search

Objective is to maximize $f(x)$, $x \in U$
X, Y, Z: multiset of solutions $\subset U$
Initialize (X);
While not finish (X) do

Begin

Y: = select (X)
Z: = create (Y)
X: = merge(X, Y, Z)

End

where

X is the initial pool of one or more potential solutions to the problem. Since X may contain multiple copies of some solutions, it is more appropriately called a multiset
Y is a selection from X
Z is created from Y

When a new solution is **created** either initially or by using the operator "**create**," the function value, f(x), is applied to determine the value of the solution. X is reconstructed from the penultimate pool of X, Y, and Z by the operator "**merge**." The process is repeated until the pool, X, is deemed satisfactory.

Integrated heuristics

Modern approaches to local search have incorporated varying degrees of intelligibility. The contribution of intelligent search techniques should not be solely viewed in terms of improved performance alone as the traditional systems engineers or analysts expect (even though very much desirable). The trust of the contribution of intelligent search techniques should be in terms of **improved intelligibility, flexibility, and transparency of these emerging computational techniques**. A synergy between intelligibility and performance is normally of utmost importance in assessing the efficiency of an intelligent heuristic.

Algorithm 4.5 Tabu Search
Tabu search is one successful variant of the neighborhood search paradigm designed to avoid the problem of becoming trapped in a local optimum. The tabu search paradigm (4.5) is as follows:

Tabu search
Objective is to maximize $f(x)$, $x \in U$

X, Z: multiset of solutions $\subset U$
Tabu set of rules of type U → {true, false}
Initialize (X);
Initialize (Tabu); {very often to Φ}
While not finish (X) do

Begin

Z: = create (X, Tabu)
Tabu = update(Tabu)
X: = merge(X, Z)

End

The difficulty in tabu search is in constructing the set of rules. Considerable expertise and experimentation is required to construct the rules and to ensure its dynamic nature is correctly controlled. If the expertise is available, the resulting search can be efficient. Aspiration criteria are often included to help the tabu search in not being too restrictive. These criteria are rules, which say that certain moves are to be preferred over others. Some form of expert rules may also serve as tabu search rules. Each rule may have an associated weight, negative if tabu and positive if an aspiration. The combined set of rules thus associates a weight to each neighbor. A large positive weight suggests it is a desirable move while a large negative weight suggests it can be discounted. Tabu search has found applications to real-world problems such as packing and scheduling problems (flow shop problems, employee scheduling problem, machine scheduling, etc.); traveling salesman; vehicle routing, and telecommunications.

Algorithm 4.6 Simulated Annealing
SA exploits an analogy between the way in which a metal cools and freezes into a minimum energy crystalline structure (the annealing process) and the search for a minimum in a more general system.

SA (Figure 4.5) is essentially a local search technique in which a move to an inferior solution is allowed with a probability that decreases as the process progresses, according to some Boltzmann-type distribution. The inspiration for SA approach is the law of thermodynamics

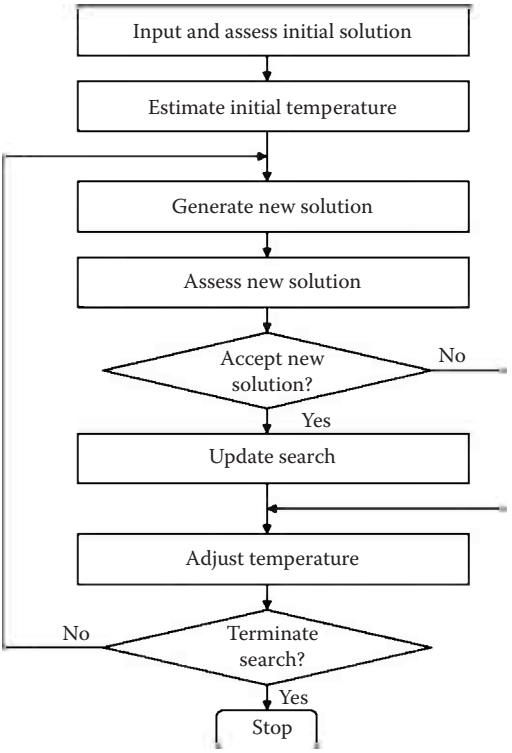


Figure 4.5 The structure of the simulated annealing algorithm.

which states that at temperature t, the probability of an increase in energy of magnitude, δE, is given by

$$P(\mathrm{d}E) = \exp\left(\frac{-\mathrm{d}E}{kt}\right) \tag{4.26}$$

where k is the physical constant known as the Boltzmann constant. The equation is applicable to a system that is cooling until it converges to a "frozen" state. The system is perturbed from current state and the resulting energy change δE is calculated. If the energy has decreased, the system moves to the new state, otherwise the new state is only accepted with the probability given earlier. The cycle can be repeated for a number of iterations at each temperature and subsequently reduced and the number of cycles repeated for the new lower temperature. This whole process is repeated until the system freezes to its steady state.

We can associate the potential solutions of an optimization problem with the system states; the cost of a solution corresponds to the concept of energy and moving to any neighbor corresponds to a change of state. A simple version of the SA paradigm (4.6) is of the following form:

Simulated annealing

```
X = x0 where x0 is some initial solution
temp = temp0 where temp0 is some initial temperature
        while not finish f(X) do
                For i = 1 to n
                Begin
                Randomly select x′, a neighbor of x
                Improvement = f(x′) − f(x)
                        If improvement > 0 then x = x′
                        Else
                                Begin
                                Generate x ∈ Q[0,1];
                                If x < exp(-improvement/temp) then x = x′
                        End {else}
                End {for}
                t = reduced (t)
        End {while}
```

A myriad of practical applications of SA include graph coloring, packing and scheduling problems, traveling salesman, vehicle routing, and quadratic assignment problems.

Genetic algorithms

GAs are search techniques based on an abstracted model of Darwinian evolution. Fixed-length strings represent solutions over some alphabet. Each such string is thought of as a "chromosome." The value of the solution then represents "fitness" of the chromosome. The concept of "survival of the fittest" is then used to allow better solutions to combine to produce offspring. The GA paradigm follows closely the same search concept exploited in tabu search and SA. The paradigm is as follows:

Algorithm 4.7 Genetic Algorithm
Objective is to maximize $f(x)$, $x \in U$

```
X, Y, Z: multiset of solutions ⊂ U
Initialize (X);
While not finish (X) do

Begin

Y: = select (X)
Z: = create (Y)
X: = merge(X, Y, Z)

End
```

Here the operators **select, create,** and **merge** correspond to the operators select, reproduction, crossover. The iteration loop of a basic GA hence looks like

Procedure GA

Begin

Generate initial population, P(0); t = 0;
Evaluate chromosomes in P(0);

Repeat

t = t + 1′
Select P(t) from P(t – 1);
Recombine chromosomes in P(t) using genetic operators;
Evaluate chromosomes in P(t);

Until termination condition is satisfied;

End

In natural evolution each species searches for beneficial adaptations in an ever-changing environment. As species evolve these new attributes are encoded in the chromosomes of individual members. This information does change by random mutation, but the real driving force behind evolutionary development is the combination and exchange of chromosomal material during breeding.

GAs differ from traditional optimization algorithms in four important respects:

- They work using an encoding of the control variables rather than the variables themselves.
- They search from one population to another rather than from individual to individual.
- They use only objective function information, not derivatives.
- They use probabilistic, not deterministic transition rules.

Genetic algorithm operators

GAs (see Figure 4.6) are rather known as evolutionary rather than genetic algorithms.

Mutation: Bit-wise change in strings at randomly selected points.

Examples

Crossover: This is a generic operator applied to two randomly selected parent solutions in the mating pool to generate two offspring solutions. Crossover operation is not performed on all pairs of parent solutions selected from the mating pool. Crossover is performed according

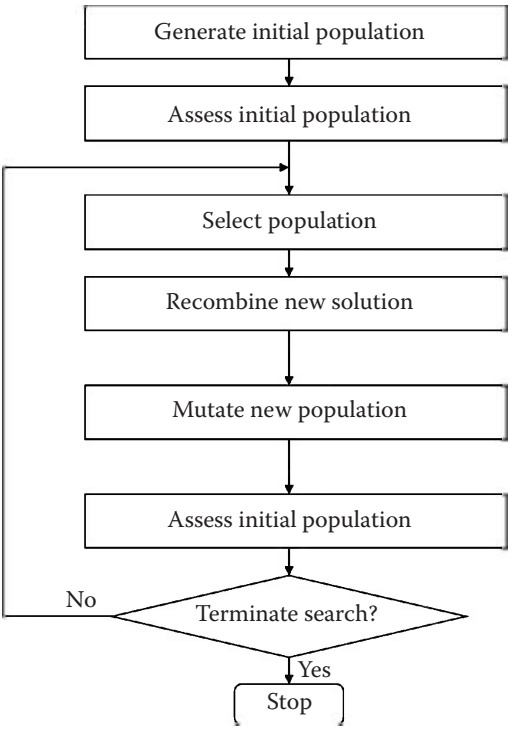


Figure 4.6 The basic structure of a genetic algorithm.

to a probability, p (usually p = 0.6). If GA decides by this probability not to perform crossover on a pair of parent solutions they are simply copied to the new population. If crossover is to take place, then one or two random splicing points are chosen in a string. The two strings are spliced and the spliced regions are mixed together to create two (potentially) new strings. These child strings are then placed in the new generation.

For example, using strings 10010111 and 00101010, suppose the GA chooses at random to perform crossover at point 5:

1 0 0 1 0 | 1 1 1 —*crossover*→ 1 0 0 1 0 0 1 0

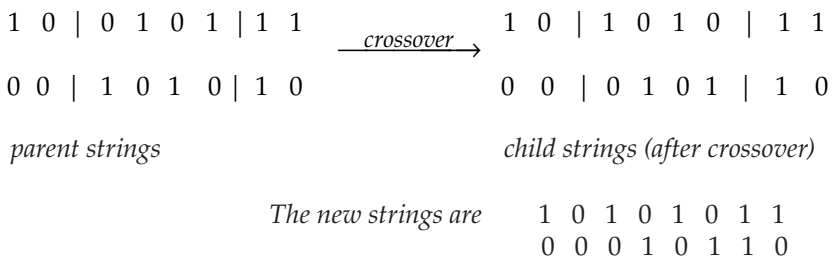
0 0 1 0 1 | 0 1 0 0 0 1 0 1 1 1 1

parent strings *child strings (after crossover)*

The new strings are 1 0 0 1 0 0 1 0

0 0 1 0 1 1 1 1

For two crossover points at points 2 and 6, the crossing looks like



Mutation: Selection and crossover alone can obviously generate a staggering amount of differing strings. However, depending on the initial population chosen, there may not be enough variety of strings to ensure GA sees the entire problem space in the event of which, GA may find itself converging on strings that are not close to the optimum it seeks due to a bad initial population. Some of these problems may be overcome by introducing a mutation operator into the GA. The GA has a mutation probability, m, which dictates the frequency at which mutation occurs. Mutation can be performed either during selection or crossover (though crossover is more usual). For each string GA checks if it should perform a mutation. If it should, it randomly changes the element value to a new one. In a binary string, 1s are changed to 01s and 0s to 1s. For example, given the string 10101111, if the GA determines to mutate this string at point 3 the 1 in that position is changed to 0, that is,

10101111 $\underrightarrow{\text{mutate}}$ **10001111**

The mutation probability is kept as low as possible (usually about 0.01) as high mutation rate will destroy fit strings and degenerate the GA into a random walk, with all the associated problems. GA has been applied successfully to a wide variety of systems. We can only highlight a small subset of the myriad of applications including packing and scheduling, design of engineering systems, robots, and transport systems.

For a number of reasons, MetaHeuristic optimization techniques differ from classical search and optimization methods.

1. MetaHeuristic optimization techniques work with a coding of decision variables, a matter which is very uncommon in classical methods. Coding discretizes the search spaces and allows MetaHeuristic optimization techniques to be applied to both discrete and discontinuous problems. MetaHeuristic optimization techniques also exploit coding similarities to make faster and parallel searches.

2. Since no gradient information is used in MetaHeuristic optimization techniques, they can also be applied to non-differentiable functions. This makes MetaHeuristic optimization techniques robust in the sense that they can be applied to a wide variety of problems.
3. Unlike many classical optimization methods, MetaHeuristic optimization techniques like GA work with a population of points. This increases the possibility of obtaining the global optimal solution even in ill-behaved problems.
4. MetaHeuristic optimization techniques use probabilistic transition rules, instead of fixed rules. For example, in early GA iterations, this randomness in GA operators makes the search unbiased toward any particular region in the search space and has an effect not making a hasty wrong decision. The use of stochastic transition rules also increases the chance of recovering from a mistake.
5. MetaHeuristic optimization techniques uses only payoff information to guide them through the problem space. Many search techniques need a variety of information to guide them. Hill climbing methods require derivatives, for example. The only information a MetaHeuristic needs is some measure of fitness about a point in the space (sometimes known as objective function value). Once the MetaHeuristic knows the current measure of "goodness" about a point, it can use this to continue searching for the optimum.
6. MetaHeuristic optimization techniques allows procedure-based function declaration. Most classical search methods do not permit such declarations. Thus, where procedures of optimization need to be declared MetaHeuristic optimization techniques proves a better optimization tool.

The coding of decision variables in MetaHeuristic optimization techniques also make it proficient in solving optimization problems involving equality and inequality constraints.

Applications of heuristics to intelligent systems

Some of these techniques have been successfully applied to model and simulate some intelligent systems. These include application of GAs to bimodal transport scheduling problem (Ibidapo-Obe and Ogunwolu, 2001). In this application, GAs are applied using a bi-level programming approach to obtain transit schedules for a bimodal transfer station in an urban transit network with multiple dispatching stations. The arrival rates of passengers are captured as fuzzy numbers in order to confer

intelligibility in the system. The model results in comparably better schedules with better levels of service (LOS) in the transit network (in terms of total waiting and transfer times for passengers) than is obtained otherwise.

Another application is on an intelligent path planner for autonomous mobile robots (Asaolu, 2002). For this robotic motion planner, a real-time optimal path planner was developed for autonomous mobile robots navigation. MetaHeuristic optimization techniques were used to guide the robot within a workspace with both static and roving obstacles. With the introduction of arbitrarily moving obstacles, the dynamic obstacle avoidance problem was recast into a dynamic graph search. The instantaneous graph is made of the connected edges of the rectangles boxing the ellipses swept out by the robot, goal, and moving obstacles. This was easily transversed in an optimal fashion.

We are presently also looking at the problems of machine translation as another application of intelligent MetaHeuristic optimization techniques.

High-performance optimization programming

Future developments in the field of optimization will undoubtedly be influenced by recent interest and rapid developments in new technologies—powerful vector and parallel machines. Indeed, their exploitation for algorithm design and solution of "grand challenge" applications is expected to bring new advances in many fields, such as computational chemistry and computational fluid dynamics. Supercomputers can provide speedup over traditional architectures by optimizing both scalar and vector computations. This can be accomplished by pipelining data as well as offering special hardware instructions for calculating intrinsic functions (e.g., $(\exp(x), \sqrt{x})$, arithmetic, and array operations. In addition, parallel computers can execute several operations concurrently. Communication among processors is crucial for efficient algorithm design so that the full parallel apparatus is exploited. These issues will only increase in significance as massively parallel networks enter into regular use. In general, one of the first steps in optimizing codes for these architectures is implementation of standard basic linear algebra subroutines (BLAS). These routines—continuously being improved, expanded, and adapted optimally to more machines—perform operations such as dot products ($x^T y$) and vector manipulations ($ax + y$), as well as matrix/vector and matrix/matrix operations.

Specific strategies for optimization algorithms have been quite recent and are not yet unified. For parallel computers, natural improvements may involve the following ideas:

1. Performing multiple minimization procedures concurrently from different starting points
2. Evaluating function and derivatives concurrently at different points (e.g., for a finite-difference approximation of gradient or Hessian or for an improved line search)
3. Performing matrix operations or decompositions in parallel for special structured systems (e.g., Cholesky factorizations of block-band preconditioned)

With increased computer storage and speed, the feasible methods for solution of very large (e.g., $O(10^5)$ or more variable) nonlinear optimization problems arising in important applications (macromolecular structure, meteorology, economics) will undoubtedly expand considerably and make possible solution of larger and far more complex problems in all fields of science and engineering.

References

Asaolu, O. S., An intelligent path planner for autonomous mobile robots, PhD thesis in Engineering Analysis, University of Lagos, Lagos, Nigeria, 2002.

Cadenas, M. and Jimenez, F., Genetic algorithm for the multi-objective solid transportation problem: A fuzzy approach. In *International Symposium on Automotive Technology and Automation, Proceedings for the Dedicated Conferences on Mechatronics and Supercomputing Applications in the Transportation Industries*, Aachen, Germany, 1994, pp. 327–334.

Darwin, C., *On the Origin of Species*. 1st edn. (facsimile—1964), Harvard University Press, Cambridge, MA, 1964.

Davis, L., *Handbook of Genetic Algorithms*. Van Nostrand Reinhold, New York, 1991.

Gill, P. E., Murray, W., and Wright, M. H., *Practical Optimization*. Academic Press, New York, 1983.

Glover, F. and Macrnillan, C., The general employee scheduling problem: An integration of management science and artificial intelligence. *Computers & Operation Research*, 15, 563–593, 1986.

Gray, P., Hart, W., Painton, L., Phillips, C., Trahan, M., and Wagner, J., A survey of global optimization methods, Sandia National Laboratories Albuquerque, NM 87185, http://www.cs.sandia.gov/opt/survey/

Haykin, S., *Neural Networks: A Comprehensive Foundation*. Prentice Hall International Inc., Upper Saddle River, NJ, 2nd, 1998, ISBN: 0-13-908385-5.

Ibidapo-Obe, O. and Ogunwolu, P. O., An optimal scheduling of a bi-modal urban transit system using genetic algorithms, Unpublished research work at the Department of Systems Engineering, University of Lagos, Lagos, Nigeria, 2001.

Ingber, L. A., Simulated annealing: Practice versus theory, *Journal of Mathematical Computer Modelling*, 18 (11), 29–57, 1993.

Luenberger, D. G., *Linear and Nonlinear Programming*. 2nd edn., Addison-Wesley, Reading, MA, 1984.

Marti, K., Stochastic optimization methods in engineering. In J. Dolezal and J. Fiedler (eds.), *System Modeling and Optimization*, Chapman and Hall, London/New York, 1996.

Powell, M. J. D., A Monte Carlo method for finding the minimum of a function of several variables without calculating derivatives, *Computer Journal*, 7, 155–162, 1964.

Rayward-Smith, V. J., *Applications of Modern Heuristic Methods*. Alfred Walter Limited Publishers in association with UNICOM, London, U.K., 1995, pp. 145–156.

Siarry, P., Berthiau, G., Durdin, F., and Haussy, J., Enhanced simulated annealing for globally minimizing functions of many-continuous variables, *ACM Transactions on Mathematical Software*, 23 (2), 209–228, June 1997.

Wasan, M. T., *Stochastic Approximation*. Cambridge University Press, Cambridge, U.K., 1969.

chapter five

Statistical control techniques

Statistical process control

Statistical process control (SPC) means controlling a process statistically. It is the use of statistical techniques to analyze a process in order to monitor, control, and improve it. The objective is to have a stable, consistent process that produces the fewest defects possible. SPC originated from the efforts of the early quality control researchers. The techniques of SPC are based on basic statistical concepts normally used for statistical quality control. In a manufacturing environment, it is known that not all products are made exactly alike. There are always some inherent variations in units of the same product. The variation in the characteristics of a product provides the basis for using SPC for quality improvement; with the help of statistical approaches, individual items can be studied and general inferences can be drawn about the process or batches of products from the process. Since 100% inspection is difficult or impractical in many processes, SPC provides a mechanism to generalize concerning process performance. SPC uses samples generated consecutively over time. The samples should be representative of the general process. SPC can be accomplished through the following steps:

- Control charts ($\bar{X}$-chart, R-chart)
- Process capability analysis (nested design, Cp, Cp_k)
- Process control (factorial design, response surface)

Control charts

Two of the most commonly used control charts in industry are the X-bar charts and the range charts (R-charts). The type of chart to be used normally depends on the kind of data collected. Data collected can be of two types: variable data and attribute data. The success of quality improvement depends on two major factors:

1. The quality of data available
2. The effectiveness of the techniques used for analyzing the data

Types of data for control charts

Variable data

The control charts for variable data are listed as follows:

- Control charts for individual data elements (X)
- Moving-range chart (MR-chart)
- Average chart ($\bar{X}$-chart)
- Range chart (R-chart)
- Median chart
- Standard deviation chart (σ-chart)
- Cumulative sum chart (CUSUM)
- Exponentially weighted moving average (EWMA)

Attribute data

The control charts for attribute data are listed as follows:

- Proportion or fraction defective chart (p-chart) (subgroup sample size can vary)
- Percent defective chart (100p-chart) (subgroup sample size can vary)
- Number defective chart (np-chart) (subgroup sample size is constant)
- Number defective (c-chart) (subgroup sample size = 1)
- Defective per inspection unit (u-chart) (subgroup sample size can vary)

The statistical theory useful to generate control limits is the same for all aforementioned charts with the exception of EWMA and CUSUM.

X-bar and range charts

The R-chart is a time plot useful in monitoring short-term process variations, while the X-bar chart monitors the longer-term variations where the likelihood of special causes is greater over time. Both charts have control lines called upper and lower control limits, as well as the central lines. The central line and control limits are calculated from the process measurements. They are not specification limits or a percentage of the specifications, or some other arbitrary lines based on experience. Therefore, they represent what the process is capable of doing when only common cause variation exists. If only common cause variation exists, then the data will continue to fall in a random fashion within the control limits. In this case, we say the process is in a state of statistical control. However, if a special cause acts on the process, one or more data points will be outside the control limits, so the process is not in a state of statistical control.

Data collection strategies

One strategy for data collection requires that about 20–25 subgroups be collected, which should adequately show the location and spread of a distribution in a state of statistical control. If it happens that due to sampling costs, or other sampling reasons associated with the process, we are unable to have 20–25 subgroups, we can still use the available samples that we have to generate the trial control limits and update these limits as more samples are made available, because these limits will normally be wider than normal control limits and will therefore be less sensitive to changes in the process. Another approach is to use run charts to monitor the process until such time as 20–25 subgroups are made available. Then, control charts can be applied with control limits included on the charts. Other data collection strategies should consider the subgroup sample size, as well as the sampling frequency.

Subgroup sample size

The subgroup samples of size n should be taken as n consecutive readings from the process and not random samples. This is necessary in order to have an accurate estimate of the process common cause variation. Each subgroup should be selected from some small period of time or small region of space or product in order to assure homogeneous conditions within the subgroup. This is necessary because the variation within the subgroup is used in generating the control limits. The subgroup sample size n can be between four or five samples. This is a good size that balances the pros and cons of using large or small sample size for a control chart as provided in the following.

Advantages of using small subgroup sample size

- Estimates of process standard deviation based on the range are as good and accurate as the estimates obtained from using the standard deviation equation which is a complex hand calculation method.
- The probability of introducing special cause variations within a subgroup is very small.
- R-chart calculation is simple and easier to compute by hand on the shop floor by operators.

Advantages of using large subgroup sample size

- The central limit theorem supports the fact that the process average will be more normally distributed with larger sample size.
- If the process is stable, the larger the subgroup size the better the estimates of process variability.
- A control chart based on larger subgroup sample size will be more sensitive to process changes.

The choice of a proper subgroup is very critical to the usefulness of any control chart. The following paragraphs explain the importance of subgroup characteristics:

- If we fail to incorporate all common cause variations within our subgroups, the process variation will be underestimated, leading to very tight control limits. Then the process will appear to go out of control too frequently even when there is no existence of a special cause.
- If we incorporate special causes within our subgroups, then we will fail to detect special causes as frequently as expected.

Frequency of sampling

The problem of determining how frequently one should sample depends on several factors. These factors include, but are not limited to the following:

- *Cost of collecting and testing samples*: The greater the cost of taking and testing samples, the less frequently we should sample.
- *Changes in process conditions*: The larger the frequency of changes to the process, the larger the sampling frequency. For example, if process conditions tend to change every 15 min, then sample every 15 min. If conditions change every 2 h, then sample every 2 h.
- *Importance of quality characteristics*: The more important the quality characteristic being charted is to the customer, the more frequently the characteristic will need to be sampled.
- *Process control and capability*: The more history of process control and capability, the less frequently the process needs to be sampled.

Stable process

A process is said to be in a state of statistical control if the distribution of measurement data from the process has the same shape, location, and spread over time. In other words, a process is stable when the effects of all special causes have been removed from a process, so that the remaining variability is only due to common causes. Figure 5.1 shows an example of a stable distribution.

Out-of-control patterns

A process is said to be unstable (*not* in a state of statistical control) if it changes from time to time because of a shifting average, or shifting variability, or a combination of shifting averages and variation. Figures 5.2 through 5.4 show examples of distributions from unstable processes.

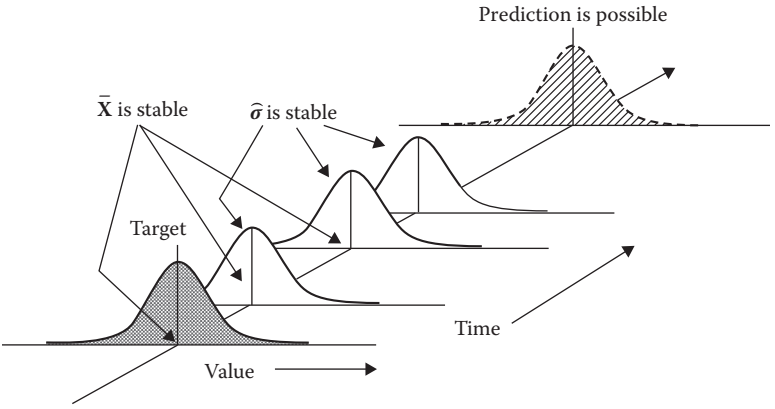


Figure 5.1 Stable distribution with no special causes.

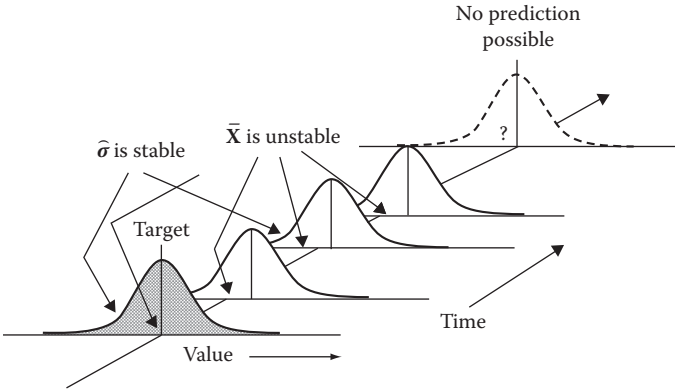


Figure 5.2 Unstable process average.

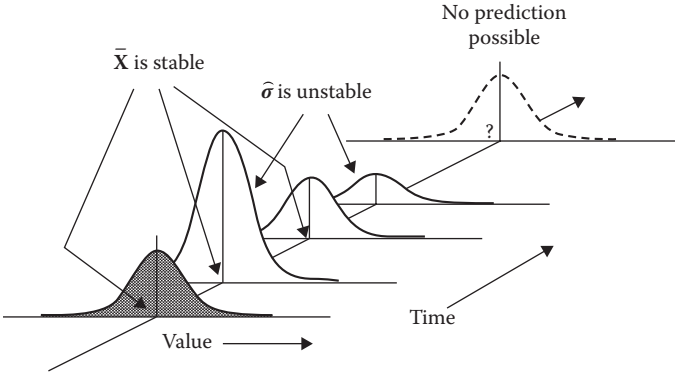


Figure 5.3 Unstable process variation.

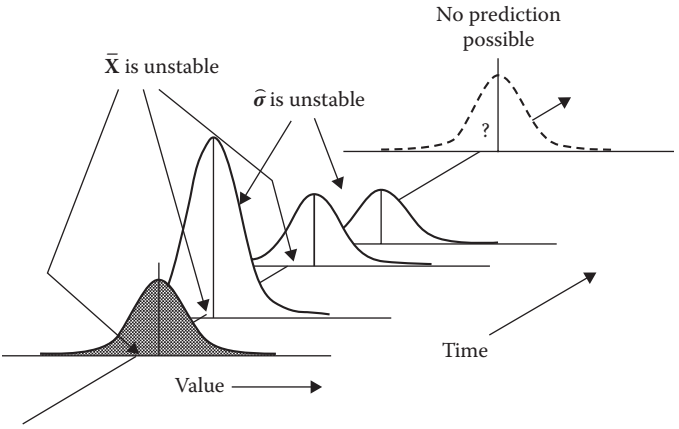


Figure 5.4 Unstable process average and variation.

Calculation of control limits

- Range (R)
  This is the difference between the highest and lowest observations:

$$R = X_{highest} - X_{lowest}$$

- Center lines
  Calculate $\bar{X}$ and $\bar{R}$

$$\bar{X} = \frac{\sum X_i}{m}$$

$$\bar{R} = \frac{\sum R_i}{m}$$

  where,
  $\bar{X}$ is the overall process average
  $\bar{R}$ is the average range
  m is the total number of subgroups
  n is within subgroup sample size

- Control limits based on R-chart

$$UCL_R = D_4\bar{R}$$

$$LCL_R = D_3\bar{R}$$

- Estimate of process variation

$$\hat{s} = \frac{\bar{R}}{d_2}$$

- Control limits based on $\bar{X}$-chart
  Calculate the upper and lower control limits for the process average:

$$UCL = \bar{X} + A_2\bar{R}$$

$$LCL = \bar{X} - A_2\bar{R}$$

  Table 5.1 shows the values of d_2, A_2, D_3, and D_4 for different values of n. Where n is within the subgroup sample size. These constants are used for developing variable control charts.

Plotting control charts for range and average charts

- Plot the R-chart first.
- If R-chart is in control, then plot X-bar chart.
- If R-chart is not in control, identify and eliminate special causes, then delete points that are due to special causes, and recompute the control limits for the range chart. If process is in control, then plot X-bar chart.
- Check to see if X-bar chart is in control, if not search for special causes and eliminate them permanently.
- Remember to perform the eight trend tests.

Table 5.1 Table of Constants for Variables Control Charts

n	d_2	A_2	D_3	D_4
2	1.128	1.880	0	3.267
3	1.693	1.023	0	2.575
4	2.059	0.729	0	2.282
5	2.326	0.577	0	2.115
6	0.534	0.483	0	2.004
7	2.704	0.419	0.076	1.924
8	2.847	0.373	0.136	1.864
9	2.970	0.337	0.184	1.816
10	3.078	0.308	0.223	1.777
11	3.173	0.285	0.256	1.744
12	3.258	0.266	0.284	1.716

Plotting control charts for moving range and individual control charts

- Plot the MR-chart first.
- If MR-chart is in control, then plot the individual chart (X).
- If MR-chart is not in control, identify and eliminate special causes, then delete special causes points, and recompute the control limits for the MR-chart. If MR-chart is in control, then plot the individual chart.
- Check to see if individual chart is in control, if not search for special causes from out-of-control points.
- Perform the eight trend tests.

Case example: Plotting of control chart

An industrial engineer in a manufacturing company was trying to study a machining process for producing a smooth surface on a torque converter clutch. The quality characteristic of interest is the surface smoothness of the clutch. The engineer then collected four clutches every hour for 30 h and recorded the smoothness measurements in micro inches. Acceptable values of smoothness lies between 0 (perfectly smooth) and 45 micro inches. The data collected by the engineer are provided in Table 5.2. Histograms of the individual and average measurements are presented in Figure 5.5.

The two histograms in Figure 5.5 show that the hourly smoothness average ranges from 27 to 32 micro inches, much narrower than the histogram of hourly individual smoothness which ranges from 24 to 37 micro inches. This is due to the fact that averages have less variability than individual measurements. Therefore, whenever we plot subgroup averages on an X-bar chart, there will always exist some individual measurements that will plot outside the control limits of an X-bar chart. The dot plots of the surface smoothness for individual and average measurements are shown in Figure 5.6.

The descriptive statistics for individual smoothness are presented in the following:

N	MEAN	MEDIAN	TRMEAN	STDEV	SEMEAN
120	29.367	29.00	29.287	2.822	0.258
MIN	MAX	Q1	Q3		
24.00	37.00	28.00	31.00		

Table 5.2 Data for Control Chart Example

Smoothness (micro inches)						
Subgroup no.	I	II	III	IV	Average	Range
1	34	33	24	28	29.75	10
2	33	33	33	29	32.00	4
3	32	31	25	28	29.00	7
4	33	28	27	36	31.00	9
5	26	34	29	29	29.50	8
6	30	31	32	28	30.25	4
7	25	30	27	29	27.75	5
8	32	28	32	29	30.25	4
9	29	29	28	28	28.50	1
10	31	31	27	29	29.50	4
11	27	36	28	29	30.00	9
12	28	27	31	31	29.25	4
13	29	31	32	29	30.25	3
14	30	31	31	34	31.50	4
15	30	33	28	31	30.50	5
16	27	28	30	29	28.50	3
17	28	30	33	26	29.25	7
18	31	32	28	26	29.25	6
19	28	28	37	27	30.00	10
20	30	29	34	26	29.75	8
21	28	32	30	24	28.50	8
22	29	28	28	29	28.50	1
23	27	35	30	30	30.50	8
24	31	27	28	29	28.75	4
25	32	36	26	35	32.25	10
26	27	31	28	29	28.75	4
27	27	29	24	28	27.00	5
28	28	25	26	28	26.75	3
29	25	25	32	27	27.25	7
30	31	25	24	28	27.00	7
Total					881.00	172

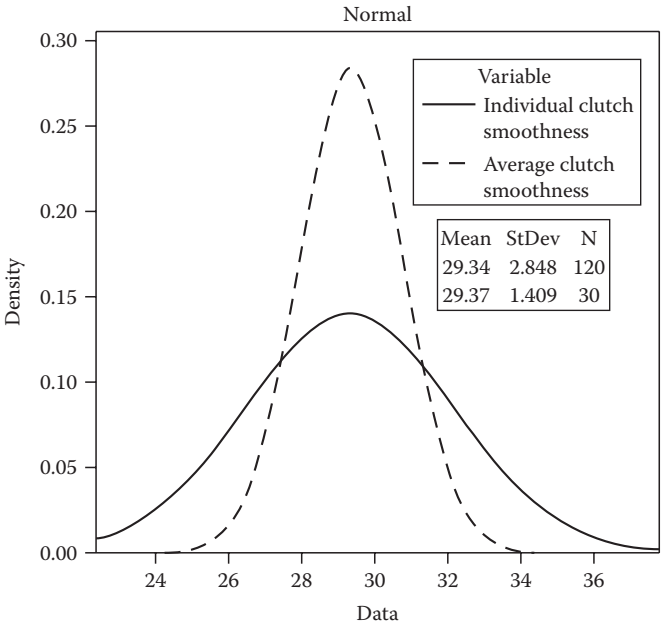


Figure 5.5 Histograms of individual measurements and averages for clutch smoothness.

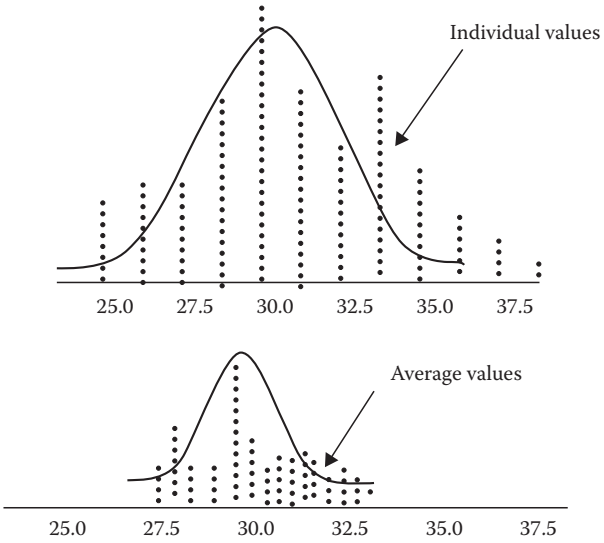


Figure 5.6 Dotplots of individual measurements and averages for clutch smoothness.

The descriptive statistics for average smoothness are presented in the following:

N	MEAN	MEDIAN	TRMEAN	STDEV	SEMEAN
30	29.367	29.375	29.246	1.409	0.257
MIN	MAX	Q1	Q3		
26.75	32.25	28.50	30.25		

Calculations

1. Natural limit of the process = $\bar{X} \pm 3s$ (based on empirical rule).
   s = estimated standard deviation of all individual samples
   Standard deviation (special and common), $s = 2.822$
   Process average, $\bar{X} = 29.367$
   Natural process limit = $29.367 \pm 3(2.822) = 29.367 \pm 8.466$
   The natural limit of the process is between 20.90 and 37.83
2. Inherent (common cause) process variability, $\hat{\sigma} = \bar{R}/d_2$
   $\bar{R}$ from the R-chart = 5.83
   d_2 (for $n = 4$) from Table 5.1 = 2.059
   $\hat{\sigma} = \bar{R}/d_2 = 5.83/2.059 = 2.83$
   Thus, the total process variation, s, is about the same as the inherent process variability. This is because the process is in control. If the process is out of control, the total standard deviation of all the numbers will be larger than $\bar{R}/d_2$.
3. Control limits for the range chart
   Obtain constants D_3, D_4 from Table 5.1 for $n = 4$.
   $D_3 = 0$
   $D_4 = 2.282$
   $\bar{R} = 172/30 = 5.73$
   UCL = $D_4^* \bar{R} = 2.282(5.73) = 16.16$
   LCL = $D_3^* \bar{R} = 0(5.73) = 0.0$
4. Control limits for the averages
   Obtain constants A_2 from Table 8.1 for $n = 4$.
   $A_2 = 0.729$
   UCL = $\bar{X} + A_2(\bar{R}) = 29.367 + 0.729(5.73) = 33.54$
   LCL = $\bar{X} - A_2(\bar{R}) = 29.367 - 0.729(5.73) = 25.19$
5. Natural limit of the process = $\bar{X} \pm 3(\bar{R})/d_2 = 29.367 \pm 3(2.83) = 29.367 \pm 8.49$
   The natural limit of the process is between 20.88 and 37.86, which is slightly different from $\pm 3s$ calculated earlier based on the empirical rule. This is due to the fact that $\bar{R}/d_2$ is used rather than the standard deviation of all the values. Again, if the process is out of control, the standard deviation of all the values will be greater than $\bar{R}/d_2$.

The correct procedure is always to use $\bar{R}/d_2$ from a process that is in statistical control.

6. Comparison with specification
   Since the specifications for the clutch surface smoothness are between 0 (perfectly smooth) and 45 micro inches, and the natural limit of the process is between 20.88 and 37.86, then the process is capable of producing within the spec limits. Figure 5.7 presents the R-chart and X-bar chart for clutch smoothness.

For this case example, the industrial engineer examined the aforementioned charts and concluded that the process is in a state of statistical control.

Process improvement opportunities

The industrial engineer realizes that if the smoothness of the clutch can be held below 15 micro inches, then the clutch performance can be

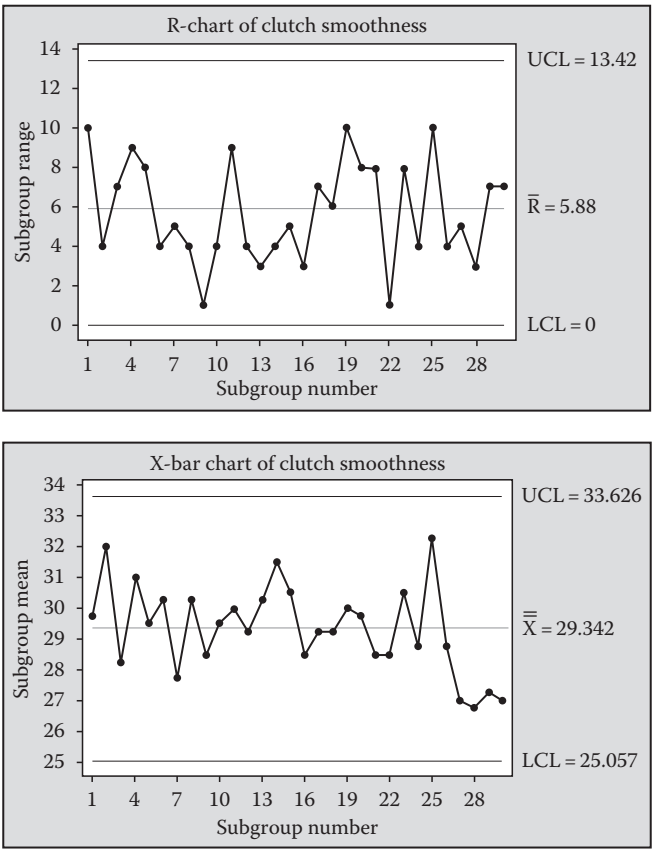


Figure 5.7 R and X-bar charts for clutch smoothness.

significantly improved. In this situation, the engineer can select key control factors to study in a two-level factorial or fractional factorial design.

Trend analysis

After a process is recognized to be out of control, zone control charting technique is a logical approach to searching for the sources of the variation problems. The following eight tests can be performed using MINITAB software or other statistical software tools. For this approach, the chart is divided into three zones. Zone A is between $\pm 3\sigma$, zone B is between $\pm 2\sigma$, and zone C is between $\pm 1\sigma$.

Test 1

Pattern: One or more points falling outside the control limits on either side of the average. This is shown in Figure 5.8.

Problem source: A sporadic change in the process due to special causes such as

- Equipment breakdown
- New operator
- Drastic change in raw material quality
- Change in method, machine, or process setting

Check: Go back and look at what might have been done differently before the out-of-control point signals.

Test 2

Pattern: A run of nine points on one side of the average (Figure 5.9).

Problem source: This may be due to a small change in the level of process average. This change may be permanent at the new level.

Check: Go back to the beginning of the run and determine what was done differently at that time or prior to that time.

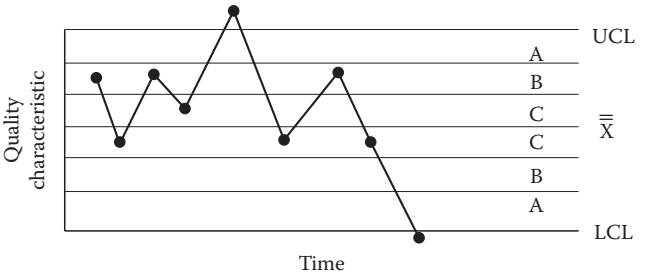


Figure 5.8 Test 1 for trend analysis.

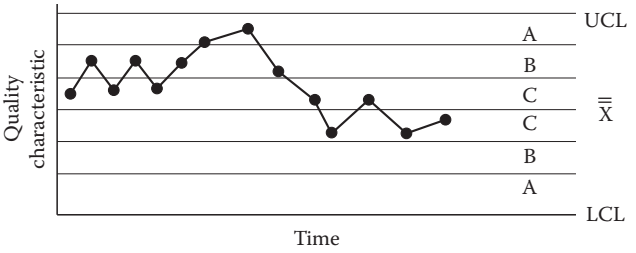


Figure 5.9 Test 2 for trend analysis.

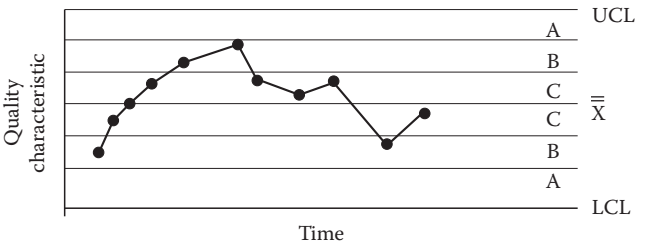


Figure 5.10 Test 3 for trend analysis.

Test 3

Pattern: A trend of six points in a row either increasing or decreasing as shown in Figure 5.10.

Problem source: This may be due to the following:

- Gradual tool wear
- Change in characteristic such as gradual deterioration in the mixing or concentration of a chemical
- Deterioration of plating or etching solution in electronics or chemical industries

Check: Go back to the beginning of the run and search for the source of the run.

The three tests mentioned earlier are useful in providing good control of a process. However, in addition to the three tests, some advanced tests for detecting out-of-control patterns can also be used. These tests are based on the zone control chart.

Test 4

Pattern: Fourteen points in a row alternating up and down within or outside the control limits as shown in Figure 5.11.

Problem source: This can be due to sampling variation from two different sources such as sampling systematically from high and low

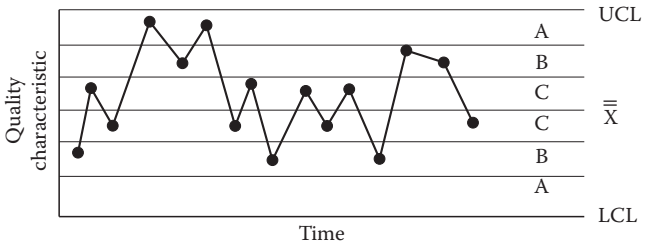


Figure 5.11 Test 4 for trend analysis.

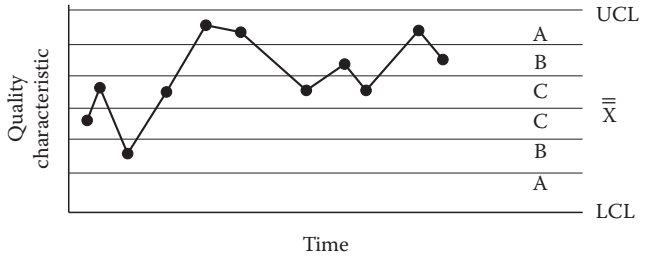


Figure 5.12 Test 5 for trend analysis.

temperatures, or lots with two different averages. This pattern can also occur if adjustment is being made all the time (over control).

Check: Look for cycles in the process, such as humidity or temperature cycles, or operator over control of process.

Test 5

Pattern: Two out of three points in a row on one side of the average in zone A or beyond. An example of this is presented in Figure 5.12.

Problem source: This can be due to a large, dramatic shift in the process level. This test sometimes provides early warning, particularly if the special cause is not as sporadic as in the case of Test 1.

Check: Go back one or more points in time and determine what might have caused the large shift in the level of the process.

Test 6

Pattern: Four out of five points in a row on one side of the average in zone B or beyond, as depicted in Figure 5.13.

Problem source: This may be due to a moderate shift in the process.

Check: Go back three or four points in time.

Test 7

Pattern: Fifteen points in a row on either side of the average in zone C as shown in Figure 5.14.

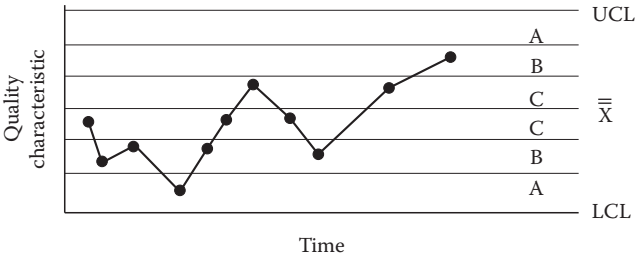


Figure 5.13 Test 6 for trend analysis.

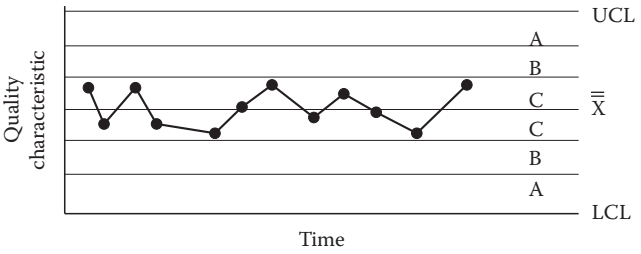


Figure 5.14 Test 7 for trend analysis.

Problem source: This is due to the following:

- Unnatural small fluctuations or absence of points near the control limits
- At first glance may appear to be a good situation, but this is not a good control
- Incorrect selection of subgroups. May be sampling from various subpopulations and combining them into a single subgroup for charting
- Incorrect calculation of control limits

Check: Look very close to the beginning of the pattern.

Test 8

Pattern: Eight points in a row on both sides of the center line with none in zone C. An example is shown in Figure 5.15.

Problem source: No sufficient resolution on the measurement system.

Check: Look at the R-chart and see if it is in control.

Process capability analysis

The capability of a process is the spread which contains almost all values of the process distribution. It is very important to note that capability is defined in terms of a distribution. Therefore, capability can only be defined for a process that is stable (has distribution) with common cause

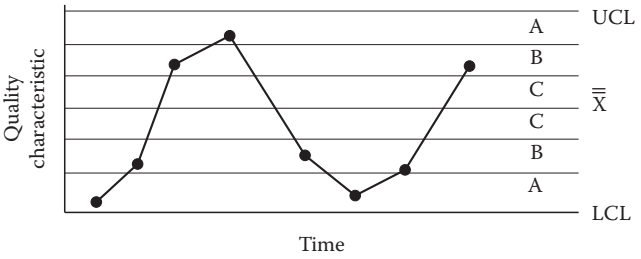


Figure 5.15 Test 8 for trend analysis.

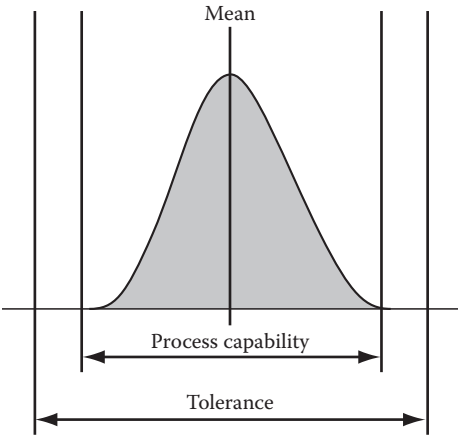


Figure 5.16 Process capability distribution.

variation (inherent variability). It cannot be defined for an out-of-control process (it has no distribution) with variation special to specific causes (total variability). Figure 5.16 shows a process capability distribution.

Capable process

A process is capable ($C_p \geq 1$) if its natural tolerance lies within the engineering tolerance or specifications. The measure of process capability of a stable process is $6\hat{\sigma}$, where $\hat{\sigma}$ is the inherent process variability estimated from the process. A minimum value of $C_p = 1.33$ is generally used for an ongoing process. This ensures a very low reject rate of 0.007% and, therefore, is an effective strategy for prevention of nonconforming items. C_p is defined mathematically as

$$C_p = \frac{USL - LSL}{6\hat{s}}$$

$$= \frac{\text{allowable process spread}}{\text{actual process spread}}$$

where

USL is the upper specification limit

LSL is the lower specification limit

C_p measures the effect of the inherent variability only. The analyst should use R-bar/d_2 to estimate $\hat{\sigma}$ from an R-chart that is in a state of statistical control, where R-bar is the average of the subgroup ranges, and d_2 can be obtained for different subgroup sizes n from Table 5.1

We do not have to verify control before performing a capability study. We can perform the study and then verify control after the study with the use of control charts. If the process is in control during the study, then our estimates of capabilities are correct and valid. However, if the process was not in control, we would have gained useful information, as well as proper insights as to the corrective actions to pursue.

Capability index

Process centering can be assessed when a two-sided specification is available. If the capability index (C_{pk}) is equal to or greater than 1.33, then the process may be adequately centered. C_{pk} can also be employed when there is only one-sided specification. For a two-sided specification, it can be mathematically defined as

$$C_{pk} = Minimum\left\{\frac{USL - \bar{\bar{X}}}{3\hat{s}}, \frac{\bar{\bar{X}} - LSL}{3\hat{s}}\right\}$$

where $\bar{\bar{X}}$ is the overall process average.

However, for a one-sided specification, the actual C_{pk} obtained is reported. This can be used to determine the percentage of observations out of specification. The overall long-term objective is to make C_p and C_{pk} as large as possible by continuously improving or reducing process variability, $\hat{\sigma}$ every time so that a greater percentage of the product is near the target value for the key quality characteristic of interest. The ideal is to center the process with zero variability.

If a process is centered but not capable, one or several courses of action may be necessary. One of the actions may be that of integrating designed experiment to gain additional knowledge on the process and in designing control strategies. If excessive variability is demonstrated, one may conduct a nested design with the objective of estimating the various sources of variability. These sources of variability can then be evaluated to determine what strategies to take in order to reduce or permanently eliminate them. Another action may be that of changing the specifications

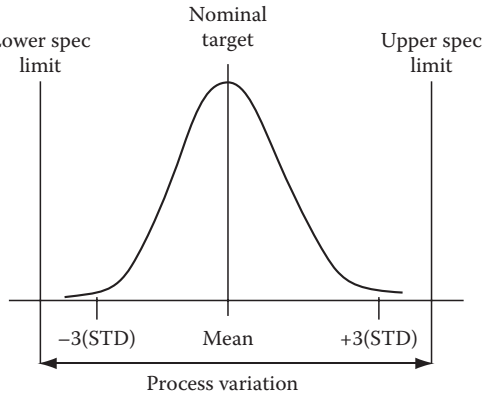


Figure 5.17 A process that is centered and capable.

or continuing production and then sorting the items. Three characteristics of a process can be observed with respect to capability:

1. The process may be centered and capable.
2. The process may be capable but not centered.
3. The process may be centered but not capable.

Figures 5.17 through 5.19 present the alternate characteristics.

Process capability example

1. Determine if the process is capable for the clutch smoothness data in Table 5.2. The engineer has determined that the process is in a state of statistical control. The specification limits are 0 (perfectly smooth)

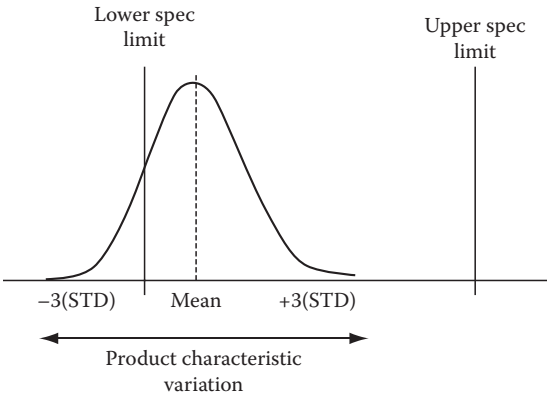


Figure 5.18 A process that is capable but not centered.

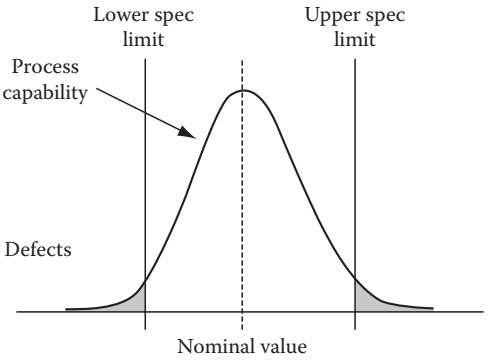


Figure 5.19 A process that is centered but not capable.

and 45 micro inches. The inherent process variability as determined from the control chart is

$$\hat{s} = \frac{\bar{R}}{d_2} = \frac{5.83}{2.059} = 2.83.$$

The capability of this process to produce within the specifications can be determined as

$$C_p = \frac{USL - LSL}{6\hat{s}} = \frac{45 - 0}{6(2.83)} = 2.650.$$

The capability of the process, $C_p = 2.65 > 1.0$, indicating that the process is capable of producing clutches that will meet the specifications of between 0 and 45. The process average is 29.367.

2. Determine if the process can be adequately centered. C_{pk} = minimum [C_l and C_u] can be used to determine if a process can be centered.

$$C_u = \frac{USL - \bar{X}}{3\hat{s}} = \frac{45 - 29.367}{3(2.83)} = 1.84$$

$$C_l = \frac{\bar{X} - LSL}{3\hat{s}} = \frac{29.367 - 0}{3(2.83)} = 3.46$$

Therefore, the C_{pk} for this process is 1.84. Since $C_{pk} = 1.84$ is greater than 1.33, the process can be adequately centered.

Applications of process capability indices:

- *Communication*: C_p and C_{pk} have been used in industry to establish a dimensionless common language useful for assessing the performance of production processes. Engineering, quality, manufacturing, etc., can communicate and understand processes with high capabilities.
- *Continuous improvement*: The indices can be used to monitor continuous improvement by observing the changes in the distribution of process capabilities. For example, if there were 20% of processes with capabilities between 1 and 1.67 in a month, and some of these improved to between 1.33 and 2.0 the next month, then this is an indication that improvement has occurred.
- *Audits*: There are so many various kinds of audits in use today to assess the performance of quality systems. A comparison of in-process capabilities with capabilities determined from audits can help establish problem areas.
- *Prioritization of improvement*: A complete printout of all processes with unacceptable C_p or C_{pk} values can be extremely powerful in establishing the priority for process improvements.
- *Prevention of nonconforming product*: For process qualification, it is reasonable to establish a benchmark capability of $C_{pk} = 1.33$ which will make nonconforming products unlikely in most cases.

Potential abuse of C_p and C_{pk}:

- *Problems and drawbacks*: C_{pk} can increase without process improvement, though repeated testing reduces test variability.

  The wider the specifications, the larger the C_p or C_{pk}, but the action does not improve the process.
- People tend to focus on number rather than on process.
- *Process control*: People tend to determine process capability before statistical control has been established. Most people are not aware that capability determination is based on process common cause variation and what can be expected in the future. The presence of special causes of variation makes prediction impossible and C_{pk} unclear.
- *Nonnormality*: Some processes result in nonnormal distribution for some characteristics. Since capability indices are very sensitive to departures from normality, data transformation may be used to achieve approximate normality.
- *Computation*: Most computer packages do not use $\bar{R}/d_2$ to calculate σ.

Time series analysis and process estimation

SPC has found widespread application in industry for monitoring and controlling manufacturing processes as well as for implementing quality and process improvement activities. The traditional Shewhart control charts have been extensively used for this purpose. The fundamental assumption in the typical application of Shewhart control charts is that the sequence of process observations is independent and uncorrelated. This assumption has been found to be reasonable in parts manufacturing industries. However, it is often not true in chemical and process industries where process data are enormous and correlated since many kinds of sensors and in-process gauges are being used with automated machines.

The presence of autocorrelation in the data can have a major impact on the expected average run lengths (ARL) performance of control charts, causing a dramatic increase in the frequency of false alarms, and the control limits of the CUSUM procedure would have to be adjusted (see Johnson and Bagshaw, 1974). In this type of situations, Alwan and Roberts (1989), Montgomery and Mastrangelo (1991), as well as other authors, have recommended fitting a time series model to track the level of the process and then using a standard control chart on the residuals to detect unusually large shocks to the process. Montgomery and Mastrangelo (1991) also presented an alternative approach based on a straightforward application of the EWMA statistics due to the practical implementation drawback of the method proposed by Alwan and Roberts (1989).

Correlated observations

Correlated observations such as those experienced by continuous process industries (petroleum, chemical, mineral processing, pulp and paper, etc.) as well as management data such as sales, profits, and so on can be monitored and controlled by fitting an appropriate time series model to the observations and then applying control charts to the stream of residuals from this model (Ermer 1979, 1980; Alwan and Roberts, 1989; Montgomery and Friedman, 1989; Yourstone and Montgomery, 1989; Montgomery and Mastrangelo, 1991). The general time series model employed is the autoregressive integrated moving average (ARIMA) model (Box and Jenkins, 1976). This model can be represented as

$$\phi_p(B)(1-B)^d Y_t = \theta(B)\, a_t$$

where

$\phi_p(B)$ is an autoregressive polynomial of order p
$\theta(B)$ is a moving average polynomial of order q
d is the dth difference of the series
B is the backward shift operator
a_t is a sequence of normally and independently distributed random "shocks" with mean zero and constant variance σ_a^2

$$\phi_p(B) = (1 - \phi_1 B - \phi_2 B^2 - \ldots\ldots\ldots\ldots - \phi_p B^p)$$

$$\theta_q(B) = (1 - \theta_1 B - \theta_2 B^2 - \ldots\ldots\ldots\ldots - \theta_q B^q)$$

An extensive application of this general ARIMA model can be found in the paper by Ayeni and Pilat (1992). If $\hat{Y}_t$ is the predicted value obtained from an appropriately identified and fitted ARIMA model, then the residuals given by

$$a_t = Y_t - \hat{Y}_t$$

will behave like independent and identically distributed random variables (Box and Jenkins, 1976). Therefore, control charts can then be applied to the set of residuals. If a shift occurs in the process average, the identified model will no longer be applicable, and this effect will be detected on the control charts applied to the residuals.

As an illustration of this approach, consider Figure 5.20, which presents 150 observations collected during a study of the moisture content of a raw material. The measurements are fiber weight scan averages. Each scan takes about 25 s.

Figure 5.20 shows a control chart for individuals with control limits based on a process standard deviation estimated from a moving range control chart. This chart shows many out-of-control signals.

Time series analysis example

MINITAB software can be used to analyze time series data. For the data in Figure 5.20, the autocorrelation and partial autocorrelation functions can be obtained as shown in Figure 5.21. The autocorrelation function of the 150 weights is shown before differencing. The inability of the autocorrelation function to die out rapidly shows that fiber weights are highly autocorrelated

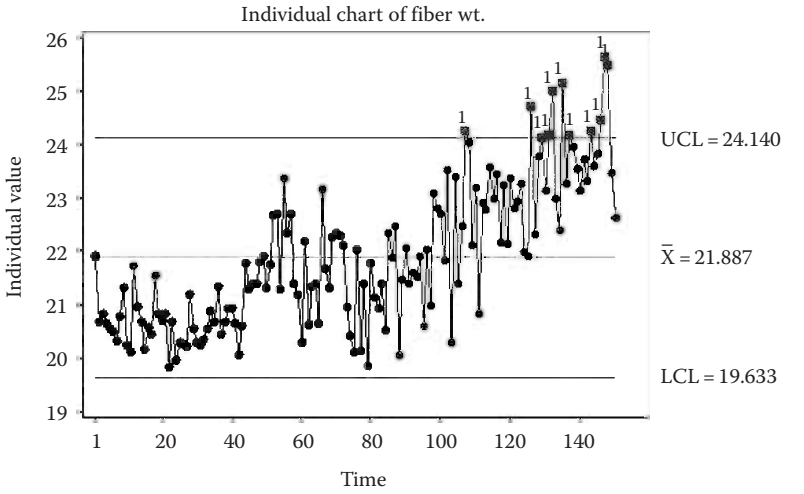


Figure 5.20 Time series plot of fiber weight data.

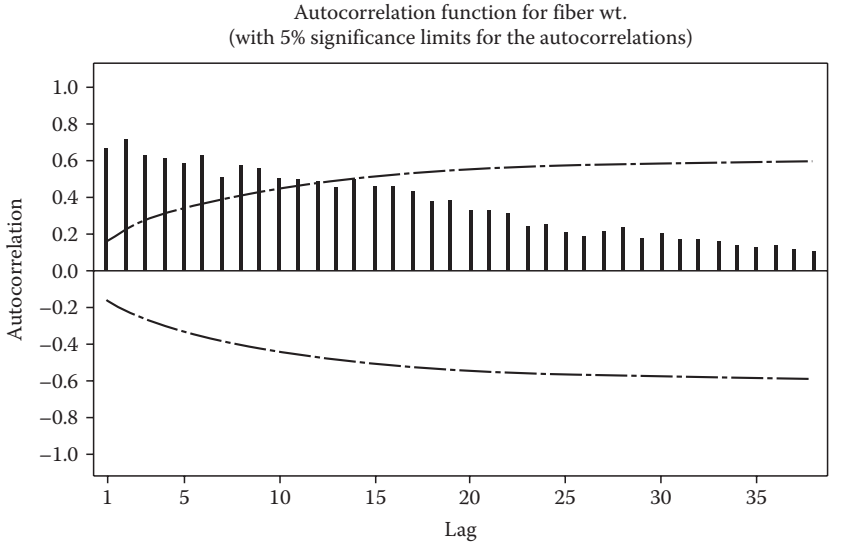


Figure 5.21 Autocorrelation function for time series data.

and differencing of the data will be necessary. After differencing with degree of differencing d = 1, the autocorrelation function dies out rapidly at lag q = 1, indicating a moving average process IMA (1). This is shown in Figure 5.22. The partial autocorrelation function (Figure 5.23) shows some exponential decay confirming that a moving average model can represent the weight data.

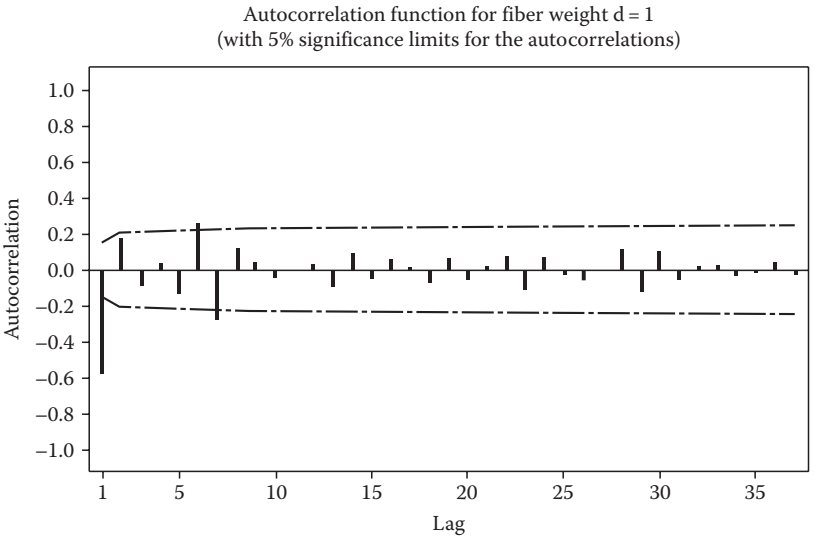


Figure 5.22 Autocorrelation function after d = 1 differencing.

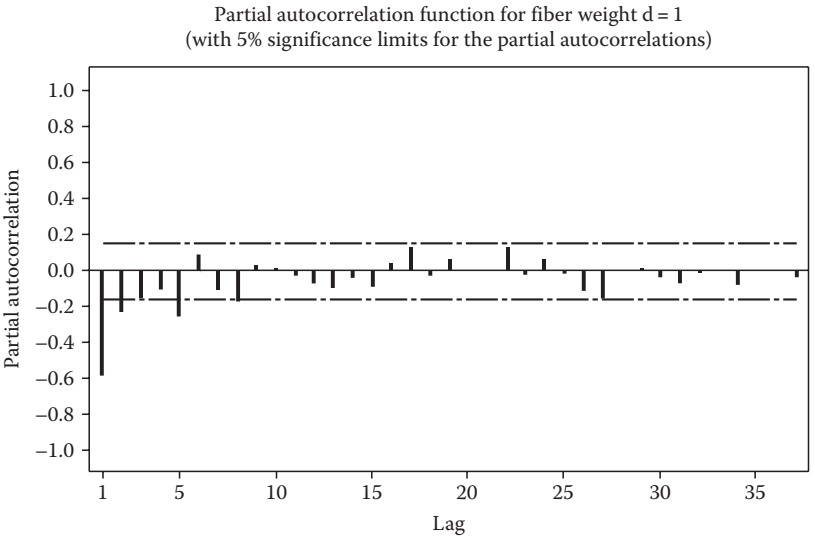


Figure 5.23 Partial autocorrelation function.

Estimation of parameters using an integrated moving average (IMA) of order 1 yields a coefficient estimate of 0.8125. Since the observations are highly autocorrelated, we need to be very suspicious about the out-of-control signals. We are not sure if they are actually due to special causes or if they are false alarms induced by the autocorrelation structure of the data.

A detailed examination of the sample autocorrelation and partial autocorrelation functions shows that the data in Figure 6.14 can be identified as an ARIMA (0,1,1), which can be represented as

$$Y_t = Y_{t-1} - q_1 a_{t-1} + a_t$$

Thus, using all the available data, the fitted model is

$$Y_t = Y_{t-1} - 0.8125\, a_{t-1} + a_t$$

The residuals from the model show that they are random (uncorrelated) indicating that the ARIMA (0,1,1) model is an adequate fit. The individual control chart shows that the process is in control and that the out-of-control signals observed previously were false alarms induced by the autocorrelation structure of the data.

Exponentially weighted moving average

The time series modeling approach illustrated in the preceding example is difficult to implement in the SPC environment. One major problem is the time that will be required to develop an ARIMA model for each quality characteristic of interest when applying control charts to several variables. Alwan (1990) extensively covered several issues regarding implementation of time series approach to SPC application. Several authors, Box, Jenkins, and MacGregor (1974), Hunter (1986), and, most recently, Montgomery and Mastrangelo (1991) have proposed the use of EWMA control chart as a possible compromise or an approximation to the general ARIMA model. The EWMA was first suggested by Roberts (1959) and can be represented as

$$Z_t = aX_t + (1-a)\, Z_{t-1}$$

where

$0 < \alpha \leq 1$ and X_t is the observation at time t
Z_t is the EWMA at time t
α is the smoothing constant

The value of Z_o is either a target or a process average. The advantages of EWMA are listed as follows:

- Applicable in certain situations where data are autocorrelated.
- Useful in approximating other members of the ARIMA family.
- Data need not be independent as in Shewhart charts.
- Applicable to processes whose means drift over time.
- Can serve as a compromise between the Shewhart and the CUSUM charts.

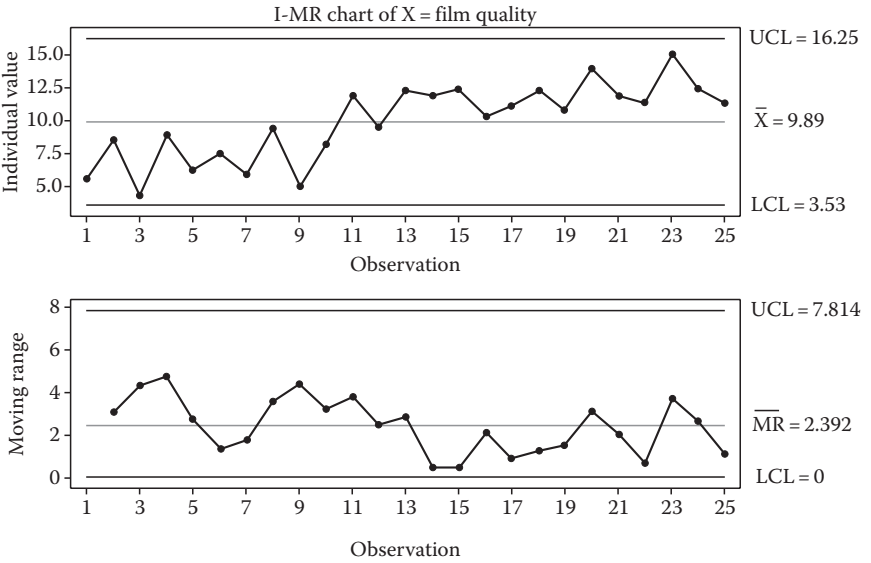


Figure 5.24 Shewhart chart for film quality example.

- EWMA uses all data in descending weights so that the most recent data are given more weight.
- Very sensitive to small shifts in the process average.
- Provides the ability to make adjustments due to its predictive capability.

Following are examples of Shewhart (Figure 5.24), EWMA (Figure 5.25), and CUSUM (Figure 5.26) charts for the data in Table 5.3. If possible one can plot all the three charts together for proper protection against both large and small process shifts. Figure 5.25 contains a linear regression model (R^2 = 92.28%) fitted to the EWMA values. Figure 5.26 also contains an exponential regression fit (R^2= 67.93%) for the CUSUM values. The results show a shift in the process level at sample number 10.

Cumulative sum chart

The CUSUM chart procedure was developed by Page (1954) and Bernard (1959) as a sequential likelihood ratio test for testing the hypothesis that the process average is equal to the target value. CUSUM uses the same assumption of independence as in the Shewhart chart. The procedure requires that we plot the following sum of deviations from target:

$$\Sigma(Y_i - \text{Target})$$

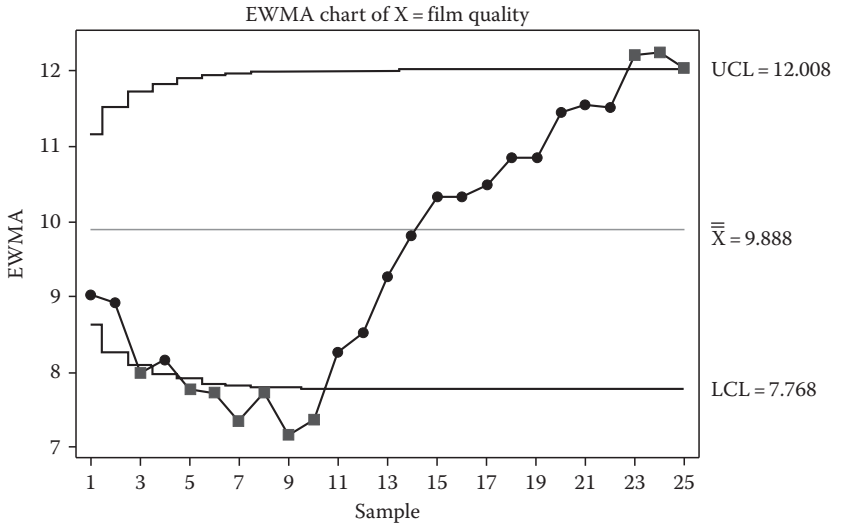


Figure 5.25 EWMA chart for film quality example.

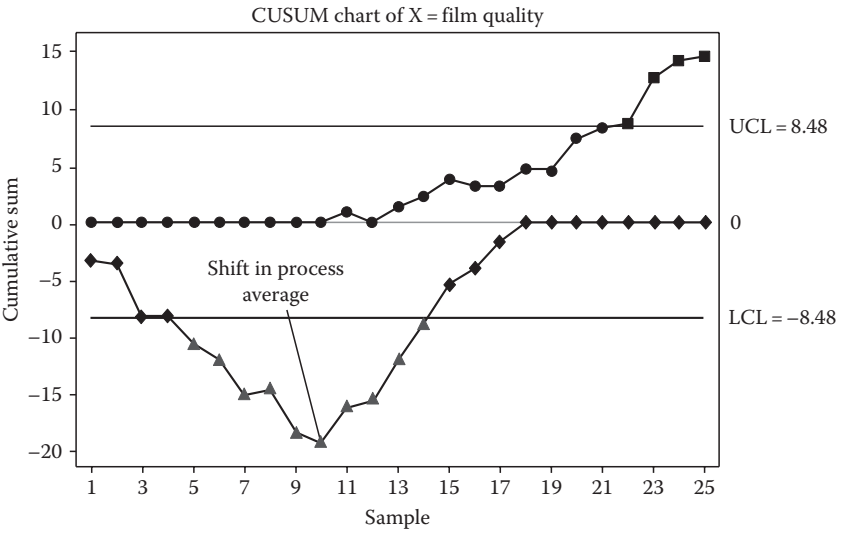


Figure 5.26 CUSUM chart for film quality example.

Any change in the process average from target will show up as a change in the slope of the CUSUM chart. This procedure has the ability to detect smaller changes in the process average more rapidly than the Shewhart chart. However, CUSUM charts are not as sensitive as Shewhart charts in detecting cycles, spikes, and trends. For these reasons, it is a good practice to use CUSUM in addition to Shewhart charts. The CUSUM

Table 5.3 Data for Plotting EWMA and CUSUM Charts

Time	X = film quality	EWMA α = 0.2	X-9.888	CUSUM
1	5.5	9.01	−4.388	−4.388
2	8.5	8.91	−1.388	−5.776
3	4.2	7.97	−5.688	−11.464
4	8.9	8.15	−0.988	−12.452
5	6.2	7.76	−3.688	−16.140
6	7.5	7.71	−2.388	−18.528
7	5.8	7.33	−4.088	−22.616
8	9.3	7.72	−0.588	−23.204
9	4.9	7.16	−4.988	−28.192
10	8.1	7.35	−1.788	−29.980
11	11.9	8.26	2.012	−27.968
12	9.5	8.51	−0.388	−28.356
13	12.3	9.26	2.412	−25.944
14	11.9	9.79	2.012	−23.932
15	12.4	10.31	2.512	−21.420
16	10.3	10.31	0.412	−21.008
17	11.1	10.47	1.212	−19.796
18	12.3	10.83	2.412	−17.384
19	10.8	10.83	0.912	−16.472
20	13.9	11.44	4.012	−12.460
21	11.9	11.53	2.012	−10.448
22	11.3	11.49	1.412	−9.036
23	15	12.19	5.112	−3.924
24	12.4	12.23	2.512	−1.412
25	11.3	12.05	1.412	0.000

chart in Figure 5.26 is based on a target of 9.88. The plot shows that a shift occurs in the process average after subgroup number 10.

Engineering feedback control

Case history: A process engineer was monitoring a production operation in a non-feedback control mode during normal production in order to observe the natural variance of the process. The following observations were what the engineer experienced:

- Using the feedback system, the process initially centered around the target weight of 21.5 g.
- Then by taking off the feedback controller at the target weight of 21.5 g, the weights displayed immediately varied slowly around the target weights of 21.5 g. Figure 5.27 shows a time series plot of the weights.

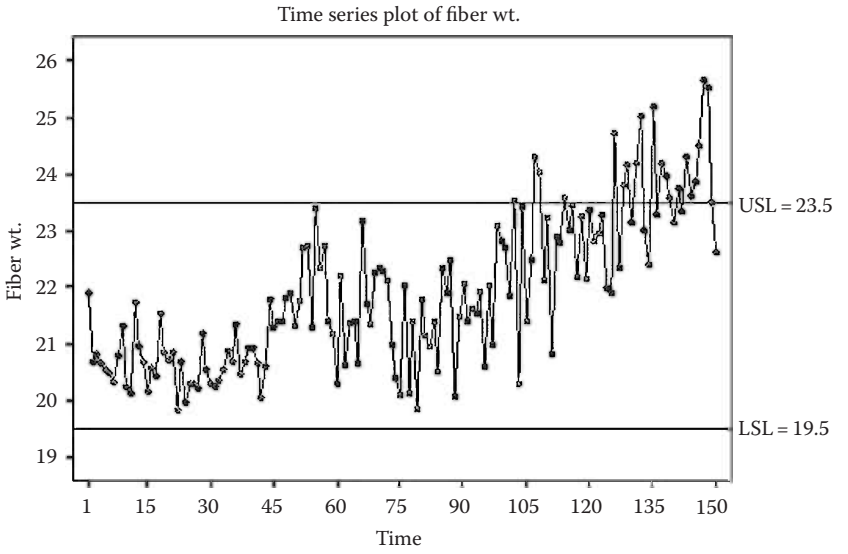


Figure 5.27 Fiber weight study.

- After a few minutes, the range had significantly increased with most of the weights falling outside the upper specification limit (USL = 23.5). All the observed packages at this time were rejected.
- The production group at this time had no choice other than to minimize cost. They, therefore, discontinued the non-feedback approach and the feedback system was installed again.
- The feedback system again brought the process back in control immediately after being installed.

The aforementioned situation often occurs in continuous process industries for processes which mix, react, and/or separate materials continuously. Chemical, pulp and paper, and petroleum refinery are a few of the continuous process industries. In these industries the automatic process control (APC) methods have dominated. The assumption behind the engineering process control approach is that the process is always being disturbed by causes that cannot be completely eliminated. In this situation, the process average level is not stable, but drifting continuously over time due to a myriad of causes such as wood variations in pulp and paper making, equipment aging, ambient temperature, quality of raw materials, and so on. Under this assumption, the logical approach has been to devise a control algorithm to be used to continuously respond to all deviations from target by compensating for the disturbances with some other variables. These continuous processes are a significant part of American industry and present unique challenges to controlling and improving product quality.

An excellent example of a feedback control problem was presented by Box, Hunter, and Hunter (1978, pp. 598–602). In the example, they considered a polymer process that was colored by adding dye at the inlet of a continuous reactor. At the outlet, the color index y was checked every 15 min and if it deviated from the target value of T = 9, the rate of the addition of dye X at the inlet could be increased or decreased. In their approach, they used a time series model to develop a controller or optimal control equation so that the deviations e_t from the target have the smallest standard deviation.

SPC versus APC

SPC and APC have been instrumental in quality improvement efforts in industry. Several papers have appeared in the literature, independently, on SPC and APC, but most recently, several techniques on integrating both methods have been published (MacGregor, 1987; Tucker et al. 1989; Box and Kramer, 1990; Palm, 1990). While the tactical approaches taken by these authors somewhat differ, several key practical considerations were presented. Some of the authors fully discussed some misconceptions concerning both methods as well as a full explanation of each approach.

Statistical process control

- Originates from manufacturing parts industry.
- Signals significant process changes from past performance.
- The philosophy is to minimize variability by finding and removing disturbances.
- Analyzes and controls infrequent or off-line measurements.
- Improves procedures, methods, and equipment.
- Monitors long-term performance.
- Expensive measurements and control actions.
- Low serial correlation.
- Monitoring may be done manually or by computer.
- The approach is to take action only when monitoring detects changes.
- Reflects Deming philosophy.

Automatic process control

- Originates from continuous process industry.
- Deals with how to adjust process to meet targets.
- The philosophy is to minimize variability by adjusting process to compensate the impact of disturbances which cannot be removed economically.
- Analyzes and controls high-speed on-line measurements.

- Accomplishes complex control strategies.
- Monitors short-term performance.
- Cheap measurements and control actions.
- High serial correlation.
- Adjustment is typically computer-based.
- The approach is to continually perform specified adjustments.
- Performs multivariate control and optimization.
- Deviates from Deming philosophy.

Criticisms of SPC and APC

There have been some aspects of controversy that sometimes arise between SPC and APC. This controversy has been fully discussed in the literature. Interested readers should refer to Box and Kramer (1990), MacGregor (1987), and Palm (1990).

SPC practitioners have sometimes criticized APC for the following:

- Overcompensating disturbances
- Compensating disturbances rather than removing them
- Concealing information rather than removing them

APC practitioners have in turn argued that Shewhart control charts are

- Inefficient for regulating a process
- Inefficient in coping well with fast system dynamics
- Misleading if sensing correlations over time

Overcompensation, disturbance removal, and information concealing

Box and Kramer (1990) and MacGregor (1991) demonstrated that in the presence of a drifting process average, by actively implementing an active optimal control strategy, a feedback scheme can provide significant reduction in variability. In addition, Box and Kramer (1990) provided a couple of examples where some disturbances cannot be removed. For example, a temperature variation in Minnesota from winter to summer may be too extreme for some people. In this case, people who cannot withstand this severe change may either relocate to Louisiana or Florida. However, if they decide to stay in Minnesota, they will have to compensate for the cold weather by using a furnace controlled by a thermostat supplying appropriate feedback control. Although automatic control conceals the nature of compensated disturbances, Box and Kramer (1990) and MacGregor and Harris (1990) also pointed out that this need not happen.

One can have the best of both worlds, if one uses SPC charts to monitor the control system. Provided the dynamics of the system are known, they show how the exact compensation can be computed and the original disturbance reconstructed. These qualities can then be displayed on charts and be subjected to routine examinations from the standpoint of generalized concept of common and special causes, where common causes are associated with the modeled process changeable only by management action, and special causes are associated with temporary deviations from the modeled process, as indicated by outliers. In this way, one can simultaneously minimize variability by active process control and at the same time have the ability to detect and eliminate special-cause disturbances.

Integration of SPC and APC

It is important to note that there are several approaches one can take for quality and process controls. Which approach or class of approaches one can take depends on the problem at hand and how realistic the assumptions are for the real problem under study. In deciding between SPC and APC, it is always important to distinguish between variations that can be eliminated at the source and those that cannot be eliminated at all. As can be seen from the preceding discussions, both SPC and APC seek to reduce variation, promoting process understanding as well as facilitating process improvement. For these reasons, majority of the aforementioned authors have proposed the integration of both SPC and APC whenever possible. Tucker, Faltin, Weil, and Doganaksoy (1989) provided an excellent concept of integrating both SPC and APC. Their concept is known as the algorithmic statistical process control (ASPC). The concept of ASPC is presented in Figure 5.28 while Figure 5.29 presents the flowchart for implementing ASPC.

Systems approach to process adjustment

In many continuous process industries and, in particular, chemical and petrochemical industries, materials are produced in batches. One major goal in these industries is to reduce batch-to-batch variability that is very prevalent. For this reason, adjustments are often made to the already produced batches so as to provide a more uniform incoming material on batch-to-batch basis for the next process step. In this industry it is common to have material produced in one part of a process become the incoming material for the next process step. It is also common that the incoming material from the first process step will be adjusted in order to provide a more uniform material on a batch-to-batch basis for the next process step. Over time, this adjustment becomes an inherent part of the process. Caffrey (1990) demonstrated that when a reasonable

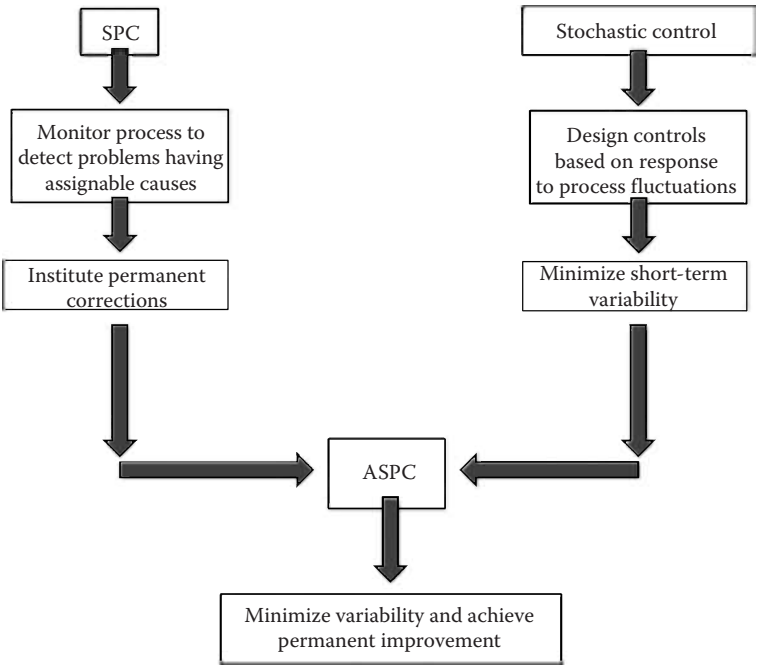


Figure 5.28 Concept of ASPC.

adjustment strategy is made to a process, say process A, very little benefit may be realized by another process, say process B, which is the next process step. But by considering the process as a whole, a more optimal strategy can be employed. The goal here is to view the whole process in an integrated manner.

ARIMA modeling of process data

Let us consider the single-input–single-output system shown in Figure 5.30.

Assume the output Y is sampled at discrete equi-spaced intervals of time, and its value at time t is represented by Y_t.

In Figure 5.30, N_t represents the total effect on the output Y of all disturbances occurring anywhere in the system. If the disturbance is not compensated for, it would cause the output Y to drift away from target. The input variable at time t, u_t, can be manipulated to affect changes in the output, Y. The general model that can be used to represent the process dynamics and the stochastic disturbances relies mainly on discrete transfer function models for the process and discrete ARIMA time series models for the disturbances. This model (Box and Jenkins, 1976) can be represented as

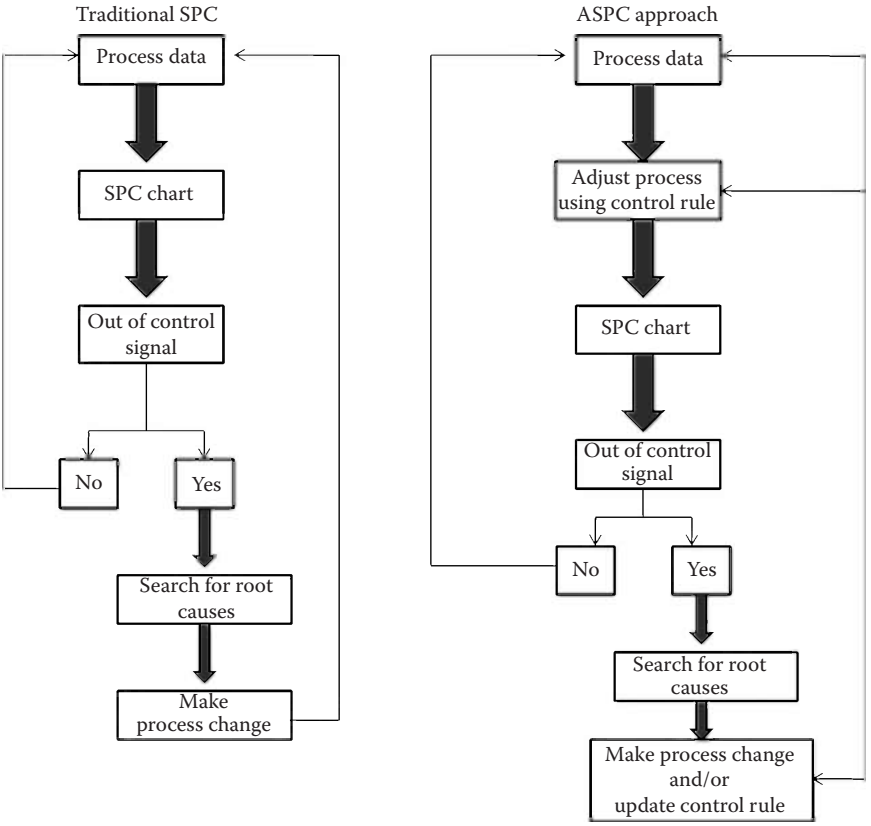


Figure 5.29 Flowchart for implementing ASPC.

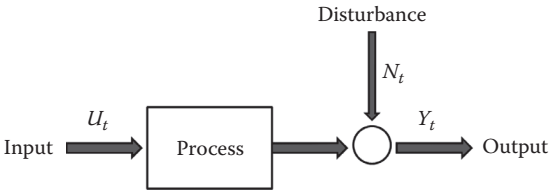


Figure 5.30 Input–output process for ARIMA example.

$$Y_t = \frac{w(B)B^b}{\delta(B)}u_t + \frac{\theta(B)}{\Phi(B)(1-B)^d}a_t$$

The first term is a transfer function model of the process relating the dynamic effect of u_t to the output Y_t and it is referred to as a discrete transfer model of order (r,s,b), while the second term represents the

independent effect of all disturbances occurring in the system on the output Y_t and referred to as a discrete ARIMA model of order (p,d,q). B is the backward shift operator, $BY_t = Y_{t-1}$; b is the whole periods of process delay. The components δ(B) and w(B) are polynomials of order r and s, respectively, in the operator B. This can be represented as

$$\delta(B) = 1 - \delta_1 B - \delta_2 B^2 - \cdots - \delta_r B^r$$

$$w(B) = w_0 - w_1 B - w_2 B^2 - \cdots - w_s B^s$$

Generally, if no manipulations were made in the input variable, u_t, the ARIMA model for the disturbance, N_t, would represent the behavior of the process output, Y_t. This can be represented as

$$N_t = \frac{\theta(B)}{\Phi(B)(1-B)^d} a_t$$

where

$$\Phi(B) = (1 - \Phi_1 B - \Phi_2 B^2 - \cdots - \Phi_p B^p)$$

is an autoregressive polynomial of order p,

$$\theta(B) = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)$$

is a moving average polynomial of order q, $(1 - B)^d$ is the backward difference operator of order d, and a_t is a sequence of independent normally distributed random shocks with mean zero and constant variance, σ_a^2. When d = 1, N_t gives rise to nonstationary disturbances in which the process average level is free to drift from time to time, while for d = 2, one obtains disturbances in which both the level and trend or slope of the disturbance drift over time.

Model identification and estimation

Detailed description of model identification and estimation procedures are fully covered in Box and Jenkins (1976) as well as MacGregor (1989) and, therefore, will not be covered in this book. MINITAB can be used for the identification and estimation for the ARIMA part of the combined model as described previously. SAS program can be used for the identification and estimation of both the transfer function and the ARIMA part of the model.

Minimum variance control

The main idea of optimal stochastic control theory is that, given a model as described previously whose combined process dynamic and disturbance model of the system has been identified, one can design a controller which will optimize some specified performance index involving the output and input variables. One case according to MacGregor (1989) is the minimum variance control (MVC). For this case, the controller is designed to compensate for the disturbances, N_t, in such a way that the variance of the difference between the output and target is minimized. Although several disturbance models can be considered from N_t, one simple case is the IMA model since it arises frequently in SPC. This model can be represented as

$$(1-B)Y_t = (1-\theta B)a_t$$

The MVC for this disturbance model (Figure 5.31) is given by MacGregor (1989) as

$$u_t = -\frac{1}{g}\cdot\frac{(1-\theta)}{(1-\theta B)}(Y_t - u_{t-1})$$

Process dynamics with disturbance

For the case where process dynamics are important, usually in the continuous process industries, we consider a simple case first-order process:

$$Y_t = \frac{w_0}{(1-\delta B)}u_{t-1} + \frac{(1-\theta B)}{(1-B)}a_t$$

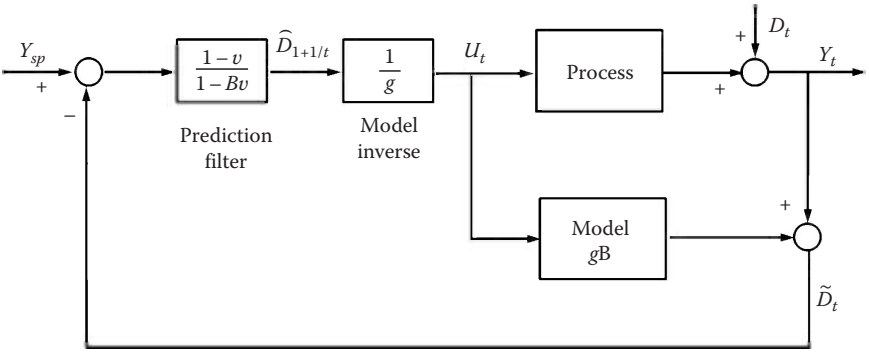


Figure 5.31 Minimum variance controller model.

The MVC is given by

$$u_t = \frac{(1-\mathrm{u})}{w_0} \cdot \frac{(1-\mathrm{d}B)}{(1-B)} Y_t$$

The specific situations for which stochastic control theory is extremely well suited are when

- Drifting stochastic disturbances are the main disturbances in the system for which control is needed.
- There is significant sampling or measurement noise.
- The process exhibits a time delay or dead time between the manipulated input and the output.

Process modeling and estimation for oil and gas production data

The oil industry has used decline curve analysis with limited success in estimating crude oil reserves and in predicting future behavior of oil and gas wells. This chapter, therefore, explores the possibility of using the ARIMA technique in forecasting and estimating crude oil reserves. The authors compare this approach with the traditional decline method using real oil production data from 12 oil wells in South Louisiana. The Box and Jenkins (1976) methodology is used to develop forecast functions for the 12 wells under study. These forecast functions are used to predict future oil productions. The forecast values generated are then used to determine the remaining crude oil reserves for each well. The accuracy of the forecasts relative to the actual values for both ARIMA and decline curve methods is determined by various statistical error analyses. The conditions, under which one method gives better results than the other, are fully investigated. In almost all the cases considered, the ARIMA method is found to perform better than the decline curve method.

Introduction

The decline curve analysis technique has been extensively used in industry to predict future oil and gas productions and to estimate crude oil reserves. These predictions are used as the basis for economic analysis to support exploration, development, unit equity, facility expansion, property sales or purchase, and loan securement; it is also used to determine if a secondary recovery project is to be carried out. The three common decline models are exponential, hyperbolic, and harmonic. These models are described in the section "Decline curve method." One problem with these models is that the graphical approach sometimes produces

large errors when extrapolated for a useful length of time. Furthermore, decline curve method is only applicable to depletion-type reservoirs, producing at capacity and cannot be used where the rate of production is constant or nearly constant with time. This will be true when the reservoir is under proration or where it has a large gas cap or an active water drive.

In this paper, the ARIMA technique is proposed because, in addition to being a good forecasting tool, this study also shows that it is applicable to situations where decline curve has failed, such as in water-drive reservoirs.

Time series approach–Box and Jenkins methodology

For many years, smoothing techniques have been the only method commonly used in forecasting time series. Then, in 1976, Box and Jenkins presented a method of identifying, fitting, and diagnosing the fit of a time series model with wide flexibility. Box and Jenkins also presented the details of forecasting with the fitted model. The techniques presented have become known collectively as the Box–Jenkins method. The method consists of four stages: (1) identification of the ARIMA model to be fitted, (2) estimation of the parameters of the ARIMA model, (3) diagnosis of the fitted model to assure its applicability, and (4) forecasting with the fitted model. These four steps are used to evaluate each of the 12 data sets under study. The ARIMA results obtained are provided in Tables 5.5 and 5.6.

The ARIMA model

In building our forecasting model to determine oil production reserves, we consider the general Box and Jenkins (1976) ARIMA model of order (*p*,*d*,*q*). This model can be represented as

$$\phi_p(B)(1-B)^d Q_t = \theta_o + \theta(B)\, a_t \tag{5.1}$$

where

$$\phi_p(B) = (1-\phi_1 B-\phi_2 B^2-\ldots\ldots\ldots\ldots-\phi_p B^p)$$

$$\theta_q(B) = (1-\theta_1 B-\theta_2 B^2-\ldots\ldots\ldots\ldots-\theta_q B^q)$$

where

Q_t is the oil production at time t
d is the degree of differencing
B denotes the backward shift operator, defined by $B^j Q_t = Q_{t-j}$,
a_t is the unobservable random disturbance which may be due to reservoir energy

The a_t's are assumed to be uncorrelated with zero means and constant variances $\mathbf{s}_a^2$. $\Phi(B)$ is the stationary autoregressive operator whose roots $\Phi(B) = 0$ lie outside the unit circle and $\theta(B)$ is the moving average operator, which is assumed to be invertible; that is, the roots of $\theta(B) = 0$ lie outside the unit circle (Box and Jenkins, 1976).

The general model in Equation 5.1 provides a range of models—stable or unstable—which can adequately be used to model oil production data from different types of reservoirs. The requirements of stationarity and invertibility apply independently and, in general, the operators $\Phi(B)$ and $\theta(B)$ will not be of the same order. The inclusion of the constant θ_o will permit automatic allowance for a deterministic polynomial trend, of degree d when θ_o is not equal to zero. Our approach, therefore, to crude oil reserve estimation will be first to identify an adequate forecasting model for each production data set under study using the Box and Jenkins (1976) approach. Once an appropriate model has been identified for each well, parameter estimation, diagnostic checking, optimal forecasting, and reserves estimation procedures follow immediately.

Methodology

This section covers the practical aspects of the method used in this chapter. Monthly oil production data were collected from 12 different oil wells from onshore production reports in South Louisiana. Each production data set is from one well. Six of the production series are suitable for decline curve analysis method, whereas the remaining six production series were obtained from water-drive reservoirs. The aim here is that, in addition to investigating ARIMA technique for decline data, we are also going to investigate the possibility of using ARIMA method for the case where decline curve method has failed or performed poorly, such as in water-drive reservoirs. The available data sets are made up of at least 60 months of production series. In order to test the accuracy of the forecasts, the original series are shortened by at least 10 months. The sum of these 10 or more months is used as an estimate of actual remaining reserves for model comparison purposes only. Using statistical errors analysis, the actual production values are compared with the forecasts obtained from ARIMA and decline methods for those 10 or more months. The results obtained are provided in Tables 5.4 through 5.7.

Table 5.4 Parameter Estimates for Decline Curve Method

Well	Decline model	Decline rate per month, D	Initial decline, Q_i	Decline exponent, m
Iberia	Exponential	0.01047	2,643.30	—
Yentzen	Exponential	0.02404	403.87	—
Brock	Exponential	0.00999	1,583.00	—
Linga	Exponential	0.01487	3,476.00	—
Merg	Hyperbolic	0.07587	13,032.00	0.75
Davie	Exponential	0.04275	3,505.20	—
WDI	Exponential	0.05156	4,178.40	—
WD2	Hyperbolic	0.01938	599.12	0.20
WD3	—	—	—	—
WD4	Exponential	0.0064	1,125.50	—
WD5	Exponential	0.03182	3,639.60	—
WD6	—	—	—	—

Table 5.5 Decline Data: Parameter Estimates and Chi-square Table for ARIMA Method

Well	Sample size	ARIMA(p,d,q) model	Parameter estimated	Standard error	Df	$X^{2(cal)}$
Iberia	80	Log Q_t	$\Phi_1 = 0.13$	0.6288	11	14.268
	(60)	(1,1,1)	$\theta_1 = 0.33$	0.5976		
			$\theta_1 = 3.63$	0.50400		
Yentzen	80	Log Q_t	$\Phi_1 = 0.98$	0.0496	12	6.43
	(60)	(1,0,1)	$\theta_1 = 0.88$	0.1169		
			$\theta_0 = 0.094$	0.2574		
Brock	70	(1,0,0)	$\Phi_1 = 0.96$	0.0262	11	10.53
	50					
Linga	80	(1,0,0)	$\Phi_1 = 0.95$	0.0272	14	15.96
	(60)					
Merg	80	(0,1,1)	$\theta_1 = 0.46$	0.1088	14	10.13
	(60)					
Davie	74	Log Q_t	$\Phi_1 = 0.39$	0.4385	11	7.544
	(60)	(1,1,1)	$\theta_1 = 0.07$	0.4366		
			$\theta_0 = 0.08$	0.0519		

Decline curve method

The decline curve technique consists of testing each individual series for exponential, specific hyperbolic or harmonic fit. The decline equations used to perform these tests are according to Arps (1945) and are given as follows:

Table 5.6 Water-Drive Data

Well	Sample size	ARIMA(p,d,q) model	Parameter estimated	Standard error	Df	$X^{2(cal)}$
WD1	60	$\log Q_t$	$\Phi_1 = -0.41$	0.1385	10	8.231
	(50)	(2,1,0)	$\theta_1 = -0.35$	0.1384		
WD2	60	$\log Q_t$	$\Phi_1 = 0.98$	0.0023	11	14.892
	(50)	(1,0,0)				
WD3	60	(0,0,1)	$\theta_1 = -0.50$	0.1222	10	4.239
	(50)					
WD4	60	$\log Q_t$	$\Phi_1 = 0.19$	0.1456	9	7.206
	(50)	(1,1,1)	$\theta_1 = 0.96$	0.0297		
			$\theta_0 = -0.01$	0.0034		
WD5	60	$\log Q_t$	$\Phi_1 = 0.75$	0.3898	9	5.411
	(50)	(1,0,1)	$\theta_1 = 0.60$	0.4627		
			$\theta_0 = 1.38$	2.8853		
WD6	60	(0,1,1)	$\theta_1 = 0.07$	0.0953	11	11.483
	(50)					

Table 5.7 ARIMA versus Decline Curve for Comparison of Reserve Estimates

Well	Sample size	Decline function	ARIMA model	Remaining reserves		
				Actual	Decline	ARIMA
Iberia	80	Exponential	$\log Q_t$	22,319	25,451	23,194
	(60)		(1,1,1)			
Yentzen	80	Exponential	$\log Q_t$	3,112	1,1516	2,331
	(60)		(1,0,1)			
Brock	70	Exponential	(1,0,0)	13.523	17,147	12,134
	(50)					
Linga	80	Exponential	$\log Q_t$	28,613	23,590	28,380
	(60)		(1,0,0)			
Merg	80	Hyperbolic	(0,1,1)	54,558	30,832	61,869
	(60)	m = 0.75				
Davie	74	Exponential	$\log Q_t$	539	2,701	634
	(60)		(1,1,1)			

Exponential decline:

$$q(t) = q_i \exp(-Dt) \tag{5.2}$$

Hyperbolic decline:

$$q(t) = q_i(1 + mDt)^{-1/m} \tag{5.3}$$

Harmonic decline:

$$q(t) = \frac{q_i}{(1 + Dt)} \tag{5.4}$$

where

$q(t)$ is the production rate at time t
q_i is the initial production at time $t = 0$
D is the decline rate
t is the time of production
m is the decline exponent

For decline curve analysis, the shortened series are tested using the most appropriate of the three decline models mentioned earlier. To determine the most appropriate decline model, we select the model that provides the lowest point error, or one that has a cumulative ratio error close to one. An adequate model has a point error close to zero or a cumulative ratio close to 1.0. The point error is calculated as

$$\text{Point error} = \sqrt{\left[\frac{\Sigma(q_p - q_a)^2}{n-1}\right]} \tag{5.5}$$

$$\text{Cumulative error} = \frac{Q_{actual}}{Q_{predicted}} \tag{5.6}$$

where

q_p is the predicted production rates
q_a is the actual production rates
Q is the cumulative production

The predicted production rates and the cumulative productions are obtained for each set of production series using exponential, hyperbolic, and harmonic decline models. These computations are done using decline curve analysis software program made by the Logic Group, Austin, Texas (The Logic Group, 1(1987)). From these three models, we select one that has the lowest point error or cumulative ratio close to one. Once an appropriate model is determined for the series, the parameters associated with the model are estimated using decline curve software program. The results obtained for each well are provided earlier in Table 5.7. The estimated parameters are then incorporated into the decline curve equation to provide a realistic forecast function (Ayeni, 1989). The forecast values generated are used to determine the remaining reserve. This remaining reserve

is compared with the remaining reserve obtained from the ARIMA method as well as to the actual remaining reserve. The remaining reserve is the sum of all the forecast values. The actual remaining reserve is the sum of all known monthly production data not used in modeling. For example, for a 60 month production data, the first 50 month data can be used for modeling, and the remaining 10 months can be added together as the actual remaining reserve for model comparison purposes only. These comparisons are provided in Tables 5.4 and 5.5.

Statistical error analysis

The statistical error analysis is used to check the performance, as well as the accuracy, of the ARIMA and decline methods. The accuracy of the forecasts relative to the actual values is determined by various statistical methods. The criteria used in this study are average relative error, average absolute error, forecast root mean square error (FRMSE), and minimum/maximum absolute error.

Average relative error

This is defined as the relative deviation of the forecast values from the actual values and is given by

$$E_r = \left(\frac{1}{n_f}\right)\sum E_i \tag{5.7}$$

where n_f is the sample size of the forecast values, and

$$E_i = \frac{q_{actual} - q_{forecast}}{q_{actual}} \tag{5.8}$$

The lower the value of E_r, the more equally distributed are the errors between positive and negative values.

Average absolute relative error

This can be defined as

$$E_a = \left(\frac{1}{n_f}\right)\sum |E_i| \tag{5.9}$$

and represents the relative absolute deviation of forecast values from actual production values.

Forecast root mean square error

This is a measure of dispersion and is expressed as

$$s = \sqrt{\left[\left(\frac{1}{n_f - 1}\right)\sum E_i^2\right]} \tag{5.10}$$

A smaller value of FRMSE indicates a better degree of fit. A model that has a perfect fit has an FRMSE of zero.

Minimum and maximum absolute relative error

After the absolute error for each data point is calculated, both the minimum and maximum values are scanned for the range of errors.

$$E_{\min} = \min|E_i| \tag{5.11}$$

and

$$E_{\max} = \max|E_i|, \quad \text{for } i = 1, 2, \ldots, n_f \tag{5.12}$$

The accuracy of the forecasts can be determined by examining the maximum absolute relative error. The lower the value of maximum absolute relative error, the higher the accuracy of the forecast.

Cumulative ratio error

The cumulative ratio error is the ratio of actual cumulative production and forecast cumulative production. It is defined as

$$E_Q = \frac{Q_{\text{actual}}}{Q_{\text{forecast}}} \tag{5.13}$$

For simplicity in interpretation, Equation 5.13 can be inverted if the forecast cumulative is less than the actual cumulative production. This will force the range of E_Q between 0 and 1. The closer the value of E_Q to 1, therefore, the closer the cumulative forecast to the cumulative actual values. This is good in our case, because the remaining reserve is calculated by adding all the forecast values together. The result obtained is equal to the cumulative forecast.

ARIMA data analysis

In this section, we shall discuss the method used in analyzing each of the production series under ARIMA method. The specific aim here is to obtain some idea of the values of order p, d, q needed in the general ARIMA model.

The Time Machine software program developed by Research Services, Utah (Research Services, 1986), is used for all our ARIMA procedures. This software is user friendly, and has the capability of modeling time series data using ARIMA (p, d, q) models. It provides information about the autocorrelation, partial autocorrelation, forecast values, residual plot, diagnostic checking, as well as parameter estimation for ARIMA (p, d, q) models.

Model identification for series WD1

In analyzing each oil production series, a plot of $q(t)$ versus t is obtained. This plot is very valuable, because qualitative features such as trend, seasonality, or discontinuities will usually be visible if present in the data. On examination of this plot for series WD1 in Figure 5.32, two facts are apparent:

1. The mean of the series is not stationary because it shows a downward trend.
2. The variability in the series is not constant (nonstationary) over the 60 month period.

The presence of trends in the data resulted in the lack of stationary in the mean. This is also confirmed by the large values of the autocorrelation function in Figure 5.33. The fact that the autocorrelation function fails to die out rapidly is a strong indication that the series is nonstationary, that is, the degree of difference d is not zero. Time series analysis requires, however, that the series be stationary, and in particular that the variance of the series be constant over time. This lack of stationary, caused by the trends, indicates that some degree of differencing is required. These trends are removed by differencing the observed series. Figure 5.34

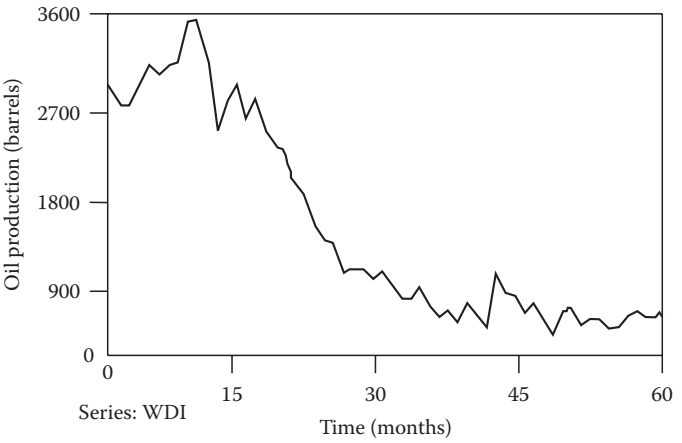


Figure 5.32 Plot of oil production versus time for series WDI.

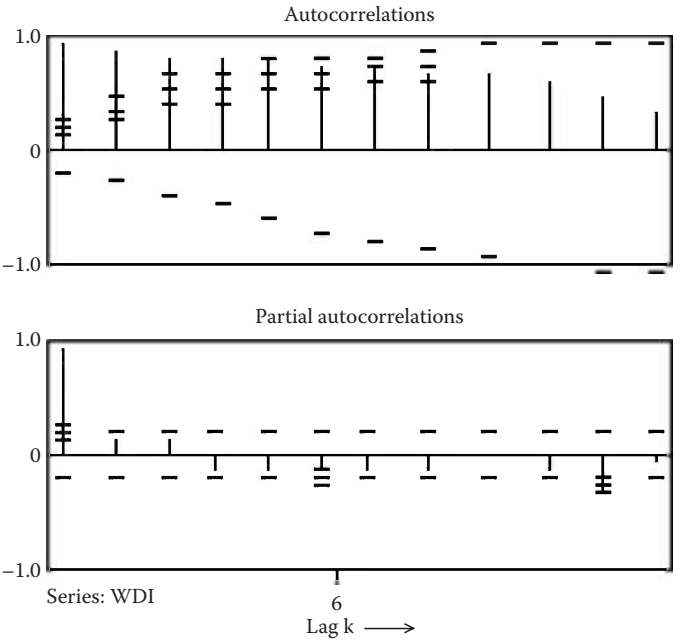


Figure 5.33 Autocorrelation and partial autocorrelation functions for series WDI, when $d = 0$.

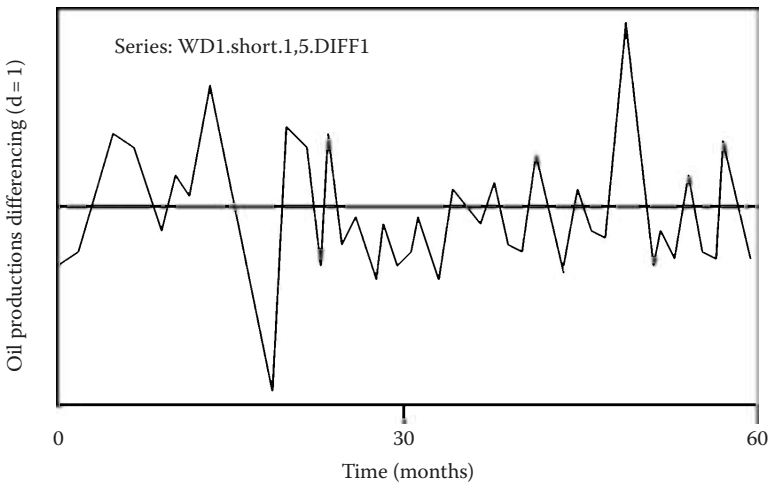


Figure 5.34 Plot of the differenced series ($d = 1$) with time.

shows a plot of the differenced series for series WD1 with degree of difference, $d = 1$. From this graph, it is clear that the mean of the series is constant or nearly constant with time. It seems, therefore, that a single difference of degree might have resolved the problem of nonstationarity in the mean. The variability, however, is still not constant over the entire time period. The relative instability of the variance with some outliers in this case shows we will probably need to transform the data. For all cases considered, where transformations are necessary, the log transformations are used. The complete removal of the trends is determined by evaluating the theoretical autocorrelation function of the transformed series in Figure 5.35. An examination of the theoretical autocorrelations and partial autocorrelations for the log transformed of series WD1 when $d = 1$, shows no sign of large autocorrelations after lag $k = 1$ or large partial autocorrelations after lag $k = 2$. This is an indication that the trends have been successfully removed. Inasmuch as the partial autocorrelation for the first two lags, $k = 2$ are nonzero, an autoregressive model of order $p = 2$ can be used to represent WD1 series. In addition, inasmuch as only the autocorrelation at lag $k = 1$ is nonzero, a moving average model of order $q = 1$ can be used to model WD1 series. This finally formed a mixed ARIMA model. The model, therefore, identified for WDI is a mixed ARIMA model of order (2, 1, 1) with logarithm transformation of the original series.

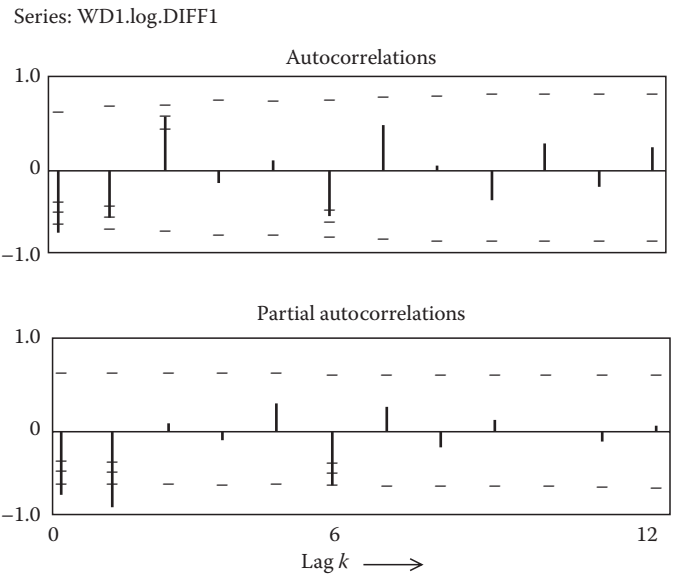


Figure 5.35 Plot of autocorrelation and partial autocorrelation functions for log transformed (d = 1), with lag *k*.

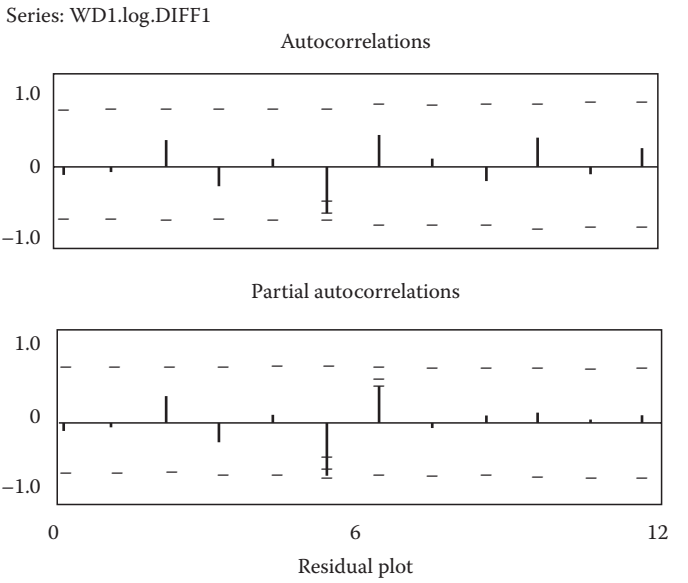


Figure 5.36 Residual autocorrelation plot.

The residual autocorrelation plot for this model is provided in Figure 5.36. This plot shows that the identified model is adequate for WDI series, because there are no large values of residual autocorrelation. The theoretical autocorrelation can be calculated using

$$r_k \frac{c_k}{c_o} \tag{5.14}$$

where

$$c_k = \frac{1}{N}\sum_{t=1}^{N-K}(Q_t - \overline{Q})(Q_{t+k} - \overline{Q})$$
$$\text{such that } K \le \frac{N}{4}, \quad k = 0,1,2,\ldots,K \quad \text{and} \quad \overline{Q} = \frac{\Sigma Q_t}{N} \tag{5.15}$$

Model identification for series Brock

In this stage, we further describe the procedures used to obtain a tentative identification of the ARIMA model using series Brock as another example; we also show how the identified model is fitted to the data; and finally,

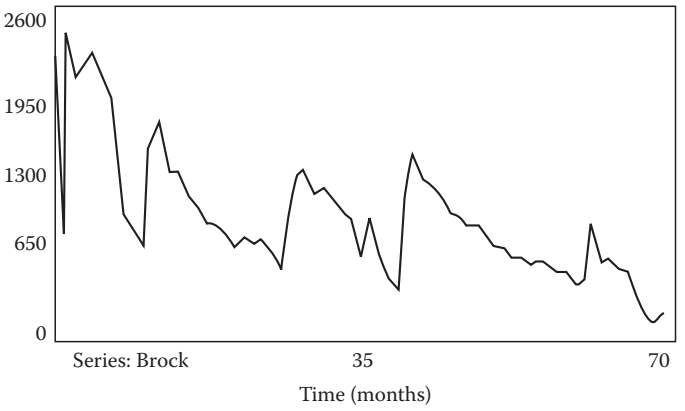


Figure 5.37 Plot of oil production versus time for series Brock.

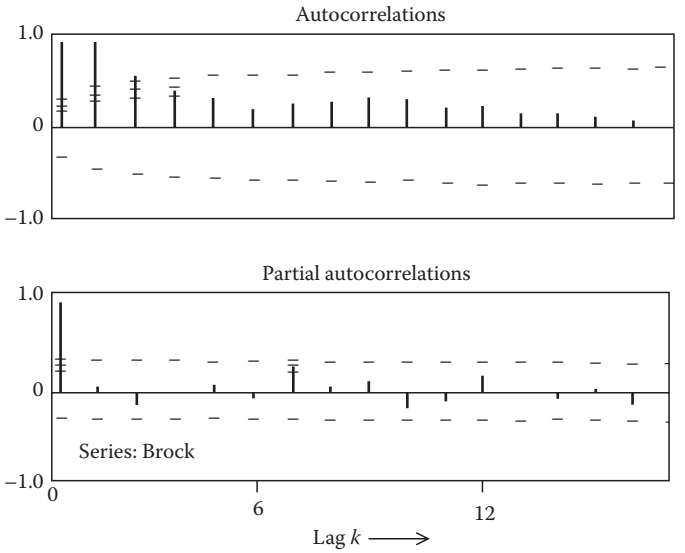


Figure 5.38 Plot of autocorrelation and partial autocorrelation functions for series Brock, (d = 2), with Lag *k*.

we test the fitted model for adequacy. In our analysis, we again make use of two important graphs in our identification process: (1) a simple graph of the data versus time (Figure 5.37), and (2) plots of autocorrelations and partial autocorrelations (Figure 5.38). The partial autocorrelations can be computed as follows:

$$\hat{\Phi}_{p+1,j} = \hat{\Phi}_{pj} - \hat{\Phi}_{p+1,p+1}\hat{\Phi}_{p,p-j+1} \tag{5.16}$$

and

$$\hat{\Phi}_{p+1,p+1} = \frac{r_{p+1} - \sum_{j=1}^{p} \overset{\circ}{\Phi}_{pj}\, r_{p+1-j}}{1 - \sum_{j=1}^{p} \overset{\circ}{\Phi}_{pj}\, r_j}, \quad j = 1, 2, \ldots, p \tag{5.17}$$

An examination of the patterns in the autocorrelations and partial autocorrelations of Figure 5.38 for series Brock with degree of differencing ($d = 0$), reveals that the autocorrelations provide rapid exponential decay from the first lag, while the partial autocorrelation rapidly dies out at *lag k* = 1. By decaying, we mean the autocorrelations are relatively large (exceeding 2 – σ limits) at the early lags, but diminish consistently to small values that are statistically indistinguishable from zero (Figure 5.38). This shows that the model identified is an ARIMA of order (1,0,0). The main horizontal line extending across the entire graph represents zero. This is an important reference point, since each autocorrelation and partial autocorrelation must lie between –1 and +1. Vertical bars, extending above the zero line or below the zero line, represent the individual autocorrelations and partial autocorrelations. The two horizontal dashes that appear with each vertical bar, one above zero and the other below, represent two standard deviations above and below zero.

When interpreting the autocorrelation and partial autocorrelation plots, we look for bars that extend beyond the horizontal dashes. If they exceed two standard deviations, we consider them to be statistically different from zero. Theoretically, this is only true for large samples, but there should be no problem since each set of series contains at least 50 observations.

Estimation and diagnostic checking

After the identification stage, we perform parameter estimation. These parameters are estimated using Time Machine software program specifically designed for ARIMA model building. Then, each model is checked for model adequacy using the diagnostic command of the software. For each model, a residual analysis test was performed. An equation that can be used to represent the residual is given as

$$\hat{a}_t = \mathfrak{q}^{-1}\hat{\Phi}(B)q_t \tag{5.18}$$

Figures 5.36 and 5.39 show the residual autocorrelations and partial autocorrelations plots for series WD1 and Brock, respectively. In both figures, it is evident that the residuals are within the 2 – σ limit, showing that the

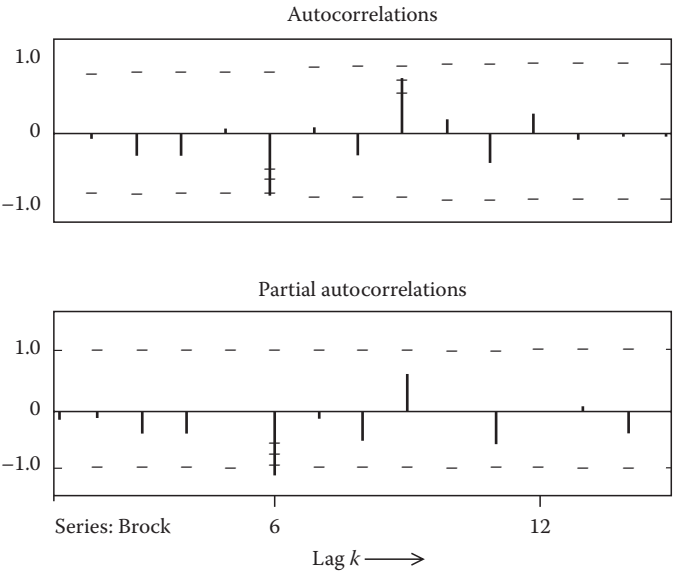


Figure 5.39 Plot of residual autocorrelation and partial autocorrelation functions for series Brock, (d = 2), with Lag *k*.

models obtained for series WD1 and Brock are adequate. Similar residual analysis test is carried out for each of the remaining models until adequacy is achieved.

For each identified model, the chi-square test is also performed. The program gives a computed Ljung–Box chi-square result. This result is compared with the value obtained from the chi-square table at a level of significance ($\alpha = 0.05$) and the corresponding degrees of freedom. The computed Ljung–Box chi-square statistics obtained are provided in Tables 5.2 and 5.3. Each model is found to be adequate at ($\alpha = 0.05$).

Then the forecast values are generated using ARIMA forecast command of the time Machine software program. These forecast values are used to determine the remaining reserves by summing all the forecast values. Figure 5.40 shows a plot of the actual and the forecast values obtained for series Brock under ARIMA and decline methods.

Comparison of results

A statistical error analysis technique is used to compare the two methods under consideration. The comparison is done by computing the average relative error, average absolute relative error, FRMSE, minimum and maximum absolute relative errors, as well as cumulative error. These computations are for each of the 12 wells using both ARIMA and decline curve methods.

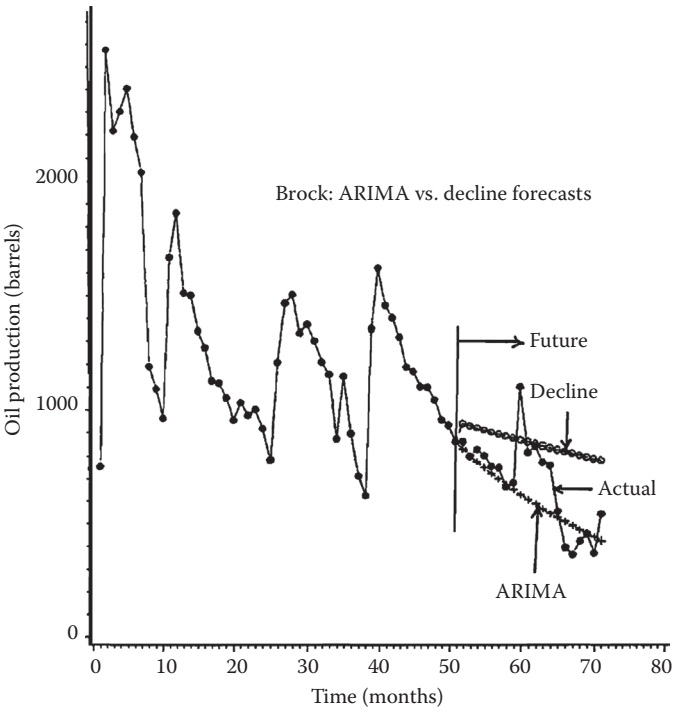


Figure 5.40 Plot of actual and forecast values (for series Brock) with time, under ARIMA and decline methods.

A comparison of the remaining reserves is provided in Tables 5.4 and 5.5. In all the cases considered, this result shows that ARIMA method provides better reserve estimates than decline curve method. This is because reserve estimates obtained under ARIMA method are more accurate than those obtained under decline curve method. For this reason, ARIMA technique should be considered as a good candidate for crude oil reserve estimation. It is interesting to note that the water-drive data for WD3 and WD6 did not pass through our numerical check for decline curve analysis for all values of decline exponent m used (Tables 5.8 through 5.10). This may be due to the fact that decline curve method rarely works with data obtained from active water-drive reservoirs.

Tables 5.6 and 5.7 represent the comparison of errors of the forecast values, relative to the actual values for each of the 12 oil wells. The forecast for this study achieved the lowest errors for the ARIMA method for the cases considered, with the highest cumulative error accuracy of 0.987, maximum FRMSE of 1.36 and minimum forecast root mean square error of 0.053. The decline curve method has the highest cumulative error accuracy of 0.877, maximum FRMSE of 8.270 and minimum FRMSE of 0.431.

Table 5.8 ARIMA versus Decline Curve for Water-Drive Data

Well	Sample size	Decline function	ARIMA model	Remaining reserves		
				Actual	Decline	ARIMA
WD1	60 (50)	Exponential	log Q_t (2,1,0)	3,686	2,225	3,574
WD2	60 (50)	Hyperbolic m = 0.2	log Q_t (1,0,0)	787	2,053	980
WD3	60 (50)	—	(0,1,1)	12,685	—	11,800
WD4	60 (50)	Exponential	(0,1,1)	5,460	7,122	7,047
WD5	60 (50)	Exponential	log Q_t (1,0,1)	13,609	5,710	12,984
WD6	60 (50)	—	(0,1,1)	22,669	—	23,625

Table 5.9 Statistical Error Analysis for Decline Data

Well	Method	Statistical errors					
		E_r	E_a	FRMSE	E_{max}	E_{min}	Q_t
Iberia	ARIMA	0.0355	0.179	1.120	3.196	0.0083	0.987
	Decline	–0.5198	0.582	1.281	3.520	0.0077	0.877
Yentzen	ARIMA	0.0401	0.354	0.622	1.442	0.0388	0.700
	Decline	0.3124	0.588	0.661	1.023	0.1942	0.487
Brock	ARIMA	0.0581	0.129	0.184	0.351	0.0020	0.897
	Decline	–0.3997	0.400	0.562	1.260	0.0663	0.789
Linga	ARIMA	–0.1469	0.338	0.627	1.670	0.0804	0.966
	Decline	0.0533	0.359	0.507	1.201	0.0111	0.858
Marg	ARIMA	–0.2070	0.225	0.365	0.786	0.0045	0.882
	Decline	0.4113	0.411	0.451	0.553	0.1963	0.565
Davie	ARIMA	–0.5650	0.659	1.366	3.000	0.0370	0.553
	Decline	–5.7110	5.711	8.270	16.600	2.6920	0.221

Therefore, for this case, as well as for well-to-well comparison of statistical errors for both methods, ARIMA performs better than decline curve method.

The preceding analyses show that a time series approach using ARIMA method is applicable for estimating crude oil reserves. Using various statistical error analysis tools, the ARIMA method provides better

Table 5.10 Statistical Error Analysis for Water-Drive Data

		Statistical errors					
Well	Method	E_r	E_a	FRMSE	E_{max}	E_{min}	Q_t
WD1	ARIMA	0.0151	0.107	0.162	0.261	0.2670	0.970
	Decline	0.3746	0.375	0.431	0.604	0.1605	0.604
WD2	ARIMA	0.2701	0.288	0.494	1.229	0.0313	0.803
	Decline	1.8570	1.857	2.100	2.690	0.6096	0.383
WD3	ARIMA	0.0627	0.063	0.080	0.097	0.0050	0.930
	Decline	—	—	—	—	—	—
WD4	ARIMA	0.3774	0.415	0.527	0.774	0.0688	0.775
	Decline	0.3934	0.427	0.427	0.799	0.0594	0.767
WD5	ARIMA	0.0422	0.102	0.125	0.176	0.0205	0.954
	Decline	0.5802	0.580	0.616	0.609	0.5048	0.420
WD6	ARIMA	0.0437	0.056	0.630	0.090	0.0198	0.960
	Decline	—	—	—	—	—	—

reserve estimates than decline curve method. ARIMA is also applicable where decline curve fails, such as in water-drive reservoirs.

The ARIMA method can be extended to a multi-well system, where data are collected from more than one well in the same area. This extension can make use of the multivariate version of the ARIMA method as presented by Lepak and Considine (1989).

References

Al-Marhoun, M. A., PVT correlations for Middle East crude oils, *Journal of Petroleum Technology*, 40 (5), 650–666, 1988.

Arps, J. J., Analysis of decline curves, *Transactions of the AIME*, 160, 228–247, 1945.

Ayeni, B. J., Parameter estimation for hyperbolic decline curve, *Journal of Energy Resources Technology*, 3 (12), 279–283, 1989.

Badiru, A. B. and B. J. Ayeni, Practitioner's Guide to Quality and Process Improvement, Chapman & Hall, London, UK, 1993.

Box, G. E. P. and Jenkins, G. M., *Time Series Analysis: Forecasting and Control*. Holden Day, San Francisco, CA, 1976, 575pp.

Cole, F. W., *Reservoir Engineering Manual*. 2nd edn., Gulf Publishing, Houston, TX, 1983, pp. 1–347.

Fetkovitch, M. J., Decline curve analysis using type curves, *Journal of Petroleum Technology*, 32 (6), 1067–1077, 1980.

Gentry, R. W., Decline curve analysis, *Journal of Petroleum Technology*, 24 (1), 38–41, 1972.

Lepak, G. M. and Considine, J. J., Testing causality between two vectors in multivariate ARMA models. Am. Stat. Assoc., Proc. Business and Economic Statistics, pp. 116–119, 1989.

Ramsey, H. J. and Guerrero, E. T., The ability of rate-time decline curves to predict production rates, *Journal of Petroleum Technology*, 21 (2), 139–141, 1969.
Research Services, *User's Manual for Time Machine*, Ogden, UT, Chapters 9 and 10, 1986.
Slider, H. C., A simplified method of hyperbolic decline curve analysis, *Journal of Petroleum Technology*, 20 (3), 38–41, 1968.
The Logic Group, *User's Manual for Decline Curve Analysis*, Austin, TX, 1987, pp. 1–20.

chapter six

Design of experiment techniques

This chapter presents additional statistical tools for quality and process improvement. These additional tools can be used to reduce variability and optimize manufacturing processes. Topics covered in this chapter include factorial designs, response surface methodology, central composite designs, and response surface optimization. The chapter is designed for practitioners who need more specialized tools for improving manufacturing processes.

Factorial designs

A factorial experiment is an experiment designed to study the effects of two or more factors, each of which is applied at two or more levels. In a balanced classical factorial experiment, all combinations of all the levels of the factors are tested. In this chapter, we consider only two-level factorial designs. In a 2k factorial design, 2 represents the number of levels and k represents the number of factors.

A factorial experiment can be used to study how a response variable is influenced by certain factors. It can be used to assess the effect of changing one factor independent of other factors. The premise of factorial experiments is that an observed response may be due to a multitude of factors. Since a dependent variable interacts with its environment, it is important to assess the simultaneous effects of more than one factor on the dependent variable. The following example shows a case where one response variable is influenced by two independent factors. Each factor is to be studied at three different levels:

Response variable: Epoxy strength
Factor 1: Temperature (75°, 80°, 85°)
Factor 2: Chemical concentration (high, medium, low)

Advantages of factorial experiment
A factorial experiment has several advantages:

Efficiency

- More robust compared to traditional single-factor experiments.
- In one-factor study, it may be difficult to identify which one factor should be studied.
- More flexibility.

Information content

- More information can be derived from factorial experiments compared to single-factor experiments.

Validity of results

- Inclusion of multiple factors increases the validity of results.
- Results can identify direction for further experiments.

Factor
This is an independent variable or condition that is likely to affect the response or quality characteristic of interest. A factor may be a continuous variable such as oven temperature, RPM, pump pressure, Webspeed, etc., or may be discrete (qualitative) variable such as catalyst type (A or B), valve (on or off), material type (A or B), cooling step (wet or dry), etc. Temperature, pressure, RPM, etc., are factors that can be controlled and measured. Therefore, they can be regarded as controllable and measurable factors. However, factors such as percent moisture going into an oven and ambient humidity are measurable but uncontrollable. These factors are known as covariates. Other factors which are uncontrollable and immeasurable are useful in defining experimental error.

Levels
These are the settings of various factors in a factorial experiment such as high and low values of temperature, pressure, etc. For example, if the range of temperature to be studied is between 120°F and 180°F, then the low level can be set at 120°F and high level set at 180°F.

Response
Response is the measurement obtained when an experiment is run at each level of the factors under study. Responses may be continuous (quantitative) variables such as adhesion, percent yield, smoothness, etc., or discrete (qualitative) variables such as good or bad tastes, corrosion or no corrosion, etc. Rating scales can be used for qualitative variables as can be seen later in this chapter.

The basic layout of a factorial design is presented in Figure 6.1 for a two-factor experiment. Factor A has a levels. Factor B has b levels. There are n replicates for each cell. Each cell in the layout is referred to as a treatment which represents a specific combination of factor levels.

There are a total of N = abn observations in the layout. Factorial designs are referred to as 2^f, 3^f, and so on. In a 2^f design, there are f factors, each having two levels. In a 3^f design, there are f factors, each having three levels.

There are three possible models for a factorial experiment depending on how the factor levels are chosen.

	Factor A levels				
Factor B levels	**(*sums, averages, etc.*)**				***Row summary***
	$i = 1$	$i = 2$	....	$i = a$	
$j = 1$	y_{111} y_{112} . . y_{11n}				α_1 α_2 . . .
$j = 2$			y_{ijk}		.
$j = 3$					. . .
$j = b$					α_b
Column summary	β_1	β_2	....	β_a	....

Figure 6.1 Layout of data collection for a two-factor factorial design.

Fixed model
In this model, all the levels of the factors in the experiment are fixed.

Random model
In this model, the levels of the factors in the experiment are chosen at random.

Mixed model
In this model, the levels of some of the factors in the experiment are fixed while the levels of some of the factors are fixed.

The statistical model for the factorial experiment in Figure 6.1 is presented as follows:

$$Y_{ijk} = \mathtt{m} + A_i + B_j + (AB)_{ij} + \mathtt{e}_{k(ij)}$$

where

i = 1, 2,..., a
j = 1, 2,..., b
k = 1, 2,..., n
A_i is the effect of the ith level of factor A
B_j is the effect of the jth level of factor B
$(AB)_{ij}$ is the effect of the interaction between A_i and B_j
Y_{ijk} is the observation for the *k*th replicate of the Ai and Bj combination
$\varepsilon_{k(ij)}$ is the random error associated with each unique combination of A_i and B_j
μ is the population mean

The error terms are assumed to be independent and identically distributed normal variates with mean of zero and variance σ_ε^2.

Experimental run

A run is when each control factor is set or fixed at a specific level and the experiment is run at those levels for the factors under study. For example, if an experimenter selects a pressure of 20 psi, a temperature of 150°F and a valve that is open, then this combination will represent a run.

One-variable-at-a-time experimentation

A one-variable-at-a-time experiment can be demonstrated by considering, say, three factors A, B, and C. Let us say initially, each factor is set at their low levels. This means that A is at low, B is at low, and C is at low. When this level is run, the response y is 45. Now to determine the effect of factor A, the level of A is changed from low to high, and factors B and C still remained at low. Under this condition, the response value y is, say, 20.

The change in the response value from 45 when all the three factors were set at their low levels to 20 when only factor A was changed from low to its high level can only be due to the effect of factor A and/or experimental error. Similarly, the investigator can determine the effect of factor B by setting A at low, B at high, and C still at low, and obtains a response value at this new level, say, 39. The investigator can now compare this 39–45 obtained for the control (run #1) to determine the effect of factor B. The setup can be described as shown in Figure 6.2. This is what is known as one-variable-at-a-time experiment.

When experimenting with more than one factor, the one-variable-at-a-time experimental approach is inefficient and can provide misleading results. This type of experimentation has several serious problems. Some of these problems are

- The effect of each factor is known at only one chosen level of each of the other factors.
- The effect of each factor is separated in time from the effect of other factors. Unknown extraneous factors which vary with time may, therefore, influence or bias the real effect of any factor under study.

	Factors				Response
Run	**A**	**B**	**C**		**Y**
1	–1	–1	–1	**(control)**	45
2	1	–1	–1		20
3	–1	1	–1		39
4	–1	–1	1		52

Figure 6.2 Design setup for one-variable-at-a-time experiment.

Let us now consider a larger picture of a one-variable-at-a-time experiment. The objective of this experiment is to minimize MAG, an undesirable tar-like by-product. The investigator considered two variables, temperature and concentration, and studied the effects of these two variables on the response MAG. By setting all factors constant including the temperature and allowing the concentration to vary between 10% and 90%, the result (Figure 6.3) shows that MAG is minimum at about 28% concentration.

Then the investigator held the concentration constant at 28% and held all other variables constant as well, but varied temperature between 20°F and 120°F. The result (Figure 6.4) shows that MAG is minimized at a

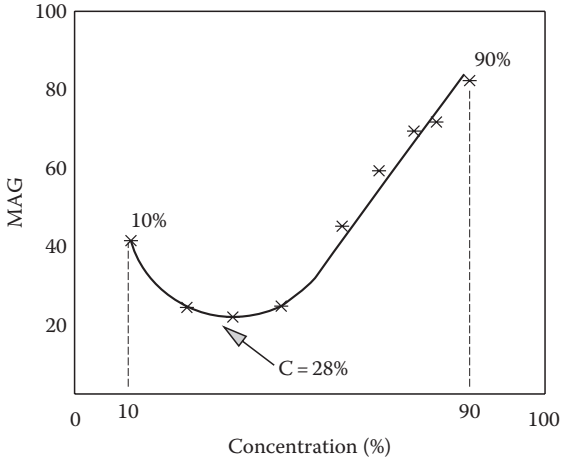


Figure 6.3 Effect of concentration on MAG.

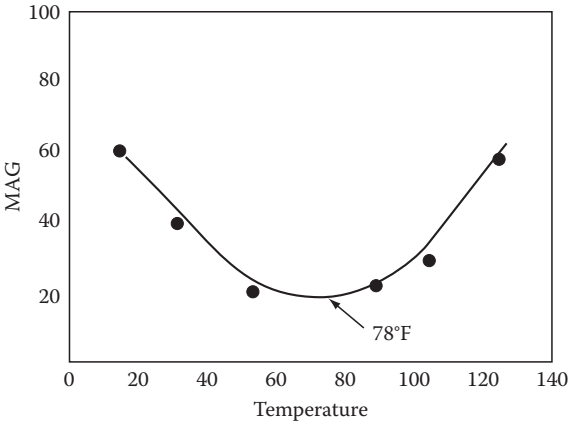


Figure 6.4 Effect of temperature on MAG.

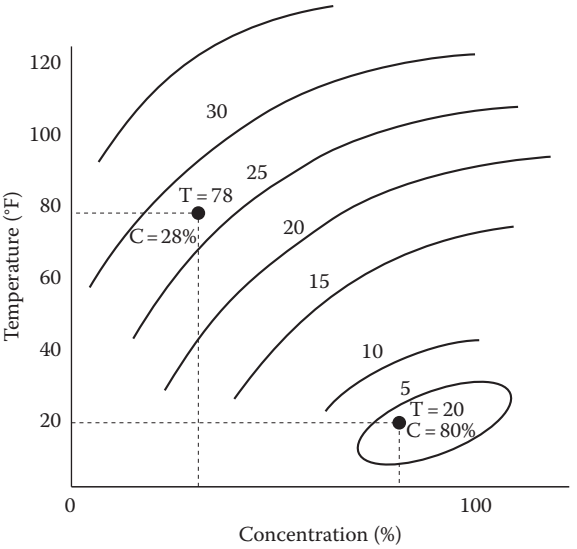


Figure 6.5 Contour plot of MAG with respect to temperature and concentration.

temperature of 76°F. From this result one can conclude that the minimum MAG we can obtain is 20 at a temperature of 76°F and concentration of 28%.

Figure 6.5 shows a contour plot of temperature and concentration with MAG values plotted within the experimental region. As can be seen, a minimum MAG value of 5 can be achieved at a region of 80% concentration and 40°F temperature. This example, therefore, demonstrates how a one-variable-at-a-time approach can fail to estimate the effects between two or more factors because the effect of temperature depends on the levels of concentration in this example. This effect between factors is known as interaction, and one-variable-at-a-time experimental approach lacks the capability of detecting it.

Factorial experiment is superior to one-variable-at-a-time experiment because of the following reasons:

- It allows the study of effects of several factors in the same set of experiment.
- It provides the ability to test for the effect of each factor at all levels of the other factors and determine if this effect changes as the other factors change.
- It is capable of providing not only estimates of the effects separately (main effects) but also the joint effects of two or more factors (interaction effects).
- It provides a complete picture of what is happening over the entire experimental region than one variable at a time.

A good factorial experiment should incorporate basic experimental concepts such as randomization, replication, and orthogonality, as well as the iterative nature of experimentation such as conjecture, design, and analysis.

Randomization
Randomization in a design means running the order of experiment in a random (nonsystematic) fashion. This eliminates or balances out the effects of undesirable systematic variation.

Replication
Replication is the running of the same set of conditions more than once. It is very important that for a true replication to occur, and be distinguished from duplication, one should run the actual set of the condition to be replicated first, record the response, then change at least one or more of the levels, run the experiment at the new levels, and record the response, then come back and run the actual set of the replication again. By running replicate conditions back to back, one would be unable to account for variations that occur due to changes in raw material, operators, etc. In the analysis of the experimental results, it is also important to have an estimate of experimental error (random error) so as to have a meaningful yardstick for determining if estimated effects are real or due to common causes of variability only. Replication runs can be used to provide the estimate of the experimental error.

Orthogonality
Orthogonality in a design implies that the estimates of the main effects and interactions are uncorrelated with each other. Designs having this property insure that if a systematic change occurs corresponding to any one of the effects, the change will be associated to that effect alone.

Experimenting with two factors: 2^2 design

Conjecture
A process engineer wants to investigate the effects of two elements, nickel and gold, on the ductility of a new product. The ranges for these variables are as follows:

	Nickel (%)	Gold (%)
Low (–)	10	5
High (+)	20	10

Table 6.1 Design Matrix for 2 × 2 Study of Ductility

Design point	Coded units			Uncoded units		
	Nickel	Gold	Strength	Nickel	Gold	Strength
1	–1	–1	52	10	5	52
2	1	–1	58	20	5	58
3	–1	1	75	10	10	75
4	1	1	64	20	10	64

The hypotheses that we will be testing are

H_o = Effects are equal to zero (no effects exist)
H_A = Effects are not equal to zero

Design
The design matrix in standard order together with response values for a 2^2 design is presented in Table 6.1.

The design point should not be confused with run order. The design point should always be randomized to obtain the run order. The geometry of the design is presented in Figure 6.6.

Analysis
Estimation of effects: To calculate an effect, we use the following equations:

$$\hat{Y} = \text{response average}$$

$$\text{Effect estimate} = \hat{Y}_{High} - \hat{Y}_{Low}$$

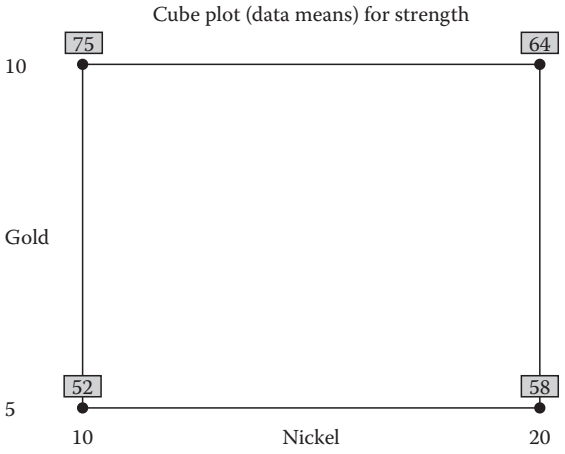


Figure 6.6 Design points geometry for ductility study.

I. To estimate the effect of nickel

$$\text{Effect of nickel} = \frac{58+64}{2} - \frac{52+75}{2}$$
$$= \hat{Y}_H - \hat{Y}_L$$
$$= 61.0 - 63.5$$
$$= -2.5$$

II. To estimate the effect of gold

$$\text{Effect of nickel} = \frac{75+64}{2} - \frac{52+58}{2}$$
$$= 69.5 - 55$$
$$= 14.5$$

Interpretation
When the amount of nickel is changed from 10% to 20%, the effect on average is a reduction of 2.5 units on the breaking strength, while changing the amount of gold from 5% to 10% increases the breaking strength on average by 14.5 units.

Interaction
The interaction effect is the extent to which the effect of a factor depends on the level of another factor.

III. To estimate the interaction effect between nickel and gold obtain the interaction column by multiplying the nickel column with gold column as in Table 6.2.

$$\text{Effect of nickel} \times \text{gold interaction} = \frac{64+52}{2} - \frac{58+75}{2}$$
$$= 58.0 - 66.5$$
$$= -8.5$$

Table 6.2 Interactions Table for Nickel and Gold Effects

	Coded units			
Design point	Nickel	Gold	Nickel* Gold	Strength
1	−1	−1	1	52
2	1	−1	−1	58
3	−1	1	−1	75
4	1	1	1	64

Interpretation
By simultaneously changing the amount of nickel and gold, the net effect on average is a reduction of 8.5 units on the breaking strength.

Replication
In order to determine if the aforementioned effects are real or statistically significant, we must have a good estimate of the experimental error. The investigator, therefore, fully replicated the aforementioned design, obtaining a total of eight runs as shown in Table 6.3.

With the new additional data, one can obtain a refined estimate of effect for each factor under study as

$$\text{Effect of nickel} = \frac{64+66+58+54}{4} - \frac{75+71+52+49}{4}$$
$$= 60.5 - 61.75$$
$$= -2.5$$

$$\text{Effect of gold} = \frac{64+66+74+71}{4} - \frac{58+54+52+49}{4}$$
$$= 69.0 - 53.25$$
$$= 15.75$$

$$\text{Effect of nickel} \times \text{gold interaction} = \frac{64+66+52+49}{4} - \frac{75+71+58+54}{4}$$
$$= 57.75 - 64.5$$
$$= -6.75$$

The interaction plot for the example is presented in Figure 6.7.

Table 6.3 Replication of Nickel and Gold Design

	Coded units			Response	
Design point	Nickel	Gold	Nickel* Gold	Replicate strength	
1	−1	−1	1	52	49
2	1	−1	−1	58	54
3	−1	1	−1	75	71
4	1	1	1	64	66

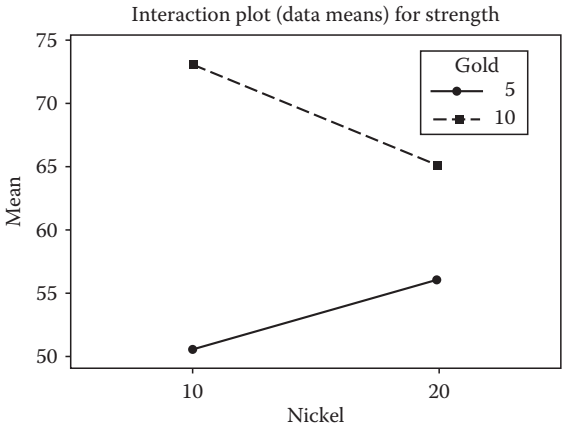


Figure 6.7 Interaction plot for example.

The final model in terms of uncoded factors is

$$\text{Strength} = 9.0 + 1.9 \times \text{Nickel} + 7.2 \times \text{Gold} - 0.27 \times \text{Nickel} \times \text{Gold}$$

The final model in terms of coded factors is

$$\text{Strength} = 61.125 - 0.625 \times \text{Nickel} + 7.875 \times \text{Gold} - 3.375 \times \text{Nickel} \times \text{Gold}$$

Estimate of the experimental error

The aforementioned estimates of the main effects and the interaction are subject to errors. Therefore, by running replicates of the experiment we will be able to estimate the experimental error, and this will provide us the opportunity to interpret the effect estimates in light of the error.

If we assume that the errors made in taking the observations are independent of one another, then we can estimate the experimental error by calculating the variances of the replicate observations within each design point. Table 6.4 shows the results obtained.

If we assume that the variance is homogeneous throughout the experimental region, then we can pool the aforementioned variances. The pooled variance can be calculated as

$$s^2_{pooled} = \frac{v_1 s_1^2 + v_2 s_2^2 + \cdots + v_k s_k^k}{v_1 + v_2 + \cdots + v_k}$$

Table 6.4 Variance of Replicate Observations

Design point	Strength		Variance (s^2)	v = d.f
1	52	49	4.5	1
2	58	54	8	1
3	75	71	8	1
4	64	66	2	1

Therefore,

$$s^2_{pooled} = \frac{(1)4.5 + (1)8 + (1)8 + (1)2}{1+1+1+1}$$

$$= \frac{22.5}{4}$$

$$= 5.625$$

Therefore, the pooled variance is equal to 5.625 with v = 4 degrees of freedom. This pooled variance will be used to construct the confidence intervals about the estimates of effects.

Confidence intervals for the effects

The 95% confidence intervals associated with an effect can be represented as

$$Effect \pm t_{v,0.025}\sqrt{2s^2/n}$$

where n = total number of observations in each average: n = 4, v = 4.

From the t-table,

$$t_{v,0.025} = t_{5,0.025} = 2.776$$

$$s^2 = s^2 \text{ pooled} = 5.625$$

The 95% confidence intervals for the error are

$$\pm 2.776\sqrt{2 \times (5.625)/4} = \pm 4.655$$

I. The confidence intervals for the effect of nickel:
The 95% confidence intervals for the true main effect of nickel are

$$\text{Effect} \pm 4.655 = -1.25 \pm 4.655 \quad \text{or} \quad (-5.905 \text{ to } 3.405)$$

II. The confidence intervals for the effect of gold:
The 95% confidence intervals for the true main effect of gold are

$$\text{Effect} \pm 4.655 = 15.75 \pm 4.655 \quad \text{or} \quad (11.095 \text{ to } 20.405)$$

III. The confidence intervals for the interaction effect:
The 95% confidence intervals for the interaction effect of nickel and gold are

$$\text{Effect} \pm 4.655 = -6.75 \pm 4.655 \quad \text{or} \quad (-11.405 \text{ to } -2.095)$$

From the aforementioned analysis, we can conclude that the main effect of gold is statistically significant at $\alpha = 0.05$. In addition, the interaction between nickel and gold is statistically significant at $\alpha = 0.05$. This is because their confidence intervals do not contain zero. However, the main effect of nickel is not statistically significant at $\alpha = 0.05$ because its confidence intervals contain zero.

Interpretation
Increasing the amount of gold from 5% to 10%, increases the breaking strength by 15.75 units. The practical significance of these 15.75 units needs to be considered further. If this is of practical significance, another experiment on % gold in a different range with some center points may be considered. This new experiment should explore the new range of the % gold at different levels of the % nickel since a significant interaction exists between nickel and gold.

Factorial design for three factors

Conjecture
We wish to study the effects of Webspeed, Voltage, and Webgap on surface roughness of a plating material. It is required that the experiment be capable of estimating all main effects as well as all interactions. Our main objective is to minimize roughness.

Design

A two-level 2^3 full factorial design with some center-point replicates will allow us to investigate the three main effects, three two-factor interactions, and one three-factor interaction. The center points will serve two purposes:

- To enable us to test for curvature effect
- To obtain experimental errors if replicated

Due to the limited resources, a full 2^3 design with four center points will be considered. These four center points will give us 3 degrees of freedom for our estimate of experimental error. For all practical purposes, 3 degrees of freedom should be considered minimum for error degrees of freedom. The conditions for the factors are shown in Table 6.5.

The design matrix in standard order is presented in Figure 6.8.

The aforementioned design is randomized and run. This design is represented geometrically as a cube shown in Figure 6.9.

Table 6.5 Factor Conditions for Plating Study

	Low	Center	High
	(−)	(0)	(+)
Webspeed	20	30	40
Voltage	20	25	30
Webgap	10	20	30

	Speed (S)	Voltage (V)	Gap (G)	S × V	S × G	V × G	S × V × G	Roughness
1.	−1	−1	−1	1	1	1	−1	30
2.	1	−1	−1	−1	−1	1	1	19
3.	−1	1	−1	−1	1	−1	1	37
4.	1	1	−1	1	−1	−1	−1	19
5.	−1	−1	1	1	−1	−1	1	21
6.	1	−1	1	−1	1	−1	−1	18
7.	−1	1	1	−1	−1	1	−1	34
8.	1	1	1	1	1	1	1	20
9.	0	0	0	0	0	0	0	24
10.	0	0	0	0	0	0	0	22
11.	0	0	0	0	0	0	0	23
12.	0	0	0	0	0	0	0	21

Figure 6.8 Design matrix for plating study.

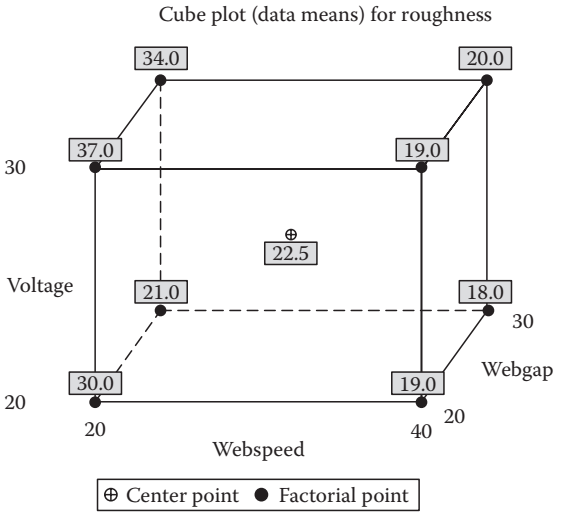


Figure 6.9 Cube plot for plating example.

Analysis

The results obtained for the example are provided in Table 6.6. The analysis of variance (ANOVA) table is shown in Tables 6.7 through 6.8.

The final equation in terms of coded variables is as follows:

$$\text{Roughness} = 24.750 - 5.750 \times A + 2.750 \times B - 1.500 \times C - 2.250 \times A \times B + 1.500 \times A \times C + 1.000 \times B \times C - 0.500 \times A \times B \times C$$

Table 6.6 Experimental Results for Plating Example Part (a)

Variable	Coefficient	Standardized effect	Sum of squares
Overall average	24.00	—	—
A	–5.75	–11.50	264.50
B	2.75	5.50	60.50
C	–1.50	–3.00	18.00
AB	–2.25	–4.50	40.50
AC	1.50	3.00	18.00
BC	1.00	2.00	8.00
ABC	–0.50	–1.00	2.00
Center point	–2.25	—	13.50

Table 6.7 ANOVA Result for Plating Example Part (b)

Source	Sum of squares	df	Square	Mean value	F Prob > F
Model	411.50	7	58.79	35.27	0.0070
Curvature	13.50	1	13.50	8.100	0.0653
Residual	5.00	3	1.67		
Pure error	5.00	3	1.67		
Corr total	430.00	11			
Root MSE = 1.291					
R-squared = 0.9880					
Adjusted R-squared = 0.9600					
C.V. = 5.38					

Table 6.8 Experimental Results for Plating Example Part (c)

Variable	Coefficient estimate	df	Standard error	t for H0 coefficient = 0	Prob > |t|
Intercept	24.750	1	0.456		
A	−5.750	1	0.456	−12.60	0.0011
B	2.750	1	0.456	6.025	0.0092
C	−1.500	1	0.456	−3.286	0.0462
AB	−2.250	1	0.456	−4.930	0.0160
AC	1.500	1	0.456	3.286	0.0462
BC	1.000	1	0.456	2.191	0.1162
ABC	−0.500	1	0.456	−1.095	0.3534
Center point	−2.250	1	0.791	−2.846	0.0653

Final equation in terms of uncoded variables:

$$\begin{aligned}\text{Roughness} = {} & 31.500 - 0.250 \times \text{Webspeed} + 0.900 \times \text{Voltage} - 1.850 \\ & \times \text{Webgap} - 0.025 \times \text{Webspeed} \times \text{Voltage} + 0.040 \times \text{Webspeed} \\ & \times \text{Webgap} + 0.050 \times \text{Voltage} \times \text{Webgap} - 0.001 \times \text{Webspeed} \\ & \times \text{Voltage} \times \text{Webgap}\end{aligned}$$

Complete analysis of the example indicates that high Webspeed, low voltage, and high webgap can be used to minimize surface roughness.

Fractional factorial experiments

For many practical situations, it may be impossible to collect all the observations required by a full factorial experiment. In such cases, fractional factorial experiments are used. In a fractional factorial experiment, only a fraction of the treatment replicates are run. The advantages of fractional factorial experiments include the following:

- Lower cost of experimentation
- Reduced time for experimentation
- Efficiency of analysis

When the number of factors, k, is greater than 5 ($k > 5$), then the number of runs required for a full factorial experiment will be impractical for most industrial applications. However, in some industries, such as semiconductor industry, where computer simulations are used in the design phase of certain products, the number of runs will be of little concern. Whereas, in most other industries, such as chemical, petrochemical, process, paper and pulp as well as parts industries, where a large number of factors must be examined, it will often be very desirable to reduce the number of runs in an experiment by taking a fraction of the full 2^k factorial design.

In a 2^{k-p} fractional factorial design there are

- Two levels of each factor under consideration
- k number of factors to be studied
- 2^{k-p} that can be estimated including the overall mean
- 2^{k-p} minimum number of experimental runs
- p number of independent generators
- 2^{k-p} number of words in defining relations (including I)

The requirement is that $2^{k-p} > k$

p = degree of fractionation
p = 1 (half fraction)
p = 2 (quarter fraction)
2^{k-p} = number of distinct conditions in the cube portion of a design

For example, a one-half of a 2^3 factorial design is referred to as 2^{3-1} fractional factorial design.

The disadvantages of fractional designs involve loss of one or more of the interaction effects that can be studied in a full factorial design. Also, the design of a fractional factorial experiment can be complicated since it may be difficult to select the treatment combinations to be used.

Fractional factorial designs are denoted as follows:

1/2 fractional design: one-half of complete factorial experiment
1/4 fractional design: one-fourth of complete factorial experiment
1/8 fractional design: one-eighth of complete factorial experiment

A 2^4 factorial design

Conjecture
We want to investigate all combinations of two levels of each of four factors, A, B, C, and D, and obtain estimates of all effects including all interactions. We may wish to include some center points or replicate points to obtain estimate of the experimental error.

Design
The design can be set up as shown in Figure 6.10.

The analysis of the aforementioned design will provide uncorrelated and independent estimates of the following:

- The overall average of the response
- The main effects due to each factor, A, B, C, D
- The estimates of 6 two-factor interaction effects, AB, AC, AD, BC, BD, CD
- The estimates of 4 three-factor interaction effects, ABC, ABD, ACD, BCD
- The estimate of one four-factor interaction effect, ABCD

Full factorial designs can be generated for any number of factors, k. However, it should be noted that the number of runs needed for a full factorial design increases rapidly with increasing values of k as shown in Figure 6.10.

Conjecture
An investigator wishes to investigate the effects of four factors (k = 4) using only eight runs.

Design
A 2^{4-1} fractional factorial design of eight runs can be set up as illustrated in the following procedure:

I. Set up a full 2^3 design of eight runs as shown in Figure 6.11.
II. Assign the fourth factor D to the highest order interaction as shown in Figure 6.12.

$$D = ABC$$

D = ABC is known as generator

III. Generate the + and – column for D by multiplying columns A, B, and C together to obtain the layout in Figure 6.12. Note that the levels of factor D are the products of the levels of factors A, B, and C.

Design point	A	B	C	D	AB	AC	AD	BC	BD	CD	ABC	ABD	ACD	BCD	ABCD
1	−	−	−	−	+	+	+	+	+	+	−	−	−	−	+
2	+	−	−	−	−	−	−	+	+	+	+	+	+	−	−
3	−	+	−	−	−	+	+	−	−	+	+	+	−	+	−
4	+	+	−	−	+	−	−	−	−	+	−	−	+	+	+
5	−	−	+	−	+	−	+	−	+	−	−	−	+	+	−
6	+	−	+	−	−	+	−	−	+	−	−	+	−	+	+
7	−	+	+	−	−	−	+	+	−	−	−	+	+	−	+
8	+	+	+	−	+	+	−	+	−	−	+	−	−	−	−
9	−	−	−	+	+	+	−	+	−	−	−	+	+	+	−
10	+	−	−	+	−	−	+	+	−	−	+	−	−	+	+
11	−	+	−	+	−	+	−	−	+	−	+	−	+	−	+
12	+	+	−	+	+	−	+	−	+	−	−	+	−	−	−
13	−	−	+	+	+	−	−	−	−	+	+	+	−	−	+
14	+	−	+	+	−	+	+	−	−	+	−	−	+	−	−
15	−	+	+	+	−	−	−	+	+	+	−	−	−	+	−
16	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Number of Factors k	Number of Experimental Runs
2	4
3	8
4	16
5	32
6	64
7	128
8	256
9	512
10	1024
11	2048
12	4096
13	8192
14	16364
15	32768

Figure 6.10 Design setup for 2^4 factorial design.

Design point	A	B	C
1	−	−	−
2	+	−	−
3	−	+	−
4	+	+	−
5	−	−	+
6	+	−	+
7	−	+	+
8	+	+	+

Figure 6.11 Setup of full 2^3 design.

Design point	A	B	C	D = ABC
1	−	−	−	−
2	+	−	−	+
3	−	+	−	+
4	+	+	−	−
5	−	−	+	+
6	+	−	+	−
7	−	+	+	−
8	+	+	+	+

Figure 6.12 Layout for design points.

IV. Obtain the defining relation (I) by multiplying both sides of the generator by D as

$$D \times D = ABC \times D$$

$$I = ABCD \text{ (A resolution IV design)}$$

The defining relation will be used to determine the resolution of a design and to generate the confounding patterns of the effects.

Design resolution

The resolution of a design is the number of letters in the smallest word of the defining relation. For example, in the aforementioned design, the defining relation I = ABCD has one word, which is ABCD, and the number of letters is equal to 4, hence a resolution IV design. In a resolution IV design, the main effects are confounded with three-factor and higher-order interactions, and two-factor interactions are confounded with each other and higher-order interactions. Similarly, a resolution III design has the main effects confounded with two-factor and higher-order interactions. A 2^{5-2} design of 8 runs is of resolution III. An extremely useful design is a 2^{5-1} design of resolution V.

This design is used to investigate five factors in only 16 runs and has the power to estimate the main effects clear of any other factors and the two-factor interactions clear of any other factors as well if three-factor and higher-order interactions are assumed to be negligible. Additional information on design resolution can be found in Box et al. (1978), Ayeni (1991), as well as in many other experimental design books and papers.

The defining relation (I) is the column of +1. Any factor multiplied by itself gives I. For example,

$$A \times A = I, \quad B \times B = I, \quad AB \times AB = AABB = II = I,$$

$$ABC \times ABC = AABBCC = III = I$$

Other operators which are not equal to I are

$$A \times AB = B = IB = B, \quad AB \times BCE = AICE = ACE$$

V. Use the defining relation I = ABCD to generate the effects or confounding patterns as follows:
   1. To obtain the confounding patterns for main effect A, multiply both sides of the defining relation by A as

$$A \times I = A \times ABCD = BCD$$

      Therefore, A = BCD. This means that when we estimate the effect of factor A, we are not only estimating the effect of A but also the effect of the three-factor interaction BCD. This is known as confounding.

      Confounding occurs when the effects of two or more factors cannot be separated. In the aforementioned example, we are really estimating the sum of two effects A + BCD.
   2. Similarly, to obtain any two-factor interaction, say, BC, multiply both sides of defining relation by BC as

$$BC \times I = BC \times ABCD = AD$$

      Therefore, BC = AD, that is, the effects of BC and AD are confounded with each other. Thus, estimating the effect of BC implies that we are really estimating the sum of BC + AD.

The complete confounding patterns are shown in Figure 6.13.

Effect	2_{IV}^{4-1}	Confounding patterns
A		A + BCD
B		B + ACD
C		C + ABD
D		D + ABC
AB		AB + CD
AC		AC + BD
BC		BC + AD

Figure 6.13 Confounding patterns for factorial design example.

From the aforementioned main effects, we can see that when we believe we are estimating the main effects, we are actually estimating the sum of the main effects and the three-factor interactions. However, since three-factor interactions and higher-order interactions are generally assumed to be negligible or nonexistence, we can obtain estimates of all the main effects clear of all other effects. Although, all the two-factor interactions are confounded with each other, this is the price we pay in running 8 rather than 16 experiments. Unless certain two-factor interactions are known not to exist, it will be necessary to run another half-fraction if each two-factor effect is to be estimated clear of all other effects. Interested readers should refer to Hicks (1982) for further details on factorial designs and fractional factorial designs.

Saturated designs

Saturated designs are designs that can be used to investigate n – 1 factors in *n* number of runs. For example, one can study the effects of 15 factors using only 16 runs. In fact, it is also possible to investigate the effects of 31 factors using 32 runs. These designs are extremely useful in screening applications as well as in situations where main effects are believed to dominate over two-factor and higher-order interactions. All saturated designs are of resolution III. This means that the main effects are confounded with two-factor and higher-order interactions. It is, therefore, extremely important to initially screen for factors that are critical to the response of interest, since only a few of these factors exist, and later conduct a more thorough investigation of those factors identified through a full factorial design or a central composite design. Examples of saturated designs are shown in Figure 6.14.

Example of a saturated design

A 2^{7-4} fractional factorial design can be used to investigate k = 7 factors with only eight runs. For this design, the number of fractions is p = 4. In any saturated design, all possible interactions are used up in building

Number of factors	Number of runs	Type of design
3	4	2^{3-1}
7	8	2^{7-4}
15	16	2^{15-11}
31	32	2^{31-26}
63	64	2^{63-57}

Figure 6.14 Examples of saturated designs.

Design point	A	B	C
1	−	−	−
2	+	−	−
3	−	+	−
4	+	+	−
5	−	−	+
6	+	−	+
7	−	+	+
8	+	+	+

Figure 6.15 Design setup for a full 2^3 design.

the generators. Since p = 4 in a 2^{7-4} design, there will be a total of four generators. The design setup is provided in Figure 6.15.

1. Procedure Step I: Set up a full 2^3 design of 8 runs (see Figure 6.15).
2. Procedure Step II: Assign the remaining four variables (D, E, F, G) to all the interactions to obtain the four generators.

   D = AB E = AC F = BC and G = ABC (these are the generators)

3. Procedure Step III: Generate the + and − columns for the generators as shown in Figure 6.16.
4. Procedure Step IV: Generate the defining relation (I) as presented in the following. Notice that the smallest word has three letters, hence a resolution III design.

$$I = ABD = ACE = BCF = ABCG = BCDE = ACDF$$
$$= CDG = ABEF = BEG = AFG = DEF = ADEG$$
$$= BDFG = CEFG = ABCDEFG$$

Design point	A	B	C	D = AB	E = AC	F = BC	G = ABC
1	−	−	−	+	+	+	−
2	+	−	−	−	−	+	+
3	−	+	−	−	+	−	+
4	+	+	−	+	−	−	−
5	−	−	+	+	−	−	+
6	+	−	+	−	+	−	−
7	−	+	+	−	−	+	−
8	+	+	+	+	+	+	+

Figure 6.16 Generation of +/− columns for a 2^{7-4} saturated design.

Effects	2^{7-4}_{III}	Confounding patterns
A		A + BD + CE + FG
B		B + AD + CF + EG
C		C + AE + BF + DG
AB		AB + D + CG + EF
AC		AC + E + BG + DF
BC		BC + F + AG + DE
ABC		CD + BE + AF + G

Figure 6.17 Confounding patterns for saturated design.

5. Procedure Step V: The confounding patterns can be generated as before by multiplying the factors on both sides of the defining relations. The confounding patterns are provided (Figure 6.17) after three-factor and higher-order interactions have been deleted.

Response surface methodology

Response surface methodology involves an analysis of the prediction equation or response surface fitted to a set of experimental data. Response surface strategies can be classified into two categories. These are single-phase and double-phase strategies.

Single-phase strategy

This strategy requires running a full factorial design plus center points and star points to fit a second-order response surface. This is shown graphically in Figure 6.18.

Double-phase strategy

This strategy requires initial running of a full factorial design with some center points. Analyze the data by fitting a first-order model, then test for lack of fit, and use the center points to test for the effects of curvature. If there is a significant lack of fit or if the quadratic effect is significant, or

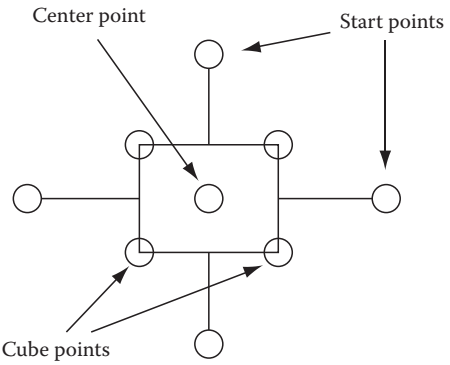


Figure 6.18 Single-phase response surface strategy.

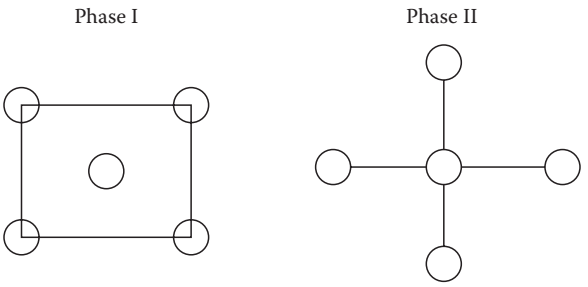


Figure 6.19 Double-phase response surface strategy.

both, then proceed further by running a second design which includes the star points and additional center points. Then analyze the data from the two designs together. A double-phase response surface strategy is depicted in Figure 6.19.

The selection of which design phase to consider depends on several factors which are discussed as follows:

- *The major goal of the experiment*. If the major goal of the experiment is to optimize the process and one is considering only two to three factors to study with no other problems as listed in the following, then one can proceed straight with a single-phase strategy.
- *The cost of running the experiment*. If the cost of running the experiment is a major concern and one is required to minimize cost as much as possible, then select a double-phase strategy. Fitting a first-order model first and further determining through a curvature test that a second-order model is not necessary will not only save you a substantial amount of money but at the same time save you a fairly large amount of experimental time.

- *The number of variables to be studied*. If the number of variables to be studied is greater than five, then select a double-phase strategy and run a fractional factorial of resolution IV or better first.
- *The time required to complete the project*. If longer time is required to complete each run in the experiment, then you may select a double-phase strategy. You will save a lot of time if it is determined that a second-degree model is not necessary.
- *Type of control variables under study*. If all the control variables are qualitative variables, then select double-phase strategy. You only need to run a full or fractional factorial design with some replicates. No star points or center points are possible for this case.
- *Prior knowledge of the experimenter*. If the experimenter (through prior experiment or any other means) knew in advance that within the range of study, the first-order model would be adequate, then select a double phase.
- *Maximum number of runs possible*. If there is a limitation on the number of runs possible, then consider double-phase strategy.

Response surface example
A process engineer has just completed a 2^{6-2} screening design where he studied 6 factors on mineral penetration of a fiber. The response of interest is the fiber thickness. The screening experiment identified three key factors, Webspeed, % solids, and Fiberweight. The engineer decided to determine the optimum operating conditions of these critical factors that can be used to achieve a minimum thickness of 0.40.

Objective
To determine control handles that will achieve a target thickness of 0.40 or better.

Design
A three-factor central composite design as shown below is selected. The ranges of interest to be studied are shown in Table 6.9. The resulting design and the response are presented in Table 6.10. Table 6.11 shows the ANOVA results for thickness in a three-factor study.

Table 6.9 Three-Factor Central Composite Design

Factors	–1.633	–1	0	1	1.633
Webspeed	43.67	50	60	70	76.33
%Solids	36.83	40	45	50	53.16
Fiberweight	16.83	20	25	30	33.16

Table 6.10 Data for Response Surface Example

Obs. no.	Run order	Block	Webspeed X1	% Solids X2	F. weight X3	Thickness (response)	Design ID
1	6	1	50.000	40.000	20.000	0.320	1
2	1	1	70.000	40.000	30.000	0.336	2
3	5	1	50.000	50.000	30.000	0.361	3
4	3	1	70.000	50.000	20.000	0.399	4
5	2	1	60.000	45.000	25.000	0.404	5
6	4	1	60.000	45.000	25.000	0.380	6
7	8	2	50.000	40.000	30.000	0.321	7
8	12	2	70.000	40.000	20.000	0.356	8
9	9	2	50.000	50.000	20.000	0.350	9
10	11	2	70.000	50.000	30.000	0.404	10
11	10	2	60.000	45.000	25.000	0.353	11
12	7	2	60.000	45.000	25.000	0.375	12
13	14	3	43.670	45.000	25.000	0.353	13
14	16	3	76.330	45.000	25.000	0.373	14
15	19	3	60.000	36.835	25.000	0.342	15
16	13	3	60.000	53.165	25.000	0.441	16
17	17	3	60.000	45.000	16.835	0.361	17
18	15	3	60.000	45.000	33.165	0.348	18
19	18	3	60.000	45.000	25.000	0.378	19
20	20	3	60.000	45.000	25.000	0.374	20

Analysis

Analysis of the data in Table 6.10 yields the following results:

- Estimated effects for thickness for a three-factor study

  Average = 0.3776
  A:X1 = 0.0264
  B:X2 = 0.0514
  C:X3 = −0.0036
  AB = 0.0102
  AC = −0.0068
  BC = 0.0088
  AA = −0.0166
  BB = 0.0048
  CC = −0.0230
  Block 1 = 0.001400
  Block 2 = −0.012166
  Block 3 = 0.010666

- Standard error estimated from total error with 8 d.f. (t = 2.30665)

Table 6.11 ANOVA for Thickness for Three-Factor Study

Independent variable	Coefficient estimate	df	Standard error	t for H0 Coeff. = 0	Prob > |t|
Intercept	0.3776	1	0.0066	57.6000	
Block 1	0.0007				
Block 2	−0.0061				
Block 3	0.0053				
A: Webspeed	0.0132	1	0.0044	2.9930	0.0173
B: % Solids	0.0257	1	0.0044	5.8380	0.0004
C: Fiberweight	−0.0018	1	0.0044	−0.4128	0.6906
AA	−0.0083	1	0.0044	−1.8850	0.0962
BB	0.0024	1	0.0044	0.5315	0.6095
CC	−0.0115	1	0.0044	−2.6050	0.0314
AB	0.0051	1	0.0057	0.9017	0.3935
AC	−0.0034	1	0.0057	−0.5938	0.5690
BC	0.0044	1	0.0057	0.7698	0.4635
Total error	0.00207	8			
Total (corr.)	0.01682	19			
R-squared = 0.8737					
R-squared (adj. for d.f.) = 0.7316					

- Regression coefficients for thickness for a three-factor response surface study

  Constant = 0.3776
  Block 1 = 0.0007
  Block 2 = −0.0061
  Block 3 = 0.0053
  A: Webspeed = 0.0132
  B: % solids = 0.0257
  C: Fiberweight = −0.0018
  AA = −0.0083
  BB = 0.0024
  CC = −0.0115
  AB = 0.0051
  AC = −0.0034
  BC = 0.0044

Figure 6.20 shows the Pareto chart for the response (thickness). The chart indicates the relative contributions of the effects from the three-factor interactions. Figure 6.21 shows the response surface with

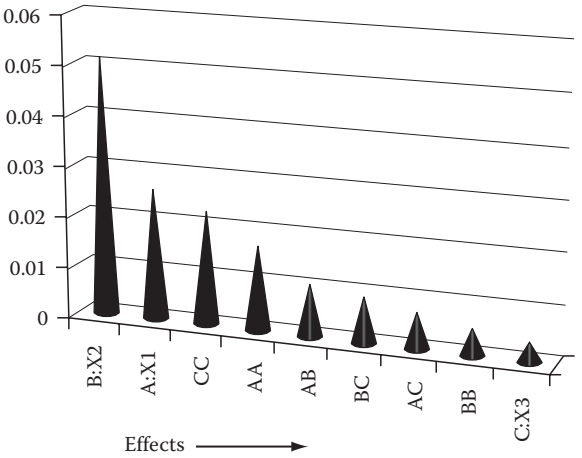


Figure 6.20 Pareto chart for response surface analysis.

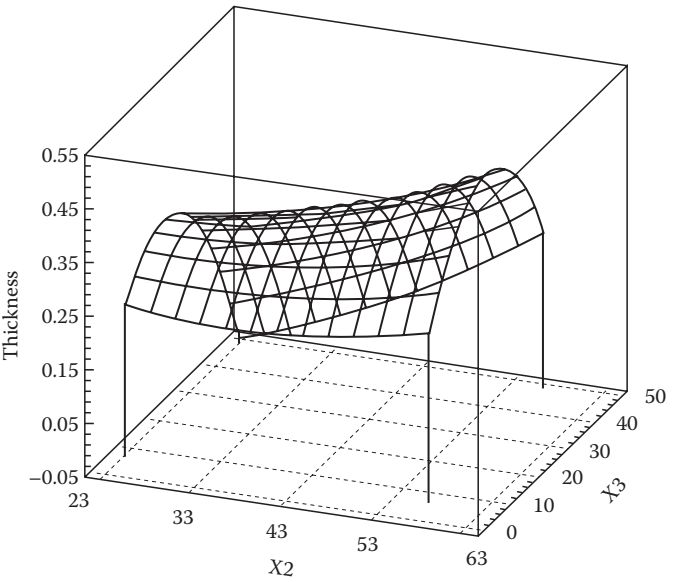


Figure 6.21 Response surface with respect to X2 and X3.

respect to factor X2 and factor X3. Figure 6.22 shows the response surface with respect to X1 and X3. Figure 6.23 shows the contour surface with respect to X1 and X2. Figure 6.24 shows the contour surface with respect to X1 and X3. Figure 6.25 shows the response surface with respect to X1 and X2.

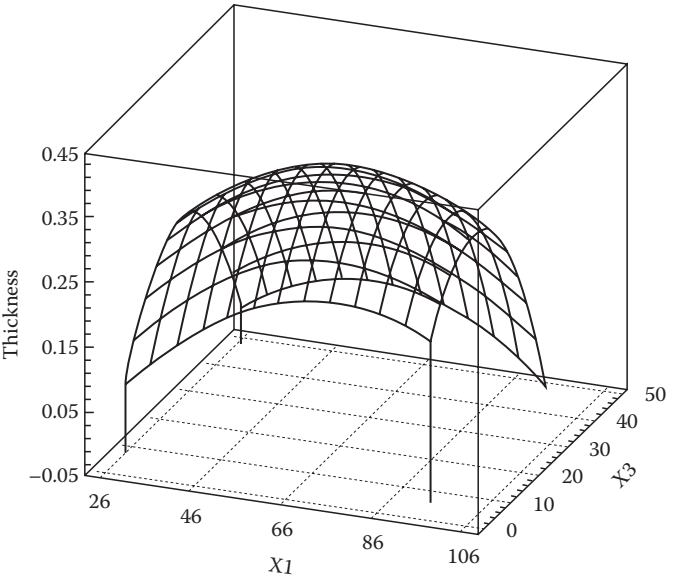


Figure 6.22 Response surface with respect to X1 and X3.

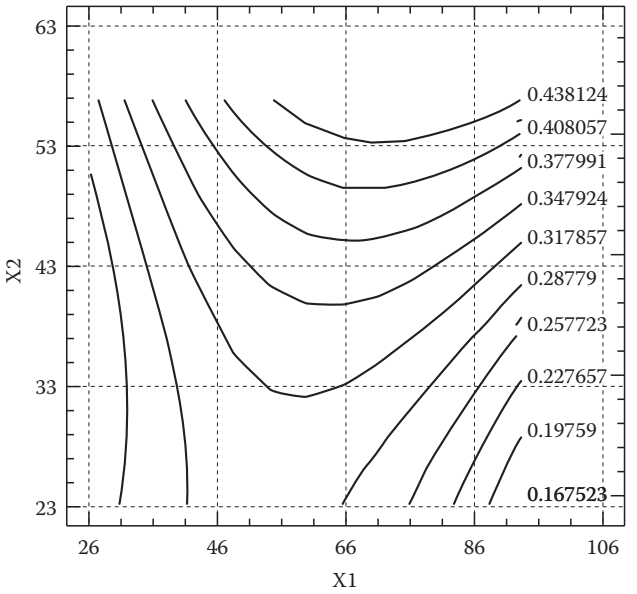


Figure 6.23 Contour surface with respect to X1 and X2.

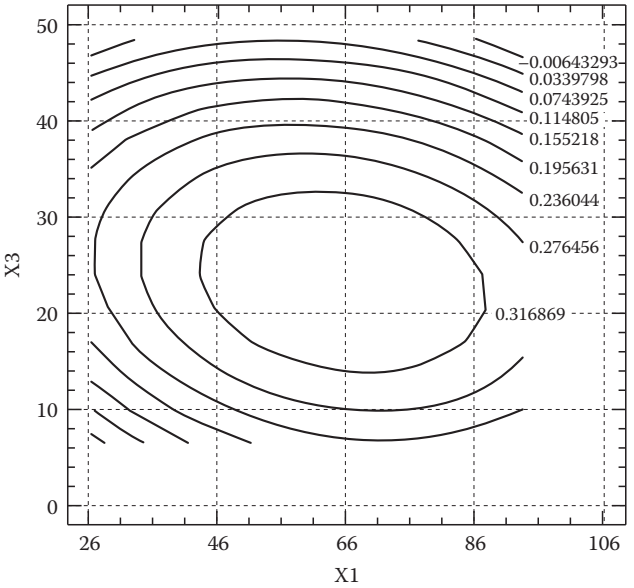


Figure 6.24 Contour surface with respect to X1 and X3.

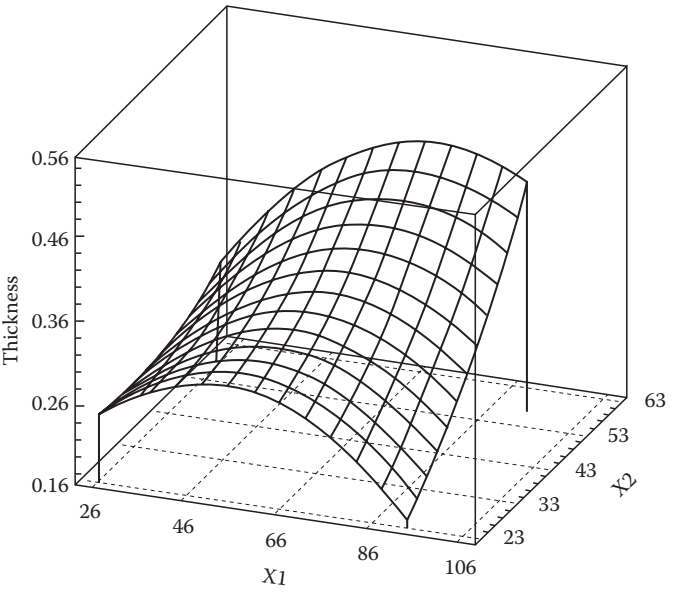


Figure 6.25 Response surface with respect to X1 and X2.

The final equation in terms of actual factors is

$$\begin{aligned}\text{Thickness} = {} & 0.0509 + 0.008396(\text{Webspeed}) - 0.01385(\%\ \text{Solids}) \\ & + 0.01886(\text{Fiberweight}) - 0.0000833(\text{Webspeed}\text{\^{}}2) \\ & + 0.00009404(\%\text{Solid}\text{\^{}}2) - 0.000461(\text{Fiberweight}\text{\^{}}2) \\ & + 0.0001025(\text{Webspeed} \times \%\text{Solids}) - 0.0000675(\text{Webspeed} \\ & \times \text{Fiberweight}) + 0.000175(\%\text{Solids} \times \text{Fiberweight}).\end{aligned}$$

Central composite designs

I. Two-factors central composite design—design orthogonally blocked and rotatable. An example of this is shown in Table 6.12.
II. Two-factors central composite design—factorial portion replicated and design orthogonally blocked and nearly rotatable as shown in Table 6.13.
III. Three-factors central composite design—orthogonally blocked and nearly rotatable as seen in the layout in Table 6.14.
IV. Four-factors central composite design—orthogonally blocked and rotatable as shown in Table 6.15. Figure 6.26 illustrates the graphical composition of a three-factor central composite design.

Response surface optimization

Controlling processes to target and minimize variation have been important issues in recent years in most industrial organizations (Ayeni, 1994). Well-designed experiments can significantly impact product and process quality. This section of the book focuses on some practical issues commonly associated with moving web processes as well as those useful in reducing

Table 6.12 Two-Factor Central Composite Design

Run	A	B	
1	−1	−1	
2	1	−1	
3	−1	1	
4	1	1	Block 1
5	0	0	
6	0	0	
7	−1.414	0	
8	1.414	0	
9	0	−1.414	Block 2
10	0	1.414	
11	0	0	
12	0	0	

Table 6.13 Central Composite Design with Three Blocks

Run	A	B	
1	–1	–1	
2	1	–1	
3	–1	1	Block 1
4	1	1	
5	0	0	
6	0	0	
7	–1	–1	
8	1	–1	
9	–1	1	Block 2
10	1	1	
11	0	0	
12	0	0	
13	–1.633	0	
14	1.633	0	
15	0	–1.633	
16	0	1.633	Block 3
17	0	0	
18	0	0	
19	0	0	
20	0	0	

variability during industrial experimentation. The fundamental principle of model building using two-level factorial and fractional factorial designs will be covered. The practice of adding center points as well as "axial" points for detection of model curvature is explored. Some examples of industrial experiments are presented, including applications involving response surface designs for moving web-type processes with machine direction (MD) and cross-direction (CD), such as paper and plastic film productions. The problems and opportunities provided by simultaneously studying dispersion effects and location effects with the objective of achieving mean on target with minimum variance are investigated with real-life examples. A numerical illustration with the use of contours is produced as a mechanism for process improvement. Multiple response optimization methods are presented as an approach to finding the common region for process optimization.

Applications for moving web processes

Factorial designs have been extensively used in industry over the last several years primarily because of their ability to efficiently provide valuable information about main effects as well as their interactions. In addition,

Table 6.14 Layout for Three-Factor Central Composite Design

Run	A	B	C	
1	−1	−1	1	
2	1	−1	−1	
3	−1	1	−1	Block 1
4	1	1	1	
5	0	0	0	
6	0	0	0	
7	−1	−1	−1	
8	1	−1	1	
9	−1	1	1	Block 2
10	1	1	−1	
11	0	0	0	
12	0	0	0	
13	−1.633	0	0	
14	1.633	0	0	
15	0	−1.633	0	
16	0	1.633	0	
17	0	0	−1.633	Block 3
18	0	0	1.633	
19	0	0	0	
20	0	0	0	

factorial designs often provide the basic foundation for response surface methodology (RSM) and mixture (formulation) experiments. Response surface method is a statistical technique useful for building and exploring the relationship between a response variable, y, and a set of independent factors, x, which can be represented as

$$y = f(x, \mathtt{b}) + \mathtt{e}$$

where

x is a vector of factor settings
β is a vector of parameters including main effects and interactions
ε is a vector of random errors which are assumed to be independent with zero mean and common variance s^2

One is usually interested in optimizing $f(x,\beta)$ over some appropriate design region. This method only approximates the true response surface with a convenient mathematical function. Therefore, for most practical situations, if an appropriate design region has been chosen, a quadratic function will usually approximate the true response surface quite well.

Table 6.15 Layout for Four-Factor Central Composite Design

Run	A	B	C	D	
1	1	1	−1	1	
2	1	−1	−1	−1	
3	−1	1	−1	−1	
4	1	1	1	−1	
5	−1	−1	1	−1	Block 1
6	1	−1	1	1	
7	−1	1	1	1	
8	−1	−1	−1	1	
9	0	0	0	0	
10	0	0	0	0	
11	0	0	0	0	
12	1	1	−1	−1	
13	1	−1	1	−1	
14	−1	1	−1	1	
15	−1	−1	−1	−1	
16	−1	−1	1	1	
17	1	−1	−1	1	Block 2
18	−1	1	1	−1	
19	1	1	1	1	
20	0	0	0	0	
21	0	0	0	0	
22	0	0	0	0	
23	−2	0	0	0	
24	2	0	0	0	
25	0	−2	0	0	
26	0	2	0	0	
27	0	0	−2	0	
28	0	0	2	0	Block 3
29	0	0	0	−2	
30	0	0	0	0	
31	0	0	0	0	
32	0	0	0	0	
33	0	0	0	0	

In industry today, engineers have experienced increasing needs to develop experimental strategies that will achieve target for a quality characteristic of interest while simultaneously minimizing the variance. The classical experimental designs described earlier have been found lacking in this area because they tend to focus solely on the mean of the quality characteristic

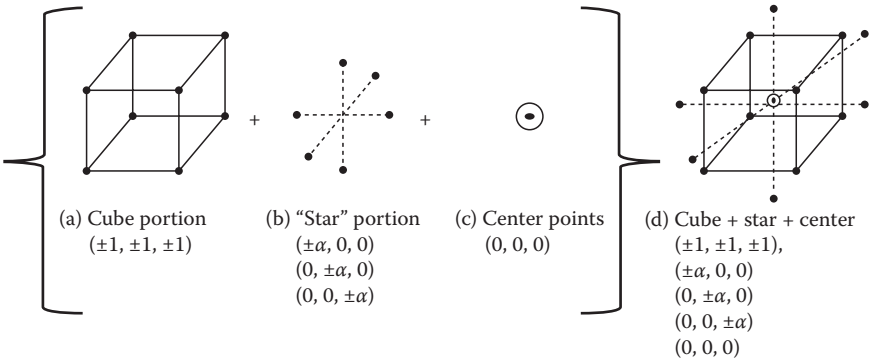


Figure 6.26 Graphical representation of central composite design.

of interest. In today's competitive manufacturing environment, engineers must be more ambitious in finding conditions where variability around the target is as small as possible in addition to meeting the target condition.

Taguchi and Wu (1985) and Taguchi (1986) have presented the need for considering the mean and variance of quality characteristic of interest. Unfortunately, Taguchi's statistical approach to this problem has drawn much criticism in the literature (Box 1985). For this reason, Vining and Myers (1990) developed a dual response approach for which one can achieve the primary goal of the Taguchi philosophy. This enables us to obtain a target condition on the mean while minimizing the variance, within a response surface methodology framework. The example used by Vining and Myers (1990) was taken from the Box et al. (1978) book. Therefore, this chapter attempts to apply this dual response approach to real-life practical examples and problems.

Dual response approach

Vining and Myers (1990) used the dual response problem formulation developed by Myers and Carter (1973). In their development, the investigator is assumed to be seeking to optimize two responses. We shall let y_p represent the response of primary interest, and y_s represent the response of secondary interest. They further assumed that these responses might be modeled by the following equations:

$$y_p = \beta_o + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum \sum_{i<j}^{k} \beta_{ij} x_i x_j + \varepsilon_p$$

$$y_s = \gamma_o + \sum_{i=1}^{k} \gamma_i x_i + \sum_{i=1}^{k} \gamma_{ii} x_i^2 + \sum \sum_{i<j}^{k} \gamma_{ij} x_i x_j + \varepsilon_s,$$

where the β_s and the γ_s represent the unknown coefficients, and ε_p and ε_s are the random errors.

The practical significance of this approach is to optimize the primary response subject to an appropriate constraint on the value of the secondary response. The decision as to which response to make as the primary response depends solely on the ultimate goal of the experiment, by considering Taguchi's three situations:

1. "Target value is best," which means keeping mean at a specified target value, while minimizing variance. This requires that the variance be the primary response.
2. "The larger the better" means making the mean as large as possible while controlling variance.
3. "The smaller the better" means making the mean as small as possible, while controlling the variance. These last two require the mean to be the primary response.

Case study of application to moving webs

In this application, a process capability experiment was performed on a long run of coated webs. This study provided the magnitude of variation arising from both down-web and cross-web effects. The sampled down-web positions were 50 yards apart in every roll, while the sampled cross-web positions were at the left, center, and right sides of the webs. These sampling points were selected for convenience. They were believed to be as representative of the process as any other position which could have been selected. There were several other possible cross-web positions that could have been selected. The process capability study showed that the major source of variability was due to cross-web positions. Therefore, reducing cross-web variability could lead to sizable impact on reducing total variation.

In order to control mineral penetration, eight factors were identified for the study. In addition to controlling the mineral penetration to some target, we were also interested in controlling the variability in the mineral penetration due to cross-web effect. Therefore, a 2^{8-4} fractional factorial design of 16 runs was set up to screen all the eight factors believed to be potentially critical to controlling mineral penetration in the web process. There were no replicates or center points run at this stage in order to minimize the cost and time of running the experiment. Each condition of the experiment was run at each of the three cross-web positions, left, center, and right sides of the web. The order of the experiment was completely randomized and the experiment was run in a random order. The design matrix and the results obtained are provided in Table 6.16.

The analysis of the screening experiment was performed using half-normal plot in Figure 6.27. The result shows that factors G and H are

Table 6.16 2^{8-4} Fractional Factorial Design for Mineral Penetration

Central composite response surface design												Mineral penetration				
Design identification				Factors								Cross-web positions			Responses	
Std order	Run order	Center point	Blocks	A	B	C	D	E	F	G	H	Left	Center	Right	Mean	Std Dev
1	3	1	1	−1	−1	−1	−1	−1	−1	−1	−1	1.7	3.1	1.8	2.2240	0.8050
2	8	1	1	1	−1	−1	−1	1	1	1	−1	3.1	5.3	5.0	4.4440	1.2120
3	12	1	1	−1	1	−1	−1	1	1	−1	1	2.2	1.3	0.9	1.4870	0.6720
4	2	1	1	1	1	−1	−1	−1	−1	1	1	6.0	3.4	7.7	5.6850	2.1940
5	6	1	1	−1	−1	1	−1	−1	1	1	1	6.2	2.8	4.0	4.3410	1.7000
6	9	1	1	1	−1	1	−1	1	−1	−1	1	2.9	2.0	1.9	2.2880	0.5530
7	1	1	1	−1	1	1	−1	1	−1	1	−1	10.2	10.8	8.0	9.6570	1.4810
8	13	1	1	1	1	1	−1	−1	1	−1	−1	1.1	6.2	1.9	3.0830	2.7710
9	14	1	1	−1	−1	−1	1	1	−1	1	1	1.9	3.2	1.2	2.0600	1.0020
10	4	1	1	1	−1	−1	1	−1	1	−1	1	2.1	0.8	1.8	1.6010	0.6960
11	16	1	1	−1	1	−1	1	−1	1	1	−1	3.0	6.4	8.0	5.8040	2.5240
12	5	1	1	1	1	−1	1	1	−1	−1	−1	1.2	5.5	4.9	3.8610	2.3160
13	11	1	1	−1	−1	1	1	1	1	−1	−1	2.4	1.8	2.3	2.1630	0.3340
14	10	1	1	1	−1	1	1	−1	−1	1	−1	5.0	7.0	4.9	5.6410	1.2100
15	15	1	1	−1	1	1	1	−1	−1	−1	1	0.7	1.4	0.7	0.9160	0.3970
16	7	1	1	1	1	1	1	1	1	1	1	2.9	6.5	3.0	4.1450	2.0320

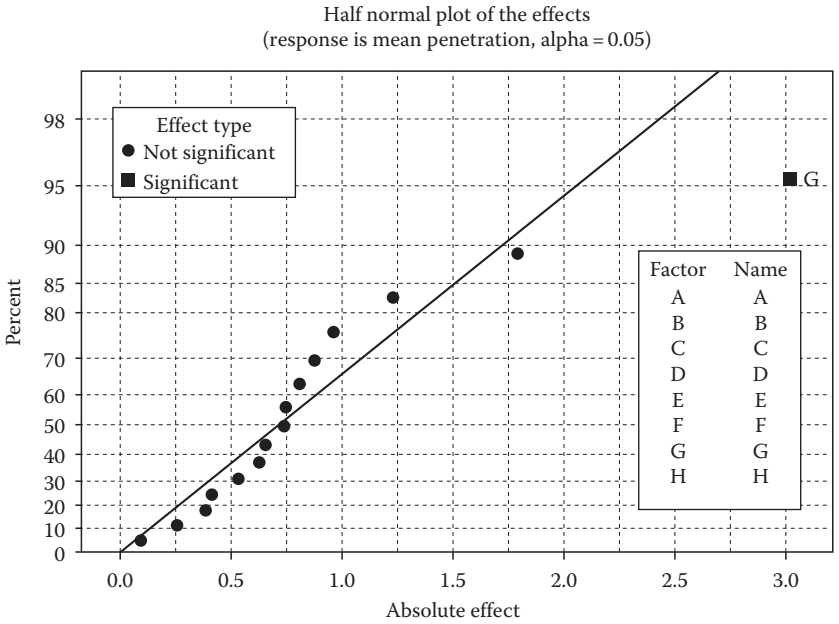


Figure 6.27 Half normal plot for moving webs application.

statistically significant to control the mean mineral penetration. Similarly, factor B is significant to controlling the variability, although the half-normal plot for the standard deviation is not provided in order to minimize space. These three factors were considered for the next phase of the experimentation.

Case application of central composite design

A three-factor central composite design (Figure 6.26) is a member of the most popular class of designs used for estimating the coefficients in the second-order model. This design consists of eight vertices of a 3D cube. The values of the coded factors in this factorial portion of the design are (B, G, H) = (+1, +1, +1). In addition, this design consists of six vertices (+1.63, 0, 0), (0, +1.63, 0), (0, 0, +1.63) of a 3D octahedron or star and six center points. If properly set up, a central composite design has the ability to possess the constant variance property of a rotatable design or may be an orthogonal design, thereby allowing an independent assessment of the three factors under study. For the illustrative study, a second-order response surface experiment was conducted for the three factors B, G, and H, previously declared statistically significant from the aforementioned screening experiment. The design setup is provided in Table 6.17.

Table 6.17 2^3 Central Composite Design for Mineral Penetration

Central composite response surface design							Mineral penetration				
Design identification				Factors			Cross-web positions			Responses	
Std order	Run order	Pt type	Blocks	B	G	H	Left	Center	Right	Mean	STD
1	20	1	1	−1	−1	−1	24.2	24.0	12.0	20.0667	6.9867
2	17	1	1	1	−1	−1	4.9	19.0	13.7	12.5333	7.1220
3	18	1	1	−1	1	−1	17.5	9.7	21.3	16.1667	5.9138
4	16	1	1	1	1	−1	21.7	13.3	30.3	21.7667	8.5002
5	10	1	2	−1	−1	1	19.5	25.7	45.7	30.3000	13.6923
6	14	1	2	1	−1	1	28.4	18.7	23.3	23.4667	4.8521
7	11	1	2	−1	1	1	52.4	32.0	44.7	43.0333	10.3016
8	9	1	2	1	1	1	39.5	44.7	48.3	44.1667	4.4242
9	19	0	1	0	0	0	21.2	27.3	34.7	27.7333	6.7604
10	15	0	1	0	0	0	31.4	19.0	24.3	24.9000	6.2217
11	13	0	2	0	0	0	13.6	34.0	10.3	19.3000	12.8371
12	12	0	2	0	0	0	18.6	13.7	17.0	16.4333	2.4987
13	4	0	3	0	0	0	38.6	25.7	19.7	28.0000	9.6576
14	5	0	3	0	0	0	30.4	23.0	12.0	21.8000	9.2585
15	8	−1	3	−1.633	0	0	33.4	20.7	26.7	26.9333	6.3532
16	3	−1	3	1.633	0	0	28.1	19.7	16.7	21.5000	5.9093
17	7	−1	3	0	−1.633	0	12.9	15.3	5.0	11.0667	5.3892
18	1	−1	3	0	1.633	0	31.5	37.0	31.7	33.4000	3.1193
19	6	−1	3	0	0	−1.633	13.7	13.0	13.0	13.2333	0.4041
20	2	−1	3	0	0	1.633	53.0	47.7	35.3	45.3333	9.0842

Analysis of variance

Tables 6.18 and 6.19 provide the corresponding ANOVA results. For the mean mineral penetration, factors G, H, H^2, and G × H interaction are statistically significant. The lack of fit (LOF) is not statistically significant (p = 0.511). Consequently, a second-order model is fitted. The resulting response surface curve for mean penetration is provided in Figure 6.28.

Table 6.18 Analysis of Variance for Mean Mineral Penetration

Response surface regression: Mean versus block, B, G, H

The analysis was done using coded units.

Estimated regression coefficients for mean

Term	Coeff.	SE Coeff.	T	P
Constant	22.9601	1.2575	18.259	0.000
Block 1	2.7460	1.2412	2.212	0.058
Block 2	−2.8590	1.2412	−2.303	0.050
B	−1.2379	0.8444	−1.466	0.181
G	5.6428	0.8444	6.682	0.000
H	10.8954	1.0902	.994	0.000
B*B	0.6572	0.8485	0.775	0.461
G*G	−0.0865	0.8485	−0.102	0.921
H*H	2.5572	0.8485	3.014	0.017
B*G	2.6375	1.0902	2.419	0.042
B*H	−0.4708	1.0902	−0.432	0.677
G*H	3.5125	1.0902	3.222	0.012

S = 3.08342 PRESS = 563.711

R-Sq = 96.11% R-Sq(pred) = 71.19% R-Sq(adj) = 90.77%

Analysis of variance for mean

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Blocks	2	238.96	56.69	28.343	2.98	0.108
Regression	9	1641.34	1641.34	182.372	19.18	0.000
Linear	3	1394.66	1394.66	464.887	48.900	.000
Square	3	90.56	90.56	30.186	3.17	0.085
Interaction	3	156.13	156.13	52.042	5.47	0.024
Residual error	8	76.06	76.06	9.507		
Lack of fit	5	48.72	48.72	9.743	1.07	0.511
Pure error	3	27.34	27.34	9.114		
Total	19	1956.36				

The fitted response surface for the mean of the mineral penetration is obtained as

$$\text{Average penetration} = 22.96 - 1.24 \times B + 5.64 \times G + 10.89 \times H + 0.66 \times B^2 - 0.09 \times G^2 + 2.56 \times H^2 + 2.64 \times B \times G - 0.47 \times B \times H + 3.51 \times G \times H$$

Table 6.19 Analysis of Variance for Standard Deviation Mineral Penetration

Response surface regression: STD versus block, B, G, H

The analysis was done using coded units.

Estimated regression coefficients for STD

Term	Coeff.	SE Coeff.	T	P
Constant	7.77642	1.4288	5.443	0.001
Block 1	1.04467	1.4103	0.741	0.480
Block 2	−0.13647	1.4103	−0.097	0.925
B	−0.95405	0.9595	−0.994	0.349
G	−0.54150	0.9595	−0.564	0.588
H	1.77350	1.2387	1.432	0.190
B*B	0.04737	0.9641	0.049	0.962
G*G	−0.65651	0.9641	−0.681	0.515
H*H	−0.47278	0.9641	−0.490	0.637
B*G	0.67672	1.2387	0.546	0.600
B*H	−2.17992	1.2387	−1.760	0.116
G*H	−0.51550	1.2387	−0.416	0.688

S = 3.50362 PRESS = 492.447

R-Sq = 52.01% R-Sq(pred) = 0.00% R-Sq(adj) = 0.00%

Analysis of variance for STD

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Blocks	2	13.110	11.419	5.710	0.47	0.644
Regression	9	93.315	93.315	10.368	0.84	0.600
Linear	3	41.209	41.209	13.736	1.12	0.397
Square	3	8.300	8.300	2.767	0.23	0.876
Interaction	3	43.806	43.806	14.602	1.19	0.373
Residual error	8	98.203	98.203	12.275		
Lack of fit	5	44.537	44.537	8.907	0.50	0.769
Pure error	3	53.666	53.666	17.889		
Total	19	204.628				

The fitted response surface for the standard deviation is obtained as

$$\text{Std. Dev} = 7.78 - 0.95 \times B - 0.54 \times G + 1.77 \times H + 0.05 \times B^2 - 0.66 \times G^2 - 0.47 \times H^2 + 0.68 \times B \times G - 2.18 \times B \times H - 0.52 \times G \times H$$

The standard deviation is modeled rather than the commonly used log s^2 because there are three cross-web positions (m = 3) under study. The theory for log s^2 transformation requires moderate to large amount of replication (m > 10). For small m, we prefer to model the standard deviation, which is the square root transformation. In addition, using the standard deviation will make it easier for engineers to correctly interpret the results from the model. In general, one would expect the true order of the variance model to be lower than the order of the model of the mean.

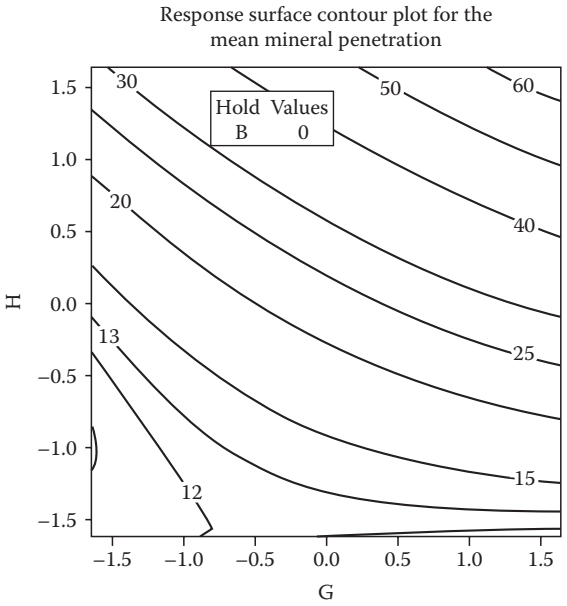


Figure 6.28 Response surface plot for the mean mineral penetration.

However, since we run a design appropriate for a full second-order model, we would fit a second order for the standard deviation. The response surface curve obtained for standard deviation is provided in Figure 6.29 for the situation when factor B is set at the center. In this case, the minimum variability occurs when factor G is at its highest level and factor H is at the lowest level.

Response surface optimization

The goal of this experiment is to find the conditions which minimize the cross-web variability while achieving a specification range of 15–20 for the mean mineral penetration. This, therefore, suggests using the standard deviation as the primary response and the mean as the secondary response. We constrain our variability between 0 and 5 for our optimization search. The result obtained is provided in Figure 6.30. The optimum operating windows are clearly provided in Figures 6.30 through 6.32. These results suggest that when B is operating at the center level, factors H should be set at its lowest level and factor G should be set at its highest level. Similarly, when factor B is set at low, then factor H should be set at its lowest level, while factor G should be at its lowest or center level. Response surface optimization plots are provided for the average and standard deviation.

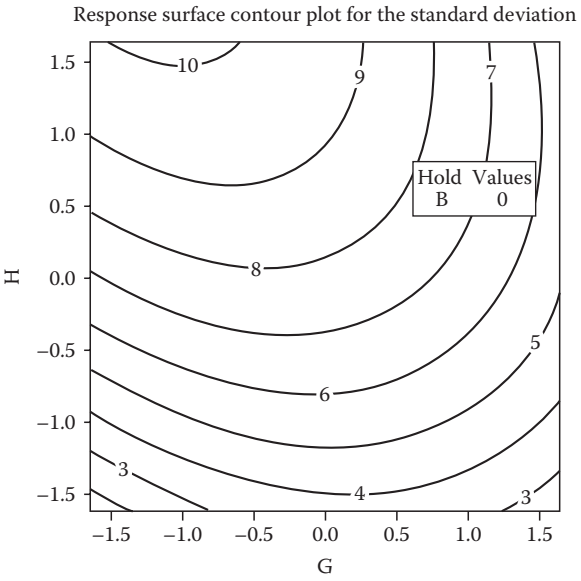


Figure 6.29 Response surface plot for mineral penetration standard deviation.

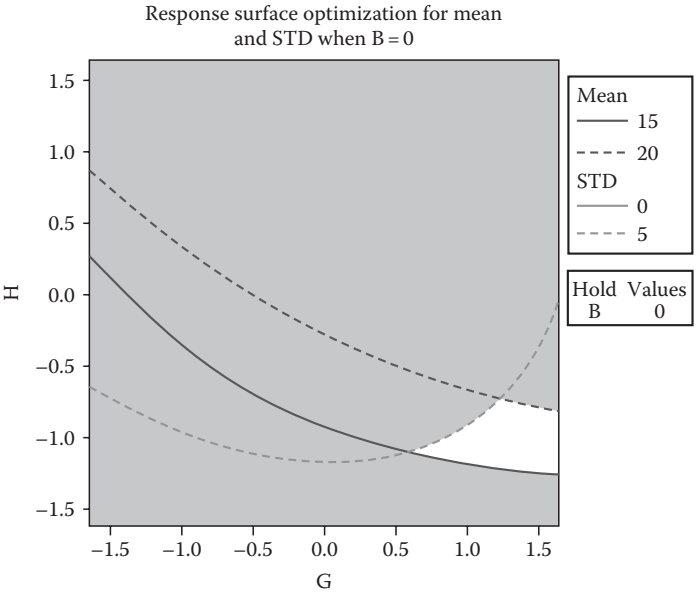


Figure 6.30 Factor B is at center level while both G and H are varied.

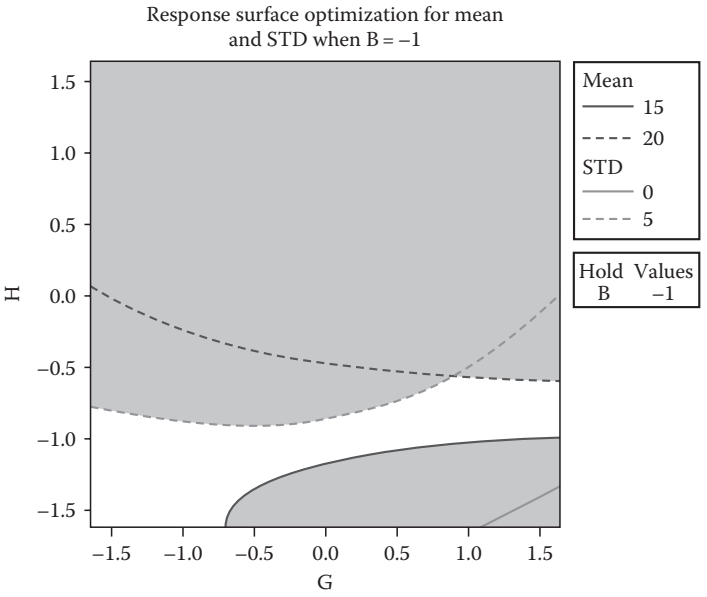


Figure 6.31 Factor B is set at low level while both factors G and H are varied.

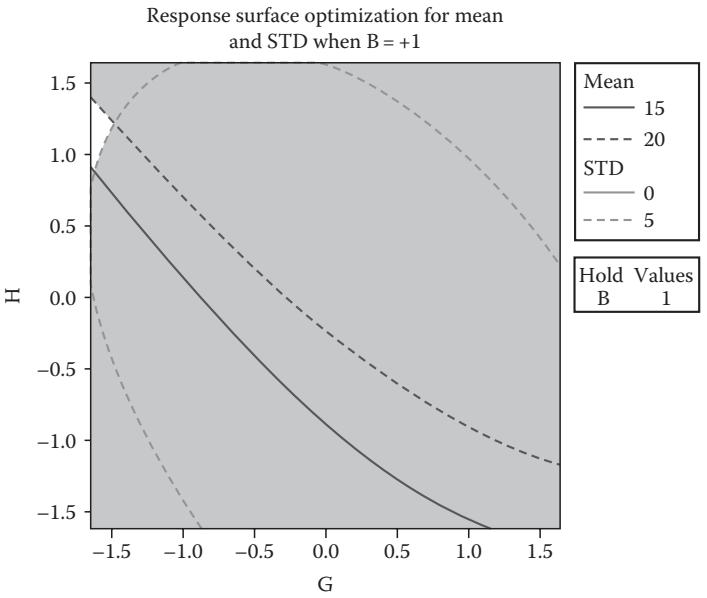


Figure 6.32 Factor B is set at high level while factors G and H are varied.

This study has significant impact on the way we control cross-web variability. Several optima (Figures 6.30 through 6.32) operating conditions with similar desirable properties are available to engineers for controlling the cross-web variability as well as achieving the mineral penetration target. In addition, the advantage of being able to plot the desirability surfaces for both mean and standard deviation to determine their sensitivities to small changes in the levels of the control factors is significant for future control of cross-web variability. This dual response optimization approach is applicable to multiple responses as well, and it is consistent with the positive aspects of the Taguchi contributions.

References

Ayeni, B. J., Design Resolution. Statistically Speaking, Internal Publication, 3MIS&DP Statistical Consulting, August 1991, P4.

Ayeni, B. J., Using response surface design to achieve target with minimum variance, *Proceedings of the 2nd Africa-USA International Conference on Manufacturing Technology*, August 1994, pp. 1–10.

Ayeni, B. J., *Application of Dual-Response Optimization for Moving Web Processes*. 3M Internal Publication, IT Statistical Consulting, 1999.

Badiru, A. B. and B. J. Ayeni, *Practitioner's Guide to Quality and Process Improvement*. Chapman & Hall, London, U.K., 1993.

Box, G. E. P., Discussion of off-line quality control, parameter design, and the Taguchi methods, *Journal of Quality Technology*, 17, 198–206, 1985.

Box, G. E. P. and N. R. Draper, *Empirical Model Building and Response Surfaces*. John Wiley & Sons, New York, 1987.

Box, G. E. P., W. G. Hunter, and J. S. Hunter, *Statistics for Experimenters*. John Wiley & Sons, New York, 1978.

Fisher, R. A., *The Design of Experiments*, 8th edn. Oliver & Boud, Edinburgh, Scotland, 1935.

Hicks, C. R., Fundamental Concepts in the Design of Experiments, 3rd Edn., Holt, Reinhart and Winston, New York, 1982.

Myers, R. H. and W. H. Carter, Response surface techniques for dual response systems, *Technometrics*, 15, 301–317, 1973.

Myers, R. H., A. I. Khuri, and G. Vining, Response surface alternatives to the Taguchi robust parameter design approach, *The American Statistician*, 46 (2), 131–139, May 1992.

Taguchi, G., *Introduction to Quality Engineering: Designing Quality into Products and Processes*. Kraus International Publications, White Plains, NY, 1986.

Taguchi, G. and Y. Wu, *Introduction to Off-line Quality Control*. Central Japan Quality Control Association, Nagaya, Japan, 1985.

Vining, G. G. and R. H. Myers, Combining Taguchi and response surface philosophies: A dual response approach, *Journal of Quality Technology*, 22, 38–45, 1990.

Yates, F., Complex experiments. Supplement to the *Journal of the Royal Statistical Society*, II (2), 181–247, 1935.

chapter seven

Risk analysis and estimation techniques

The problem considered in this chapter is that of estimating the nature of the size distribution of petroleum reservoirs and their values.

The method used consists of a Bayesian technique for parameter estimation. A sampling procedure based on minimizing the mean square error (MSE) of the posterior Bayesian estimator is developed using the beta density function to model the prior distribution. This Bayesian approach provides several typical representative distributions which are broad enough in scope, to provide a satisfactory economic analysis. These distributions reflect general patterns of similar regions, and include dry holes, as well as several representative class sizes of discoveries, and of course the probabilities associated with each of these classes. Mathematical expressions are provided for the probability estimates for the three-category case, as well as for the general case of k samples. This Bayesian method permits a more detailed economic analysis than is possible by the use of binomial distribution, where wells are simply classified as good or bad.

Bayesian estimation procedure

This section addresses Bayesian estimation procedure for petroleum discoveries of an exploratory well. Early exploratory efforts in a newly recognized geologic area provide an ideal application for the binomial probability distribution. In using the binomial, no consideration is given to how big or how small a discovery might prove to be. It is taken to be either good or bad, acceptable or unacceptable, dry or producer, with classification restricted to one or the other of two groups. However, if an exploratory well is grouped into three general classes: (Ayeni, B. J. and Ayeni, F. O., 1993): (1) $y_i = 0$, zero reserves for a dry well; (2) discovery y_2 barrels of reserves; and (3) discovery of y_3 barrels of reserves, where y_2 and y_3 are barrels of reserves discovered for two different groups of nondry wells. The probabilities p_i, $i = 1, 2, 3$ of a well discovering y_i barrels of reserves are not known and in most cases are assumed. In this chapter, however, we will estimate these probabilities for the three-category case, as well as for the general case of k-category.

In addition, we will also provide a procedure for estimating the probabilities of discovering various total reserves. For the three-category case, the binomial distribution is not adequate, because the population of interest is divided into more than two categories. Therefore, a multi-nominal distribution will be considered for this extended case. The parameter estimation procedure will be based on Bayesian methodology. The sampling procedure will be based on minimizing the MSE of the posterior Bayesian estimator using the beta density function to model the prior distribution (Pore and Dennis, 1980). The estimated probabilities are then used in estimating total reserves.

Formulation of the oil and gas discovery problem

Let us consider a certain region, where an oil company has grouped the possible outcomes of an exploratory well into three general classes:

1. Discovery of $y_1 = 0$ barrels of reserves (a dry well).
2. Discovery of y_2 barrels of reserves (a nondry well).
3. Discovery of y_3 barrels of reserves (a nondry well), and p_i, $i = 1, 2, 3$ are the corresponding probabilities of having x_i number of wells discovering y_i barrels of reserves. Let us further represent the following: x_1 = number of wells labeled as discovering $y_1 = 0$ barrels of reserves (dry well); x_2 = number of wells labeled as discovering y_2 barrels of reserves; x_3 = number of wells labeled as discovering y_3 barrels of reserves; p_1 = probability of a well discovering $y_1 = 0$ barrels of reserves; p_2 = probability of a well discovering y_2 barrels of reserves; and p_3 = probability of a well discovering y_3 barrels of reserves. For this case, $p_1 + p_2 + p_3 = 1$, and the conditional distribution of x_1, x_2, and x_3 is given in the next section.

Computational procedure

The Bayesian procedure considered focuses on the Bayesian technique for the parameter estimation and a sampling procedure based on minimizing the MSE of the posterior Bayesian estimator (Pore and Dennis, 1980). From the previous section, the conditional distribution of x_1, x_2, and x_3 can be represented as

$$f(x_1, x_2, x_3 \mid p_1, p_2, p_3) = \frac{(x_1 + x_2 + x_3)!}{x_1!\,x_2!\,x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3}$$

$$= A_0 p_1^{x_1} p_2^{x_2} (1 - p_1 - p_2)^{x^3} \tag{7.1}$$

where $p_1 \in (0,1), x_1 \in (0,1,2,\ldots)$, and

$$A_0 = \frac{(x_1 + x_2 + x_3)!}{x_1!\,x_2!\,x_3!}$$

This is a multi-nominal model which is a generalization of the binomial model. If we assume a prior distribution of the form

$$g(p_1, p_2, p_3) = B_0 p_1^{\mathtt{a}_1} p_2^{\mathtt{a}_2} p_3^{\mathtt{a}_3} = B_0 p_1^{\mathtt{a}_1} p_2^{\mathtt{a}_2} (1 - p_1 - p_2)^{\mathtt{a}_3}, \tag{7.2}$$

where $\mathtt{a}_1, \mathtt{a}_2, \mathtt{a}_3 > -1$; $p_1 \in [0,1]$, $p_2 \in [0, 1 - p_1]$ and

$$B_0 = \frac{\Gamma[\mathtt{a}_1 + \mathtt{a}_2 + \mathtt{a}_3 + 3]}{\Gamma[\mathtt{a}_1 + 1]\ \Gamma[\mathtt{a}_2 + 1]\ \Gamma[\mathtt{a}_3 + 1]}$$

Then, since data are available in discrete form, the posterior conditional probability can be represented as

$$w(p_1, p_2, p_3 | x_1, x_2, x_3) = \frac{g(p_1, p_2, p_3) f(x_1, x_2, x_3 \mid p_1, p_2, p_3)}{\sum g(p_1, p_2, p_3) f(x_1, x_2, x_3 \mid p_1, p_2, p_3)} \tag{7.3}$$

and, for the continuous case, we have

$$W(p_1, p_2, p_3 | x_1, x_2, x_3) = \frac{g(p_1, p_2, p_3) f(x_1, x_2, x_3 \mid p_1, p_2, p_3)}{p(x_1, x_2, x_3)} \tag{7.4}$$

where

$$P(x_1, x_2, x_3) = \int_0^1 \int_0^{1-p_2} g(p_1, p_2, p_3)\, f(x_1, x_2, x_3 \mid p_1, p_2, p_3)\, dp_1 dp_2$$

$$= A_0 B_0 \int_0^1 \int_0^{1-p_2} p_1^{(\mathtt{a}_1 + x_1)} p_2^{(\mathtt{a}_2 + x_2)} (1 - p_1 - p_2)^{(\mathtt{a}_3 + x_3)} dp_1 dp_2$$

$$P(x_1, x_2, x_3) = \frac{A_0 B_0 \Gamma[x_1 + \mathtt{a}_1 + 1] \Gamma[x_2 + \mathtt{a}_2 + 1] \Gamma[x_3 + \mathtt{a}_3 + 1]}{\Gamma[x_1 + x_2 + x_3 + \mathtt{a}_1 + \mathtt{a}_2 + \mathtt{a}_3 + 3]} \tag{7.5}$$

Then, to estimate the probabilities, P_i, $i = 1, 2, 3$ of a well discovering y_i barrels of reserves, we have

$$\hat{p}_i = E[p_i \mid x_1, x_2, x_3]$$

which implies that

$$\hat{p}_1 = \int_0^1 \int_0^{1-p_2} p_1 W(p_1, p_2, p_3 \mid x_1, x_2, x_3) dp_1 dp_2$$

$$\hat{p}_1 = \int_0^1 \int_0^{1-p_2} \frac{p_1 g(p_1, p_2, p_3) f(x_1, x_2, x_3 \mid p_1, p_2, p_3) dp_1 dp_2}{p(x_1, x_2, x_3)}$$

$$\hat{p}_1 = \frac{x_1 + \mathtt{a}_1 + 1}{x_1 + x_2 + x_3 + \mathtt{a}_1 + \mathtt{a}_2 + \mathtt{a}_3 + 3} \tag{7.6}$$

Similarly,

$$\hat{p}_2 = E[p_2 \mid x_1, x_2, x_3]$$

$$\hat{p}_2 = \frac{x_2 + \mathtt{a}_2 + 1}{x_1 + x_2 + x_3 + \mathtt{a}_1 + \mathtt{a}_2 + \mathtt{a}_3 + 3} \tag{7.7}$$

$$\hat{p}_3 = \frac{x_3 + \mathtt{a}_3 + 1}{x_1 + x_2 + x_3 + \mathtt{a}_1 + \mathtt{a}_2 + \mathtt{a}_3 + 3} \tag{7.8}$$

$$Q = \sum x_i y_i \tag{7.9}$$

$$E(Q) = \sum W_i(P \mid X) Q_i \tag{7.10}$$

The k-category case

The k-category case is a generalization of the three-category case. Pore and Dennis (1980) obtained their k-category pixel expression using Bayesian estimation procedure. Their result is, therefore, extended here to

multi-category reservoir systems. Then, the probabilities (P_i), $i = 1, 2, 3, \ldots, k$ of a well discovering y_i barrels of reserves can be written as

$$\hat{p}_i = \frac{x_i + a_i + 1}{\sum(x_i + a_i + 1)} \tag{7.11}$$

Discussion of results

Table 7.1 is the result obtained by McCray (1975) for various values of x_1, x_2 and x_3 which can appear in a sample. Each line in the table represents one possible sample, and the probability of that particular sample is calculated using Equation 7.1. McCray (1975) assumed $p_1 = 0.5$, $p_2 = 0.3$, and $p_3 = 0.2$ and their corresponding reserves are, respectively, $y_1 = 0$, $y_2 = 15$, and $y_3 = 60$.

Based on the assumed probabilities, the Bayesian approach with $\alpha_1 = 0$, $\alpha_2 = 0$, and $\alpha_3 = 0$ provides the total reserves (Q), as well as the expected reserves that correspond to each sample as presented in columns 11 and 12 of Table 7.1. The results from this table are exactly the same as those obtained by McCray (1975) based on the assumed probabilities. The total expected reserve in this case is 49.5 MM barrels. However, since different sampling arrangements should result in different probabilities, the probabilities, p_1, p_2, and p_3 can, therefore, be estimated from the given samples x_1, x_2, and x_3 using Equations 7.6 through 7.8 rather than assuming these probabilities. These estimates are based on various values of αs associated with the prior distribution. For the special case of $\alpha_1 = \alpha_2 = \alpha_3 = 0$, the estimated probabilities, as well as the expected reserves, is provided in Table 7.2. These probabilities, which are functions of the sampling arrangements (x_1, x_2, x_3), provide reserve estimates that do not correspond to McCray's (1975) results. Figure 7.1 shows a histogram plot of the probabilities for the reserve levels.

The total expected reserve in this case is 75 MM barrels. This is only due to the fact that the probabilities used by McCray in Table 7.1 are based on assumed probabilities. Table 7.7 is arranged according to the total reserves of each individual sample of three wells for the case of assumed probabilities as well as when probabilities are estimated from the sample. This allows the calculation of the cumulative probability so that it can be interpreted as a probability of not less than certain amounts of discovered crude oil reserves. These results are plotted in Figure 7.1, where one starts with 100% probability of at least 0 (zero) barrels discovered, progresses stepwise to 69.5% chance of discovering at least 45 MM barrels, 40% chance of discovering at least 90 MM barrels, and 11.35% chance of discovering at least 180 MM barrels of crude oil reserves. For other values of a (Tables 7.2 through 7.6), different values

Table 7.1 Calculations of Expected Well Reserves (Run 1)

Expected reserves calculation											
	Sample			Probabilities			Joint			Reserves	Expected reserves
OBS	X_1	X_1	X_1	P_1	P_1	P_1	Prob.	$f(x\|p)$	$w(p\|x)$	(MM bbl)	(MM bbl)
1	3	0	0	0.5	0.3	0.2	0.250	0.125	0.125	0	0.000
2	2	1	0	0.5	0.3	0.2	0.450	0.225	0.225	15	3.375
3	2	0	1	0.5	0.3	0.2	0.300	0.150	0.150	60	9.000
4	1	2	0	0.5	0.3	0.2	0.270	0.135	0.135	30	4.050
5	1	1	1	0.5	0.3	0.2	0.360	0.180	0.180	75	13.500
6	1	0	2	0.5	0.3	0.2	0.120	0.060	0.060	120	7.200
7	0	3	0	0.5	0.3	0.2	0.054	0.027	0.027	45	1.215
8	0	2	1	0.5	0.3	0.2	0.102	0.054	0.054	90	4.860
9	0	1	2	0.5	0.3	0.2	0.072	0.036	0.036	135	4.860
10	0	0	3	0.5	0.3	0.2	0.016	0.008	0.008	180	1.440
											Total = 49.500

$\alpha_1 = 0$, $\alpha_2 = 0$, and $\alpha_3 = 0$ with assumed probabilities $p_1 = 0.5$, $p_2 = 0.3$, $p_3 = 0.2$.

Table 7.2 Expected Reserves Calculations (Run 2)

Expected reserves calculation											
	Sample			Probabilities			Joint			Reserves	Expected reserves
OBS	X_1	X_1	X_1	P_1	P_1	P_1	Prob.	*f*(x|p)	w(p|x)	(MM bbl)	(MM bbl)
1	3	0	0	0.66667	0.1667	0.16667	0.593	0.296	0.11348	0	0.000
2	2	1	0	0.50000	0.33333	0.16667	0.500	0.250	0.09574	15	1.4362
3	2	0	1	0.50000	0.1667	0.33333	0.500	0.250	0.09574	60	5.7447
4	1	2	0	0.33333	0.50000	0.1667	0.500	0.250	0.09574	30	2.8723
5	1	1	1	0.33333	0.33333	0.33333	0.444	0.222	0.08511	75	6.3830
6	1	0	2	0.33333	0.1667	0.50000	0.500	0.250	0.09574	120	11.4894
7	0	3	0	0.16667	0.66667	0.16667	0.593	0.296	0.11348	45	5.1064
8	0	2	1	0.16667	0.50000	0.33333	0.500	0.250	0.09574	90	8.6170
9	0	1	2	0.16667	0.33333	0.50000	0.500	0.250	0.09574	135	12.9255
10	0	0	3	0.16667	0.16667	0.66667	0.593	0.296	0.11348	180	20.4255
								s			Total = 75.0000

$\alpha_1 = 0$, $\alpha_2 = 0$, and $\alpha_3 = 0$ with probabilities calculated from Equations 7.6 through 7.8.

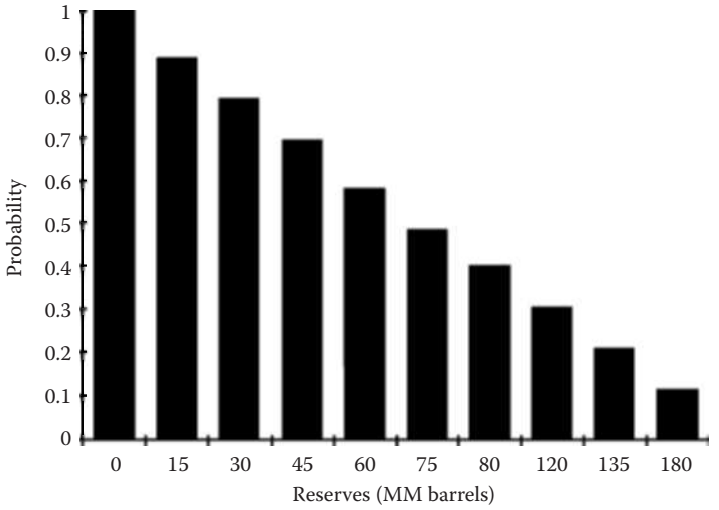


Figure 7.1 Histogram of probabilities versus reserves.

of the expected reserves can be generated. Table 7.7 shows the probabilities arranged by reserve levels. Table 7.8 contains total expected reserves for various combinations of α values. This result shows that the total expected reserves can vary from a low of 40 MM barrels to a high of 124 MM barrels of reserves.

This approach allows the estimation of the probability of a well discovering certain barrels of reserves from the sample rather than assuming this probability. The results show that the total expected reserve is 49 MM barrels when the probabilities are assumed. However, when the probabilities are estimated from a Bayesian standpoint, the total expected reserves can vary (with different levels of α) from a low of 40 MM barrels of reserves to as high as 124 MM barrels of reserves. This Bayesian approach provides several typical representative distributions which are broad enough in scope, to provide a satisfactory economic analysis. These distributions reflect general patterns of similar regions, and include dry holes, as well as several representative class sizes of discoveries, and of course the probabilities associated with each of these classes. In addition, this method permits a more detailed economic analysis than is possible by the use of binomial distribution, where wells are simply classified as good or bad.

Table 7.3 Expected Reserves Calculations (Run 3)

Expected reserves calculation											
	Sample			Probabilities			Joint			Reserves	Expected reserves
OBS	X_1	X_1	X_1	P_1	P_1	P_1	prob.	$f(x\|p)$	w(p|x)	(MM bbl)	(MM bbl)
1	3	0	0	0.64286	0.21429	0.14286	0.753	0.265	0.12468	0	0.000
2	2	1	0	0.50000	0.35714	0.14286	0.864	0.267	0.14313	15	2.146
3	2	0	1	0.50000	0.21429	0.28571	0.535	0.214	0.08869	60	5.321
4	1	2	0	0.35714	0.50000	0.14286	0.864	0.267	0.14343	30	4.293
5	1	1	1	0.35714	0.35714	0.28571	0.593	0.218	0.09875	75	7.406
6	1	0	2	0.35714	0.21429	0.42857	0.415	0.196	0.06884	120	8.261
7	0	3	0	0.21429	0.64286	0.14286	0.753	0.265	0.12468	45	5.611
8	0	2	1	0.21429	0.50000	0.28571	0.535	0.214	0.08869	90	7.982
9	0	1	2	0.21429	0.35714	0.42857	0.415	0.196	0.06884	135	9.293
10	0	0	3	0.21429	0.21429	0.57143	0.305	0.186	0.05056	180	9.100
											Total = 59.413

$\alpha_1 = 0.5$, $\alpha_2 = 0.5$, and $\alpha_3 = 0$.

Table 7.4 Expected Reserves Calculations (Run 4)

Expected reserves calculation											
	Sample			Probabilities			Joint			Reserves	Expected reserves
OBS	X_1	X_1	X_1	P_1	P_1	P_1	prob.	$f(x\|p)$	w(p|x)	(MM bbl)	(MM bbl)
1	3	0	0	0.65574	0.18033	0.16393	0.548	0.282	0.10368	0	0.0000
2	2	1	0	0.49180	0.34426	0.16393	0.518	0.243	0.09199	15	1.4699
3	2	0	1	0.49180	0.18033	0.32787	0.548	0.254	0.10361	30	5.2490
4	1	2	0	0.32787	0.50820	0.16393	0.548	0.254	0.10361	30	3.1083
5	1	1	1	0.32787	0.34426	0.32787	0.461	0.222	0.08710	75	6.5329
6	1	0	2	0.32787	0.18033	0.49180	0.463	0.238	0.08748	120	10.4980
7	0	3	0	0.16393	0.67213	0.16393	0.674	0.304	0.12736	45	5.7310
8	0	2	1	0.16393	0.50820	0.32787	0.548	0.254	0.10361	90	9.3249
9	0	1	2	0.16393	0.34486	0.49180	0.518	0.249	0.09799	135	13.2291
10	0	0	3	0.16393	0.18033	0.65574	0.548	0.282	0.10368	180	18.6630
											Total = 65.9661

$\alpha_1 = 0$, $\alpha_2 = 1$, and $\alpha_3 = 0$.

Table 7.5 Expected Reserves Calculations (Run 5)

Expected reserves calculation											
	Sample			Probabilities			Joint			Reserves	Expected reserves
OBS	X_1	X_1	X_1	P_1	P_1	P_1	prob.	$f(x\|p)$	w(p|x)	(MM bbl)	(MM bbl)
1	3	0	0	0.500	0.250	0.250	0.188	0.125	0.02641	0	0.0000
2	2	1	0	0.375	0.375	0.250	0.356	0.158	0.05014	15	0.7522
3	2	0	1	0.375	0.250	0.375	0.356	0.158	0.05014	60	3.0087
4	1	2	0	0.250	0.500	0.250	0.563	0.188	0.07324	30	2.3772
5	1	1	1	0.250	0.375	0.375	0.712	0.211	0.10029	75	7.5217
6	1	0	2	0.250	0.250	0.500	0.563	0.188	0.07924	120	9.5089
7	0	3	0	0.125	0.625	0.250	0.916	0.244	0.12897	45	5.8038
8	0	2	1	0.125	0.500	0.375	1.266	0.281	0.17829	90	16.0462
9	0	1	2	0.125	0.375	0.500	1.266	0.281	0.17829	135	24.0693
10	0	0	3	0.125	0.250	0.625	0.916	0.244	0.12897	180	23.2150
											Total = 92.3000

$\alpha_1 = 0$, $\alpha_2 = 1$, and $\alpha_3 = 1$.

Table 7.6 Expected Reserves Calculations (Run 6)

Expected reserves calculation											
	Sample			Probabilities			Joint			Reserves	Expected reserves
OBS	X_1	X_1	X_1	P_1	P_1	P_1	prob.	$f(x\|p)$	w(p|x)	(MM bbl)	(MM bbl)
1	3	0	0	0.500	0.125	0.375	0.211	0.125	0.02572	0	0.0000
2	2	1	0	0.375	0.250	0.375	0.178	0.105	0.02170	15	0.3256
3	2	0	1	0.375	0.125	0.500	0.633	0.211	0.07717	60	4.6302
4	1	2	0	0.250	0.375	0.375	0.178	0.105	0.02170	30	0.6511
5	1	1	1	0.250	0.250	0.500	0.563	0.188	0.06860	75	5.1447
6	1	0	2	0.250	0.125	0.625	1.373	0.293	0.16747	120	20.0965
7	0	3	0	0.125	0.500	0.375	0.211	0.125	0.02572	45	1.1576
8	0	2	1	0.125	0.375	0.500	0.633	0.211	0.07717	90	6.9453
9	0	1	2	0.125	0.250	0.625	1.373	0.293	0.16747	135	22.6085
10	0	0	3	0.125	0.125	0.750	2.848	0.422	0.34727	180	62.5080
											Total = 124.0675

$\alpha_1 = 0$, $\alpha_2 = 0$, and $\alpha_3 = 2$.

Table 7.7 Tabulation of Probabilities from Three Exploratory Wells

Reserves (MM bbl)	From Table 7.1 $w(p\|x)$	Probability of at least these reserves	From Table 7.2 $w(p\|x)$	Probability of at least these reserves
0	0.125	1.00	0.11348	1.0000
15	0.225	0.875	0.09574	0.8865
30	0.135	0.650	0.09574	0.7908
45	0.027	0.515	0.11348	0.6950
60	0.150	0.488	0.09574	0.5815
75	0.180	0.338	0.08511	0.4858
90	0.054	0.158	0.09574	0.4007
120	0.060	0.104	0.09574	0.3049
135	0.036	0.044	0.09574	0.2092
180	0.008	0.008	0.11348	0.1135

Table 7.8 Total Expected Reserves for Various Combinations of Alpha Values

α_1	α_2	α_3	Total expected reserves
0.0	0.0	0.0	49.50
0.5	0.0	0.0	61.89
0.0	0.5	0.0	69.76
0.0	0.0	0.5	93.36
0.5	0.5	0.0	59.42
0.0	0.5	0.5	86.13
0.5	0.5	0.5	75.00
1.0	0.0	0.0	52.42
0.0	1.0	0.0	65.97
0.0	0.0	1.0	106.62
1.0	1.0	0.0	50.75
0.0	1.0	1.0	92.30
2.0	0.0	0.0	40.00
0.0	2.0	0.0	60.98
0.0	0.0	2.0	124.06
2.0	2.0	2.0	75.00

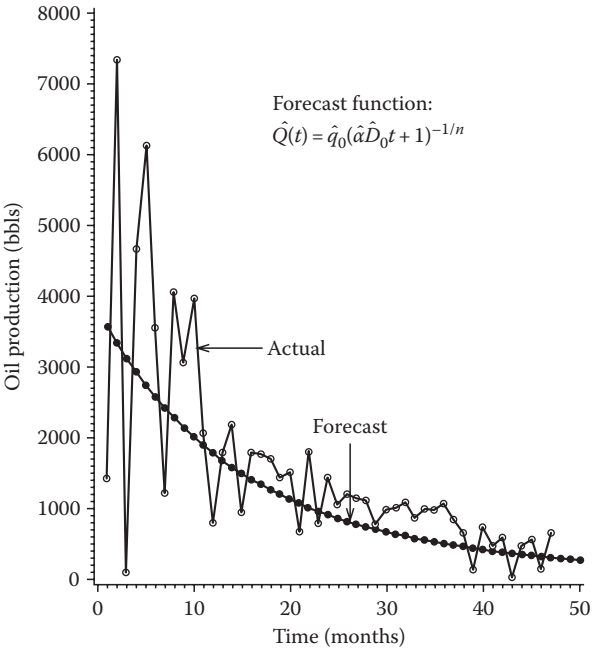


Figure 7.2 Oil production versus time.

Parameter estimation for hyperbolic decline curve

The problem of estimating the nonlinear parameters associated with the hyperbolic decline curve equation is presented in this chapter. Estimation equations are developed to estimate these nonlinear parameters. The condition under which the results can be used to predict future oil productions is examined using actual field data. An approximate linear term is obtained from the nonlinear hyperbolic equation through Taylor's series expansion, and the optimum parameter values are determined by employing the method of least squares through an iterative process. The estimated parameters are incorporated into the original hyperbolic decline equation to provide realistic forecast function. This method does not require any straight-line extrapolation, shifting, correcting, and/or adjusting scales in order to estimate future oil and gas predictions. The method has been successfully applied to actual oil production data from a West Cameron Block 33 Field in South Louisiana. The results obtained are provided in Figure 7.2.

Robustness of decline curves

Over the years, the decline curve technique has been extensively used by the oil industry to evaluate future oil and gas predictions, Arps (1945), Gentry (1972), Slider (1968), Fetkovitch (1980). These predictions are used

as the basis for economic analysis to support development, property sale or purchase, industrial loan provisions, and also to determine if a secondary recovery project should be carried out. The graphical solution of the hyperbolic equation is through the use of a log–log paper which sometimes provides a straight line that can be extrapolated for a useful length of time to predict future oil and gas productions. This technique, however, sometimes failed to produce the straight line needed for extrapolation for some oil and gas wells. Furthermore, the graphical method usually involves some manipulation of data, such as shifting, correcting, and/or adjusting scales, which eventually introduce bias into the actual data. In order to avoid the foregoing graphical problems and to accurately predict future performance of a producing well, a nonlinear least-squares technique is considered. This method does not require any straight-line extrapolation for future predictions.

Mathematical analysis

The general hyperbolic decline equation for oil production rate (q) as a function of time (t), Arps (1945), can be represented as

$$q(t) = q_0(1 + mD_0t)^{-1/m} \tag{7.12}$$

$$0 < m < 1$$

where

$q(t)$ is oil production at time t
q_0 is initial oil product
D_0 is initial decline
m is decline exponent

Also, the cumulative oil production at time t, $Q(t)$, Arps (1945), can be written as

$$Q(t) = \frac{q_0}{(m-1)D_0}\left[(1 + mD_0t)^{\frac{m-1}{m}} - 1\right] \tag{7.13}$$

By combining Equations 7.1 and 7.2 and performing some algebraic manipulations, see Arps (1945), it can be shown that

$$q(t)^{1-m} = q_0{}^{1-m} + (m-1)D_0q_0{}^{-m}Q(t) \tag{7.14}$$

Equation 7.14 shows that the oil production at time t is a nonlinear function of its cumulative oil production. By rewriting Equation 7.14 in terms of cumulative oil production, we have

$$Q(t) = \frac{q_0}{(1-m)D_0} + q(t)^{1-m}\frac{{q_0}^m}{(m-1)D_0} \tag{7.15}$$

Statistical analysis

For any given oil well, lease or property, the oil production at any time t can be observed. The observed production values are always available at discrete equi-spaced time intervals; this will, therefore, make Equation 7.15 not be satisfied exactly, due to the continuity assumption used in deriving Equation 7.12; hence, it will only define the cumulative production (Q_t) plus the residuals (ε_t) as follows (Ayeni, B. J., 1989):

$$Q_t = \frac{q_0}{(1-m)D_0} + {q_t}^{1-m}\frac{{q_0}^m}{(m-1)D_0} + \mathbf{e}_t \tag{7.16}$$

The general assumption for ε_t (Draper and Smith 1981) is that the residuals are assumed to be statistically independent and normally distributed with mean zero and constant variance, σ^2, that is, the expected value of ε_t, denoted by $E[\varepsilon_t] = 0$, and variance of ε_t, $Var(\varepsilon_t) = \sigma^2$. This normality assumption can be checked after the model has been fitted using a residual analysis test or histogram. If this assumption fails due to lack of fit, it may be that there is an outlier or an extreme value in the original data. Then, for this situation, the actual data can first be transformed to stabilize the variability in the data and then use the transformed data in the equations developed in this chapter. The most common transformation method is the log transformation. Other useful transformation techniques are reciprocal, square root, and inverse square root transformations. For more information about outliers and how to handle them, interested readers should see Box et al. (1978).

Now let

$$\begin{aligned} \mathbf{a}_1 &= \frac{q_0}{(1-m)D_0} \\ \mathbf{a}_2 &= \frac{{q_0}^m}{(1-m)D_0} \end{aligned} \tag{7.17}$$

and

$$\mathbf{a}_3 = 1 - m$$

By substituting Equation 7.7 into Equation 7.6, we obtain

$$Q_t = \alpha_1 + \alpha_2 q_t^{\alpha_3} + \varepsilon_t \tag{7.18}$$

A close examination of Equation 7.18 shows that it is completely nonlinear in parameters. This is because Equation 7.18 is nonlinear in α_3, which is controlled by the exponent m, just as m controls α_1 and α_2; therefore, both α_1 and α_2 depend on α_3.

Parameter estimation

In order to estimate the parameters in Equation 7.18, we chose to minimize the sum of squares of the residuals given as

$$SS(\alpha) = \sum_{i=1}^{n} \left(Q_t - \alpha_1 - \alpha_2 q_t^{\alpha_3}\right)^2 \tag{7.19}$$

Since the model is nonlinear in α, the normal equations will be nonlinear. Also, since α_1 and α_2 depend on α_3, an iterative technique will be used to solve the normal equations.

Optimization technique

Let us rewrite Equation 7.18 as

$$Q_t = f(q_t, \alpha) + \varepsilon_t \tag{7.20}$$

where

$$f(q_t, \alpha) = \alpha_1 + \alpha_2 q_t^{\alpha_3}$$

Let α_{10}, α_{20}, α_{30}, be the initial values for the parameters α_1, α_2, and α_3, respectively. If we carry out a Taylor series expansion of $f(q_t, \alpha)$ about the point α_0, where $\alpha_0 = (\alpha_{10}, \alpha_{20}, \alpha_{30})$, and truncate the expansion, we can say that, approximately, when α is close to α_0,

$$f(q_t, \alpha) = f(q_t, \alpha_0) + \sum_{i=1}^{3} \left[\frac{df}{d\alpha_1}(q_t, \alpha)\right]_{\alpha=\alpha_0} (\alpha_1 - \alpha_{10}) \tag{7.21}$$

If we set

$$\begin{aligned} f_t^0 &= f(q_t, \mathbf{a}_0) \\ \mathbf{b}_i^0 &= \mathbf{a}_1 - \mathbf{a}_{10} \\ Z_{it}^0 &= \left[\frac{df(q_t, \mathbf{a})}{d\mathbf{a}_1} \right]_{\mathbf{a}=\mathbf{a}_0} \end{aligned} \tag{7.22}$$

we can see that Equation 7.10 is approximately

$$Q_t = f_t^0 + \sum_{i=1}^{3} \mathbf{b}_i^0 Z_{it}^0 + \mathbf{e}_t \tag{7.23}$$

$$Q_t - f_t^0 = \sum_{i=1}^{3} \mathbf{b}_i^0 Z_{it}^0 + \mathbf{e}_t \tag{7.24}$$

which is of linear form. Therefore, we can now estimate the parameters $\mathbf{b}_i^0, i = 1, 2, 3$ by applying the linear least-squares theory with the assumptions that $E[\varepsilon_t] = 0$ and var $(\varepsilon_t) = \sigma^2$ (Draper and Smith 1981). This is achieved by minimizing the sum of squares of the residual in Equation 7.17.

If we write

$$Z_0 = \begin{bmatrix} \frac{\partial f(q_1, \mathbf{a})}{\partial \mathbf{a}_1} & \frac{\partial f(q_1, \mathbf{a})}{\partial \mathbf{a}_2} & \frac{\partial f(q_1, \mathbf{a})}{\partial \mathbf{a}_3} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \frac{\partial f(q_n, \mathbf{a})}{\partial \mathbf{a}_1} & \frac{\partial f(q_n, \mathbf{a})}{\partial \mathbf{a}_2} & \frac{\partial f(q_n, \mathbf{a})}{\partial \mathbf{a}_3} \end{bmatrix} \tag{7.25}$$

$$\mathbf{b} = \begin{bmatrix} \mathbf{b}_1^0 \\ \mathbf{b}_2^0 \\ \mathbf{b}_3^0 \end{bmatrix} \quad \text{and} \quad Q_0 = \begin{bmatrix} Q_1 - f_1^0 \\ \cdot \quad \cdot \\ \cdot \quad \cdot \\ \cdot \quad \cdot \\ Q_n - f_n^0 \end{bmatrix} = Q - f^0$$

then, the estimate $\hat{\mathbf{b}}_0 = (\mathbf{b}_1^0, \mathbf{b}_2^0, \mathbf{b}_3^0)$ is given by

$$\hat{\mathbf{b}} = (Z_0'Z_0)^{-1}Z_0'(Q - f^0) \tag{7.26}$$

The vector $\hat{\beta}_0$ will, therefore, minimize the sum of squares $SS(\alpha)$ of the residual with respect to $\mathbf{b}_i^0$ for $i = 1, 2, 3$ where

$$SS(\mathbf{a}) = \sum_{t=1}^{n}\left[Q_t - f(q_t, \mathbf{a}_0) - \sum_{i=1}^{3}\mathbf{b}_1^0 Z_{it}^0\right]^2 \quad \text{and} \quad \mathbf{b}_i^0 = \mathbf{a}_i - \mathbf{a}_{i0} \tag{7.27}$$

Iterative procedure

Let $\mathbf{b}_i^0 = \mathbf{a}_{i1} - \mathbf{a}_{i0}$, then $\mathbf{a}_{i1}$, $i = 1, 2, 3$ can be thought of as the revised best estimates of α. We can now place the values α_{i1}, the revised estimates in the same roles as were played in the foregoing by α_{i0}, and to through exactly the same procedure as already described, but replacing all zero subscripts by ones. This will lead to another set of revised estimates, α_{i2} and so on. In vector form, extending the previous notation, we can write

$$\mathbf{a}_{J+1} = a_J + \left(Z_J' Z_J\right)^{-1} Z_J' (Q - f^J) \tag{7.28}$$

where

$$Z_J = Z_{it}{}^J$$

$$f^J = \left(f_1{}^J, f_2{}^J, \ldots, f_n{}^J\right)'$$

$$\mathbf{a}^J = (\mathbf{a}_{1J}, \mathbf{a}_{2J}, \mathbf{a}_{3J})'$$

The foregoing iterative process is continued until the solution converges, that is, until in successive iterations, J, $J + 1$, such that

$$\left[\frac{\mathbf{a}_{i,J+1} - \mathbf{a}_{iJ}}{\mathbf{a}_{iJ}}\right] < \mathbf{d} \quad \text{for } i = 1, 2, 3 \tag{7.29}$$

where δ is some prespecified amount, for example, 0.0001. Also at each stage of the iterative procedure, $SS(\alpha_J)$ can be evaluated to check if a reduction in its value has actually been achieved. For rapid convergence, if $SS(\alpha_{J+1})$ is greater than $SS(\alpha_J)$, the vector β in Equation 7.17 can be amended by having it. But if $SS(\alpha_{J+1})$ is less than $SS(\alpha_J)$, we can double the vector β_J. This halving and/or doubling process is continued until three points between α_J, α_{J+1} are

found, which include between them a local minimum of $SS(\alpha)$. A quadratic interpolation can be used to locate the minimum and the iterative cycle begins again. Figure 7.4 shows a plot of residuals versus normal scores.

After convergence, the optimum parameters estimated can be used to determine the estimates $\hat{m}$, $\hat{q}_0$, $\hat{D}_0$, as follows: From Equation 7.7, we have

$$\begin{aligned} \hat{m} &= 1 - \hat{\alpha}_3 \\ \hat{q}_0 &= \left(-\hat{\alpha}_1/\hat{\alpha}_2\right)^{1/\hat{\alpha}_3} \quad 0 < \hat{\alpha}_3 < 1 \\ \hat{D}_0 &= \frac{\hat{q}_0}{\hat{\alpha}_1\hat{\alpha}_3} \qquad \hat{\alpha}_2 < 0 \\ & \qquad\qquad \hat{\alpha}_1 > 0 \end{aligned} \tag{7.30}$$

where $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\alpha}_3$ are optimum parameters estimated with the minimum residual sum of squares. We can now incorporate Equation 7.20 into Equation 7.1 to give the optimum forecast function that can be used to generate future oil production forecast, as follows:

$$\hat{q}(t) = \hat{q}_0(1 + \hat{m}_0\hat{D}t)^{-1/\hat{m}} \tag{7.31}$$

This technique has been applied to actual oil field data from a West Cameron Block 33 Field in South Louisiana. The results obtained are provided in Figure 7.2. The method described in this chapter can probably be improved if the Taylor's series expansion is carried out further; however, this problem is not explored in this chapter.

Residual analysis test

A simple residual analysis test is carried out in this section to check the normality assumption used in the development of the estimating equations. A plot of the residuals versus the normal scores shows that the residual is random (Figure 7.3). This is because there is an approximately linear relationship with a correlation coefficient of $r = 0.90$ between the residual and the normal score. In addition, a plot of the histograms of the residual shows that the residual is normally distributed with the mean centered around zero (Figure 7.4). These tests substantiate the normality assumption used to define the residual in the earlier section.

This section has presented a technique for estimating parameters associated with the hyperbolic decline curve equation. The method used provides optimum parameter values with minimum residual sum of squares. The estimated parameters eventually produce the optimum

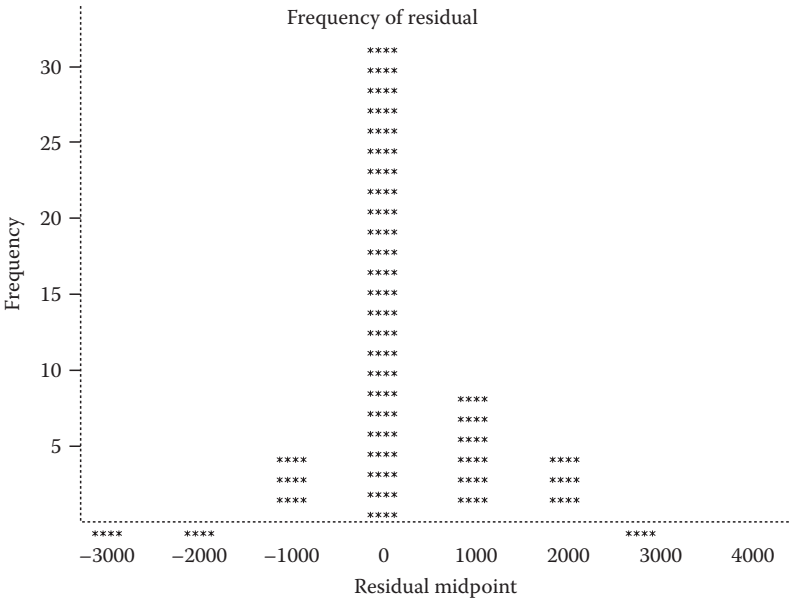


Figure 7.3 Histogram of residuals.

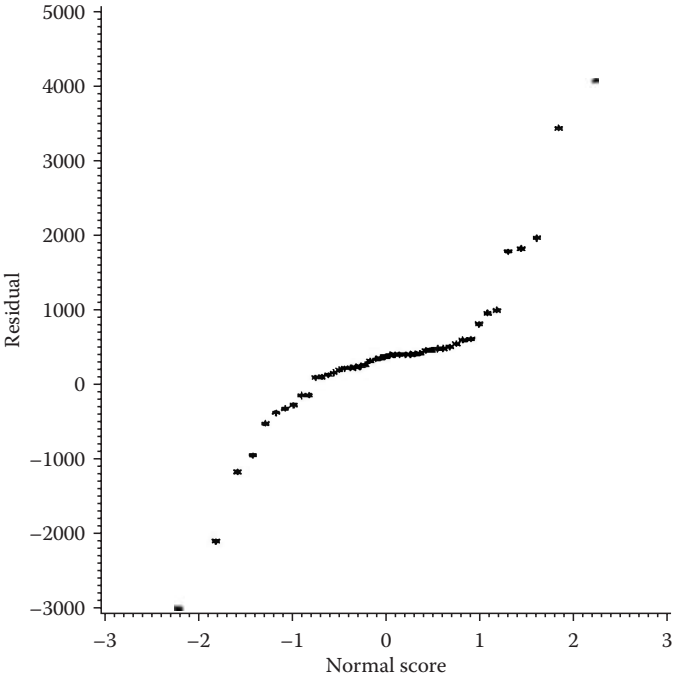


Figure 7.4 Residual versus normal scores.

forecast function when incorporated into the original hyperbolic decline equation. This forecast function can be used to generate accurate forecast values needed for estimating future reserves.

Simplified solution to the vector equation

Let $\alpha_0 = (\alpha_{10}, \alpha_{20}, \alpha_{30})^T$ be the initial guess. Then, by using Equation 7.23, set the following:

$$f_t^0 = \mathtt{a}_{10} + \mathtt{a}_{20} q_t^{\mathtt{a}_{30}}$$

$$\mathtt{b}_1^0 = \mathtt{a}_1 - \mathtt{a}_{10}$$

$$\mathtt{b}_2^0 = \mathtt{a}_2 - \mathtt{a}_{20}$$

$$\mathtt{b}_3^0 = \mathtt{a}_3 - \mathtt{a}_{30}$$

$$Z_{1_t}^0 = \left.\frac{\partial f(q_t, \mathtt{a})}{\partial \mathtt{a}_1}\right|_{\mathtt{a}=\mathtt{a}_0} = \frac{\partial}{\partial \mathtt{a}_1}(\mathtt{a}_1 + \mathtt{a}_2 q_t^{\mathtt{a}_{30}})$$

$$Z_{2_t}^0 = \left.\frac{\partial f(q_t, \mathtt{a})}{\partial \mathtt{a}_2}\right|_{\mathtt{a}=\mathtt{a}_0} = q_t^{\mathtt{a}_{30}}$$

$$Z_{3_t}^0 = \left.\frac{\partial f(q_t, \mathtt{a})}{\partial \mathtt{a}_3}\right|_{\mathtt{a}=\mathtt{a}_0} = \mathtt{a}_{20} e^{\mathtt{a}_{30} \ln q_t} \cdot (\ln\, q_t)$$

$$Z_{3_t}^0 = \mathtt{a}_{20} e^{\mathtt{a}_3} \cdot \ln\, q_t$$

where $t = 1, 2, \ldots, n$ observations. Therefore, the n × 1 matrix $\mathbf{Q} - \mathbf{f}^0$ can be expressed as

$$\mathbf{Q} - \mathbf{f}^0 = \begin{bmatrix} Q_1 \\ Q_2 \\ \cdot \\ \cdot \\ \cdot \\ Q_n \end{bmatrix} - \begin{bmatrix} \mathtt{a}_{10} + \mathtt{a}_{20} q_1^{\mathtt{a}_{30}} \\ \cdot \quad \cdot \\ \cdot \quad \cdot \\ \cdot \quad \cdot \\ \cdot \quad \cdot \\ \mathtt{a}_{10} + \mathtt{a}_{20} q_n^{\mathtt{a}_{30}} \end{bmatrix} \tag{7.32}$$

Therefore, the n × 3 matrix $\mathbf{Z}_0$ can be written as

$$Z_0 = \begin{bmatrix} 1 & q_1^{\mathbf{a}_{30}} & \mathbf{a}_{20} q_1^{\mathbf{a}_{30}} \ln q_1 \\ 1 & q_2^{\mathbf{a}_{30}} & \mathbf{a}_{20} q_2^{\mathbf{a}_{30}} \ln q_2 \\ . & . & . \\ & & \\ . & . & . \\ . & . & . \\ 1 & q_n^{\mathbf{a}_{30}} & \mathbf{a}_{20} q_n^{\mathbf{a}_{30}} \ln q_n \end{bmatrix} \tag{7.33}$$

Then, the estimate $\hat{\beta}_0$ is given by

$$\hat{\mathbf{b}}_0 = \begin{bmatrix} n & \sum_{t=1}^{n} q_t^{\mathbf{a}_{30}} & \mathbf{a}_{20} \sum_{t=1}^{n} q_t^{\mathbf{a}_{30}} \ln q_t \\ \sum_{t=1}^{n} q_t^{\mathbf{a}_{30}} & \sum_{t=1}^{n} q_t^{2\mathbf{a}_{30}} & \mathbf{a}_{20} \sum_{t=1}^{n} q_t^{2\mathbf{a}_{30}} \cdot \ln q_t \\ \mathbf{a}_{20} \sum_{t=1}^{n} q_t^{\mathbf{a}_{30}} \ln q_t & \mathbf{a}_{20} \sum_{t=1}^{n} q_t^{2\mathbf{a}_{30}} \ln q_t & \mathbf{a}_{20}^2 \sum_{t=1}^{n} \left(q_t^{\mathbf{a}_{30}} \ln q_t \right)^2 \end{bmatrix}^{-1}$$

$$\times \begin{bmatrix} \sum_{t=1}^{n} \left(Q_t - \mathbf{a}_{10} - \mathbf{a}_{20} q_t^{\mathbf{a}_{30}} \right) \\ \sum_{t=1}^{n} \left(Q_t - \mathbf{a}_{10} - \mathbf{a}_{20} q_t^{\mathbf{a}_{30}} \right) \cdot q_t^{\mathbf{a}_{30}} \\ \mathbf{a}_{20} \sum_{t=1}^{n} \left(Q_t - \mathbf{a}_{10} - \mathbf{a}_{20} q_t^{\mathbf{a}_{30}} \right) \cdot q_t^{\mathbf{a}_{30}} \cdot \ln q_t \end{bmatrix} \tag{7.34}$$

Since $\alpha_{j+1} = \alpha_j + \beta_j$ from Equation 7.28, and by using Equation 7.34, we have

$$\mathbf{a}_{j+1} = \mathbf{a}_j + \begin{bmatrix} n & \sum_{t=1}^{n} q_t^{\mathbf{a}_3^{(j)}} & \mathbf{a}_2^{(j)} \sum_{t=1}^{n} q_t^{\mathbf{a}_3^{(j)}} \ln q_t \\ \sum_{t=1}^{n} q_t^{\mathbf{a}_3^{(j)}} & \sum_{t=1}^{n} q_t^{2\mathbf{a}_3^{(j)}} & \mathbf{a}_2^{(j)} \sum_{t=1}^{n} q_t^{2\mathbf{a}_{30}^{(j)}} \ln q_t \\ \mathbf{a}_{20}^{(j)} \sum_{t=1}^{n} q_t^{\mathbf{a}_3^{(j)}} & \mathbf{a}_2^{(j)} \sum_{t=1}^{n} q_t^{2\mathbf{a}_{30}^{(j)}} \ln q_t & \left(\mathbf{a}_2^2\right)^{(j)} \sum_{t=1}^{n} \left(q_t^{2\mathbf{a}_{30}^{(j)}} \ln q_t\right)^2 \end{bmatrix}^{-1}$$

$$\times \begin{bmatrix} \sum_{t=1}^{n} \left(Q_t - \mathbf{a}_1^{(j)} - \mathbf{a}_2^{(j)} q_t^{\mathbf{a}_3^{(j)}}\right) \\ \sum_{t=1}^{n} \left(Q_t - \mathbf{a}_1^{(j)} - \mathbf{a}_2^{(j)} q_t^{\mathbf{a}_3^{(j)}}\right) q_t^{\mathbf{a}_3^{(j)}} \\ \mathbf{a}_2^{(j)} \sum_{t=1}^{n} \left(Q_t - \mathbf{a}_1^{(j)} - \mathbf{a}_1^{(j)} q_t^{\mathbf{a}_3^{(j)}}\right) q_t^{\mathbf{a}_3^{(j)}} \ln q_t \end{bmatrix} \tag{7.35}$$

Equation 7.35 is very easy to program and can be used for the iterative process. The process continues until the solution converges, that is, until successive iterations, j, $j + 1$, such that

$$\left| \frac{\mathbf{a}_i^{(j+1)} - \mathbf{a}_i^{(j)}}{\mathbf{a}_i^{(j)}} \right| < \delta \quad i = 1, 2, 3, \tag{7.36}$$

where δ is some prespecified small number.

Integrating neural networks and statistics for process control

This section explores the opportunities available using a neural network (NN) back-propagation technique to model process data. The effects of using different transfer functions (linear, sigmoid, hyperbolic tangent, etc.) as well as multiple hidden nodes are extensively explored for monitoring and controlling manufacturing processes. We emphasize some practical issues useful in developing process control strategies for process optimization. Some actual examples from industrial experiments are presented.

These examples are based on statistically designed experiments employing two-level factorial and fractional factorial designs, as well as a central composite response surface design. Experiments from a large system consisting of more than 50 independent variables are also considered in the study. The results obtained from the various NN architectures are compared with those obtained from the statistical linear regression method. Strengths and weaknesses of each method are identified. The conditions under which one method performs better than the other are fully explored. We also investigate the circumstances under which the NN and regression approaches can be integrated with the objective of maximizing the benefits from both methods.

Fundamentals of neural network

An NN is a system that is modeled after the human brain and arranged in patterns similar to biological neural nets. The unit analogous to the biological neuron is referred to as a "processing element" (PE). A PE has many input paths representing individual neurons. These artificial neurons receive and sum information from other neurons or external inputs, perform a transformation on the inputs and then pass on the transformed information to other neurons or external outputs. The output path of a processing element can be connected to input paths of other processing elements through connection weights which correspond to the strength of neural connections. The information passed from neuron to neuron can be thought of as a way to activate a response from certain neurons based on the information received.

The most important characteristic of NNs is that they can learn to simulate any behavior and can be used to generate the action necessary to produce a given response. Learning is the process of adapting or modifying the connection weights in response to stimuli being presented at the input buffer and optionally at the output buffer. Back-propagation network is the most common form of NN. This network has an input layer and at least one hidden layer. There is no theoretical limit on the number of hidden layers but typically there will be one or two. Each layer is fully connected to the succeeding layer.

The input function

The simplest input function is a simple weighted input:

$$I = \sum_{i=1}^{n} w_i x_i$$

A back-propagation element transfers its inputs as

$$X_j = f(I),$$

where f is traditionally the sigmoid function but can be any differentiable function.

Transfer functions

(a) Linear transfer function

$$f(z) = Z$$

(b) Hyperbolic tangent

$$f(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

(c) Sigmoid transfer function

$$f(z) = \frac{1}{1 + e^{-z}}$$

Statistics and neural networks predictions

Tables 7.9 through 7.11 provide the actual as well as the predicted values of the statistical regression method with NNs for various transfer functions. These results are based respectively on experiments specifically designed for a 2^3 replicated factorial design with center points; a 2^3 central composite design with center points; as well as a 2^{8-4} fractional factorial design. Plots of the various NN methods are shown in Figures 7.5 through 7.10. Figures 7.11 through 7.13 provide the graphical comparisons of the various methods.

Statistical error analysis

Statistical error analysis is used to compare the performance, as well as the accuracy, of the statistical regression and NN methods (Ayeni and Pilat, 1992). The accuracy of the predicted values relative to the actual is determined by various statistical methods. The criteria used in this study are average relative error (ARE), forecast root mean square error (FRMSE), minimum error as well as maximum error. The average relative error is the

Table 7.9 Prediction Comparisons for a 2^3 Full Factorial Design

Actual	Statistics prediction	Linear prediction	Hyperbolic tangent	Sigmoid prediction
5.4083	5.8087	6.4648	5.9992	6.1625
13.263	13.484	13.4754	13.4031	13.5137
16.921	16.5432	16.572	16.4019	16.3156
24.57	24.2212	23.5832	24.1862	23.8225
17.094	17.2289	16.7375	17.1906	16.7648
24.272	23.8769	23.7487	23.7327	13.7379
26.667	27.026	26.8453	26.9531	27.1393
32.787	33.1473	33.8565	33.3524	33.3696
19.268	20.167	20.1603	20.2921	20.5331
21.322	20.167	20.1603	20.2921	20.5331
6.1463	5.8087	6.4642	5.9992	6.1625
13.643	13.484	13.4754	13.4031	13.5137
16.103	16.5432	16.572	16.4019	16.3156
23.81	24.2212	23.5832	24.1862	23.8225
17.301	17.2289	16.7375	17.1906	16.7648
23.419	23.8769	23.7487	23.7327	23.7379
27.322	27.026	26.8453	26.9531	27.1393
33.445	33.1473	33.8565	33.3524	33.3696
20.661	20.167	20.1603	20.2921	20.5331
19.92	20.167	20.1603	20.2921	20.5331

relative deviation of the predicted values from the actual. The lower the average relative error, the more equally distributed is the error between positive and negative values. The FRMSE is a measure of the dispersion. A smaller value of FRMSE indicates a better degree of fit.

Tables 7.12 through 7.15 show the results obtained from the error analysis. The results show that statistical linear regression performs better in some cases while the hyperbolic tangent transfer function of the NN performs better than any other transfer function.

Integration of statistics and neural networks

In process control, experiments are frequently performed primarily to measure the effects of one or more variables on a response. From these effects, one needs to determine significant variables that can be used to develop control strategies. However, most NNs' approaches do not provide information about variables that are significant to the response of interest. In addition, the coefficients or weights obtained from NN are difficult to interpret (Ayeni et al., 1993; Ayeni and Koval, 1995).

Table 7.10 Prediction Comparisons for a 2^3 Central Composite Design

Actual	Statistics prediction	Linear prediction	Hyperbolic tangent	Sigmoid prediction
34.2	37.51	30.5013	34.4385	33.4627
12.2	11.25	16.1799	11.0012	11.4857
19.2	14.95	17.7643	19.3777	14.2051
13.85	13.84	3.44269	13.1522	11.9871
42.2	42.71	32.2846	42.1012	40.786
9.8	14.91	17.9633	10.0964	13.2724
12.5	14.3	19.5477	13.1515	14.8784
14.1	11.65	5.2263	12.8587	12.4026
31	30.55	29.3209	30.7354	31.2331
8.3	1.42	6.4067	8.0917	9.7927
41.6	31.01	28.0534	41.1906	36.6313
13.1	16.35	7.6742	13.6192	13.9306
9	11.7	16.9721	10.2448	12.4473
13.2	12.99	17.8638	13.1686	13.0293
18.2	12.99	17.8638	13.1686	13.0293
10.4	12.99	17.8638	13.1686	13.0293
17.3	12.99	17.8638	13.1686	13.0293
8.6	12.99	17.8638	13.1686	13.0293
12.7	12.99	17.8638	13.1686	13.0293
11.6	12.99	17.8638	13.1686	13.0293

Also, most NN packages are lacking in providing response surface curves needed for process optimization. In this section, we explore the opportunities available in performing Yates' analysis on NN-trained data. The results obtained are compared with the Yates' analysis of the actual data as well as the Yates' analysis from the predicted values of the statistical regression method. The approach used here can be used in identifying significant variables as well as generating response surface curves for process optimization.

Figures 7.5 through 7.9 show the main effects plots from the 2^3 central composite experiment with the actual data, statistical predictions, as well as plots of NN predictions for various NN transfer functions. These effects were obtained using Yates' algorithm. They are the average effects for factor A at five levels (–2, –1, 0, 1, 2), factor B at five levels (–2, –1, 0, 1, 2), and factor C at three levels (–1, –1, 1). Factors A and B are quantitative factors while factor C is a qualitative factor. The results obtained show that NN method provides similar results as compared to actual Yates' analysis as well as the results obtained under statistical regression method.

Table 7.11 Prediction Comparisons for a 2^{8-4} Fractional Factorial Design

Actual	Statistics prediction	Neural network linear	Linear prediction	Hyperbolic tangent	Sigmoid prediction
2.67	2.67	3.63503	2.66989	2.685	2.73969
2.17	2.17	2.33152	2.16992	2.155	2.11396
1.33	1.33	2.32557	1.3301	1.39	1.71533
1.17	1.17	1.30798	1.71004	1.163	1.18535
4.83	4.86	5.05969	4.82995	4.817	4.80184
2.83	2.83	3.75591	2.82991	2.853	2.90004
5.83	5.83	6.08295	5.83005	5.822	5.83877
4	4	5.06537	4.00005	4.038	4.27219
2.17	2.17	−0.1753	2.17012	2.172	1.8063
2.17	2.17	4.36818	2.16999	2.174	2.14609
3.5	3.5	4.20293	3.50008	3.466	3.41812
3.33	3.33	3.0058	3.32992	3.33	3.32857
6.67	6.67	6.02276	6.66999	6.666	6.66445
9.67	9.67	7.56678	9.67015	9.664	9.42287
4.67	4.67	4.38513	4.66992	4.67	4.67029
2.5	2.5	3.18801	2.50012	2.449	2.39139

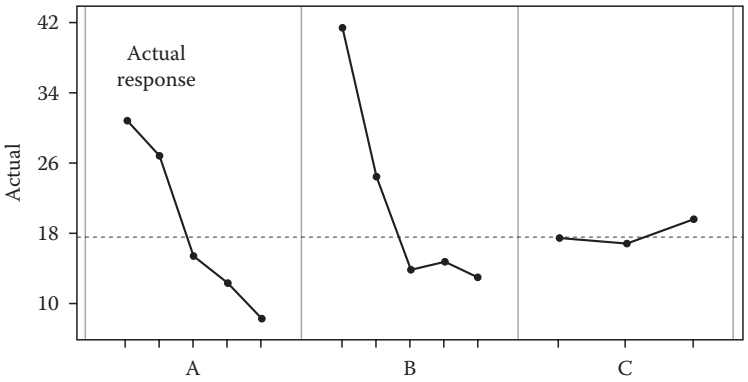


Figure 7.5 Actual effects using Yates' algorithm.

These results indicate that the Yates' analysis can be used on NN-trained data to determine the effects of variables as well as identifying statistically significant factors. Figures for interaction plots as well as for other designs provide similar results, but are not provided here due to limited space.

It is recommended that whenever it is necessary to obtain information on significant effects as well as generating response surface curves, Yates' algorithm can be performed on NNs-trained data. This integration

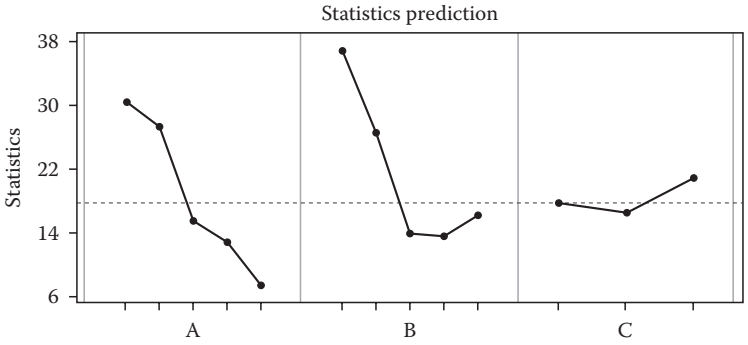


Figure 7.6 Predicted effects using Yates' algorithm: statistics prediction.

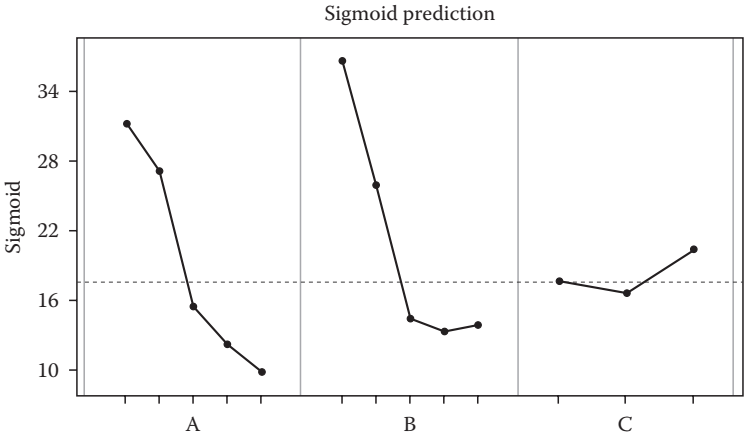


Figure 7.7 Predicted effects using Yates' algorithm: sigmoid prediction function of neural networks.

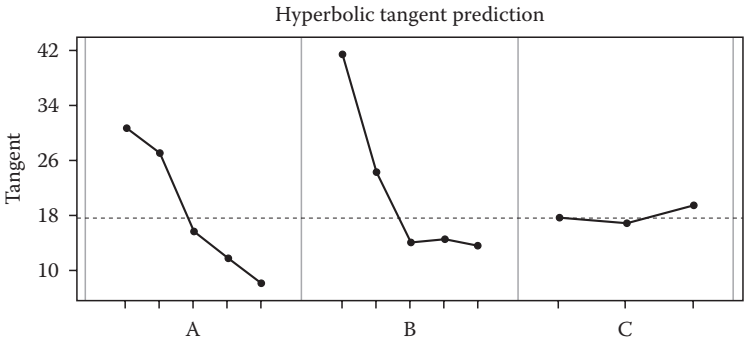


Figure 7.8 Predicted effects using Yates' algorithm: hyperbolic tangent function of neural networks.

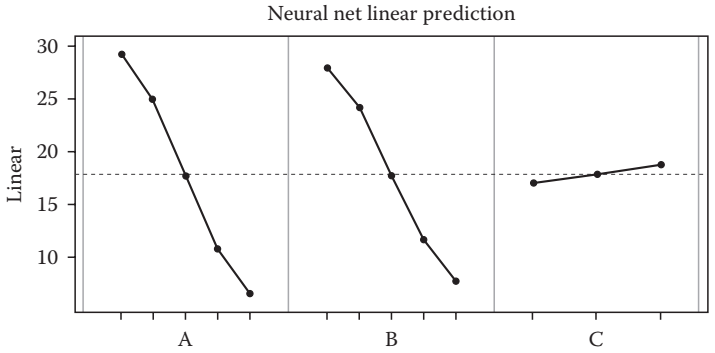


Figure 7.9 Predicted effects using Yates' algorithm: linear prediction neural networks.

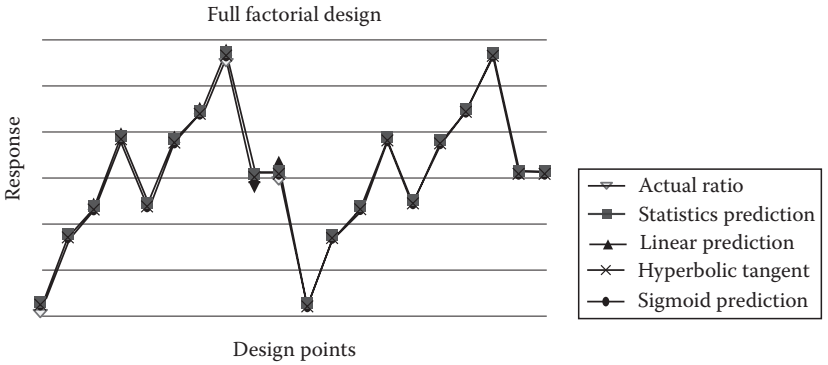


Figure 7.10 Graphical comparison of neural networks and statistical method for a three-factor factorial design.

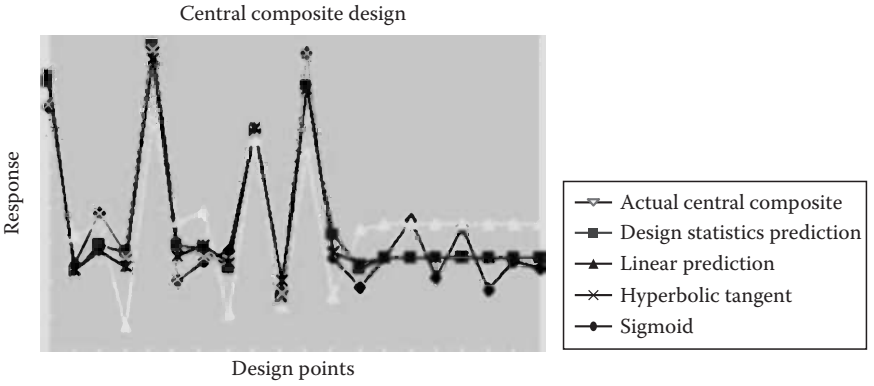


Figure 7.11 Graphical comparison of neural networks and statistical method for a three-factor central composite design.

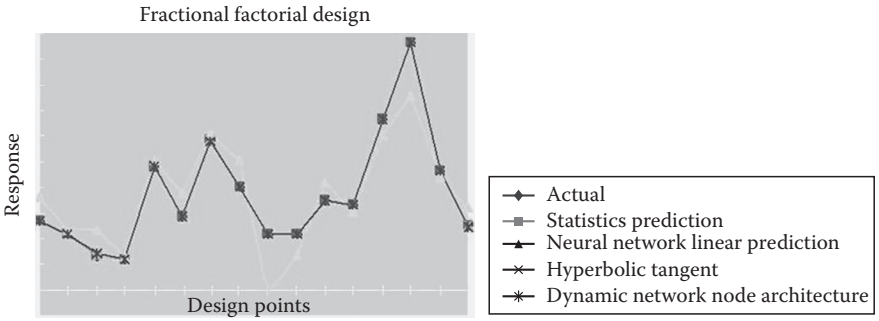


Figure 7.12 Graphical comparison of neural networks and statistical method for 2^{8-4} fractional factorial design.

Table 7.12 Central Composite Design: Statistical Error Comparison of Methods

Parameter	Linear regression statistics	Neural net sigmoid	Neural net linear	Neural Net hyperbolic tangent
Average relative error (ARE)	–0.04972	–0.048602	–0.145163	–0.02388
Forecast root mean square error (FRMSE)	0.23451	0.224877	0.53887	0.17005
Minimum error	–0.52140	–0.51500	–1.0770	–0.53120
Maximum error	0.28630	0.284100	0.74500	0.27650

Table 7.13 A 2^3 Factorial Design: Statistical Error Comparison of Methods

Parameter	Linear regression statistics	Neural net sigmoid	Neural net linear	Neural net hyperbolic tangent
Average relative error (ARE)	–0.00125	–0.006015	–0.00812	–0.00371
Forecast root mean square error (FRMSE)	0.03059	0.040799	0.053185	0.033937
Minimum error	–0.07403	–0.139450	–0.19520	–0.10926
Maximum error	0.05493	0.037000	0.054500	0.048300

of neural nets and statistical methods would enable engineers and statisticians to obtain additional information from NNs approach.

Figures 7.10 through 7.12 provide the graphical comparisons of the various methods (Ayeni, 1995). The number of hidden nodes used for each NN approach was equal to the number of input variables. For example, in

Table 7.14 A 2^{8-4} Fractional Factorial Design: Statistical Error Comparison of Methods

Parameter	Linear regression statistics	Neural net sigmoid	Neural net linear	Neural net hyperbolic tangent	DNNA
Average relative error (ARE)	0.0000	–0.007478	–0.0337	–0.000007	–0.0014
Forecast root mean square error (FRMSE)	0.0000	0.090165	0.39863	0.000035	–0.01377
Minimum error	0.0000	–0.289700	–0.7485	–0.000080	–0.04520
Maximum error	0.0000	0.167300	1.08100	0.000040	0.02037

Table 7.15 Statistical Error Comparison of Methods for 50 Factors Large System

Parameter	Linear regression statistics	Neural net sigmoid	Neural net linear	Neural net hyperbolic tangent	DNNA
Average relative error (ARE)	–0.01421	0.001216	0.93529	0.017745	–0.019631
Forecast root Mean square error (FRMSE)	0.13756	0.128879	0.93739	0.101538	0.139777
Minimum error	–1.23053	–1.14555	0.84756	–0.847010	–1.105730
Maximum error	0.41139	0.424750	0.97797	0.355280	0.454540

a three-factor experiment, we used three hidden nodes. Similarly, for the eight-factor fractional factorial experiment, we used eight hidden nodes. We found that too many hidden nodes can cause overtraining which might lead to predictions that look great. However, predictions on a new data set are generally bad.

The aforementioned examples confirm that NN approach does work if carefully applied. The results obtained under different types of orthogonally designed experiments show that the statistical linear regression method provides as good results as NNs. Except for the linear transfer function of the NNs that failed in the situation where the quadratic effect is significant, the results of other transfer functions compare favorably with the linear statistical method. The results obtained for the large experiment with 50 factors are consistent with those of smaller factors. The tabulated results for different transfer functions are not provided due to limited manuscript space. NNs-trained data can be integrated with statistical methods to obtain additional information about model adequacy, significant factors, as well as generating response surface curves for process optimization.

References

Arps, J. J., Analysis of decline curves, *Transactions of AIME*, 160, 228–247, 1945.

Attili, J., Techniques for synthesizing piecewise linear and quadratic neural network classifiers, *IEEE International Joint Conference on Neural Networks*, Washington, DC, June 22, 1989.

Ayeni, B. J., Parameter estimation for hyperbolic decline curve, *Journal of Energy Resources Technology*, 3 (12), 279–283, December 1989.

Ayeni, B. J., Comparison of results from statistics and neural nets, PTX Video Conference presentation to about 40 3M plants during the *1995 3M Fall Technical Forum Meeting*, November 1995.

Ayeni, B. J. and F. O. Ayeni, Bayesian estimation procedure for petroleum discoveries of an exploratory well, *Journal of Petroleum Science and Engineering*, 9, 281–288, 1993.

Ayeni, B. J., C. J. Hall, and H. H. Koval, Integrating statistics and neural networks, Presented at the *1993 Fall Meeting of 3M Statistical Practitioners Forum*, 3M Internal Publication, 1993.

Ayeni, B. J. and H. H. Koval, Integration of statistics and neural nets for process control application, Presented at the *1995 Annual Meeting of the American Statistical Association*, Orlando, FL. *Proceedings of the Section on Physical and Engineering Sciences of the American Statistical Association*, 1995, pp. 201–205.

Ayeni, B. L. and R. Pilat, Crude oil reserve estimation: An application of the autoregressive integrated moving average (ARIMA) model. 1. *Petroleum Science and Engineering*, 8, 13–28, 1992.

Box, G. E. P., W. G. Hunter, and J. S. Hunter, *Statistics for Experimenters*. John Wiley & Sons, Inc., New York, 1978.

Chung, L. and D. Rowland, Determining the constants of hyperbolic production decline by a linear graphic method, SPE New Papers, Order No. 11329, 1982.

Cockcroft, P., Reserves and probabilities—Synergism or anachronism, *Journal of Petroleum Technology*, 10, 1258–1264, 1991.

De Veaux, R. D., D. C. Psichogios, and L. H. Ungar, A comparison of two nonparametric estimation schemes: MARS and neural networks, *Computers and Chemical Engineering*, 17 (8), 819–837, 1993.

De Veaux, R. D. and L. H. Ungar, Multicollinearity: A tale of two nonparametric regressions. In P. Cheeseman and R. W. Oldford (eds.), *Selecting Models from Data: Al and Statistics IV*. Springer-Verlag, New York, 1994, pp. 393–401.

Draper, N. K. and H. Smith, *Applied Regression Analysis*, 2nd edn. John Wiley & Sons, New York, 1981.

Fetkovitch, M. J., Decline curve analysis using type curves, *Journal of Petroleum Technology*, June, 1067–1077, 1980.

Forsythe, G., M. Malcolm, and C. Moler, *Computer Methods for Mathematical Computations*. Prentice-Hall, Englewood Cliffs, NJ, 1972.

Gentry, R. W., Decline curve analysis, *Journal of Petroleum Technology*, January, 38–41, 1972.

Grayson, C. L., Bayesian analysis—A new approach to statistical decision making, *Petroleum Technology*, 14 (6), 603–607, 1962.

McCray, A. W., *Petroleum Evaluations and Economic Decisions*. Prentice-Hall, Englewood Cliffs, NJ, 1975, 448pp.

McKelvey, V. E., Concepts of reserves and resources, methods of estimating the volume of undiscovered oil and gas resources. In J. D. Haun (ed.), AAPG, Tulsa, Okla, 1975.

Pore, M. D., Bayesian techniques in stratified pro-portion estimation. *Proc. Business Economic Statistics Section*, Am. Stat. Assoc., LEC 13940, Alexandria, Virginia, 1979.

Pore, M. D. and T. B. Dennis, The multicategory case of the sequential Bayesian pixel selection and estimation procedure, Technical report, Lockheed Martin Engineering, LEMSCO-14807, Orlando, Florida, 1980.

Ramsey, H. J. and E. T. Guerrero, The ability of rate-time decline curves to predict production rates, *Journal of Petroleum Technology*, 139–141, February 1969.

Rezayat, F. and J. Holton, Neural network controller and quality improvement: An application, *1994 ASA Proceedings, Section on Quality and Productivity*, Birmingham, Alabama, 1994.

Slider, H. C., A simplified method of hyperbolic decline curve analysis, *Journal of Petroleum Technology*, March, 38–41, 1968.

chapter eight

Mathematical modeling and control of multi-constrained projects

The premise of this chapter is the development of a generic project scheduling tool that incorporates (1) resource characteristics, such as preferences, time-effective capabilities, costs, and availability of project resources, and (2) performance interdependencies among different resource groups, and proceeds to map the most adequate resource units to each newly scheduled project activity. The chapter is based on the work of Milatovic and Badiru (2004). The principal challenge in this generic model development is to make it applicable to realistic project environments, which often involve multifunctional resources, whose capabilities or other characteristics may cross activities, as well as within a single activity relative to specific interactions among resources themselves. The scope of this research challenge further increases when the actual duration, cost, and successful completion of a project activity is assumed as resource driven and dependent on the choice of particular resource units assigned to it.

The proposed methodology dynamically executes two alternative procedures: the *activity scheduler* and *resource mapper*. The *activity scheduler* prioritizes and schedules activities based on their attributes, and may also attempt to centralize selected resource loading graphs based on activity resource requirements. The *resource mapper* considers resource characteristics, incorporates interdependencies among resource groups or types, and maps the available resource units to newly scheduled activities according to a project manager's prespecified mapping (objective) function.

Introduction

Project resources, generally limited in quantity, are the most important constraints in scheduling of activities. In cases when resources have prespecified assignments and responsibilities toward one or more activities, their allocation is concurrently performed with the scheduling of applicable activities. In other cases, an activity may only require a certain number of (generic) resource units of particular type(s), which are assigned after

the scheduling of the particular activity. These two approaches coarsely represent the dominant paradigms in project scheduling. The objective of this research is to propose a new strategy that will shift these paradigms to facilitate a more refined guidance for allocation and assignment of project resources. In other words, there is a need for tools which will provide for more effective resource tracking, control, interaction, and, most importantly, resource-activity mapping.

The main assumption in the methodology of this chapter is that project environments often involve multi-capable resource units with different characteristics. This is especially the case in knowledge-intensive settings and industries, which are predominantly staffed with highly trained personnel. The specific characteristics considered were resource preferences, time-effective capabilities, costs, and availability. Each resource unit's characteristics may further vary across project activities, but also within a single activity relative to interaction among resource units. Finally, resource preferences, cost, and time-effective capabilities may also independently vary with time due to additional factors, such as learning, forgetting, weather, type of work, etc. Therefore, although we do not exclude a possibility that an activity duration is independent of resources assigned to it, in this research, we assume that it is those resource units assigned to a particular activity that determine how long it will take for the activity to be completed.

The scheduling strategy as earlier illustrated promotes a more balanced and integrated activity-resource mapping approach. Mapping the most qualified resources to each project activity, and thus preserving the values of resource, is achieved by proper consideration or resource time-effective capabilities and costs. By considering resource preferences and availability, which may be entered in either crisp or fuzzy form, the model enables consideration of personnel's voice and its influence on a project schedule and quality. Furthermore, resource interactive dependencies may also be evaluated for each of the characteristics and their effects incorporated into resource-activity mapping. Finally, by allowing flexible and dynamic modifications of scheduling objectives, the model permits managers or analysts to incorporate some of their tacit knowledge and discretionary input into project schedules.

Literature review

Literature presents extensive work on scheduling projects by accounting for worker (resource) preferences, qualifications, and skills, as decisive factors to their allocation. Yet, a recent survey of some 400 top contractors in construction showed that 96.2% of them still use basic Critical Path Method (CPM) for project scheduling (Mattila and Abraham, 1998). Roberts (1992) argued that information sources for project planners and

schedulers are increasingly nonhuman, and stressed that planners must keep computerized tools for project management and scheduling in line and perspective with human resources used by projects. In other words, the author warns that too much technicalities may prompt and mislead the managers into ignoring human aspects of management.

Franz and Miller (1993) considered a problem of scheduling medical residents to rotations, and approached it as a large-scale multi-period staff assignment problem. The objective of the problem was to maximize residents' schedule preferences while meeting hospital's training goals and contractual commitments for staffing assistance.

Gray et al. (1993) discussed the development of an expert system to schedule nurses according to their scheduling preferences. Assuming consistency in nurses' preferences, an expert system was proposed and implemented to produce feasible schedules considering nurses' preferences, but also accounting for overtime needs, desirable staffing levels, patient acuity, etc. A similar problem was also addressed by Yura (1994), where the objective was to satisfy worker's preferences for time off as well as overtime, but under due-date constraints.

Campbell (1999) further considered allocation of cross-trained resources in multi-departmental service environment. Employers generally value more resource units with various skills and capabilities for performing greater number of jobs. It is in those cases when managers face challenges of allocating these workers such that the utility of their assignment to a department is maximized. The results of experiments showed that the benefits of cross-training utilization may be significant. In most cases only a small degree of cross-training captured the most benefits, and tests also showed that beyond a certain amount, the additional cross-training adds little additional benefits.

Finally, many companies face problems of continuous reassignment of people in order to facilitate new projects (Cooprider, 1999). Cooprider suggests a seven-step procedure to help companies consider a wide spectrum of parameters when distributing people within particular projects or disciplines.

Badiru (1993) proposed *Critical Resource Diagramming* (CRD), which is a simple extension to traditional CPM graphs. In other words, criticalities in project activities may also be reflected on resources. Different resource types or units may vary in skills, supply, or be very expensive. This discrimination in resource importance should be accounted for when carrying out their allocation in scheduling activities.

Unlike activity networks, the CRDs use nodes to represent each resource units. Also, unlike activities, a resource unit may appear more than once in a CRD network, specifying all different tasks for which a particular unit is assigned to. Similar to CPM, the same backward and forward computations may be performed to CRDs.

Methodology

The methodology of this chapter represents an analytical extension of CRDs discussed in Chapter 7. As previously mentioned, the design considerations of the proposed model consist of two distinct procedures: activity scheduling and resource mapping. At each decision instance during a scheduling process, the *activity scheduler* prioritizes and schedules some or all candidate activities, and then the *resource mapper* iteratively assigns the most adequate resource units to each of the newly scheduled activities.

Representation of resource interdependencies and multifunctionality

This study is primarily focused on renewable resources. In addition, resources are not necessarily categorized into types or groups according to their similarities (i.e., into personnel, equipment, space, etc.), but more according to the hierarchy of their interdependencies. In other words, we assume that time-effective capabilities, preferences, or even cost of any particular resource unit assigned to work on an activity may be dependent on other resource units also assigned to work on the same activity. Some or all of these other resource units may, in similar fashion, be also dependent on a third group of resources, and so on. Based on the assumptions mentioned earlier, we model competency of project resources in terms of following four resource characteristics: time-effective capabilities, preferences, cost, and availability. Time-effective capability of a resource unit with respect to a particular activity is the amount of time the unit needs to complete its own task if assigned to that particular activity. Preferences are relative numerical weights that indicate personnel's degree of desire to be assigned to an activity, or manager's perception on assigning certain units to particular activities. Similarly, each resource unit may have different costs associated with it, relative to which activities it gets assigned to. Finally, not all resource units may be available to some or all activities at all times during project execution. Thus, the times during which a particular unit is available to some or all activities are also incorporated into the mapping methodology. Each of the characteristics described may vary across different project activities. In addition, some or all of these characteristics (especially time-effective capabilities and preferences) may also vary within a particular activity relative to resource interaction with other resources that are also assigned to work on the same activity.

Those resources whose performance is totally independent of their interaction with other units are grouped together and referred to as the

type or group "one" and allocated first to scheduled activities. Resource units whose performance or competency is affected by their interaction with the type or group "one" units are grouped into type or group "two" and assigned (mapped) next. Resource units whose competency or performance is a function of type "two" or both types "one" and "two" are grouped into type "three" and allocated to scheduled activities after the units of the first two types have been assigned to them.

A project manager may consider one, more than one, or all of the four characteristics when performing activity-resource mapping. For example, a manager may wish to keep project costs as low as possible, while at the same time attempting to use resources with the best time-effective capabilities, consider their availability, and even incorporate their voice (in case of humans) or his/her own perception (in cases of human or nonhuman resources) in the form of preferences. This objective may be represented as follows:

$$U_i^{j,k} = f(t_i^{j,k}, c_i^{j,k}, p_i^{j,k}, \mathbf{a}_i^{j,k}(t_c))$$

(See Appendix A for a detailed notation.)

Mapping units of all resource types according to the same mapping function may often be impractical and unrealistic. Cost issues may be of greater importance in mapping some, while inferior to time-effective capabilities of other resource types. To accommodate the need for a resource-specific mapping function as mapping objective, we formulated the mapping function as additive utility function (Keeney and Raiffa, 1993). In such a case, each of its components pertains to a particular resource type and is multiplied by a *Kronecker's delta* function (Bracewell, 1978). Kronecker's delta then detects resource type whose units are currently being mapped and filters out all mapping function components, except the one that pertains to the currently mapped resource type.

As an example, consider again a case where all resource types would be mapped according to their time-effective capabilities, except in the case of resource types "two" and "three" where costs would also be of consideration, and in the case of type "five," resource preferences and availabilities would be considered:

$$U_i^{j,k} = f(t_i^{j,k}) + f_2(c_i^{j,k}) \cdot \mathbf{d}(j,2) + f_3(c_i^{j,k}) \cdot \mathbf{d}(j,3) + f_5(p_i^{j,k}, \mathbf{a}_i^{j,k}(t_c)) \cdot \mathbf{d}(j,5)$$

The aforementioned example illustrates a case where mapping of resource units is performed according to filtered portions of a manager's objective (mapping) function, which may in turn be dynamically adaptive and varying with project scheduling time. As previously indicated, some

resource characteristics may be of greater importance to a manager in the early scheduling stages of a project rather than in the later stages. Such a mapping function may be modeled as follows:

$$\mathrm{U}_i^{j,k} = f_g(t_i^{j,k}, c_i^{j,k}, p_i^{j,k}, a_i^{j,k}(t_c)) + \sum_{s \in T} f_s(t_i^{j,k}, c_i^{j,k}, p_i^{j,k}, a_i^{j,k}(t_c)) \cdot w(t_{LO}^s, t_{HI}^s, t_c)$$

where

f_g is the component of the mapping function that is common to all resource types

f_s is the component of the mapping function that pertains to a *specific* project scheduling interval

t_{LO}^s, t_{HI}^s are specific time interval during which resource mapping must be performed according to a unique function

T is the set of earlier-defined time intervals for a particular project

$w(t_{LO}^s, t_{HI}^s, t_c)$ is the window function with a value of one if t_c falls within the interval $[t_{LO}^s, t_{HI}^s)$, and zero otherwise

Finally, it is also possible to map different resource types according to different objectives and at different times simultaneously, by simply combining the two concepts mentioned earlier. For example, assume again that a manager forms his objective in the early stage of the project based on resources' temporal capabilities, costs, and preferences. Then at a later stage, the manager wishes to drop the costs and preferences and consider only resource capabilities, with the exception of resource type "three" whose costs should still remain in consideration for mapping. An example of a mapping function that would account for this scenario may be as follows:

$$\mathrm{U}_i^{j,k} = f(c_i^{j,k}, p_i^{j,k}, t_i^{j,k}) \cdot w(0, 30, t_c) + \left(f(t_i^{j,k}) + f(c_i^{j,k}) \cdot \mathrm{d}(j, 3) \right) \cdot w(30, 90, t_c)$$

Modeling of resource characteristics

For resource units whose performance on a particular activity is independent of their interaction with other units, that is, for the *drivers*, $t_i^{j,k}$ is defined as the time, t, it takes kth unit of resource type j to complete its own task or process when working on activity i. Thus, different resource units, if multi-capable, can be expected to perform differently on different activities. Each *dependent* unit, on the other hand, instead of $t_i^{j,k}$, generally has a set of interdependency functions associated with it.

In this research we consider two types of interactive dependencies among resources, which, due to their simplicity, are expected to be the most commonly used ones: *additive* and *percentual*. *Additive* interaction between a *dependent* and each of its *driver* resource unit indicates the amount of time

that the *dependent* will need to complete its own task if assigned to work in conjunction with a particular driver. This is in addition to the time the driver itself needs to spend working on the same activity:

$$(T_i^{j,k})_z \equiv (t_i^{j_D,k_D} + \tilde{t}_i^{j,k}) \cdot y_i^{j_D,k_D}$$

where

- $<j_D,k_D> \in D^{j,k}$, where $D^{j,k}$ is a set of *driver* units (each defined by an indexed pair $<j_D,k_D>$) for a particular resource unit $<j,k>$
- $(T_i^{j,k})_z$ is the zth interactive time-effective dependency of kth unit of type j on its *driver* $<j_D,k_D>$, $z = 1,\ldots,$ size($D^{j,k}$). The actual number of these dependencies will depend on a manager's knowledge and familiarity with his/her resources
- $\tilde{t}_i^{j,k}$ is the time needed in addition to $t_i^{j_D,k_D}$ for kth *dependent* unit of type j to complete its task on activity i if it interacts with its *driver* unit j_D,k_D
- $y_i^{j_D,k_D}$ is the binary (zero-one) variable indicating mapping status of the *driver* unit $<j_D,k_D>$. It equals one if the unit $<j_D,k_D>$ is assigned to activity i, and zero if the unit $<j_D,k_D>$ has been assigned to activity i. Therefore, each $(T_i^{j,k})_z$ will have a nonzero value only if $y_i^{j_D,k_D}$ is also nonzero (i.e., if the *driver* resource unit $<j_D,k_D>$ has been previously assigned to activity i)

The percentual interactive dependency is similarly defined as

$$(T_i^{j,k})_z = t_i^{j_D,k_D} \cdot (1 + \tilde{t}_i^{j,k}\%) \cdot y_i^{j_D,k_D}$$

where $\tilde{t}_i^{j,k}\%$ is the percentage of time by which $t_i^{j_D,k_D}$ will be prolonged if the unit k of type j interacts with its *driver* $<j_D,k_D>$.

Modeling cost characteristics follows a similar logic used for representation of temporal capabilities and interdependencies. In place of $t_i^{j,k}$, we now define a variable $c_i^{j,k}$, which represents the cost (say, in dollars) of kth unit of resource type j if it gets assigned to work on activity i. This value of $c_i^{j,k}$ may be invariant regardless of a unit's interaction with other resources, or it may vary relative to interaction among resources, and thus, implying cost interdependencies, which need to be evaluated before any mapping is performed (provided that the cost considerations are, indeed, a part of a manager's utility or objective for mapping).

In cases when a cost of a resource unit for an activity varies depending on its interaction with units of other (lower indexed) types, we define cost dependencies as

$$(C_i^{j,k})z = \tilde{c}_i^{j,k} \cdot y_i^{j_D,k_D}$$

where

$y_i^{j_D,k_D}$ is a binary variable indicating the status of the particular *driver* resource unit $< j_D, k_D >$, as defined in the previous section

$\tilde{c}_i^{j,k}$ is the interactive cost of kth unit of type j on its *driver* $< j_D, k_D >$, with respect to activity i

$(\boldsymbol{C}_i^{j,k})z$ is the zth evaluated interactive cost dependency of kth unit of type j on its *driver* $< j_D, k_D >$, $z = 1, \ldots,$ size$(D^{j,k})$. The values of each $(\boldsymbol{C}_i^{j,k})_z$ equals $\tilde{c}_i^{j,k}$ when $y_i^{j_D,k_D}$ equals one, and zero otherwise. The actual number of these interactive cost dependencies will again depend on a manager's knowledge and information about available resources

Given a set of cost dependencies, we compute the overall $c_i^{j,k}$ as a sum of all evaluated $(\boldsymbol{C}_i^{j,k})$zs as follows:

$$c_i^{j,k} = \sum_{z=1}^{|D^{j,k}|} (C_i^{j,k})_z$$

In many instances, due to political, environmental, safety, or community standards, aesthetics, or other similar nonmonetary reasons, pure monetary factors may not necessarily prevail in decision making. It is those other nonmonetary factors that we wish to capture by introducing preferences in resource mapping to newly scheduled activities. The actual representation of preferences is almost identical to those of the costs:

$$(P_i^{j,k})z = \tilde{p}_i^{j,k} \cdot y_i^{j_D,k_D}$$

where $\tilde{p}_i^{j,k}$ is an interactive preference of kth unit of type j on its *driver* $< j_D, k_D >$, with respect to activity i. $(P_i^{j,k})z$ is zth evaluated interactive preference dependency of kth unit of type j, with respect to activity i. Finally, again identically to modeling costs, $p_i^{j,k}$ is computed as

$$p_i^{j,k} = \sum_{z=1}^{|D^{j,k}|} (P_i^{j,k})_z$$

Having certain number of resource units of each type available for a project does not necessarily imply that all of the units are available all the time for the project or any of its activities in particular. Due to transportation, contracts, learning, weather conditions, logistics, or other factors, some units may only have *time preferences* for when they are available to start working on a project activity or the project as a whole. Others may have *strict time*

intervals during which they are allowed to start working on a particular activity or the project as a whole. This latter, strictly constrained availability may be easily accommodated by the previously considered *window* function, $w(t_{LO}, t_{HI}, t_c)$.

In many cases, especially for humans, resources may have a desired or "ideal" time when to start their work or be available in general. This flexible availability can simply be represented by fuzzifying the specified desired times using the following function:

$$\mathtt{a}_i^{j,k}(t_c) = \frac{1}{1 + a(t_c - \mathtt{t}_i^{j,k})^b}$$

where

$\mathtt{t}_i^{j,k}$ is the desired time for kth unit of resource type j to start its task on activity i. This desirability may either represent the voice of project personnel (as in the case of preferences), or manager's perception on resource's readiness and availability to take on a given task

$\mathtt{a}_i^{j,k}(t_c)$ is the fuzzy membership function indicating a degree of desirability of $<j,k>$th unit to start working on activity i, at the decision instance t_c

a is the parameter that adjusts for the width of the membership function

b is the parameter that defines the extent of *start time* flexibility

Resource mapper

At each scheduling time instance, t_c, available resource units are mapped to newly scheduled activities. This is accomplished by solving J number of zero-one linear integer problems (i.e., one for each resource type), where the coefficients of the decision vector correspond to evaluated mapping function for each unit of the currently mapped resource type:

$$max \sum_{h \in \Omega(t_c)} \sum_{k=1}^{R_j} U_h^{j,k} \cdot y_h^{j,k} \quad \text{for } j = 1, \ldots, J$$

where

$y_h^{j,k}$ is the binary variable of the decision vector

$\Omega(t_c)$ is the set of newly scheduled activities at decision instance t_c

A $y_i^{j,k}$ resulting in a value of one would mean that kth unit of resource type j is mapped to ith ($i \in \Omega(t_c)$) newly scheduled activity at t_c. The aforementioned

objective in each of J number of problems is subjected to four types of constraints, as illustrated in the following:

1. The first type of constraints ensures that each newly scheduled activity receives its required number of units of each project resource type:

$$\sum_{k=1}^{R_j} y_i^{j,k} = \mathrm{r}_i^j \quad \text{for } i \in \Omega(t_c) \quad \text{for } j = 1,\ldots,J$$

2. The second type of constraints prevents mapping of any resource units to more than one activity at the same time at t_c:

$$\sum_{i \in \Omega(t_c)} y_i^{j,k} \le 1 \quad \text{for } k = 1,\ldots,R_j \quad \text{for } j = 1,\ldots,J$$

3. The third type of constraints prevents mapping of those resource units that are currently in use by activities in progress at time t_c:

$$\sum_{k=1}^{R_j} u_{t_c}^{j,k} \cdot y_i^{j,k} = 0 \quad \text{for } i \in \Omega(t_c) \quad \text{for } j = 1,\ldots,J$$

4. The fourth type of constraints ensures that the variables in the decision vector $y_i^{j,k}$ take on binary values:

$$y_i^{j,k} = 0 \text{ or } 1 \quad \text{for } k = 1,\ldots,R_j, \quad i \in \Omega(t_c), \quad \text{for } j = 1,\ldots,J$$

Therefore, in the first of the total of J runs at each decision instance t_c, available units of resource type "one" compete (based on their characteristics and prespecified mapping function) for their assignments to newly scheduled activities. In the second run, resources of type "two" compete for their assignments. Some of their characteristics, however, may vary depending on the "winners" from the first run. Thus, the information from the first run is used to refine the mapping of type or group "two" resources. Furthermore, the information from either or both of the first two runs is then used in tuning the coefficients of the objective function for the third run when resources of type "three" are mapped.

Due to the nature of linear programming, zeros in the coefficients of the objective do not imply that corresponding variables in the solution will also take the value of zero. In our case, that would mean that although we flagged off a resource unit as unavailable, the solution may still map it to an activity. Thus, we need to strictly enforce the interval (un)

availability by adding information into constraints. Thus, we perturbed the third mapping constraint which was previously set to prohibit mapping of resource units at time t_c which are in use by activities in progress at that time. The constraint was originally defined as

$$\sum_{k=1}^{R_j} u_{t_c}^{j,k} \cdot y_i^{j,k} = 0 \quad \text{for } i \in \Omega(t_c) \quad \text{for } j = 1,\ldots,J$$

To now further prevent mapping of resource units whose $\mathbf{a}_i^{j,k}(t_c)$ equals zero at t_c, we modify the aforementioned constraint as follows:

$$\sum_{k=1}^{R_j} \left(u_{t_c}^{j,k} + \left(1 - \mathbf{a}_i^{j,k}(t_c)\right)\right) \cdot y_i^{j,k} = 0 \quad \text{for } i \in \Omega(t_c) \quad \text{for } j = 1,\ldots,J$$

This modified constraint now not only filters out those resource units that are engaged in activities in progress at t_c, but also those units which were flagged as unavailable at t_c due to any other reasons.

Activity scheduler

Traditionally, a project manager estimates duration of each project activity first, and then assigns resources to it. In this study, although we do not exclude a possibility that an activity duration is independent of resources assigned to it, we assume that it is those resource units and their skills or competencies assigned to a particular activity that determine how long it will take for the activity to be completed. Normally, more capable and qualified resource units are likely to complete their tasks faster, and vice versa. Thus, activity duration in this research is considered a *resource-driven activity attribute*.

At each decision instance t_c (in resource-constrained non-preemptive scheduling as investigated in this study), activities whose predecessors have been completed enter the set of qualifying activities, $Q(t_c)$. In cases of resource conflicts we often have to prioritize activities in order to decide which ones to schedule. In this methodology we prioritize activities based on two (possibly conflicting) objectives:

1. Basic *activity attributes*, such as the *current amount of depleted slack*, number of successors, and initially estimated optimistic activity duration, d_i
2. Degree of manager's desire to *centralize* (or balance) *the loading* of one or more preselected project resource types

Amount of Depleted Slack, $S_i(t_c)$, is defined in this research as a measure of how much total slack of an activity from unconstrained CPM computations has been depleted each time the activity is delayed in resource-constrained

scheduling due to lack of available resource units. The larger the $S_i(t_c)$ of an activity, the more it has been delayed from its unconstrained schedule, and the greater probability that it will delay the entire project.

Before resource-constrained scheduling of activities (as well as resource mapping which is performed concurrently) starts, we perform a single run of CPM computations to determine initial unconstrained *Latest Finish Time*, LFT_i of each activity. Then, as the resource-constrained activity scheduling starts, at each decision instance t_c, we calculate $S_i(t_c)$ for each candidate activity (from the set $Q(t_c)$) as follows:

$$S_i(t_c) = \frac{t_c + d_i}{LFT_i} = \frac{t_c + d_i}{LST_i + d_i} \quad i \in Q(t_c)$$

$S_i(t_c)$, as a function of time, is always a positive real number. The value of its magnitude is interpreted as follows:

- When $S_i(t_c) < 1$, the activity i still has some slack remaining and it may be safely delayed.
- When $S_i(t_c) = 1$, the activity i has depleted all of its resource-unconstrained slack and any further delay to it will delay its completion as initially computed by conventional unconstrained CPM.
- When $S_i(t_c) > 1$, the activity i has exceeded its slack and its completion will be delayed beyond its unconstrained CPM duration.

Once calculated at each t_c, the current *amount of depleted*, $S_i(t_c)$, is then used in combination with the other two activity attributes for assessing activity priority for scheduling. (These additional attributes are *the number of activity successors*, as well as its *initially estimated duration* d_i.) The number of successors is an important determinant in prioritizing, because if an activity with many successors is delayed, chances are that any of its successors will also be delayed, thus eventually prolonging the entire project itself. Therefore, the prioritizing weight, w_p^t, pertaining to basic activity attributes is computed as follows:

$$w_i^p = S_i(t_c) \cdot \left(\frac{\varsigma_i}{\max(\varsigma_i)} \right) \cdot \left(\frac{d_i}{\max(d_i)} \right)$$

where

w_i^p is the activity prioritizing weight that pertains to basic activity attributes

ς_i is the number of successors activities of current candidate activity i

$max(\varsigma_i)$ is the maximum number of activity successors in project network

$max(d_i)$ is the maximum of the most optimistic activity durations in a project network

The second objective that may influence activity prioritizing is a manager's desire for a somewhat centralized (i.e., balanced) resource loading graph for one or more resource groups or types. This is generally desirable in cases when a manager does not wish to commit all of the available project funds or resources at the very beginning of the project (Dreger, 1992), or to avoid frequent hiring and firing of project resources (Badiru and Simin Pulat, 1995).

In this research, we attempt to balance (centralize) loading of prespecified resources by scheduling those activities whose resource requirements will minimize the increase in loading graph's stair step size of the early project stages, and then minimize the decrease in the step size in the later stages. A completely balanced resource loading graph contains no depression regions, as defined by Konstantinidis (1998), that is, it is a nondecreasing graph up to a certain point at which it becomes nonincreasing.

The activity prioritizing weight that pertains to attempting to centralize resource loading is computed in this research as follows:

$$w_i^r = \sum_{j=1}^{J} \frac{\mathtt{r}_i^j}{R_j}$$

where

w_i^r is the prioritizing weight that incorporates activity resource requirements
$\mathtt{r}_i^j$ is the number of resource type j units required by activity i
R_j is the total number of resource type j units required for the project

Notice that w_i^p and w_i^r are weights of possibly conflicting objectives in prioritization of candidate activities for scheduling.

To further limit the range of w_i^r between zero and one, we scale it as follows:

$$w_i^r = \frac{w_i^r}{\max(w_i^r)}$$

With the two weights w_i^p and w_i^r defined and computed, we further use them as the coefficients of activity scheduling objective function:

$$max\left(\sum_{i\in Q(t_c)} w_i^p \cdot x_i\right) + W\left(\sum_{i\in Q(t_c)} (1 - w_i^r \cdot x_i)\right)$$

where

x_i is the binary variable whose value becomes one if a candidate activity $i \in Q(t_c)$ is scheduled at t_c, and zero if the activity i is not scheduled at t_c

W is the decision maker's supplied weight that conveys the importance of resource centralization (balancing) in project schedule

Notice that W is a parameter that allows a manager to further control the influence of w_i^p. Large values of W will place greater emphasis on the importance of resource balancing. However, to again localize the effect of W to the early stages of a project, we dynamically decrease its value at each subsequent decision instance, t_c according to the following formula:

$$W_{new} = W_{old}\left(\frac{\sum_{i=1}^{I} d_i - \sum_{i \in H(t_c)} d_i}{\sum_{i=1}^{I} d_i}\right)$$

where

$\sum_{i=1}^{I} d_i$ is the sum of all the most optimistic activity durations (as determined by conventional resource-unconstrained CPM computations) for all activities in project network

$H(t_c)$ is the set of activities that have been so far scheduled by the time t_c

Figure 8.1 shows a Gantt chart and resource loading graphs of sample project with seven activities and two resource types. The bottom plot in the figure is the Gantt chart, the middle plot is the resource loading for resource type one, and the top plot is the resource loading for resource type two. Clearly, neither of the two resource types is balanced. The same project has been rerun using the aforementioned reasoning, and shown in Figure 8.2. Notice that the loading of resource type two is now fully balanced. The loading of resource type one still contains depression regions, but to a considerably lesser extent than in Figure 8.1.

With the two weights w_i^p and w_i^r defined and computed, we further use them as the coefficients of activity scheduling objective function:

$$max\left(\sum_{i \in Q(t_c)} w_i^p \cdot x_i\right) + W\left(\sum_{i \in Q(t_c)} (1 - w_i^r \cdot x_i)\right)$$

where

x_i is the binary variable whose value becomes one if a candidate activity $i \in Q(t_c)$ is scheduled at t_c, and zero if the activity i is not scheduled at t_c

W is the decision maker's supplied weight that conveys the importance of resource centralization (balancing) in project schedule

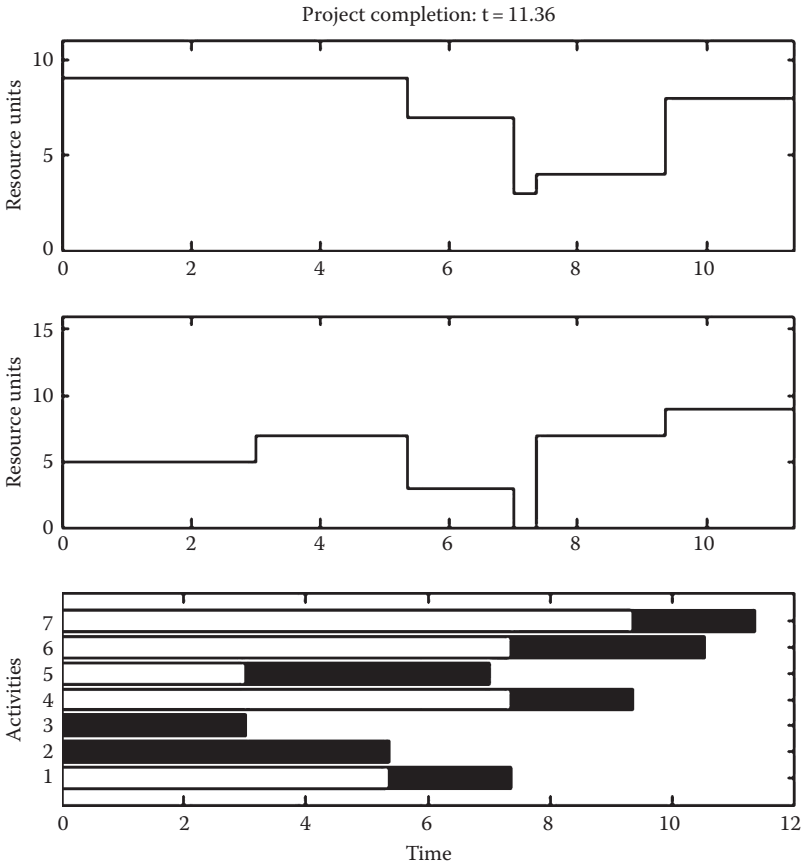


Figure 8.1 Gantt chart and resource loading graphs.

Large values of W will place greater emphasis on the importance of resource balancing. However, to again localize the effect of W to the early stages of a project, we dynamically decrease its value at each subsequent decision instance, t_c according to the following formula:

$$W_{new} = W_{old}\left(\frac{\sum_{i=1}^{I} d_i - \sum_{i \in H(t_c)} d_i}{\sum_{i=1}^{I} d_i}\right)$$

where

$\sum_{i=1}^{I} d_i$ is the sum of all the most optimistic activity durations (as determined by conventional resource-unconstrained CPM computations) for all activities in project network

$H(t_c)$ is the set of activities that have been so far scheduled by the time t_c

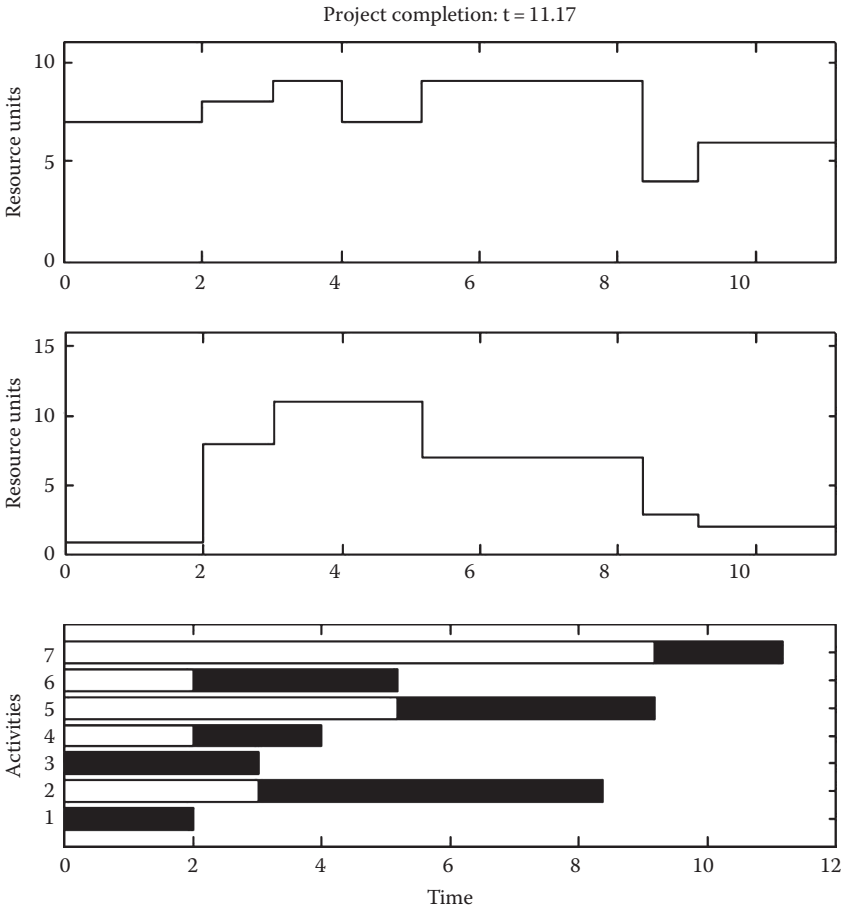


Figure 8.2 Rerun of Gantt chart and resource loading graphs.

Previously, it was proposed that one way of balancing resource loading was to keep minimizing the increase in the stair step size of the loading graph in the early project stages, and then minimize the decrease in the step size in the later stages. The problem with this reasoning is that a continuous increase in the loading graph in early stages may eventually lead to scheduling infeasibility due to limiting constraints in resource availability. Therefore, an intelligent mechanism is needed that will detect the point when resource constraints become binding and force the scheduling to proceed in a way that will start the decrease in resource loading, as shown in Figure 8.1. In other words, we need to formulate a linear programming model whose constraints will drive the increase in resource stair step–shaped loading function up to a point when limits in resource availability are reached. At that point, the model

must adjust the objective function and modify (relax) the constraints, in order to start minimizing the stair step decrease of resource loading.

To ensure this, the constraints are formulated such that at each decision instance t_c, maximal number of candidate activities are scheduled, while satisfying activity precedence relations, preventing the excess of resource limitations, and most importantly, flag off the moment when resource limitations are reached. To facilitate a computer implementation and prevent the strategy from crashing, we introduce an auxiliary zero–one variable, $\hat{x}$, in this study referred to as the *peak flag*. The value of $\hat{x}$ in the decision vector is zero as long as current constraints are capable of producing a feasible solution. Once that is impossible, all variables in the decision vector must be forced to zero, except $\hat{x}$, which will then take a value of one and indicate that the peak of resource loading is reached. At that moment, the constraints that force the increase in resource loading are relaxed (eliminated).

The *peak flag* is appended to the previous objective function as follows:

$$max\left(\sum_{i\in Q(t_c)} w_i^p \cdot x_i\right) + \mathrm{W}\left(\sum_{i\in Q(t_c)} \left(1 - w_i^r \cdot x_i\right)\right) - b\hat{x}$$

where b is the arbitrary large positive number (in computer implementation of this study, b was taken as $b = \sum_{i=1}^{I} d_i$).

There are two types of constraints associated with the aforementioned objective of scheduling project activities. The first type simply serves to prevent scheduling of activities, which would overuse available resource units:

$$\sum_{i\in Q(t_c)} \mathrm{r}_i^j \cdot x_i + \left(R_j - \sum_{i\in G(t_c)} \mathrm{r}_i^j\right)\hat{x} \le \left(R_j - \sum_{i\in G(t_c)} \mathrm{r}_i^j\right), \quad j = 1,\ldots,J$$

where

x_i is the candidate activity qualified to be scheduled at t_c

$G(t_c)$ is the set of activities that are in progress at time t_c

$\left(R_j - \sum_{i\in G(t_c)} \mathrm{r}_i^j\right)$ is the difference between the total available units of resource type j (denoted as R_j) and the number of units of the same resource type being currently consumed by the activities in progress during the scheduling instant t_c

The second type of constraints serves to force the gradual increase in the stair step resource loading graphs. In other words, at each scheduling

instant t_c, this group of constraints will attempt to force the model to schedule those candidate activities whose total resource requirements are greater than or equal to the total requirements of the activities that have just finished at t_c. The constraints are formulated as follows:

$$\sum_{i\in Q(t_c)} r_i^j x_i + \left(\sum_{i\in F(t_c)} r_i^j\right)\hat{x} \geq \left(\sum_{i\in F(t_c)} r_i^j\right), \quad j \in \mathcal{D}$$

where

$F(t_c)$ is the set of activities that have been just completed at t_c

$\mathcal{D}$ is the set of manager's preselected resource types whose loading graphs are to be centralized (i.e., balanced)

$\left(\sum_{i\in F(t_c)} r_i^j\right)$ is the total resource type j requirements by all activities that have been completed at the decision instance t_c

Finally, to ensure an integer zero–one solution, we impose the last type of constraints as follows:

$$x_i = 0 \text{ or } 1, \quad \text{for } i \in Q(t_c)$$

As previously discussed, once $\hat{x}$ becomes unity, we adjust the objective function and modify the constraints that will, from that point on, allow a decrease in resource loading graph(s). Objective function for activity scheduling is modified such that the product $w_i^r \cdot x_i$ is not being subtracted from one any more, while the second type of constraints is eliminated completely:

$$min\left(-\sum_{i\in Q(t_c)} w_i^t \cdot x_i\right) - W\left(\sum_{i\in Q(t_c)} w_i^r \cdot x_i\right)$$

subject to

$$\sum_{i\in Q(t_c)} r_i^j \cdot x_i \leq \left(R_j - \sum_{i\in G(t_c)} r_i^j\right), \quad j = 1,\ldots,J$$

$x_i = 0$ or 1.

Since the second type of constraints is eliminated, resource loading function is now allowed to decrease. The first type of constraints still remains in place to prevent any overuse of available resources.

Model implementation and graphical illustrations

The model as described previously has been implemented in a software prototype *Project Resource Mapper* (PROMAP), with its code, input format, and sample outputs illustrated in the appendices. The output consists of five types of charts. The more traditional ones include project *Gantt chart* (Figure 8.3), and *resource loading graphs* (Figure 8.4) for all resource groups or types involved in a project. More specific graphs include *resource-activity mapping grids* (Figure 8.5), *resource utilization* (Figure 8.6) and *resource cost* (Figure 8.7) bar charts. Based on the imported resource characteristics, their interdependencies, and the form of the objective, the *resource-activity mapping grid* provides a decision support in terms of which units of each specified resource group should be assigned to which particular project activity. Therefore, the *resource-activity grids* are, in effect, the main contributions of this study. *Unit utilization charts* track the resource assignments and provide a relative resource usage of each unit relative to the total project duration. The bottom (darker shaded) bars indicate the total time it

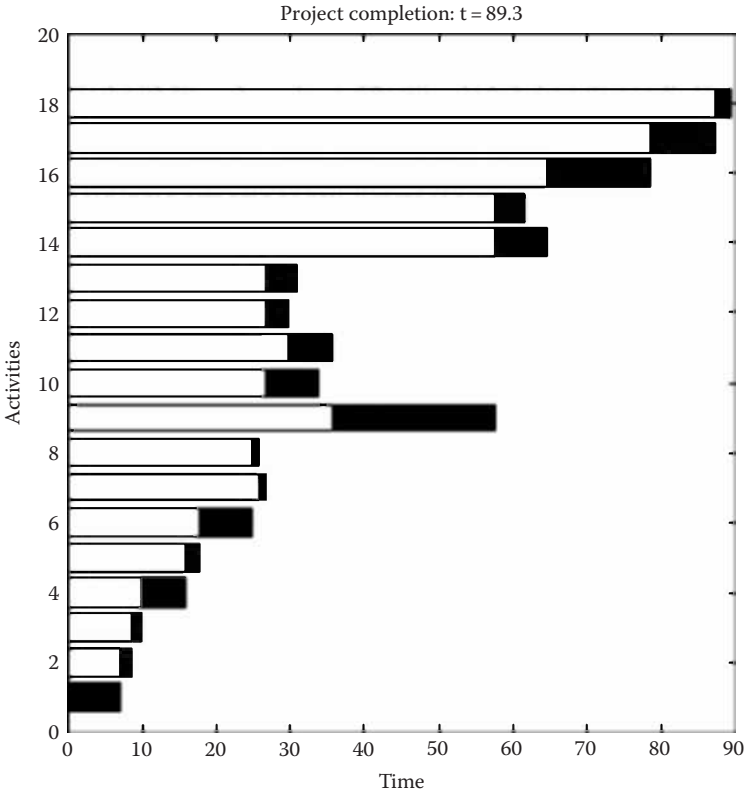


Figure 8.3 Project completion at time 89.3.

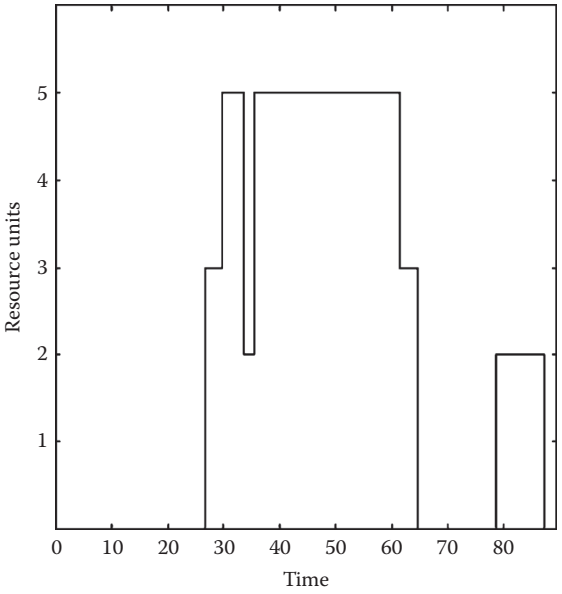


Figure 8.4 Resource type 3 loading graph.

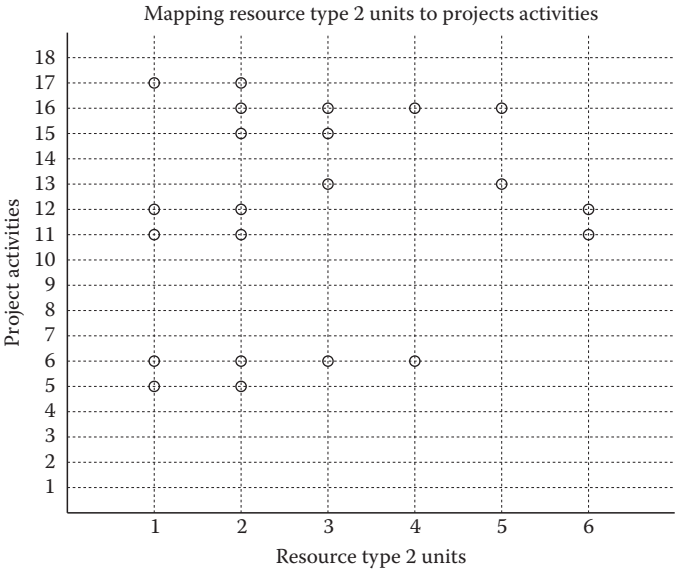


Figure 8.5 Mapping of resource type 2 units to project activities.

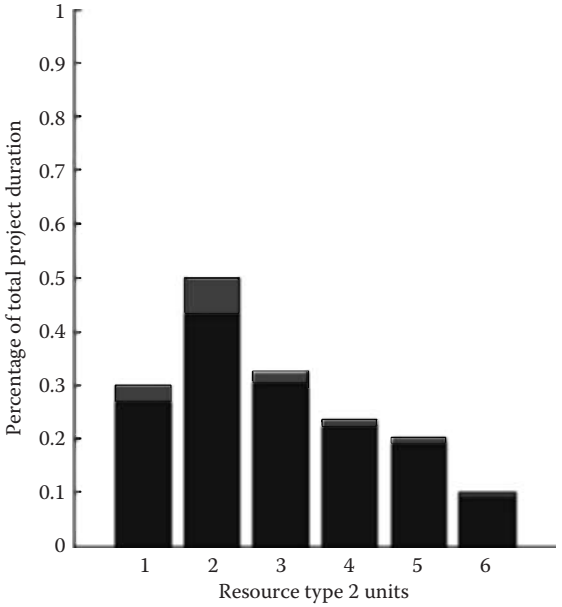


Figure 8.6 Time percentage of resource type 2 units engagement versus total project duration.

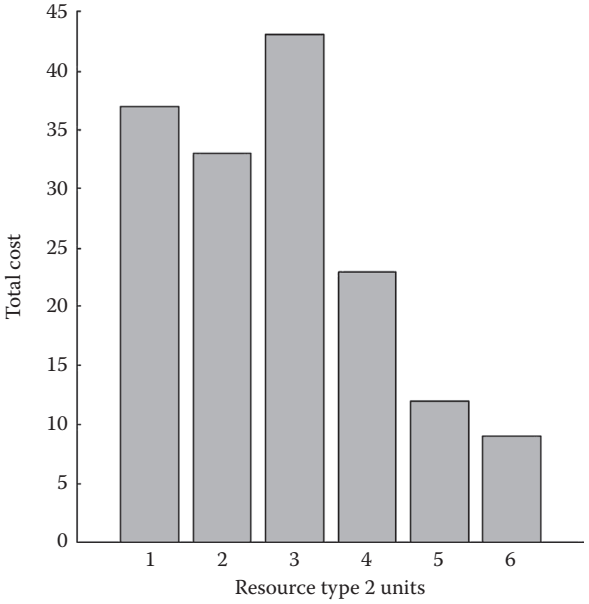


Figure 8.7 Project cost for type 2 resource units.

takes each unit to complete all of its own project tasks. The upper (lighted shaded) bars indicate the total additional time a unit may be locked in or engaged in an activity by waiting for other units to finish their tasks. In other words, the upper bars indicate the total possible resource idle time during which it cannot be reassigned to other activities because it is blocked waiting for other units to finish their own portions of work. This information is very useful in non-preemptive scheduling as assumed in this study, as well as in contract employment of resources. *Resource cost* charts compare total project resource expenditures for each resource unit.

The model developed in this chapter represents an initial step toward a more comprehensive resource-activity integration in project scheduling and management. It provides for both, effective activity scheduling based on dynamically updated activity attributes, as well as intelligent iterative mapping of resources to each activity based on resource characteristics and preselected shape of project manager's objectives. The model consists of two complementary procedures: an *activity scheduler* and *resource mapper*. The procedures are alternatively being executed throughout the scheduling process at each newly detected decision instance, such that the final output is capable of providing decision support and recommendations with respect to both, scheduling project activities and resource assignments. This approach allows human, social, as well as technical resources to interact and be utilized in value creating ways, while facilitating effective resource tracking and job distribution control.

Notations

i	Project activity i, such that $i = 1,\ldots, I$.
I	Number of activities in project network.
t_c	Decision instance, that is, time moment at which one or more activities qualify to be scheduled since their predecessor activities have been completed.
$PR(i)$	Set of predecessor activities of activity i.
$Q(t_c)$	Set of activities qualifying to be scheduled at t_c, i.e., $Q(t_c) = \{i \mid PR(i) = \emptyset\}$.
j	Resource type j, $j = 1,\ldots, J$.
J	Number of resource types involved in the project.
R_j	Number of units of resource type j available for the project.
<j,k>	Notation for kth unit of type j.
r_i^j	Number of resource units type j required by activity i.
$u_{t_c}^{j,k}$	A binary variable with a value of one if kth unit of type j is engaged in one of the project activities that are in progress at the decision instance t_c, and zero otherwise. All $u_{t_c}^{j,k}$'s are initially set to zero.
$t_i^{j,k}$	Time-effective executive capability of kth unit of resource type j if assigned to work on activity i.

$p_i^{j,k}$ Preference of kth unit of resource type j to work on activity i.

$c_i^{j,k}$ Estimated cost of kth unit of resource type j if assigned to work on activity i.

$\mathbf{a}_i^{j,k}(t_c)$ Desired start time or interval availability of kth unit of type j to work on activity i at the decision instance t_c. In many cases this parameter is invariant across activities, and the subscript i may often be dropped.

References

Badiru, A. B., Activity resource assignments using critical resource diagramming, *Project Management Journal*, 14 (3), 15–21, 1993.

Badiru, A. B. and P. Simin Pulat, *Comprehensive Project Management: Integrating Optimization Models, Management Principles, and Computers*. Prentice Hall, Upper Saddle River, NJ, 1995, pp. 162–209.

Bracewell, R. N., *The Fourier Transform and Its Applications*. McGraw-Hill, Inc., New York, 1978, p. 97.

Campbell, G. M., Cross-utilization of workers whose capabilities differ, *Management Science*, 45 (5), 722–732, 1999.

Cooprider, C., Solving a skill allocation problem, *Production and Inventory Management Journal*, Third Quarter, 1–6, 1999.

Dreger, J. B., *Project Management: Effective Scheduling*. Van Nostrand Reinhold, New York, 1992, pp. 202–231.

Franz, L. S. and J. L. Miller, Scheduling medical residents to rotations: Solving the large scale multiperiod staff assignment problem, *Operations Research*, 41 (2), 269–279, 1993.

Gray, J. J., D. McIntire, and H. J. Doller, Preferences for specific work schedulers: Foundation for an expert-system scheduling program, *Computers in Nursing*, 11 (3), 115–121, 1993.

Keeney, R. L. and H. Raiffa, *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Cambridge University Press, Cambridge, NY, 1993.

Konstantinidis, P. D., A model to optimize project resource allocation by construction of a balanced histogram, *European Journal of Operational Research*, 104, 559–571, 1998.

Mattila, K. G. and D. M. Abraham, Resource leveling of linear schedules using integer linear programming, *Journal of Construction Engineering and Management*, 124 (3), 232–244, 1998.

Milatovic, M. and A. B. Badiru, Applied mathematics modeling of intelligent mapping and scheduling of interdependent and multifunctional project resources, *Applied Mathematics and Computation*, 149 (3), 703–721, 2004.

Roberts, S. M., Human skills—Keys to effectiveness, *Cost Engineering*, 34 (2), 17–19, 1992.

Yura, K., Production scheduling to satisfy worker's preferences for days off and overtime under due-date constraints, *International Journal of Production Economics*, 33, 265–270, 1994.

chapter nine

*Online support vector regression with varying parameters for time-dependent data**

Support vector regression (SVR) is a machine-learning technique that continues to receive interest in several domains including manufacturing, engineering, and medicine (Vapnik, 1998). In order to extend its application to problems in which data sets arrive constantly and in which batch processing of the data sets is infeasible or expensive, an accurate online support vector regression (AOSVR) technique was proposed. The AOSVR technique efficiently updates a trained SVR function whenever a sample is added to or removed from the training set without retraining the entire training data. However, the AOSVR technique assumes that the new samples and the training samples are of the same characteristics; hence, the same value of SVR parameters is used for training and prediction. This assumption is not applicable to data samples that are inherently noisy and nonstationary such as sensor data. As a result, we propose AOSVR with varying parameters (AOSVR-VP) that uses varying SVR parameters rather than fixed SVR parameters and hence accounts for the variability that may exist in the samples. To accomplish this objective, we also propose a generalized weight function to automatically update the weights of SVR parameters in online monitoring applications. The proposed function allows for lower and upper bounds for SVR parameters. We tested our proposed approach and compared results with the conventional AOSVR approach using two benchmark time series data and sensor data from a nuclear power plant. The results show that using varying SVR parameters is more applicable to time-dependent data.

Introduction

The advances in the various automatic data acquisition and sensor systems continue to create tremendous opportunities for collecting valuable process and operational data for several enterprises including automobile

* Adapted and reprinted from Omitaomu, O. A., Jeong, M. K., and Badiru, A. B., Online support vector regression with varying parameters for time-dependent data, *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, 41 (1), 191–197, January 2011. © 2011 IEEE.

manufacturing, semiconductor manufacturing, nuclear power plants, and transportation. These technologies have also made it possible to infer the operating conditions of critical system parameters using data from correlated sensors. This approach is called inferential sensing. *Inferential sensing* is the prediction of a system variable through the use of correlated system variables. Most online monitoring systems produce inferred values that are used to determine the status of critical system variable. The values can also be used to monitor drift or other failures in the system, thereby reducing unexpected system breakdown. Several approaches have been used for inferential sensing including regularization methods (Hines et al., 2000) and support vector regression (Omitaomu et al., 2007). However, these methods assume that the training data are collected in a single batch. Therefore, each time a new sample is added to the training set, a new model is obtained by retraining the entire training data. This approach could be very expensive for online monitoring applications.

Several online monitoring applications require decisions to be made within minutes of an anticipated problem. For example, in the electric grid application in which synchrophasors such as phasor measurement units (PMUs) are used for monitoring the conditions of the transmission lines and in which data samples are collected every 1/20th of a second (Bank et al., 2009), decisions may be required within few minutes of an impending problem. In other applications such as nuclear power plants, decisions may be required within several minutes—for example, 30 min (Hines et al., 2000). In both examples, the idea of retraining the entire training set every time new samples are added is not appropriate. The AOSVR technique (Ma et al., 2003) is a better alternative to SVR because it uses incremental algorithm that efficiently and accurately updates the SVR parameters each time a new sample is added to the training set without retraining from scratch. However, the AOSVR technique assumes that the distribution of the samples is constant over time. This assumption is not valid for sensor data that are inherently noisy and nonstationary. The nonstationary characteristic implies that the distribution of the data changes over time, which leads to gradual changes in the dependency between the input and output variables. Thus, this feature should be taken into consideration by the prediction technique. Our solution, then, is to find a way to incorporate the time-dependent characteristic of the new samples into the AOSVR technique.

Fundamentally, the performance of AOSVR, like SVR, depends on three training parameters (kernel, C, and ε). However, for any particular type of kernel, the values of C and ε are what affect the performance of the final model. A small value of C underfits the training data points and a large value of C overfits the training data points. In addition, the number of support vectors—a subset of training data points used for prediction—is related to the tube size defined by the value of ε. A large ε reduces

the number of converged support vectors, thus causing the solution to be very sparse. Several approaches for computing SVR parameters have been proposed (Cherkassky and Mulier, 1998; Smola et al., 1998; Mattera and Haykin, 1999; Schölkopf et al., 1999; Kwok, 2001; Cao and Tay, 2003; Cherkassky and Ma, 2004). One approach that has been found effective is using resampling methods. The resampling methods (Cherkassky and Mulier, 1998; Schölkopf et al., 1999) show good performance for tuning SVR parameters in off-line applications. However, for online applications, resampling methods could become computationally expensive, and may not be appropriate for problems in which both the SVR coefficients and parameters are updated periodically (Schölkopf et al., 1999). Our approach is to update, in an automated manner, the respective value of C and ε, as new samples are added to the training set. Therefore, in this chapter, we present procedures for integrating varying weights into the AOSVR algorithm. To achieve this objective and enhance the performance of AOSVR, we present a weight function for updating SVR parameters for online regression problems. In addition, we present a generalized accurate online support vector regression algorithm (AOSVR-VP) that uses fixed or varying regression parameters.

The outline of the chapter is as follows. In "Modified Gompertz weight function for varying SVR parameters" section, we introduce a weight function for updating SVR parameters and discuss the general behaviors of this function. We present a modification of the AOSVR algorithm called the AOSVR-VP algorithm in "Accurate online SVR with varying parameters" section. In "Experimental results" section, we demonstrate the performance of the AOSVR-VP algorithm using two benchmark time series data and nuclear power plant data. We provide some conclusions about this study in "Conclusion" section.

Modified Gompertz weight function for varying SVR parameters

For online predictions, one approach for selecting C and ε will be to vary their values with respect to the relative importance of the training samples. In some applications such as equipment maintenance, recent data points may provide more quality information about the condition of the system than past data points especially when sensors are used to monitor the condition of the system. Process condition information increases monotonically over time starting possibly from a zero or near-zero level. Therefore, recent data about a possible failure or the deteriorating condition of a system should be given more weight in predicting failures than distant data points. In line with this idea, Cao and Tay (2003) proposed ascending regularization constant and descending tube for batch SVR

applications in financial data. To extend their idea of using varying SVR parameters to AOSVR, we propose a simple online weight function for updating SVR parameters.

One of the popular classical asymmetric functions is the Gompertz function. It is a double exponential function that has wide applications in several areas including engineering, natural sciences, and statistics. However, the standard form of Gompertz function is not flexible in setting lower and upper bounds on weights. For time-dependent data, it is reasonable to have a fixed lower and upper bounds, so that the weight varies between these extremes. Therefore, we present Modified Gompertz Weight Function (MGWF) equations as a weight function for SVR parameters. The MGWF equation for adaptive regularization constant (C_i) is defined as

$$C_i = C_{\min} + C_{\max}\left(\exp\left(-\exp\left(-g\times(i-m_c)\right)\right)\right) \tag{9.1}$$

and the MGWF function for adaptive accuracy parameter (ε_i) is defined as

$$\varepsilon_i = \varepsilon_{\min} + \varepsilon_{\max}\left(\exp\left(-\exp\left(g\times(i-m_c)\right)\right)\right) \tag{9.2}$$

where $i = 1,\ldots, m$, m is the number of training samples, m_c is the changing point, $C_{\min}$ and $\varepsilon_{\min}$ are the desired lower bound for the regularization constant and the accuracy parameter, respectively, $C_{\max}$ and $\varepsilon_{\max}$ are the desired upper bound for the regularization constant and the accuracy parameter, respectively, and g is a constant that controls the curvature (slope) of the function; that is, it represents the factor for the relative importance of the samples. The essence of the lower bound is to avoid underestimation (underfitting) and the upper bound avoids overestimation (overfitting) of the parameters. The value of g could range from zero to infinity depending on applications but we consider only four special cases in this chapter. The four cases considered are summarized as follows and their pictorial representations are shown in Figure 9.1a and b.

1. Constant weight: When $g = 0$, $C_i \cong C_{\min} + C_{\max}/e$ and $\varepsilon_i \cong \varepsilon_{\min} + \varepsilon_{\max}/e$. That is, a fixed value is used for all data points.
2. Linear weight: When $g = 0.005$, the value of C_i is a linearly increasing relationship and the value of ε_i is a linearly decreasing relationship.
3. Sigmoidal weight: When $g = 0.03$, the weight function follows a sigmoidal pattern.

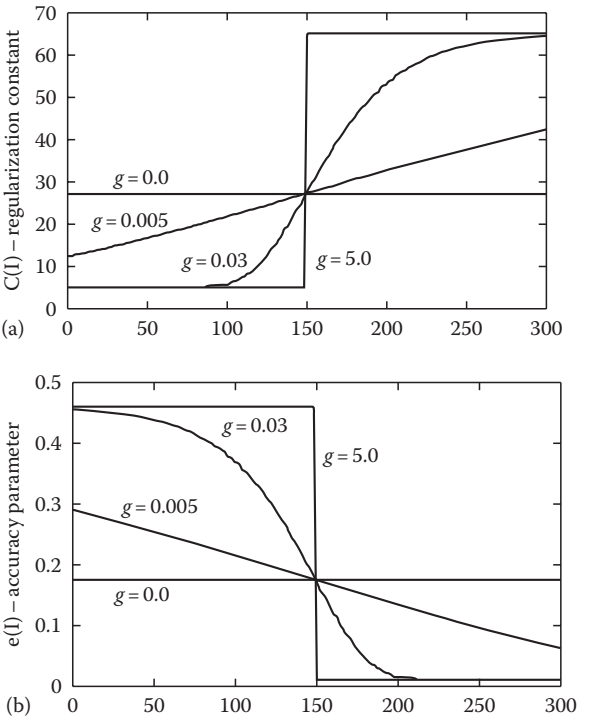


Figure 9.1 The pictorial representations of MGWF with different values of g.

4. Two distinct weights: When $g = 5$,

$$C_i \cong \begin{cases} C_{\min}, & i < m_c \\ C_{\min} + C_{\max}, & i \geq m_c \end{cases} \quad \text{and} \quad \mathbf{e}_i \cong \begin{cases} \mathbf{e}_{\min} + \mathbf{e}_{\max}, & i < m_c \\ \mathbf{e}_{\min}, & i \geq m_c \end{cases}$$

The plots in Figure 9.1 show that the MGWF profile is asymmetric around the midpoint (m_c) of the total training set. For the plots in Figure 9.1, the $C_{\min}$ and $C_{\max}$ are set to 5.0 and 60.0, respectively, and $\varepsilon_{\min}$ and $\varepsilon_{\max}$ are set to 0.01 and 0.45, respectively, for $m = 300$ and $m_c = 150$.

Based on some experiments, if the recent samples provide more important information than past samples, g must be greater than zero ($g > 0$). Otherwise, g must be less than zero ($g < 0$); these cases are not considered in this chapter. In this chapter, the lower and the upper bounds are set using the approach proposed by Cherkassky and Ma (2004), which is a data-dependent approach and found to be robust to outliers.

Accurate online SVR with varying parameters

In order to use the weight function proposed in "Modified Gompertz weight function for varying SVR parameters" section for computing adaptive regularization constant and adaptive accuracy parameter, we propose online SVRAOSVR-VP. This is achieved by modifying both the empirical error (risk), which is measured by the ε-insensitive loss function, and the constraints of the AOSVR formulation, which will lead to a new set of KKT conditions. Therefore, the regularized constant adopts adaptive regularization constant C_i and every training sample uses adaptive accuracy parameter (different tube size) ε_i. The modified algorithm will compute the SVR parameters, (C_i and ε_i), as explained in "Modified Gompertz weight function for varying SVR parameters" section; while it avoids retraining of the training set.

Given a set of data points $(\mathbf{x}_1, y_1)$, $(\mathbf{x}_2, y_2)$, …, $(\mathbf{x}_m, y_m)$ for online learning, where $x_i \in X \subseteq R^n$, $y_i \in Y \subseteq R$, m is the total number of training samples, a linear regression function can be stated as

$$f(\mathbf{x}) = \mathbf{w}^{\mathbf{T}}\Phi(\mathbf{x}_i) + b \tag{9.3}$$

in a feature space ***F***, where $\mathbf{w}$ is a vector in ***F*** and $\Phi(\mathbf{x}_i)$ maps the input $\mathbf{x}$ to a vector in ***F***. Assuming an ε-insensitive loss function (Cherkassky and Mulier, 1998), the $\mathbf{w}$ and b in Equation 9.3 are obtained by solving an optimization problem:

$$\text{minimize} \quad \frac{1}{2}\mathbf{w}^{\mathbf{T}}\mathbf{w} + \sum_{i=1}^{m} C_i\left(\mathbf{x}_i^{+} + \mathbf{x}_i^{-}\right)$$

$$\text{subject to:} \quad \begin{cases} y_i - \mathbf{w}^{\mathbf{T}}\Phi(\mathbf{x}_i) - b \le \mathbf{e}_i + \mathbf{x}_i^{+} \\ \mathbf{w}^{\mathbf{T}}\Phi(\mathbf{x}_i) + b - y_i \le \mathbf{e}_i + \mathbf{x}_i^{-} \\ \mathbf{x}_i^{+},\ \mathbf{x}_i^{-} \ge 0 \end{cases} \tag{9.4}$$

where

$\varepsilon_i(\ge 0)$ is the maximum deviation allowed during each training step
$C_i(>0)$ is the associated penalty for excess deviation during each training step

The slack variables, $\mathbf{x}_i^{+}$ and $\mathbf{x}_i^{-}$, correspond to the size of this excess deviation for positive and negative deviations, respectively. The first term of Equation 9.4, $\mathbf{w}^{\mathbf{T}}\mathbf{w}$, is the regularized term; thus, it controls the function capacity. The second term $\left(\sum_{i=1}^{m}\left(\mathbf{x}_i^{+} + \mathbf{x}_i^{-}\right)\right)$ is the empirical error measured by the ε-insensitive loss function.

The Lagrange formulation of the dual becomes

$$\begin{aligned} L_D = {} & \frac{1}{2}\sum_{i=1}^{m}(\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-) + \sum_{i=1}^{m}\varepsilon_i(\alpha_i^+ + \alpha_i^-) \\ & - \sum_{i=1}^{m} y_i(\alpha_i^+ - \alpha_i^-) - \sum_{i=1}^{m}(\delta_i^+\alpha_i^+ + \delta_i^-\alpha_i^-) \\ & + \sum_{i=1}^{m}\left[u_i^+(\alpha_i^+ - C_i) + u_i^-(\alpha_i^- - C_i)\right] + b\sum_{i=1}^{m}(\alpha_i^+ - \alpha_i^-) \end{aligned} \tag{9.5}$$

the KKT conditions for AOSVR-VP are

$$\begin{aligned} \frac{\partial L_D}{\partial \alpha_i^+} &= \sum_{i,j=1}^{m} K(x_i, x_j)(\alpha_j^+ - \alpha_j^-) + \varepsilon_i - y_i + b = 0 \\ \frac{\partial L_D}{\partial \alpha_i^-} &= -\sum_{i,j=1}^{m} K(x_i, x_j)(\alpha_j^+ - \alpha_j^-) + \varepsilon_i + y_i - b = 0 \\ \frac{\partial L_D}{\partial b} &= \sum_{i=1}^{m}(\alpha_i^+ - \alpha_i^-) = 0 \end{aligned} \tag{9.6}$$

Using the following definitions (Ma et al., 2003),

$$\begin{aligned} Q_{ij} &\equiv \Phi(\mathbf{x}_i)^T\Phi(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j) \\ \beta_i &= \alpha_i^+ - \alpha_i^- \\ \beta_j &= \alpha_j^+ - \alpha_j^- \\ h(x_i) &\equiv f(x_i) - y_i = \sum_{j=1}^{m} Q_{ij}\beta_j - y_i + b \end{aligned} \tag{9.7}$$

where $h(x_i)$ is the error of the target value for vector i. The KKT conditions in Equation 9.6 can be rewritten as

$$\begin{aligned} \frac{\partial L_D}{\partial \alpha_i^+} &= h(x_i) + \varepsilon_i = \gamma_i^+ = 0 \\ \frac{\partial L_D}{\partial \alpha_i^-} &= -h(x_i) + \varepsilon_i = \gamma_i^- = -\gamma_i^+ + 2\varepsilon_i = 0 \\ \frac{\partial L_D}{\partial b} &= \sum_{i=1}^{m}\beta_i = 0 \end{aligned} \tag{9.8}$$

where $\boldsymbol{\gamma}_i^{(*)}$ is the adaptive margin function and can be described as threshold for error on both sides of the adaptive ε-tube. Modifying the approach by Ma et al. (2003), these KKT conditions lead to five new conditions for AOSVR-VP:

$$\begin{aligned}
&2\varepsilon_i < \boldsymbol{\gamma}_i^{+} \rightarrow \boldsymbol{\gamma}_i^{-} < 0, && \boldsymbol{\alpha}_i = -C_i && i \in E^{-} \\
&\boldsymbol{\gamma}_i^{+} = 2\varepsilon_i \rightarrow \boldsymbol{\gamma}_i^{-} = 0, && -C_i < \boldsymbol{\alpha}_i < 0 && i \in S \\
&0 < \boldsymbol{\gamma}_i^{+} < 2\varepsilon_i \rightarrow 0 < \boldsymbol{\gamma}_i^{-} < 2\varepsilon_i, && \boldsymbol{\alpha}_i = 0 && i \in R \\
&\boldsymbol{\gamma}_i^{+} = 0 \rightarrow \boldsymbol{\gamma}_i^{-} = 2\varepsilon_i, && 0 < \boldsymbol{\alpha}_i < C_i && i \in S \\
&\boldsymbol{\gamma}_i^{+} < 0 \rightarrow \boldsymbol{\gamma}_i^{-} > 2\varepsilon_i, && \boldsymbol{\alpha}_i = C_i, && i \in E^{+}
\end{aligned} \tag{9.9}$$

These conditions can be used to classify the training set into three subsets defined as follows:

$$\begin{aligned}
&\text{The } E \text{ set: Error support vectors: } E = \left\{ i \middle| |\boldsymbol{\alpha}_i| = C_i \right\} \\
&\text{The } S \text{ set: Margin support vectors: } S = \left\{ i \middle| 0 < |\boldsymbol{\alpha}_i| < C_i \right\} \\
&\text{The } R \text{ set: Remaining samples: } R = \left\{ i \middle| \boldsymbol{\alpha}_i = 0 \right\}
\end{aligned} \tag{9.10}$$

Based on these conditions, we modify the AOSVR algorithm appropriately and incorporate the algorithms for computing adaptive SVR parameters for the online training as described in "Modified Gompertz weight function for varying SVR parameters" section. We follow the same approach proposed in Ma et al. (2003) for initializing and updating the algorithm.

Experimental results

In this section, we apply the proposed AOSVR-VP to two benchmark time series data and feed-water flow rate data. For the implementations, we used a typical online time series prediction scenario as presented by Tashman (2000) and used a prediction horizon of one time step. The procedure used is, consider given a time series $\{x(t),\ t = 1,\ 2,\ldots\}$ and prediction origin O, time from which the prediction is generated, we construct a set of training samples, $\mathbf{A}_{O,B}$, from the segment of time series $\{x(t),\ t = 1,\ldots, O\}$ as

$$\mathbf{A}_{O,B} = \left\{ \mathbf{X}(t), y(t), t = B, \ldots, 0-1 \right\}$$

where $\mathbf{X}(t) = [x(t), \ldots, x(t - B + 1)]^T$, $y(t) = x(t + 1)$, and B is the embedding dimension of the training set $\mathbf{A}_{O,B}$, which in this chapter is taken to be five. We train the predictor $P(\mathbf{A}_{O,B}; \mathbf{X})$ from the training set $\mathbf{A}_{O,B}$. Then, predict $x(O + 1)$ using $\hat{x}(O + 1) = P(\mathbf{A}_{O,B}; \mathbf{X}(O))$. When $x(O + 1)$ becomes available, we update the prediction origin; that is, $O = O + 1$ and repeat the procedure. As the origin increases, the training set keeps growing and this can become very expensive. However, online predictions take advantage of the fact that the training set is augmented one sample at a time and continues to update and improve the model as more data arrive.

Application to time series data

We also implement the AOSVR-VP algorithm based on the proposed weight function for time series predictions. The performance of AOSVR-VP is compared to the existing AOSVR using two benchmark time series data: Mackey–Glass equation with $\tau = 17$ (Mackey and Glass, 1977) and the Santa Fe Institute Competition time series A (Weigend and Gershenfeld, 1994). The Mackey–Glass equation (mg17) data has 1500 data points; whereas the Santa Fe Institute Competition time series A (SFIC) data has 1000 data points. Both data sets are shown in Figure 9.2. The measures of prediction performance are the mean squared error (MSE) and the mean absolute error (MAE).

Tables 9.1 and 9.2 summarize the results (test error) of the experiments for mg17 and SFIC data, respectively. In this experiment, we set the value of $C_{\min} = 5.0$, $C_{\max} = 60.0$, $\varepsilon_{\min} = 0.01$, $\varepsilon_{\max} = 0.45$, and use a Gaussian radial basis function (RBF) kernel, $\exp(-p|\mathbf{x}_i - \mathbf{x}_j|^2)$, with $p = 1$ based on our experience in implementing the SVR technique. We also implemented the algorithm for the four cases of weight patterns described in "Modified Gompertz weight function for varying SVR parameters" section. The g values for these cases are 0.0 for constant weight, 0.005 for linear weight, 0.3 for sigmoidal weight, and 5.0 for two distinct weights. The plots of the original and predicted data for two benchmark data are shown in Figures 9.3 and 9.4, respectively.

As shown in Tables 9.1 and 9.2, AOSVR-VP performs better than AOSVR for both data sets, which confirms that using varying parameters capture more of the properties of the data than using fixed parameters. The plots in Figures 9.3 and 9.4 further confirm this statement. The error between AOSVR and AOSVR-VP is more obvious in Figure 9.4 than in Figure 9.3 as evidenced from Tables 9.1 and 9.2 respectively. One insight from this experiment is that both methods yield comparative results for more stationary data such as the mg17 data set. However, for non-stationary data set such as the SFIC data, the AOSVR-VP yields better results. Furthermore, there is no significant difference in results when g is set to 0.3 or 0.5.

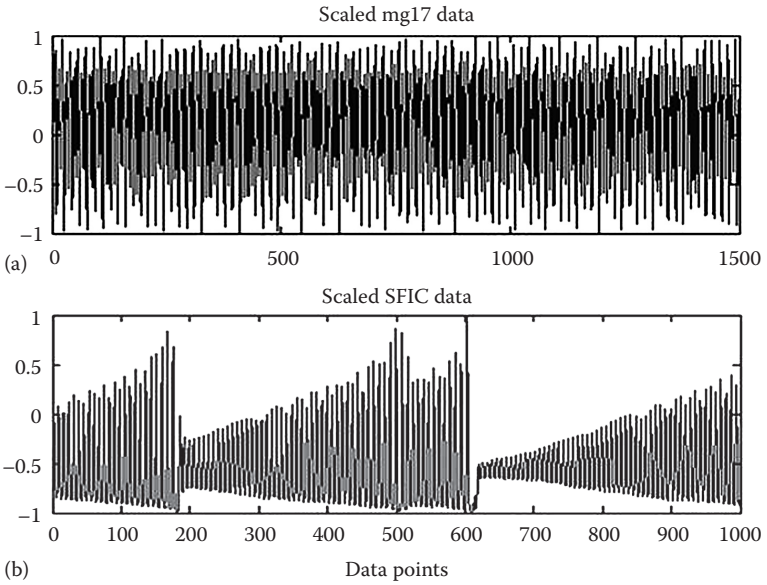


Figure 9.2 The scaled data for the benchmark time series data.

Table 9.1 Performance Comparison for AOSVR-VP and AOSVR for the Mackey–Glass Equation Data

Technique	*g* value	MSE (MAE)
AOSVR	0.0	0.0121 (0.0959)
AOSVR-VP	0.005	0.0011 (0.0294)
	0.3	4.65*E*−05 (0.0058)
	5.0	4.63*E*−05 (0.0058)

Table 9.2 Performance Comparison for AOSVR-VP and AOSVR for the Santa Fe Institute Competition Time Series A Data

Technique	*g* value	MSE (MAE)
AOSVR	0.0	0.0164 (0.0985)
AOSVR-VP	0.005	0.0149 (0.0937)
	0.3	0.0037 (0.0186)
	5.0	0.0037 (0.0186)

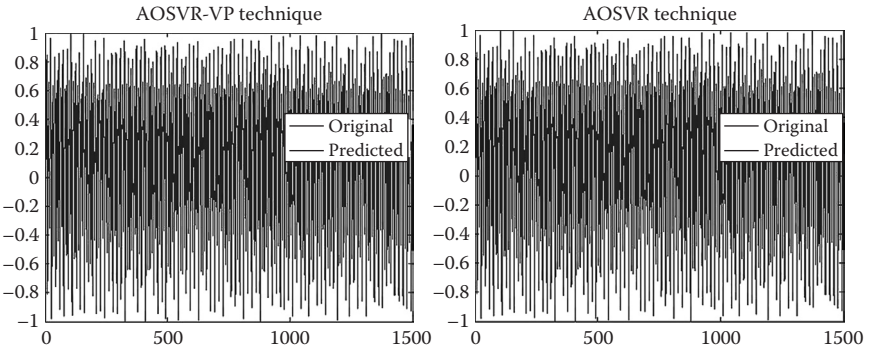


Figure 9.3 A visual comparison between our proposed AOSVR-VP (left) and AOSVR (right) for the Mackey–Glass time series data. The AOSVR-VP technique uses MGWF with g = 0.3. For both approaches, only the first 75 data points are used for training.

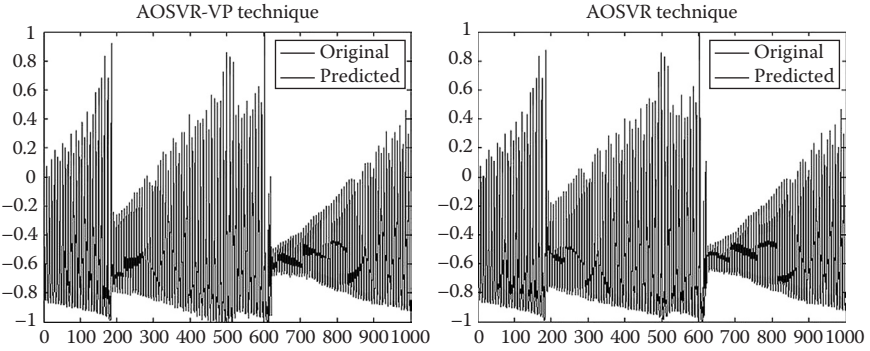


Figure 9.4 A visual comparison of our proposed AOSVR-VP (left) and AOSVR (right) for the Santa Fe Institute time series data. The AOSVR-VP technique uses MGWF with g = 0.3. For both approaches, only the first 50 data points are used for training.

Application to feed-water flow rate data

In a nuclear power plant, the accurate prediction of an important variable, such as feed-water flow rate, can reduce periodic monitoring. Such prediction can be used to assess sensor performance, thereby reducing maintenance costs and increasing reliability of the instrument. Feed-water flow rate directly estimates the thermal power of a reactor. Nuclear power plants use venturi meters to measure feed-water flow rate. These meters are sensitive to measurement degradation due to corrosion products in the feed water (Gribok et al., 2000). Therefore, measurement error due to feed-water

fouling results in feed-water flow rate overestimation. As a result, the thermal power of the reactor is also overestimated and the reactor must be adjusted to stay within regulatory limits, which is an unnecessary action and involves unnecessary costs.

To overcome this problem, several online inferential sensing systems have been developed to infer the "true" feed-water flow rate (Kavaklioglu and Upadhyaya, 1994; Gross et al., 1997; Gribok et al., 1999, 2000; Hines et al., 2000). Inferential sensing is the use of correlated variables for prediction. Inferential measurement is different from conventional prediction where a parameter value is estimated at time t_{n+1}, based on information about other parameters at time t_n. In inferential measurements, a parameter is estimated at time t_n based on information about other parameter also at time t_n. A detailed description of this problem is available in Gribok et al. (1999, 2000).

Inferential sensing is an ill-posed problem and SVR has been found useful for ill-posed problems. For online monitoring of a nuclear power plant, the application of the AOSVR-VP approach can further enhance prediction accuracy of inferential sensing problems. To infer the feed-water flow rate, 24 variables were selected as predictors based on engineering judgment and on their high correlation with the measured feed-water flow rate (Gribok et al., 1999). The plots of the first 12 predictors are shown in Figure 9.5. We limited the plots to the first 12 predictors due to space constraint. The measured flow rate is shown in Figure 9.6. The difference between the estimated (inferred) flow rate and the measured flow rate is called drift and the mean of the drift is used to quantify the prediction performance.

The objective in this application is to determine if we can estimate the feed-water flow rate at any point in the power cycle. In other words, is it possible to predict recent data points based on training data points collected in the past? To answer this question, we used the first 100 data samples from the 24 predictors to train the model and predict the flow rate from 8001 to 8700 data points (700 data points) shown in Figure 9.7.

For this implementation, we set $C_{\min} = 2.0$, $C_{\max} = 20.0$, $\varepsilon_{\min} = 0.01$, $\varepsilon_{\max} = 0.45$, and use the RBF kernel with $p = 1$. We implemented the algorithm for the four cases of weight patterns described in "Modified Gompertz weight function for varying SVR parameters" section. The g values for these cases are 0.0 for constant weight, 0.005 for linear weight, 0.3 for sigmoidal weight, and 5.0 for two distinct weights. The results of the experiment are shown in Table 9.3.

The results show that AOSVR-VP performs better than AOSVR for these data. The plots of the measured and predicted feed-water flow rate are shown in Figure 9.8. We observe significant difference in prediction performance between AOSVR and AOSVR-VP; this is

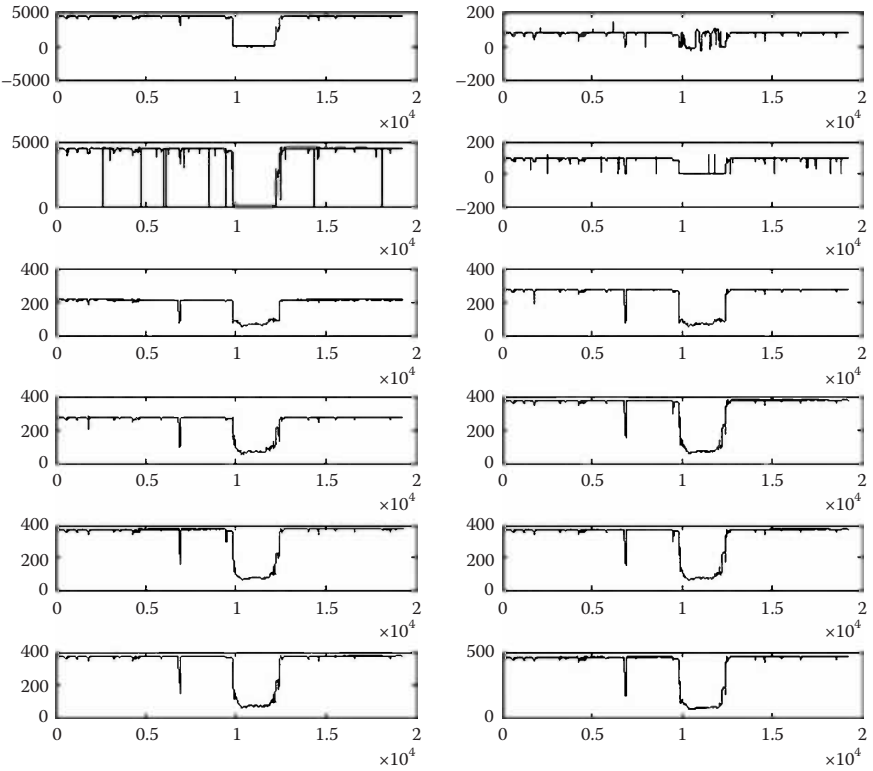


Figure 9.5 Plots of the raw data for the first 12 predictors.

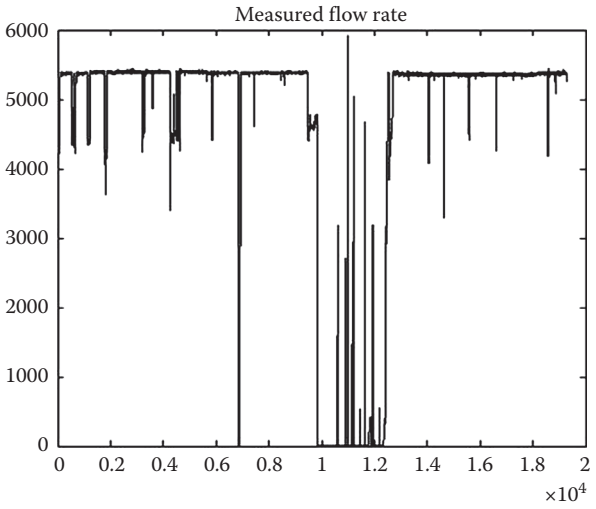


Figure 9.6 A plot of the raw data for the measured flow rate.

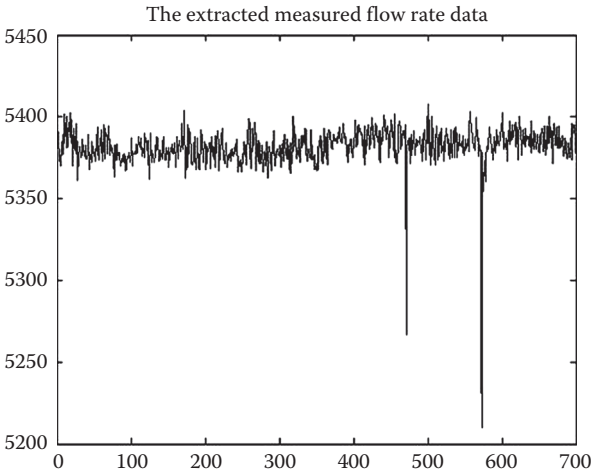


Figure 9.7 The testing data for the AOSVR-VP model.

Table 9.3 Drift Performance for the Feed-Water Flow Rate Data

Technique	*g* value	Mean drift (klb/h)
AOSVR	0.0	67.7703
AOSVR-VP	0.005	28.3342
	0.3	4.0732
	5.0	4.0713

shown visually in Figure 9.8. Both sigmoidal weight and two distinct weights models achieve the smallest mean drift; using linear weight also achieves smaller mean drift than AOSVR but its value is on the high side, which indicates that the linear weight is not very useful for this application.

Based on this major difference in performance, we decided to repeat the experiment using the same first 100 samples for training but predicting the data point from 101 to 8700. The results are shown in Figure 9.9. The mean drift using AOSVR-VP is 8.9634 klb/h, while the mean drift for the AOSVR is 27.2027 klb/h.

One insight from these results is that, due to its performance, it may be difficult to use the AOSVR technique for prediction at any point in the power cycle, which was our objective. Another insight is that, compared to the AOSVR-VP technique, it takes the AOSVR technique longer

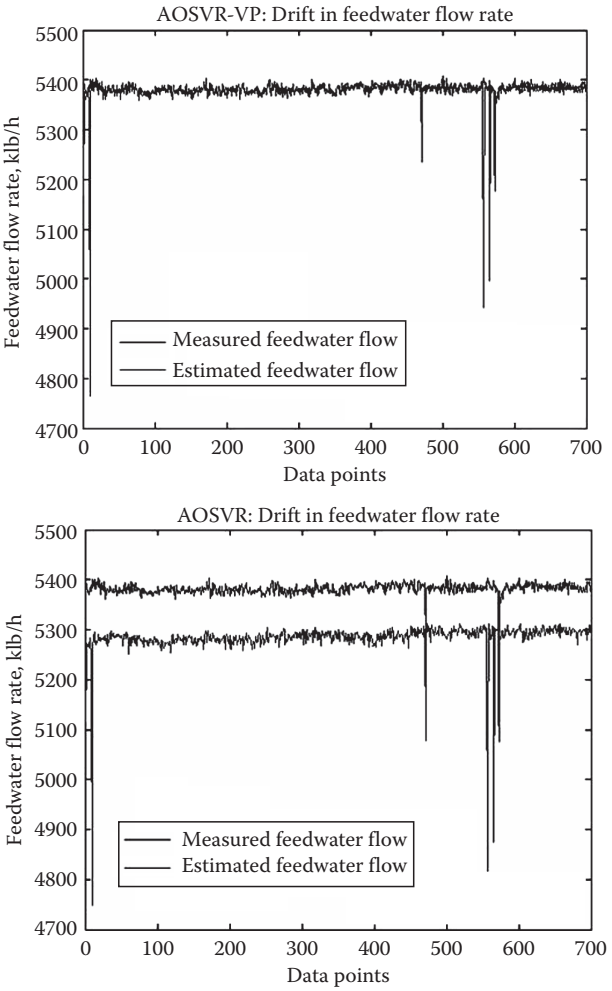


Figure 9.8 A visual comparison of the measured and predicted feed-water flow rate using AOSVR-VP and AOSVR techniques. The AOSVR-VP technique uses MGWF with $g = 5.0$.

time (or more data samples) before it starts to predict more accurately (i.e., before it becomes more stable). Looking at Figure 9.9, we can say that AOSVR starts accurate prediction around data point 7000; whereas the AOSVR-VP technique starts accurate prediction around data point 4300. It must be noted that both techniques are implemented in an online mode. The difference in implementation is whether they respectively use constant or varying values for the SVR parameters.

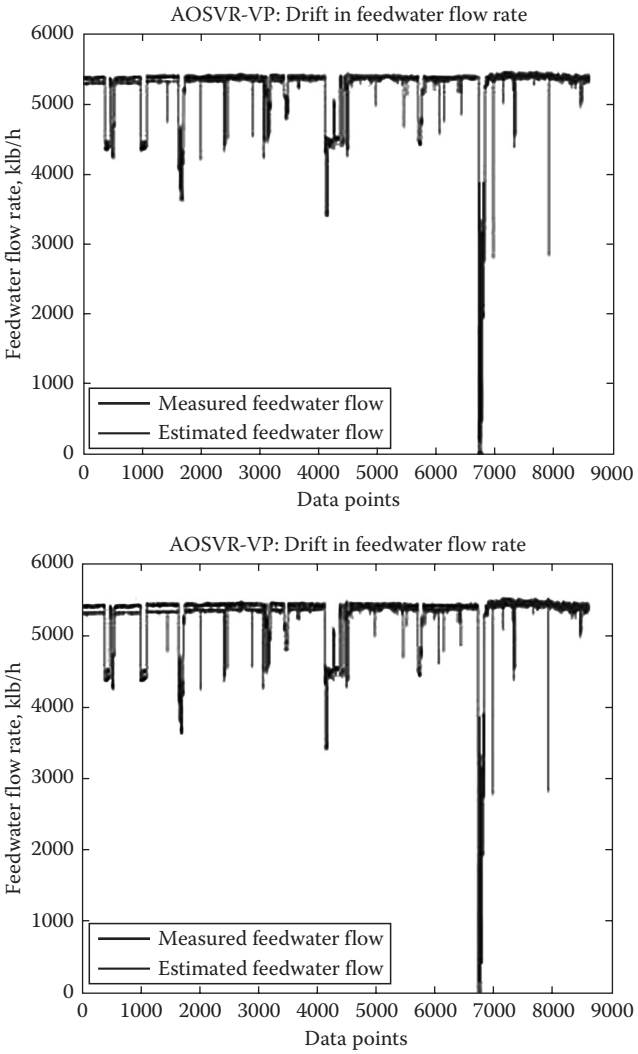


Figure 9.9 A visual comparison of the measured and predicted feed-water flow rate using AOSVR-VP and AOSVR techniques. The AOSVR-VP technique uses MGWF with $g = 0.005$.

Conclusion

We have proposed and implemented a weight function, MGWF, for updating SVR parameters in online regression applications. Based on the experimental results, the weight function is suitable for integrating the properties of time series data into online SVR predictions. We also presented accurate online AOSVR-VP, based on the proposed weight

function. We compared the performance of the proposed procedure with conventional AOSVR based on two benchmark time series data and a nuclear power plant data for feed-water flow rate prediction. As demonstrated in the experiments, AOSVR-VP predicts better than AOSVR in all cases in which the data sets are nonstationary. The AOSVR-VP is a generalized procedure that can be used for both fixed and varying properties as demonstrated with the experiments.

References

Bank, J. N., O. A. Omitaomu, S. J. Fernandez, and Y. Liu, Visualization and classification of power system frequency data streams, *Proceedings of the IEEE ICDM Workshop on Spatial and Spatiotemporal Data Mining*, Miami, FL, December 6, 2009, pp. 650–655.

Cao, L. and F. E. H. Tay, Support vector machine with adaptive parameters in financial time series forecasting, *IEEE Transactions on Neural Networks*, 14 (6), 1506–1518, November 2003.

Cherkassky, V. and Y. Ma, Practical selection of SVM parameters and noise estimation for SVM regression, *Neural Networks*, 17 (1), 113–126, 2004.

Cherkassky, V. and F. Mulier, *Learning from Data: Concepts, Theory, and Methods*. John Wiley, New York, 1998.

Gribok, A. V., I. Attieh, J. W. Hines, and R. E. Uhrig, Regularization of feedwater flow rate evaluation for venture meter fouling problem in nuclear power plants, *Ninth International Meeting on Nuclear Reactor Thermal Hydraulics* (*NURETH-9*), San Francisco, CA, October 3–8, 1999.

Gribok, A. V., J. W. Hines, and R. E. Uhrig, Use of kernel based techniques for sensor validation in nuclear power plants, *International Topical Meeting on Nuclear Plant Instrumentations, Controls, and Human–Machine Interface Technologies*, Washington, DC, 2000.

Gross, K. C., R. M. Singer, S. W. Wegerich, J. P. Herzog, R. V. Alstine, and F. K. Bockhorst, Application of a model-based fault detection system to nuclear plant signals, *Proceedings of the International Conference on Intelligent System Application to Power Systems*, Seoul, Korea, 1997, pp. 60–65.

Hines, J. W., A. V. Gribok, I. Attieh, and R. E. Uhrig, Regularization methods for inferential sensing in nuclear power plants. In D. Ruan (ed.), *Fuzzy Systems and Soft Computing in Nuclear Engineering*. Springer-Verlag, New York, 2000, pp. 285–310.

Kavaklioglu, K. and B. R. Upadhyaya, Monitoring feedwater flow rate and component thermal performance of pressurized water reactors by means of artificial neural networks, *Nuclear Technology*, 107, 112–123, 1994.

Kwok, J. T., Linear dependency between ε and the input noise in ε-support vector regression. In G. Dorffner, H. Bishof, and K. Hornik (eds.), *ICANN 2001*, LNCS 2130, 2001, pp. 405–410.

Ma, J., T. James, and P. Simon, Accurate on-line support vector regression, *Neural Computation*, 15, 2683–2703, 2003.

Mackey, M. C. and L. Glass, Oscillation and chaos in physiological control systems, *Science*, 197, 287–289, 1977.

Mattera, D. and S. Haykin, Support vector machines for dynamic reconstruction of a chaotic system. In B. Schölkopf, J. Burges, A. Smola (eds.), *Advances in Kernel Methods: Support Vector Machine*. MIT Press, Cambridge, MA, 1999.

Omitaomu, O. A., M. K. Jeong, and A. B. Badiru, Online support vector regression with varying parameters for time-dependent data, *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, 41 (1), 191–197, January 2011.

Omitaomu, O. A., M. K. Jeong, A. B. Badiru, and J. W. Hines, On-line support vector regression approach for the monitoring of motor shaft misalignment and feedwater flow rate, *IEEE Transactions on Systems, Man, and Cybernetics: Part C—Applications and Reviews*, 37 (5), 962–970, 2007.

Schölkopf, B., J. Burges, and A. Smola, *Advances in Kernel Methods: Support Vector Machine*. MIT Press, Cambridge, MA, 1999.

Smola, A., N. Murata, B. Schölkopf, and K. Muller, Asymptotically optimal choice of ε-loss for support vector machines. In L. Niklasson, M. Boden, T. Ziemke (eds.), *Proceedings of the International Conference on Artificial Neural Networks* (*ICANN 1998*), Perspectives in Neural Computing. Springer, Berlin, Germany, 1998, pp. 105–110.

Tashman, L. J., Out-of-sample tests of forecasting accuracy: An analysis and review, *International Journal of Forecasting*, 16, 437–450, 2000.

Vapnik, V., *Statistical Learning Theory*. John Wiley, New York, 1998.

Weigend, A. S. and N. A. Gershenfeld, *Time-Series Prediction: Forecasting the Future and Understanding the Past*. Addison-Wesley, Reading, MA, 1994.

Appendix: Mathematical and engineering formulae

Quadratic equation

$$ax^2 + bx + c = 0$$

Solution

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

If $b^2 - 4ac < 0$, the roots are complex.
If $b^2 - 4ac > 0$, the roots are real.
If $b^2 - 4ac = 0$, the roots are real and repeated.

Derivation of the solution
Dividing both sides of equation by "a," ($a \neq 0$)

$$x^2 + \frac{b}{a}x + \frac{c}{a} = 0$$

Note: If a = 0, the solution to $ax^2 + bx + c = 0$ is $x = -\frac{c}{b}$.

Rewrite the equation as

$$\left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a^2} + \frac{c}{a} = 0$$

$$\left(x + \frac{b}{2a}\right)^2 = \frac{b^2}{4a^2} - \frac{c}{a} = \frac{b^2 - 4ac}{4a^2}$$

$$x+\frac{b}{2a}=\pm\sqrt{\frac{b^2-4ac}{4a^2}}=\pm\frac{\sqrt{b^2-4ac}}{2a}$$

$$x=-\frac{b}{2a}\pm\sqrt{\frac{b^2-4ac}{4a^2}}$$

$$x=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$$

Overall mean

$$\bar{x}=\frac{n_1\bar{x}_1+n_2\bar{x}_2+n_3\bar{x}_3+\cdots+n_k\bar{x}_k}{n_1+n_2+n_3+\cdots+n_k}=\frac{\sum n\bar{x}}{\sum n}$$

Chebyshev's theorem

$$1-1/k^2$$

Permutations

A permutation of *m* elements from a set of n elements is any arrangement, without repetition, of the *m* elements. The total number of all the possible permutations of *n* distinct objects taken *m* times is

$$P(n,m)=\frac{n!}{(n-m)!},\quad (n\geq m)$$

Example
Find the number of ways a president, vice-president, secretary, and a treasurer can be chosen from a committee of eight members.

Solution

$$P(n,m)=\frac{n!}{(n-m)!}=P(8,4)=\frac{8!}{(8-4)!}=\frac{8.7.6.5.4.3.2.1}{4.3.2.1}=1680$$

There are 1680 ways of choosing the four officials from the committee of eight members.

Combinations

The number of combination of n distinct elements taken is given by

$$C(n,m) = \frac{n!}{m!(n-m)!}, \quad (n \geq m)$$

Example
How many poker hands of 5 cards can be dealt from a standard deck of 52 cards?

Solution
Note: The order in which the five cards are dealt is not important.

$$C(n,m) = \frac{n!}{m!(n-m)!} = C(52,5) = \frac{52!}{5!(52-5)!} = \frac{52!}{5!47!}$$

$$= \frac{52.51.50.49.48}{5.4.3.2.1} = 2,598,963$$

Failure

$$q = 1 - p = \frac{n-s}{n}$$

Probability

$$P(X \leq x) = F(x) = \int_{-\infty}^{x} f(x)dx$$

Expected value

$$\mathtt{m} = \sum (xf(x))$$

Variance

$$\mathtt{s}^2 = \sum (x-\mathtt{m})^2 f(x) \quad \text{or} \quad \mathtt{s}^2 = \int_{-\infty}^{\infty} (x-\mathtt{m})^2 f(x)dx$$

Binomial distribution

$$f(x) = {}^{n}c_{x} p^{x} (1-p)^{n-x}$$

Poisson distribution

$$f(x) = \frac{(np)^x e^{-np}}{x!}$$

Mean of a binomial distribution

$$\mu = np$$

Variance

$$\sigma^2 = npq$$

where $q = 1 - p$ and is the probability of obtaining x failures in the n trials.

Normal distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Cumulative distribution function

$$F(x) = P(X \le x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx$$

Population mean

$$\mu_{\bar{x}} = \mu$$

Standard error of the mean

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

t-Distribution

$$\bar{x} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \le \mu \le \bar{x} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

where

$\bar{x}$ is the sample mean
μ is the population mean
s is the sample standard deviation

Chi-squared distribution

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

Definition of set and notation

A set is a collection of object called elements. In mathematics, we write a set by putting its elements between the curly brackets { }.

Set A containing numbers 3, 4, and 5 is written

$$A = \{3, 4, 5\}$$

a. Empty set: A set with no elements is called an empty set and it is denoted by

$$\{\ \} = \Phi$$

b. Subset: Sometimes every element of one set also belongs to another set:

$$A = \{3, 4, 5\} \quad \text{and} \quad B = \{1, 2, 3, 4, 5, 6, 7\},$$

A set A is a subset of a set B because every element of set A is also an element of set B, and it is written as

$$A \square\ B$$

c. Set equality: The sets A and B are equal if and only if they have exactly the same elements, and the equality is written as

$$A = B$$

d. Set union: The union of a set A and set B is the set of all elements that belong to either A or B or both, and is written as

$$A \cup b = \left\{ x \middle| x \in A \quad \text{or} \quad x \in B \text{ or both} \right\}$$

Set terms and symbols

{ } set braces
$\in$ is an element of
$\notin$ is not an element of
$\subseteq$ is a subset of
$\not\subset$ is not a subset of
A' complement of set A
$\cap$ set intersection
$\cup$ set union

Operations on sets

If A, B, and C are arbitrary subsets of universal set U, then the following rules govern the operations on sets:

1. Commutative law for union

$$A \cup B = B \cup A$$

2. Commutative law for intersection

$$A \cap B = B \cap A$$

3. Associative law for union

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

4. Associative law for intersection

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

5. Distributive law for union

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

6. Distributive law for intersection

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

De Morgan's laws

$$(A \cup B)' = A' \cap B' \tag{A.1}$$

$$(A \cap B)' = A' \cup B' \tag{A.2}$$

The complement of the union of two sets is equal to the intersection of their complements. The complement of the intersection of two sets is equal to the union of their complements.

Counting the elements is a set

The number of the elements in a finite set is determined by simply counting the elements in the set.

If A and B are disjoint sets, then

$$n(A \cup B) = n(A) + n(B)$$

In general, A and B need not to be disjoint, so

$$n(A \cup B) = n(A) + n(B) - n(A \cap B)$$

where n is the number of the elements in a set.

Permutations

A permutation of m elements from a set of n elements is any arrangement, without repetition, of the m elements. The total number of all the possible permutations of n distinct objects taken m times is

$$P(n,m) = \frac{n!}{(n-m)!}, \quad (n \geq m)$$

Example
Find the number of ways a president, vice-president, secretary, and a treasurer can be chosen from a committee of eight members.

Solution

$$P(n,m) = \frac{n!}{(n-m)!} = P(8,4) = \frac{8!}{(8-4)!} = \frac{8.7.6.5.4.3.2.1}{4.3.2.1} = 1680$$

There are 1680 ways of choosing the four officials from the committee of eight members.

Combinations

The number of combination of n distinct elements taken is given by

$$C(n,m) = \frac{n!}{m!(n-m)!}, \quad (n \geq m)$$

Example
How many poker hands of five cards can be dealt from a standard deck of 52 cards?

Solution
Note: The order in which the five cards are dealt is not important.

$$C(n,m) = \frac{n!}{m!(n-m)!} = C(52,5) = \frac{52!}{5!(52-5)!} = \frac{52!}{5!47!}$$

$$= \frac{52.51.50.49.48}{5.4.3.2.1} = 2,598,963$$

Probability terminology

A number of specialized terms are used in the study of probability.

- Experiment: An experiment is an activity or occurrence with an observable result.
- Outcome: The result of the experiment.
- Sample point: An outcome of an experiment.
- Event: An event is a set of outcomes (a subset of the sample space) to which a probability is assigned.

Basic probability principles

Consider a random sampling process in which all the outcomes solely depend on chance, that is, each outcome is equally likely to happen. If S is a uniform sample space and the collection of desired outcomes is E, the probability of the desired outcomes is

$$P(E) = \frac{n(E)}{n(S)}$$

where

$n(E)$ is the number of favorable outcomes in E
$n(S)$ is the number of possible outcomes in S

Since E is a subset of S,

$$0 \leq n(E) \leq n(S)$$

the probability of the desired outcome is

$$0 \leq P(E) \leq 1$$

Random variable

A random variable is a rule that assigns a number to each outcome of a chance experiment.

Example

1. A coin is tossed six times. The random variable X is the number of tails that are noted. X can only take the values 1, 2,…, 6, so X is a discrete random variable.
2. A light bulb is burned until it burns out. The random variable Y is its lifetime in hours. Y can take any positive real value, so Y is a continuous random variable.

Mean value $\hat{x}$ or expected value μ

The mean value or expected value of a random variable indicates its average or central value. It is a useful summary value of the variable's distribution.

1. If random variable X is a discrete mean value,

$$\hat{x} = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n = \sum_{i=1}^{n} x_1 p_1$$

   where p_i are probability densities.
2. If X is a continuous random variable with probability density function f(x), then the expected value of X is

$$\mathrm{m} = E(X) = \int_{-\infty}^{+\infty} x f(x) dx$$

   where $f(x)$ are probability densities.

Series expansions

a. Expansions of common functions

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

$$a^x = 1 + x \ln a + \frac{(x \ln a)^2}{2!} + \frac{(x \ln a)^3}{3!} + \cdots$$

$$e^{-x^2} = 1 - x^2 + \frac{x^4}{2!} - \frac{x^6}{3!} + \frac{x^8}{4!} - \cdots$$

$$\ln x = (x-1) - \frac{1}{2}(x-1)^2 + \frac{1}{3}(x-1)^3 - \cdots, \quad 0 < x \le 2$$

$$\ln x = \frac{x-1}{x} + \frac{1}{2}\left(\frac{x-1}{x}\right)^2 + \frac{1}{3}\left(\frac{x-1}{x}\right)^3 + \cdots, \quad x > \frac{1}{2}$$

$$\ln x = 2\left[\frac{x-1}{x+1} + \frac{1}{3}\left(\frac{x-1}{x+1}\right)^3 + \frac{1}{5}\left(\frac{x-1}{x+1}\right)^5 + \cdots\right], \quad x > 0$$

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots, \quad |x| \le 1$$

$$\ln(a+x) = \ln a + 2\left[\frac{x}{2a+x} + \frac{1}{3}\left(\frac{x}{2a+x}\right)^3 + \frac{1}{5}\left(\frac{x}{2a+x}\right)^5 + \cdots\right],$$

$$a > 0, \quad -a < x < +\infty$$

$$\ln\left(\frac{1+x}{1-x}\right) = 2\left(x + \frac{x^3}{3} + \frac{x^5}{5} + \frac{x^7}{7} + \cdots\right), \quad x^2 < 1$$

$$\ln\left(\frac{1+x}{1-x}\right) = 2\left[\frac{1}{x} + \frac{1}{3}\left(\frac{1}{x}\right)^3 + \frac{1}{5}\left(\frac{1}{x}\right)^5 + \left(\frac{1}{x}\right)^7 + \cdots\right], \quad x^2 > 1$$

$$\ln\left(\frac{1+x}{x}\right) = 2\left[\frac{1}{2x+1} + \frac{1}{3(2x+1)^3} + \frac{1}{5(2x+1)^5} + \cdots\right], \quad x > 0$$

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots$$

$$\tan x = x + \frac{x^3}{3} + \frac{2x^5}{15} + \frac{17x^7}{315} + \frac{62x^9}{2835} + \cdots, \quad x^2 < \frac{\mathrm{p}^2}{4}$$

$$\sin^{-1} x = x + \frac{x^3}{6} + \frac{1}{2}\cdot\frac{3}{4}\cdot\frac{x^3}{5} + \frac{1}{2}\cdot\frac{3}{4}\cdot\frac{5}{6}\cdot\frac{x^7}{7} + \cdots, \quad x^2 < 1$$

$$\tan^{-1} x = x - \frac{1}{3}x^3 + \frac{1}{5}x^5 - \frac{1}{7}x^7 + \cdots, \quad x^2 < 1$$

$$\tan^{-1} x = \frac{\mathrm{p}}{2} - \frac{1}{x} + \frac{1}{3x^3} - \frac{1}{5x^5} + \cdots, \quad x^2 > 1$$

$$\sinh x = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} + \cdots$$

$$\cosh x = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \frac{x^6}{6!} + \cdots$$

$$\tanh x = x - \frac{x^3}{3} + \frac{2x^5}{15} - \frac{17x^7}{315} + \cdots$$

$$\sinh^{-1} x = x - \frac{1}{2}\cdot\frac{x^3}{3} + \frac{1\cdot 3}{2\cdot 4}\cdot\frac{x^5}{5} - \frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}\cdot\frac{x^7}{7} + \cdots, \quad x^2 < 1$$

$$\sinh^{-1} x = \ln 2x + \frac{1}{2}\cdot\frac{1}{2x^2} - \frac{1\cdot 3}{2\cdot 4}\cdot\frac{1}{4x^4} + \frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}\cdot\frac{1}{6x^6} - \cdots, \quad x > 1$$

$$\cosh^{-1} x = \ln 2x - \frac{1}{2}\cdot\frac{1}{2x^2} - \frac{1\cdot 3}{2\cdot 4}\cdot\frac{1}{4x^4} - \frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}\cdot\frac{1}{6x^6} - \cdots$$

$$\tanh^{-1} x = x + \frac{x^3}{3} + \frac{x^5}{5} + \frac{x^7}{7} + \cdots, \quad x^2 < 1$$

b. Binomial theorem

$$(a+x)^n = a^n + na^{n-1}x + \frac{n(n-1)}{2!}a^{n-2}x^2 + \frac{n(n-1)(n-2)}{3!} \times a^{n-3}x^3 + \cdots, \quad x^2 < a^2$$

c. Taylor series expansion: A function $f(x)$ may be expanded about $x = a$ if the function is continuous, and its derivatives exist and are finite at $x = a$.

$$f(x) = f(a) + f'(a)\frac{(x-a)}{1!} + f''(a)\frac{(x-a)^2}{2!} + f'''(a)\frac{(x-a)^3}{3!} + \cdots + f^{n-1}(a)\frac{(x-a)^{n-1}}{(n-1)!} + R_n$$

d. Maclaurin series expansion: The Maclaurin series expansion is a special case of the Taylor series expansion for $a = 0$.

$$f(x) = f(0) + f'(0)\frac{x}{1!} + f''(0)\frac{x^2}{2!} + f'''(0)\frac{x^3}{3!} + \cdots + f^{(n-1)}(0)\frac{x^{n-1}}{(n-1)!} + R_n$$

e. Arithmetic progression: The sum to n terms of the arithmetic progression

$$S = a + (a+d) + (a+2d) + \cdots + [a + (n-1)d]$$

is (in terms of the last number l)

$$S = \frac{n}{2}(a + l)$$

where $l = a + (n - 1)d$.

f. Geometric progression: The sum of the geometric progression to n terms is

$$S = a + ar + ar^2 + \cdots + ar^{n-1} = a\left(\frac{1 - r^n}{1 - r}\right)$$

g. Sterling's formula for factorials

$$n! \approx \sqrt{2\pi}\, n^{n+1/2} e^{-n}$$

Mathematical signs and symbols

± (∓)	plus or minus (minus or plus)
:	divided by, ratio sign
::	proportional sign
<	less than
≮	not less than
>	greater than
≯	not greater than
≅	approximately equals, congruent
∽	similar to
≡	equivalent to
≠	not equal to
≐	approaches, is approximately equal to
∝	varies as
∞	Infinity
∴	Therefore
$\sqrt{\ }$	square root
$\sqrt[3]{\ }$	cube root
$\sqrt[n]{\ }$	nth root
∠	Angle
⊥	perpendicular to
∥	parallel to
$\lvert x \rvert$	numerical value of x
log or $\log_{10}$	common logarithm or Briggsian logarithm
$\log_e$ or ln	natural logarithm or hyperbolic logarithm or Napierian logarithm

e	base (2.718) of natural system of logarithms
a^0	an angle a degrees
a'	a prime, an angle a minutes
a''	a double prime, an angle a seconds, a second
sin	Sine
cos	Cosine
tan	Tangent
ctn or cot	Cotangent
sec	Secant
csc	Cosecant
vers	versed sine
covers	coversed sine
exsec	Exsecant
$\sin^{-1}$	anti sine or angle whose sine is
sinh	hyperbolic sine
cosh	hyperbolic cosine
tanh	hyperbolic tangent
$\sinh^{-1}$	anti hyperbolic sine or angle whose hyperbolic sine is
$f(x)$ or $\phi(x)$	function of x
Δx	increment of x
Σ	summation of
dx	differential of x
dy/dx or y'	derivative of y with respect to x
d^2y/dx^2 or y''	second derivative of y with respect to x
d^ny/dx^n	nth derivative of y with respect to x
$\partial y/\partial x$	partial derivative of y with respect to x
$\partial^n y/\partial x^n$	nth partial derivative of y with respect to x
$\dfrac{\partial^n y}{\partial x \partial y}$	nth partial derivative with respect to x and y
$\int$	integral of
$\int_a^b$	integral between the limits a and b
$\dot{y}$	first derivative of y with respect to time
$\ddot{y}$	second derivative of y with respect to time
Δ or ∇^2	the "Laplacian" $\left(\dfrac{\partial^2}{\partial x^2}+\dfrac{\partial^2}{\partial y^2}+\dfrac{\partial^2}{\partial x^2}\right)$
δ	sign of a variation
ξ	sign of integration around a closed path

Greek alphabet

Alpha = A, α = A, a
Beta = B, β = B, b
Gamma = Γ, γ = G, g
Delta = Δ, δ = D, d
Epsilon = E, e = E, e
Zeta = Z, ζ = Z, z
Eta = H, η = E, e
Theta = Θ, θ = Th, th
Iota = I, ι = I, i
Kappa = K, κ = K, k
Lambda = $\wedge$, λ = L, 1
Mu = M, μ = M, m
Nu = N, ν = N, n
Xi = Ξ, ξ = X, x
Omicron = O, o = O, o
Pi = Π, π = P, p
Rho = P, p = R, r
Sigma = Σ, $\mathbf{s}$ = S, s
Tau = T, τ = T, t
Upsilon = T, ν = U, u
Phi = Φ, ϕ = Ph, ph
Chi = X, x = Ch, ch
Psi = Ψ, ψ = Ps, ps
Omega = Ω, ω = O, o

Algebra

Laws of algebraic operations

a. Commutative law: $a + b = b + a$, $ab = ba$
b. Associative law: $a + (b + c) = (a + b) + c$, $a(bc) = (ab)c$
c. Distributive law: $c(a + b) = ca + cb$

Special products and factors

$$(x+y)^2 = x^2 + 2xy + y^2$$

$$(x-y)^2 = x^2 - 2xy + y^2$$

$$(x+y)^3 = x^3 + 3x^2y + 3xy^2 + y^3$$

$$(x-y)^3 = x^3 - 3x^2y + 3xy^2 - y^3$$

$$(x+y)^4 = x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4$$

$$(x-y)^4 = x^4 - 4x^3y + 6x^2y^2 - 4xy^3 + y^4$$

$$(x+y)^5 = x^5 + 5x^4y + 10x^3y^2 + 10x^2y^3 + 5xy^4 + y^5$$

$$(x-y)^5 = x^5 - 5x^4y + 10x^3y^2 - 10x^2y^3 + 5xy^4 - y^5$$

$$(x+y)^6 = x^6 + 6x^5y + 15x^4y^2 + 20x^3y^3 + 15x^2y^4 + 6xy^5 + y^6$$

$$(x-y)^6 = x^6 - 6x^5y + 15x^4y^2 - 20x^3y^3 + 15x^2y^4 - 6xy^5 + y^6$$

The results mentioned earlier are special cases of the binomial formula.

$$x^2 - y^2 = (x-y)(x+y)$$

$$x^3 - y^3 = (x-y)(x^2 + xy + y^2)$$

$$x^3 + y^3 = (x+y)(x^2 - xy + y^2)$$

$$x^4 - y^4 = (x-y)(x+y)(x^2 + y^2)$$

$$x^5 - y^5 = (x-y)(x^4 + x^3y + x^2y^2 + xy^3 + y^4)$$

$$x^5 + y^5 = (x+y)(x^4 - x^3y + x^2y^2 - xy^3 + y^4)$$

$$x^6 - y^6 = (x-y)(x+y)(x^2 + xy + y^2)(x^2 - xy + y^2)$$

$$x^4 + x^2y^2 + y^4 = (x^2 + xy + y^2)(x^2 - xy + y^2)$$

$$x^4 + 4y^4 = (x^2 + 2xy + 2y^2)(x^2 - 2xy + 2y^2)$$

Some generalization of the above are given by the following results where n is a positive integer:

$$\begin{aligned} x^{2n+1} - y^{2n+1} &= (x-y)(x^{2n} + x^{2n-1}y + x^{2n-2}y^2 + \cdots + y^{2n}) \\ &= (x-y)\left(x^2 - 2xy\cos\frac{2\pi}{2n+1} + y^2\right)\left(x^2 - 2xy\cos\frac{4\pi}{2n+1} + y^2\right) \\ &\quad \cdots\left(x^2 - 2xy\cos\frac{2n\pi}{2n+1} + y^2\right) \end{aligned}$$

$$x^{2n+1}+y^{2n+1}=(x+y)\left(x^{2n}-x^{2n-1}y+x^{2n-2}y^2-\cdots+y^{2n}\right)$$

$$=(x+y)\left(x^2+2xy\cos\frac{2\pi}{2n+1}+y^2\right)\left(x^2+2xy\cos\frac{4\pi}{2n+1}+y^2\right)$$

$$\cdots\left(x^2+2xy\cos\frac{2n\pi}{2n+1}+y^2\right)$$

$$x^{2n}-y^{2n}=(x-y)(x+y)\left(x^{n-1}+x^{n-2}y+x^{n-3}y^2+\cdots\right)\left(x^{n-1}-x^{n-2}y+x^{n-3}y^2-\cdots\right)$$

$$=(x-y)(x+y)\left(x^2-2xy\cos\frac{\pi}{n}+y^2\right)\left(x^2-2xy\cos\frac{2\pi}{n}+y^2\right)$$

$$\cdots\left(x^2-2xy\cos\frac{(n-1)\pi}{n}+y^2\right)$$

$$x^{2n}+y^{2n}=\left(x^2+2xy\cos\frac{\pi}{2n}+y^2\right)\left(x^2+2xy\cos\frac{3\pi}{2n}+y^2\right)$$

$$\cdots\left(x^2+2xy\cos\frac{(2n-1)\pi}{2n}+y^2\right)$$

Powers and roots

$$a^x\times a^y=a^{(x+y)} \qquad a^0=1\ [\text{if } a\neq 0] \qquad (ab^x)=a^xb^x$$

$$\frac{a^x}{a^y}=a^{(x-y)} \qquad a^{-x}=\frac{1}{a^x} \qquad \left(\frac{a}{b}\right)^x=\frac{a^x}{b^x}$$

$$(a^x)^y=a^{xy} \qquad a^{\frac{1}{x}}=\sqrt[x]{a} \qquad \sqrt[x]{ab}=\sqrt[x]{a}\sqrt[x]{b}.$$

$$\sqrt[x]{\sqrt[y]{a}}=\sqrt[xy]{a} \qquad a^{\frac{x}{y}}=\sqrt[y]{a^x} \qquad \sqrt[x]{\frac{a}{b}}=\frac{\sqrt[x]{a}}{\sqrt[x]{b}}$$

Proportion

If $\frac{a}{b}=\frac{c}{d}$, then $\frac{a+b}{b}=\frac{c+d}{d}$

$$\frac{a-b}{b}=\frac{c-d}{d} \qquad \frac{a-b}{a+b}=\frac{c-d}{c+d}$$

Sum of arithmetic progression to n terms

$$a+(a+d)+(a+2d)+\cdots+\big(a+(n-1)d\big)$$

$$=na+\frac{1}{2}n(n-1)d=\frac{n}{2}(a+l),$$

last term in series = $l = a + (n - 1)d$

Sum of geometric progression to n terms

$$s_n=a+ar+ar^2+\cdots+ar^{n-1}=\frac{a(1-r^n)}{1-r}$$

$$\lim_{n\to\infty}8_n=a!(1-r)\quad(-1<r<1)$$

Arithmetic mean of n quantities A

$$\mathrm{A}=\frac{a_1+a_2+\cdots+a_n}{n}$$

Geometric mean of n quantities G

$$G=(a_1a_2\ldots a_n)^{1/n}$$

$$(a_k>0,\,k=1,2,\ldots,n)$$

Harmonic mean of n quantities H

$$\frac{1}{H}=\frac{1}{n}\left(\frac{1}{a_1}+\frac{1}{a_2}+\cdots+\frac{1}{a_n}\right)$$

$$(a_k>0,\,k=1,2,\ldots,n)$$

Generalized mean

$$M(t)=\left(\frac{1}{n}\sum_{k=1}^{n}a_k^l\right)^{1/t}$$

$$M(t)=0\ (t<0,\text{ some } a_k \text{ zero})$$

$$\lim_{t\to\infty}M(t)=\max.\quad(a_1,a_2,\ldots,a_n)=\max.\,a$$

$$\lim_{t\to-\infty}M(t)=\min.\quad(a_1,a_2,\ldots,a_n)=\min.\,a$$

$$\lim_{t \to 0} M(t) = G$$

$$M(1) = \mathrm{A}$$

$$M(-1) = H$$

Solution of quadratic equations

Given $az^2 + bz + c = 0$

$$z_{1,2} = -\left(\frac{b}{2a}\right) \pm \frac{1}{2a} q^{\frac{1}{2}}, \quad q = b^2 - 4ac,$$

$$z_1 + z_2 = -b/a, \quad z_1 z_2 = c/a$$

If $q > 0$, two real roots
$q = 0$, two equal roots
$q < 0$, pair of complex conjugate roots

Solution of cubic equations

Given $z^2 + a_2 z^2 + a_1 z + a_0 = 0$, let

$$q = \frac{1}{3} a_1 - \frac{1}{9} a_2^2$$

$$r = \frac{1}{6}(a_1 a_2 - 3a_0) - \frac{1}{27} a_2^3$$

If $q^3 + r^2 > 0$, one real root and a pair of complex conjugate roots
$q^3 + r^2 = 0$, all roots real and at least two are equal
$q^3 + r^2 < 0$, all roots real (irreducible case)

Let

$$s_1 = \left[r + (q^3 + r^2)^{\frac{1}{2}} \right]^{\frac{1}{2}}$$

$$s_2 = \left[r - (q^3 + r^2)^{\frac{1}{2}} \right]^{\frac{1}{2}}$$

then

$$z_1 = (s_1 + s_2) - \frac{a_2}{3}$$

$$z_2 = -\frac{1}{2}(s_1 + s_2) - \frac{a_2}{3} + \frac{i\sqrt{3}}{2}(s_1 - s_2)$$

$$z_3 = -\frac{1}{2}(s_1 + s_2) - \frac{a_2}{3} - \frac{i\sqrt{3}}{2}(s_1 - s_2)$$

If z_1, z_2, z_3 are the roots of the cubic equation

$$z_1 + z_2 + z_3 = -a_2$$

$$z_1 z_2 + z_1 z_3 + z_2 z_3 = a_1$$

$$z_1 z_2 z_3 = a_0$$

Trigonometric solution of the cubic equation

The form $x^3 + ax + b = 0$ with $ab \neq 0$ can always be solved by transforming it to the trigonometric identity

$$4\cos^3 \mathrm{q} - 3\cos \mathrm{q} - \cos(3\mathrm{q}) \equiv 0$$

Let $x = m \cos \theta$, then

$$x^3 + ax + b = m^3 \cos^3 \mathrm{q} + am \cos \mathrm{q} + b = 4\cos^3 \mathrm{q} - 3\cos \mathrm{q} - \cos(3\mathrm{q}) \equiv 0$$

Hence,

$$\frac{4}{m^3} = -\frac{3}{am} = \frac{-\cos(3\mathrm{q})}{b}$$

from which follows that

$$m = 2\sqrt{-\frac{a}{3}}, \quad \cos(3\mathrm{q}) = \frac{3b}{am}$$

Any solution θ_1 which satisfies $\cos(3\mathrm{q}) = \dfrac{3b}{am}$ will also have the solutions

$$\mathrm{q}_1 + \frac{2\mathrm{p}}{3} \quad \text{and} \quad \mathrm{q}_1 + \frac{4\mathrm{p}}{3}$$

The roots of the cubic $x^3 + ax + b = 0$ are

$$2\sqrt{-\frac{a}{3}}\cos\theta_1$$

$$2\sqrt{-\frac{a}{3}}\cos\left(\theta_1 + \frac{2\pi}{3}\right)$$

$$2\sqrt{-\frac{a}{3}}\cos\left(\theta_1 + \frac{4\pi}{3}\right)$$

Given $z^4 + a_3z^3 + a_2z^2 + a_1z + a_0 = 0$, find the real root u_1 of the cubic equation

$$u^3 - a_2u^2 + (a_1a_3 - 4a_0)u - (a_1^2 + a_0a_3^2 - 4a_0a_2) = 0$$

and determine the four roots of the quadric as solutions of the two quadratic equations

$$\theta^2 + \left[\frac{a_3}{2} \mp \left(\frac{a_3^2}{4} + u_1 - a_2\right)^{\frac{1}{2}}\right]\theta + \frac{u_1}{2} \mp \left[\left(\frac{u_1}{2}\right)^2 - a_0\right]^{\frac{1}{2}} = 0$$

If all roots of the cubic equation are real, use the value of u_1 which gives real coefficients in the quadratic equation and select signs so that if

$$z^4 + a_3z^3 + a_2z^3 + a_1z + a_0 = (z^2 + p_1z + q_1)(z^2 + p_2z + q_2)$$

then

$$p_1 + p_2 = a_3, \quad p_1p_2 + q_1 + q_2 = a_2, \quad p_1q_2 + p_2q_1 = a_1, \quad q_1q_2 = a_0$$

If z_1, z_2, z_3, z_4 are the roots,

$$\sum z_t = -a_3, \quad \sum z_tz_jz_k = -a_1$$

$$\sum z_tz_j = a_2, \quad z_1z_2z_3z_4 = a_0$$

Partial fractions

This section applies only to rational algebraic fractions with numerator of lower degree than the denominator. Improper fractions can be reduced to proper fractions by long division.

Every fraction may be expressed as the sum of component fractions whose denominators are factors of the denominator of the original fraction.

Let N(x) = numerator, a polynomial of the form

$$N(x) = n_0 + n_1 x + n_2 x^2 + \cdots + n_1 x^1$$

Non-repeated linear factors

$$\frac{N(x)}{(x-a)G(x)} = \frac{A}{x-a} + \frac{F(x)}{G(x)}$$

$$A = \left[\frac{N(x)}{G(x)}\right]_{x=a}$$

F(x) determined by methods discussed in the following sections.

Repeated linear factors

$$\frac{N(x)}{x^m G(x)} = \frac{A_0}{x^m} + \frac{A_1}{x^{m-1}} + \cdots + \frac{A_{m-1}}{x} + \frac{F(x)}{G(x)}$$

$$N(x) = n_o + n_1 x + n_2 x^2 + n_3 x^3 + \cdots$$

$$F(x) = f_0 + f_1 x + f_2 x^2 + \cdots,$$

$$G(x) = g_0 + g_1 x + g_2 x^2 + \cdots$$

$$A_0 = \frac{n_0}{g_0}, \quad A_1 = \frac{n_1 - A_0 g_1}{g_0}$$

$$A_2 = \frac{n_2 - A_0 g_2 - A_1 g_1}{g_0}$$

General terms

$$A_0 = \frac{n_0}{g_0}, \quad A_k = \frac{1}{g_0}\left[n_k - \sum_{t=0}^{k-1} A_t g_k - t\right] k \geqq 1$$

$$m^* = 1 \begin{cases} f_0 = n_1 - A_0 g_1 \\ f_1 = n_2 - A_0 g_2 \\ f_1 = n_{j+1} - A_0 g_{t+1} \end{cases}$$

$$m = 2 \begin{cases} f_0 = n_2 - A_0 g_2 - A_1 g_1 \\ f_1 = n_3 - A_0 g_3 - A_1 g_2 \\ f_1 = n_{j+2} - \left[A_0 g_{1+2} + A_1 g_1 + 1 \right] \end{cases}$$

$$m = 3 \begin{cases} f_0 = n_3 - A_0 g_3 - A_1 g_2 - A_2 g_1 \\ f_1 = n_3 - A_0 g_4 - A_1 g_3 - A_2 g_2 \\ f_1 = n_{j+3} - \left[A_0 g_{j+3} + A_1 g_{j+2} + A_2 g_{j+1} \right] \end{cases}$$

$$\text{any } m\text{: } f_1 = n_{m+1} - \sum_{i=0}^{m-1} A_1 g_{m+j-1}$$

$$\frac{N(x)}{(x-a)^m G(x)} = \frac{A_0}{(x-a)^m} + \frac{A_1}{(x-a)^{m-1}} + \cdots + \frac{A_{m-1}}{(x-a)} + \frac{F(x)}{G(x)}$$

Change to form $\dfrac{N'(y)}{y^m G'(y)}$ by substitution of $x = y + a$. Resolve into partial fractions in terms of y as described earlier. Then express in terms of x by substitution $y = x - a$.

Repeated linear factors
Alternative method of determining coefficients:

$$\frac{N(x)}{(x-a)^m G(x)} = \frac{A_0}{(x-a)^m} + \cdots + \frac{A_k}{(x-a)^{m-k}} + \cdots + \frac{A_{m-1}}{x-a} + \frac{F(x)}{G(x)}$$

$$A_k = \frac{1}{k!} \left\{ D_x^k \left[\frac{N(x)}{G(x)} \right] \right\}_{x-G}$$

where D_x^k is the differentiating operator, and the derivative of zero order is defined as

$$D_x^0 u = u.$$

Factors of higher degree
Factors of higher degree have the corresponding numerators indicated:

$$\frac{N(x)}{(x^2 + h_1 x + h_0) G(x)} = \frac{a_1 x + a_0}{x^2 + h_1 x + h_0} + \frac{F(x)}{G(x)}$$

$$\frac{N(x)}{(x^2+h_1x+h_0)^2G(x)} = \frac{a_1x+a_0}{(x^2+h_1x+h_0)^2} + \frac{b_1x+b_0}{(x^2+h_1x+h_0)} + \frac{F(x)}{G(x)}$$

$$\frac{N(x)}{(x^3+h_2x^2+h_1x+h_0)G(x)} = \frac{a_2x^2+a_1x+a_0}{x^3+h_2x^2+h_1x+h_0} + \frac{F(x)}{G(x)} \text{ etc.}$$

Problems of this type are determined first by solving for the coefficients due to linear factors as shown earlier, and then determining the remaining coefficients by the general methods given in the following.

Geometry

Mensuration formulas are used for measuring angles and distances in geometry. Examples are presented as follows.

Triangles

Let K = area, r = radius of the inscribed circle, R = radius of circumscribed circle.

Right triangle

$$A+B=C=90^\circ$$

$$c^2=a^2+b^2 \text{ (Pythagorean relations)}$$

$$a=\sqrt{(c+b)(c-b)}$$

$$K=\frac{1}{2}ab$$

$$r=\frac{ab}{a+b+c}, \quad R=\frac{1}{2}c$$

$$h=\frac{ab}{c}, \quad m=\frac{b^2}{c}, \quad n=\frac{a^2}{c}$$

Equilateral triangle

$$A=B=C=60^\circ$$

$$K=\frac{1}{4}a^2\sqrt{3}$$

$$r=\frac{1}{6}a\sqrt{3}, \quad R=\frac{1}{3}a\sqrt{3}$$

$$h=\frac{1}{2}a\sqrt{3}$$

General triangle

Let $s = \frac{1}{2}(a+b+c)$, $h_c =$ length of altitude on side c, t_c = length of bisector of angle C, m_c = length of median to side c.

$$A + B + C = 180^\circ$$

$$c^2 = a^2 + b^2 - 2ab\cos C \quad \text{(law of cosines)}$$

$$K = \frac{1}{2}h_c c = \frac{1}{2}ab\sin C$$

$$= \frac{c^2 \sin A \sin B}{2\sin C}$$

$$= rs = \frac{abc}{4R}$$

$$= \sqrt{s(s-a)(s-b)(s-c)} \quad \text{(Heron's formula)}$$

$$r = c\sin\frac{A}{2}\sin\frac{B}{2}\sec\frac{C}{2} = \frac{ab\sin C}{2s} = (s-c)\tan\frac{C}{2}$$

$$= \sqrt{\frac{(s-a)(s-b)(s-c)}{s}} = \frac{K}{s} = 4R\sin\frac{A}{2}\sin\frac{B}{2}\sin\frac{C}{2}$$

$$R = \frac{c}{2\sin C} = \frac{abc}{4\sqrt{s(s-a)(s-b)(s-c)}} = \frac{abc}{4K}$$

$$h_c = a\sin B = b\sin A = \frac{2K}{c}$$

$$t_c = \frac{2ab}{a+b}\cos\frac{C}{2} = \sqrt{ab\left\{1 - \frac{c^2}{(a+b)^2}\right\}}$$

$$m_c = \sqrt{\frac{a^2}{2} + \frac{b^2}{2} - \frac{c^2}{4}}$$

Menelaus' theorem

A necessary and sufficient condition for points D, E, F on the respective side lines BC, CA, AB of a triangle ABC to be collinear is that

$$BD \cdot CE \cdot AF = -DC \cdot EA \cdot FB,$$

where all segments in the formula are directed segments.

Ceva's theorem
A necessary and sufficient condition for AD, BE, CF, where D, E, F are points on the respective side lines BC, CA, AB of a triangle ABC, to be concurrent is that

$$BD \cdot CE \cdot AF = +DC \cdot EA \cdot FB,$$

where all segments in the formula are directed segments.

Quadrilaterals
Let K = area, p and q are diagonals.

Rectangle

$$A = B = C = D = 90^\circ$$

$$K = ab, \quad p = \sqrt{a^2 + b^2}$$

Parallelogram

$$A = C, \quad B = D, \quad A + B = 180^\circ$$

$$K = bh = ab \sin A = ab \sin B$$

$$h = a \sin A = a \sin B$$

$$p = \sqrt{a^2 + b^2 - 2ab \cos A}$$

$$q = \sqrt{a^2 + b^2 - 2ab \cos B} = \sqrt{a^2 + b^2 + 2ab \cos A}$$

Rhombus

$$p^2 + q^2 = 4a^2$$

$$K = \frac{1}{2} pq$$

Trapezoid

$$m = \frac{1}{2}(a + b)$$

$$K = \frac{1}{2}(a + b)h = mh$$

General quadrilateral

Let $s = \frac{1}{2}(a+b+c+d)$.

$$K = \frac{1}{2}pq\sin\theta$$

$$= \frac{1}{4}(b^2+d^2-a^2-c^2)\tan\theta$$

$$= \frac{1}{4}\sqrt{4p^2q^2-(b^2+d^2-a^2-c^2)^2}$$

(Bretschneider's formula)

$$= \sqrt{(s-a)(s-b)(s-c)(s-d)-abcd\cos^2\left(\frac{A+B}{2}\right)}$$

Theorem
The diagonals of a quadrilateral with consecutive sides a, b, c, d are perpendicular if and only if $a^2 + c^2 = b^2 + d^2$.

Regular polygon of n sides each of length b

$$\text{Area} = \frac{1}{4}nb^2\cot\frac{\pi}{n} = \frac{1}{4}nb^2\frac{\cos(\pi/n)}{\sin(\pi/n)}$$

Perimeter = nb

Circle of radius r

$$\text{Area} = \pi r^2$$

$$\text{Perimeter} = 2\pi r$$

Regular polygon of n sides inscribed in a circle of radius r

$$\text{Area} = \frac{1}{2}nr^2\sin\frac{2\pi}{n} = \frac{1}{2}nr^2\sin\frac{360^\circ}{n}$$

$$\text{Perimeter} = 2nr\sin\frac{\pi}{n} = 2nr\sin\frac{180^\circ}{n}$$

Regular polygon of n sides circumscribing a circle of radius r

$$\text{Area} = nr^2\tan\frac{\pi}{n} = nr^2\tan\frac{180^\circ}{n}$$

$$\text{Perimeter} = 2nr\tan\frac{\pi}{n} = 2nr\tan\frac{180^\circ}{n}$$

Cyclic quadrilateral
Let R = radius of the circumscribed circle.

$$A + C = B + D = 180^\circ$$

$$K = \sqrt{(s-a)(s-b)(s-c)(s-d)} = \frac{\sqrt{(ac+bd)(ad+bc)(ab+cd)}}{4R}$$

$$p = \sqrt{\frac{(ac+bd)(ab+cd)}{ad+bc}}$$

$$q = \sqrt{\frac{(ac+bd)(ad+bc)}{ab+cd}}$$

$$R = \frac{1}{2}\sqrt{\frac{(ac+bd)(ad+bc)(ab+cd)}{(s-a)(s-b)(s-c)(s-d)}}$$

$$\sin\theta = \frac{2K}{ac+bd}$$

Prolemy's theorem
A convex quadrilateral with consecutive sides a, b, c, d and diagonals p and q is cyclic if and only if $ac + bd = pq$.

Cyclic-inscriptable quadrilateral
Let r be the radius of the inscribed circle.
R is the radius of the circumscribed circle.
m is the distance between the centers of the inscribed and the circumscribed circles.

$$A + C = B + D = 180^\circ$$

$$a + c = b + d$$

$$K = \sqrt{abcd}$$

$$\frac{1}{(R-m)^2} + \frac{1}{(R+m)^2} = \frac{1}{r^2}$$

$$r = \frac{\sqrt{abcd}}{s}$$

$$R = \frac{1}{2}\sqrt{\frac{(ac+bd)(ad+bc)(ab+cd)}{abcd}}$$

Sector of circle of radius r

$$\text{Area} = \frac{1}{2}r^2\theta \quad [\theta \text{ in rad}]$$

$$\text{Arc length } s = r\theta$$

Radius of circle inscribed in a triangle of sides a, b, c

$$r = \frac{\sqrt{8(8-a)(8-b)(8-c)}}{8}$$

where $s = \frac{1}{2}(a+b+c) = \text{semiperimeter.}$

Radius of circle circumscribing a triangle of sides a, b, c

$$R = \frac{abc}{4\sqrt{8(8-a)(8-b)(8-c)}}$$

where $s = \frac{1}{2}(a+b+c) = \text{semiperimeter.}$

Segment of circle of radius r

$$\text{Area of shaded part} = \frac{1}{2}r^2(\theta - \sin\theta)$$

Ellipse of semi-major axis a and semi-minor axis b

$$\text{Area} = \pi ab$$

$$\text{Perimeter} = 4a\int_0^{\pi/2}\sqrt{1-k^2\sin^2\theta\, d\theta}$$

$$= 2\pi\sqrt{\frac{1}{2}(a^2+b^2)} \quad \text{[approximately]}$$

where $k = \sqrt{a^2-b^2}/a$.

Segment of a parabola

$$\text{Area} = \frac{2}{8}ab$$

$$\text{Arc length ABC} = \frac{1}{2}\sqrt{b^2+16a^2} + \frac{b^2}{8a}\ln\left(\frac{4a+\sqrt{b^2+16a^2}}{b}\right)$$

Planar areas by approximation

Divide the planar area K into n strips by equidistant parallel chords of lengths y_0, y_1, y_2,…, y_n (where y_0 and/or y_n may be zero), and let h denote the common distance between the chords.

Then, approximately,

Trapezoidal rule

$$K = h\left(\frac{1}{2}y_0 + y_1 + y_2 + \cdots + y_{n-1} + \frac{1}{2}y_n\right)$$

Durand's rule

$$K = h\left(\frac{4}{10}y_0 + \frac{11}{10}y_1 + y_2 + y_3 + \cdots + y_{n-2} + \frac{11}{10}y_{n-1} + \frac{4}{10}y_n\right)$$

Simpson's rule (*n even*)

$$K = \frac{1}{3}h(y_0 + 4y_1 + 2y_2 + 4y_3 + 2y_4 + \cdots + 2y_{n-2} + 4y_{n-1} + y_n)$$

Weddle's rule ($n = 6$)

$$K = \frac{3}{10}h(y_0 + 5y_1 + y_2 + 6y_3 + y_4 + 5y_5 + y_6)$$

Solids bounded by planes

In the following: S = lateral surface, T = total surface, V = volume.

Cube

Let a = length of each edge

$$T = 6a^2, \quad \text{diagonal of face} = a\sqrt{2}$$

$$V = a^3, \quad \text{diagonal of cube} = a\sqrt{3}$$

Rectangular parallelepiped (or box)

Let a, b, c, be the lengths of its edges.

$$T = 2(ab + bc + ca), \quad V = abc$$

$$\text{diagonal} = \sqrt{a^2 + b^2 + c^2}$$

Prism
S = (perimeter of right section) × (lateral edge)
V = (area of right section) × (lateral edge)
= (area of base) × (altitude)

Truncated triangular prism

$$\text{V} = (\text{area of right section}) \times \frac{1}{3}(\text{sum of the three lateral edges})$$

Pyramid
S of regular pyramid = $\frac{1}{2}$ (perimeter of base) × (slant height)

V = $\frac{1}{3}$ (area of base) × (altitude)

Frustum of pyramid
Let B_1 = area of lower base, B_2 = area of upper base, h = altitude.

S of regular figure = $\frac{1}{2}$ (sum of perimeters of base) × (slant height)

$$V = \frac{1}{3}h\left(B_1 + B_2 + \sqrt{B_1 B_2}\right)$$

Prismatoid
A prismatoid is a polyhedron having for bases two polygons in parallel planes, and for lateral faces triangles or trapezoids with one side lying in one base, and the opposite vertex or side lying in the other base, of the polyhedron. Let B_1 = area of lower base, M = area of midsection, B_2 = area of upper base, h = altitude.

$$V = \frac{1}{6}h(B_1 + 4M + B_2) \quad \text{(the prismoidal formula)}$$

Note: Since cubes, rectangular parallelepipeds, prisms, pyramids, and frustums of pyramids are all examples of prismatoids, the formula for the volume of a prismatoid subsumes most of the aforementioned volume formulae.

Regular polyhedra

Let
v = number of vertices
e = number of edges
f = number of faces
α = each dihedral angle

a = length of each edge
r = radius of the inscribed sphere
R = radius of the circumscribed sphere
A = area of each face
T = total area
V = volume

$$v - e + f = 2$$

$$T = fA$$

$$V = \frac{1}{3}rfA = \frac{1}{3}rT$$

Name	Nature of surface	T	V
Tetrahedron	4 equilateral triangles	$1.73205a^2$	$0.11785a^3$
Hexahedron (cube)	6 squares	$6.00000a^2$	$1.00000a^3$
Octahedron	8 equilateral triangles	$3.46410a^2$	$0.47140a^3$
Dodecahedron	12 regular pentagons	$20.64573a^2$	$7.66312a^3$
Icosahedron	20 equilateral triangles	$8.66025a^2$	$2.18169a^2$

Name	v	e	f	α	a	r
Tetrahedron	4	6	4	70° 32′	1.633R	0.333R
Hexahedron	8	12	6	90°	1.155R	0.577R
Octahedron	6	12	8	190° 28′	1.414R	0.577R
Dodecahedron	20	30	12	116° 34′	0.714R	0.795R
Icosahedron	12	30	20	138° 11′	1.051R	0.795R

Name	A	r	R	V
Tetrahedron	$\frac{1}{4}a^2\sqrt{3}$	$\frac{1}{12}a\sqrt{6}$	$\frac{1}{4}a\sqrt{6}$	$\frac{1}{12}a^3\sqrt{2}$
Hexahedron (cube)	a^2	$\frac{1}{2}a$	$\frac{1}{2}a\sqrt{3}$	a^3
Octahedron	$\frac{1}{4}a^2\sqrt{3}$	$\frac{1}{6}a\sqrt{6}$	$\frac{1}{2}a\sqrt{2}$	$\frac{1}{3}a^3\sqrt{2}$
Dodecahedron	$\frac{1}{4}a^2\sqrt{25+10\sqrt{5}}$	$\frac{1}{20}a\sqrt{250+110\sqrt{5}}$	$\frac{1}{4}a\left(\sqrt{15}+\sqrt{3}\right)$	$\frac{1}{4}a^3\left(15+7\sqrt{5}\right)$
Icosahedron	$\frac{1}{4}a^2\sqrt{3}$	$\frac{1}{12}a\sqrt{42+18\sqrt{5}}$	$\frac{1}{4}a\sqrt{10+2\sqrt{5}}$	$\frac{5}{12}a^3\left(3+\sqrt{5}\right)$

Sphere of radius r

$$\text{Volume} = \frac{3}{4}\pi r^3$$

$$\text{Surface area} = 4\pi r^2$$

Right circular cylinder of radius r and height h

$$\text{Volume} = \pi r^2 h$$

$$\text{Lateral surface area} = 2\pi rh$$

Circular cylinder of radius r and slant height ℓ

$$\text{Volume} = \pi r^2 h = \pi r^2 \ell \sin\theta$$

$$\text{Lateral surface area} = p\ell$$

Cylinder of cross-sectional area A and slant height ℓ

$$\text{Volume} = Ah = A\ell \sin\theta$$

$$\text{Lateral surface area} = p\ell$$

Right circular cone of radius r and height h

$$\text{Volume} = \frac{1}{3}\pi r^2 h$$

$$\text{Lateral surface area} = \pi r\sqrt{r^2 + h^2} = \pi rl$$

Spherical cap of radius r and height h

$$\text{Volume (shaded in figure)} = \frac{1}{3}\pi h^2(3r - h)$$

$$\text{Surface area} = 2\pi rh$$

Frustum of right circular cone of radii a, b, and height h

$$\text{Volume} = \frac{1}{3}\pi h(a^2 + ab + b^2)$$

$$\text{Lateral surface area} = \pi(a+b)\sqrt{h^2 + (b-a)^2}$$

$$= \pi(a+b)l$$

Zone and segment of two bases

$$S = 2\pi Rh = \pi Dh$$

$$V = \frac{1}{6}\pi h(3a^2 + 3b^2 + h^2)$$

Lune

$S = 2R^3\theta$, θ in rad

Spherical sector

$$V = \frac{2}{3}\pi R^2 h = \frac{1}{6}\pi D^2 h$$

Spherical triangle and polygon
Let A, B, C be the angles, in radians, of the triangle; let θ = sum of angles, in radians, of a spherical polygon on n sides.

$$S = (A + B + C - \pi)R^2$$

$$S = [\theta - (n-2)\pi]R^2$$

Spheroids

Ellipsoid
Let a, b, c be the lengths of the semi-axes.

$$V = \frac{4}{3}\pi abc$$

Oblate spheroid
An oblate spheroid is formed by the rotation of an ellipse about its minor axis. Let a and b be the major and minor semi-axes, respectively, and ϵ the eccentricity, of the revolving ellipse.

$$S = 2\pi a^2 + \pi \frac{b^2}{\epsilon} \log_e \frac{1+\epsilon}{1-\epsilon}$$

$$V = \frac{4}{3}\pi a^2 b$$

Prolate spheroid

A prolate spheroid is formed by the rotation of an ellipse about its major axis. Let a and b be the major and minor semi-axes, respectively, and ϵ the eccentricity, of the revolving ellipse.

$$S = 2\pi b^2 + 2\pi \frac{ab}{\epsilon} \sin^{-1} \epsilon$$

$$V = \frac{3}{4}\pi ab^2$$

Circular torus

A circular torus is formed by the rotation of a circle about an axis in the plane of the circle and not cutting the circle. Let r be the radius of the revolving circle and let R be the distance of its center from the axis of rotation.

$$S = 4\pi^2 Rr$$

$$V = 2\pi^2 Rr^2$$

Formulas from plane analytic geometry

Distance d between two points

$$P_1(x_1, y_1) \quad \text{and} \quad P_2(x_2, y_2)$$

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Slope m of line joining two points

$$P_1(x_1, y_1) \quad \text{and} \quad P_2(x_2, y_2)$$

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \tan\theta$$

Equation of line joining two points

$$P_1(x_1, y_1) \quad \text{and} \quad P_2(x_2, y_2)$$

$$\frac{y - y_1}{x - x_1} = \frac{y_2 - y_1}{x_2 - x_1} = m \quad \text{or} \quad y - y_1 = m(x - x_1)$$

$$y = mx + b$$

where $b = y_1 - mx_1 = \dfrac{x_2 y_1 - x_1 y_2}{x_2 - x_1}$ is the intercept on the y axis, that is, the y intercept.

Equation of line in terms of x intercept $a \neq 0$ and y intercept $b \neq 0$

$$\frac{x}{a} + \frac{y}{b} = 1$$

Normal form for equation of line

$$x\cos\alpha + y\sin\alpha = p$$

where

p is the perpendicular distance from origin O to line
α is the angle of inclination of perpendicular with positive x axis

General equation of line

$$Ax + By + C = 0$$

Distance from point (x_1, y_1) to line $Ax + By + C = 0$

$$\frac{Ax_1 + By_1 + C}{\pm\sqrt{A^2 + B^2}}$$

where the sign is chosen so that the distance is nonnegative.

Angle ψ between two lines having slopes m_1 and m_2

$$\tan\psi = \frac{m_2 - m_1}{1 + m_1 m_2}$$

Lines are parallel or coincident if and only if $m_1 = m_2$.
Lines are perpendicular if and only if $m_2 = -1/m_1$.

Area of triangle with verticles
At (x_1, y_1), (x_2, y_2), (x_3, y_3)

$$\text{Area} = \pm\frac{1}{2}\begin{vmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{vmatrix}$$

$$= \pm\frac{1}{2}(x_1y_2 + y_1x_3 + y_3x_2 - y_2x_3 - y_1x_2 - x_1y_3)$$

where the sign is chosen so that the area is nonnegative.
If the area is zero the points all lie on a line.

Transformation of coordinates involving pure translation

$$\begin{cases} x = x' + x_0 \\ y = y' + y_0 \end{cases} \quad \text{or} \quad \begin{cases} x' = x + x_0 \\ y' = y + y_0 \end{cases}$$

where

x, y are old coordinates [i.e., coordinates relative to xy system]
(x', y') are new coordinates [relative to $x'y'$ system]
$(x_0\ y_0)$ are the coordinates of the new origin O' relative to the old xy coordinate system

Transformation of coordinates involving pure rotation

$$\begin{cases} x = x'\cos\alpha - y'\sin\alpha \\ y = x'\sin\alpha + y'\cos\alpha \end{cases} \quad \text{or} \quad \begin{cases} x' = x\cos\alpha + y\sin\alpha \\ y' = y\cos\alpha - x\sin\alpha \end{cases}$$

where the origins of the old [xy] and new [$x'y'$] coordinate systems are the same but the x' axis makes an angle α with the positive x axis.

Transformation of coordinates involving translation and rotation

$$\begin{cases} x = x'\cos\alpha - y'\sin\alpha + x_0 \\ y = x'\sin\alpha + y'\cos\alpha + y_0 \end{cases}$$

$$\text{or} \quad \begin{cases} x' = (x - x_0)\cos\alpha + (y - y_0)\sin\alpha \\ y' = (y - y_0)\cos\alpha - (x - x_0)\sin\alpha \end{cases}$$

where the new origin O' of $x'y'$ coordinate system has coordinates (x_0, y_0) relative to the old xy coordinate system and the x' axis makes an angle α with the positive x axis.

Polar coordinates (r, θ)

A point P can be located by rectangular coordinates (x, y) or polar coordinates (r, θ). The transformation between these coordinates is

$$\begin{cases} x = r\cos\theta \\ y = r\sin\theta \end{cases} \quad \text{or} \quad \begin{cases} r = \sqrt{x^2 + y^2} \\ \theta = \tan^{-1}(y/x) \end{cases}$$

Plane curves

$$(x^2 + y^2)^2 = ax^2y$$

$$r = a\sin\theta\cos^2\theta$$

Catenary, hyperbolic cosine

$$y = \frac{a}{2}(e^{x/e} + e^{-x/e}) = a\cosh\frac{x}{a}$$

Cardioid

$$(x^2 + y^2 - ax)^2 = a^2(x^2 + y^2)$$

$$r = a(\cos\theta + 1)$$

or

$$r = a(\cos\theta - 1)$$

$$\left[P'A = AP = a\right]$$

Circle

$$x^2 + y^2 = a^2$$

$$r = a$$

Cassinian curves

$$x^2 + y^2 = 2ax$$

$$r = 2a\cos\theta$$

$$x^2 + y^2 = ax + by$$

$$r = a\cos\theta + b\sin\theta$$

Cotangent curve

$$y = \cot x$$

Cubical parabola

$$y = ax^3, \quad a > 0$$

$$r^2 = \frac{1}{a}\sec^2\theta\tan\theta, \quad a > 0$$

Cosecant curve

$$y = \csc x$$

Cosine curve

$$y = \cos x$$

Ellipse

$$x^2/a^2 + y^2/b^2 = 1$$

$$\begin{cases} x = a\cos \mathtt{f} \\ y = b\sin \mathtt{f} \end{cases}$$

Gamma function

$$\Gamma(n) = \int_0^\infty x^{n-1}e^{-x}dx \quad (n > 0)$$

$$\Gamma(n) = \frac{\Gamma(n+1)}{n} \quad (0 > n \neq -1, -2, -3, \ldots)$$

Hyperbolic functions

$$\sinh x = \frac{e^x - e^{-x}}{2} \quad \csc hx = \frac{2}{e^x - e^{-x}}$$

$$\cosh x = \frac{e^x - e^{-x}}{2} \quad \csc hx = \frac{2}{e^x - e^{-x}}$$

$$\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad \coth x = \frac{e^x + e^{-x}}{e^x - e^{-x}}$$

Inverse cosine curve

$$y = \arccos x$$

Inverse sine curve

$$y = \arcsin x$$

Inverse tangent curve

$$y = \arctan x$$

Logarithmic curve

$$y = \log_a x$$

Parabola

$$y = x^2$$

Cubical parabola

$$y = x^3$$

Tangent curve

$$y = \tan x$$

Ellipsoid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$$

Elliptic cone

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 0$$

Elliptic cylinder

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

Hyperboloid of one sheet

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1$$

Elliptic paraboloid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = cz$$

Hyperboloid of two sheets

$$\frac{z^2}{c^2} - \frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$$

Hyperbolic paraboloid

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = cz$$

Sphere

$$x^2 + y^2 + z^2 = a^2$$

Distance d between two points

$$P_1(x_1, y_1, z_1) \quad \text{and} \quad P_2(x_2, y_2, z_2)$$

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

Equations of line joining $P_1(x_1, y_1, z_1)$ and $P_2(x_2, y_2, z_2)$ in standard form

$$\frac{x - x_1}{x_2 - x_1} = \frac{y - y_1}{y_2 - y_1} = \frac{z - z_1}{z_2 - z_1} \quad \text{or}$$

$$\frac{x - x_1}{l} = \frac{y - y_1}{m} = \frac{z - z_1}{n}$$

Equations of line joining $P_1(x_1, y_1, z_1)$ and $P_2(x_2, y_2, z_2)$ in parametric form

$$x = x_1 + lt, \quad y = y_1 + mt, \quad z = z_1 + nt$$

Angle ϕ between two lines with direction cosines l_1, m_1, n_1 and l_2, m_2, n_2

$$\cos \mathtt{f} = l_1 l_2 + m_1 m_2 + n_1 n_2$$

General equation of a plane

$$Ax + By + Cz + D = 0$$

where A, B, C, D are constants.

Equation of plane passing through points

$$(x_1, y_1, z_1), \quad (x_2, y_2, z_2), \quad (x_3, y_3, z_3)$$

$$\begin{vmatrix} x - x_1 & y - y_1 & z - z_1 \\ x_2 - x_1 & y_2 - y_1 & z_2 - z_1 \\ x_3 - x_1 & y_3 - y_1 & z_3 - z_1 \end{vmatrix} = 0$$

or

$$\begin{vmatrix} y_2 - y_1 & z_2 - z_1 \\ y_3 - y_1 & z_3 - z_1 \end{vmatrix} (x - x_1) + \begin{vmatrix} z_2 - z_1 & x_2 - x_1 \\ z_3 - z_1 & x_3 - x_1 \end{vmatrix} (y - y_1)$$

$$+ \begin{vmatrix} x_2 - x_1 & y_2 - y_1 \\ x_3 - x_1 & y_3 - y_1 \end{vmatrix} (z - z_1) = 0$$

Equation of plane in intercept form

$$\frac{x}{a}+\frac{y}{b}+\frac{z}{c}=1$$

where a, b, c are the intercepts on the x, y, z axes, respectively.

Equations of line through (x_0, y_0, z_0) *and perpendicular to plane*

$$Ax+By+Cz+D=0$$

$$\frac{x-x_0}{A}=\frac{y-y_0}{B}=\frac{z-z_0}{C}$$

$$\text{or} \quad x=x_0+At, \quad y=y_0+Bt, \quad z=z_0+Ct$$

Distance from point (x, y, z) *to plane* $Ax + By + D = 0$

$$\frac{Ax_0+By_0+Cz_0+D}{\pm\sqrt{A^2+B^2+C^2}}$$

where the sign is chosen so that the distance is nonnegative.

Normal form for equation of plane

$$x\cos\alpha+y\cos\beta+z\cos\gamma=p$$

where

p is the perpendicular distance from O to plane at P

α, β, γ are angles between OP and positive x, y, z axes

Transformation of coordinates involving pure translation

$$\begin{cases} x=x'+x_0 \\ y=y'+y_0 \\ z=z'+z_0 \end{cases} \quad \text{or} \quad \begin{cases} x'=x+x_0 \\ y'=y+y_0 \\ z'=z+z_0 \end{cases}$$

where (x, y, z) are old coordinates [i.e., coordinates relative to xyz system], (x', y', z') are new coordinates [relative to (x', y', z') system and (x_0, y_0, z_0) are the coordinates of the new origin O' relative to the old xyz coordinate system.

Transformation of coordinates involving pure rotation

$$\begin{cases} x=l_1x'+l_2y'+l_3z' \\ y=m_1x'+m_2y'+m_3z' \\ z=n_1x'+n_2y'+n_3z' \end{cases} \quad \text{or} \quad \begin{cases} x'=l_1x+m_1y+n_1z \\ y'=l_2x+m_2y+n_3z \\ z'=l_3x+m_3y+n_3z \end{cases}$$

where the origins of the xyz and x', y', z' systems are the same and l_1, m_1, n_1; l_2, m_2, n_2; l_3, m_3, n_3 are the direction cosines of the x', y', z' axes relative to the x, y, z axes respectively.

Transformation of coordinates involving translation and rotation

$$\begin{cases} x = l_1x' + l_2y' + l_3z' + x_0 \\ y = m_1x' + m_2y' + m_3z' + y_0 \\ z = n_1x' + n_2y' + n_3z' + z_0 \end{cases}$$

or

$$\begin{cases} x' = l_1(x - x_0) + m_1(y - y_0) + n_1(z - z_0) \\ y' = l_2(x - x_0) + m_2(y - y_0) + n_2(z - z_0) \\ z' = l_3(x - x_0) + m_3(y - y_0) + n_3(z - z_0) \end{cases}$$

where the origin O' of the $x'y'z'$ system has coordinates (x_0, y_0, z_0) relative to the xyz system and l_1, m_1, n_1; l_2, m_2, n_2; l_3, m_3, n_3 are the direction cosines of the $x'y'z'$ axes relative to the x, y, z axes respectively.

Cylindrical coordinates (r, θ, z)
A point P can be located by cylindrical coordinates (r, θ, z) as well as rectangular coordinates (x, y, z). The transformation between these coordinates is

$$\begin{cases} x = r\cos\theta \\ y = r\sin\theta \\ z = z \end{cases} \quad \text{or} \quad \begin{cases} r = \sqrt{x^2 + y^2} \\ \theta = \tan^{-1}(y/x) \\ z = z \end{cases}$$

Spherical coordinates (r, θ, ϕ)
A point P can be located by cylindrical coordinates (r, θ, ϕ) as well as rectangular coordinates (x, y, z). The transformation between these coordinates is

$$\begin{cases} x = r\cos\theta\cos\phi \\ y = r\sin\theta\sin\phi \\ z = r\cos\theta \end{cases} \quad \text{or} \quad \begin{cases} r = \sqrt{x^2 + y^2 + z^2} \\ \phi = \tan^{-1}(y/x) \\ \theta = \cos^{-1}\left(z/\sqrt{x^2 + y^2 + z^2}\right) \end{cases}$$

Equation of sphere in rectangular coordinates

$$(x-x_0)^2+(y-y_0)^2+(z-z_0)^2=R^2$$

where the sphere has cent (x_0, y_0, z_0) and radius R.

Equation of sphere in cylindrical coordinates

$$r^2-2r_0r\cos(\theta-\theta_0)+r_0^2+(z-z_0)^2=R^2$$

where the sphere has center (r_0, θ_0, z_0) in cylindrical coordinates and radius R.

If the center is at the origin, the equation is

$$r^2+z^2=R^2$$

Equation of sphere in spherical coordinates

$$r^2+r_0^2-2r_0r\sin\theta\sin\theta_0\cos(\phi-\phi_0)=R^2$$

where the sphere has center (r_0, θ_0, ϕ_0) in spherical coordinates and radius R.

If the center is at the origin, the equation is

$$r=R$$

Logarithmic identities

$$\mathrm{Ln}(z_1z_2)=\mathrm{Ln}\,z_1+\mathrm{Ln}\,z_2.$$

$$\ln(z_1z_2)=\ln z_1+\ln z_2 \quad (-\pi<\arg z_1+\arg z_2\le\pi)$$

$$\mathrm{Ln}\,\frac{z_1}{z_2}=\mathrm{Ln}\,z_1-\mathrm{Ln}\,z_2$$

$$\ln\frac{z_1}{z_2}=\ln z_1-\ln z_2 \quad (-\pi<\arg z_1-\arg z_2\le\pi)$$

$$\mathrm{Ln}\,z^n=n\,\mathrm{Ln}\,z \quad (n \text{ integer})$$

$$\ln z^n=n\ln z \quad (n\,\text{integer}, -\pi<n\arg z\le\pi)$$

Special values

$$\ln 1=0$$

$$\ln 0=-\infty$$

$$\ln(-1) = \pi i$$

$$\ln(\pm i) = \pm \frac{1}{2}\pi i$$

$\ln e = 1$, e is the real number such that

$$\int_1^e \frac{dt}{t} = 1$$

$$e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = 2.71828\ 18284\ldots$$

Logarithms to general base

$$\log_a z = \ln z / \ln a$$

$$\log_a z = \frac{\log_b z}{\log_b a}$$

$$\log_a b = \frac{1}{\log_b a}$$

$$\log_e z = \ln z$$

$$\log_{10} z = \ln z / \ln 10 = \log_{10} e \ln z = (.43429\quad 44819\ldots)\ln z$$

$$\ln z = \ln 10 \log_{10} z = (2.30258\quad 50929\ldots)\log_{10} z$$

$$\left(\begin{array}{l}\log_e x = \ln x\text{, called natural, Napierian, or hyperbolic logarithms;}\\ \log_{10} x\text{, called common or Briggs logarithms.}\end{array}\right)$$

Series expansions

$$\ln(1+z) = z - \frac{1}{2}z^2 + \frac{1}{3}z^3 - \cdots \quad \left(|z| \le 1 \text{ and } z \ne -1\right)$$

$$\ln z = \left(\frac{z-1}{z}\right) + \frac{1}{2}\left(\frac{z-1}{z}\right)^2 + \frac{1}{3}\left(\frac{z-1}{z}\right)^3 + \cdots \quad \left(\Re z \ge \frac{1}{2}\right)$$

$$\ln z = (z-1) - \frac{1}{2}(z-1)^2 + \frac{1}{3}(z-1)^3 - \cdots \quad \left(|z-1| \le 1,\ z \ne 0\right)$$

$$\ln z = 2\left[\left(\frac{z-1}{z+1}\right) + \frac{1}{3}\left(\frac{z-1}{z+1}\right)^3 + \frac{1}{5}\left(\frac{z-1}{z+1}\right)^5 + \cdots\right] \quad (\Re z \ge 0,\ z \ne 0)$$

$$\ln\left(\frac{z+1}{z-1}\right) = 2\left(\frac{1}{z} + \frac{1}{3z^3} + \frac{1}{5z^5} + \cdots\right) \quad \left(|z| \ge 1,\ z \ne \pm 1\right)$$

$$\ln(z+a) = \ln a + 2\left[\left(\frac{z}{2a+z}\right) + \frac{1}{3}\left(\frac{z}{2a+z}\right)^3 + \frac{1}{5}\left(\frac{z}{2a+z}\right)^5 + \cdots\right]$$

$$(a > 0,\ \Re z \ge -a \ne z)$$

Limiting values

$$\lim_{x\to\infty} x^{-\mathtt{a}} \ln x = 0 \quad (\mathtt{a} \text{ constant}, \Re \mathtt{a} > 0)$$

$$\lim_{x\to 0} x^{\mathtt{a}} \ln x = 0 \quad (\mathtt{a} \text{ constant}, \Re \mathtt{a} > 0)$$

$$\lim_{m\to\infty}\left(\sum_{k=1}^{m} \frac{1}{k} - \ln m\right) = \mathtt{g} \quad \text{(Euler's constant)} = .57721\ 56649\ldots$$

Inequalities

$$\frac{x}{1+x} < \ln(1+x) < x \quad (x > -1,\ x \ne 0)$$

$$x < -\ln(1-x) < \frac{x}{1+x} \quad (x < 1,\ x \ne 0)$$

$$\left|\ln(1-x)\right| < \frac{3x}{2} \quad (0 < x \le .5828)$$

$$\ln x \le x - 1 \quad (x > 0)$$

$$\ln x \le n(x^{1/n} - 1) \quad \text{for any positive } n\ (x > 0)$$

$$\left|\ln(1-z)\right| \le -\ln\left(1 - |z|\right) \quad \left(|z| < 1\right)$$

Continued fractions

$$\ln(1+z) = \frac{z}{1+}\frac{z}{2+}\frac{z}{3+}\frac{4z}{4+}\frac{4z}{5+}\frac{9z}{6+}\cdots$$

(z in the plane cut from -1 to $-\infty$)

$$\ln\left(\frac{1+z}{1-z}\right) = \frac{2z}{1-}\frac{z^2}{3-}\frac{4z^2}{5-}\frac{9z^2}{7-}\cdots$$

Polynomial approximations

$$\frac{1}{\sqrt{10}} \le x \le \sqrt{10}$$

$$\log_{10} x = a_1 t + a_3 t^3 + \epsilon(x), \quad t = (x-1)/(x+1)$$

$$|\epsilon(x)| \le 6 \times 10^{-4}$$

$$a_1 = .86304 \quad a_3 = .36415$$

$$\frac{1}{\sqrt{10}} \le x \le \sqrt{10}$$

$$\log_{10} x = a_1 t + a_3 t^3 + a_5 t^5 + a_7 t^7 + a_9 t^9 + \epsilon(x)$$

$$t = (x-1)/(x+1)$$

$$|\epsilon(x)| \le 10^{-7}$$

$$a_1 = .86859\ 1718$$

$$a_3 = .28933\ 5524$$

$$a_5 = .17752\ 2071$$

$$a_7 = .09437\ 6476$$

$$a_9 = .19133\ 7714$$

$$0 \le x \le 1$$

$$\ln(1+x) = a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 + a_5 x^5 + \epsilon(x)$$

$$|\epsilon(x)| \le 1 \times 10^{-5}$$

$$a_1 = .99949\ 556$$

$$a_2 = .49190\ 896$$

$$a_3 = .28947\ 478$$

$$a_4 = .13606\ 275$$

$$a_5 = .03215\ 845$$

$$0 \le x \le 1$$

$$\ln(1+x) = a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5 + a_6x^6 + a_7x^7 + a_8x^8 + \epsilon(x)$$

$$|\epsilon(x)| \le 3 \times 10^{-8}$$

$$a_1 = .99999\ 64239$$

$$a_2 = -.49987\ 41238$$

$$a_3 = .33179\ 90258$$

$$a_4 = -.24073\ 38084$$

$$a_5 = .16765\ 40711$$

$$a_6 = -.09532\ 93897$$

$$a_7 = .03608\ 84937$$

$$a_8 = -.00645\ 35442$$

Exponential function series expansion

$$e^z = \exp z = 1 + \frac{z}{1!}\frac{z^2}{2!} + \frac{z^3}{3!} + \cdots \quad (z = x + iy)$$

Fundamental properties

$$\mathrm{Ln}(\exp z) = z + 2k\pi i \quad (k \text{ any integer})$$

$$\ln(\exp z) = z\left(-\pi < \oint z \le \pi\right)$$

$$\exp(\ln z) = \exp(\operatorname{Ln} z) = z$$

$$\frac{d}{dz}\exp z = \exp z$$

Definition of general powers

$$\text{If } N = a^z, \quad \text{then } z = \log_a N$$

$$a^z = \exp(z \ln a)$$

$$\text{If } a = |a|\exp(i \arg a) \quad (-\pi < \arg a \le \pi)$$

$$\left|a^z\right| = |a|^x e^{-v \arg a}$$

$$\arg(a^z) = y \ln|a| + x \arg a$$

$$\operatorname{Ln} a^z = z \ln a \quad \text{for one of the values of Ln } a^z$$

$$\ln a^x = x \ln a \quad (a \text{ real and positive})$$

$$\left|e^z\right| = e^x$$

$$\arg(e^z) = y$$

$$a^{z_1} a^{z_2} = a^{z_1 + z_2}$$

$$a^z b^z = (ab)^z \quad (-\pi < \arg a + \arg b \le \pi)$$

Logarithmic and exponential functions

Periodic property

$$e^{z + 2\pi ki} = e^z \quad (k \text{ any integer})$$

$$e^x < \frac{1}{1-x} \quad (x < 1)$$

$$\frac{1}{1-x} < (1 - e^{-x}) < x \quad (x > -1)$$

$$x < (e^x - 1) < \frac{1}{1-x} \quad (x < 1)$$

$$1 + x > e^{\frac{x}{1+x}} \quad (x > -1)$$

$$e^x > 1 + \frac{x^n}{n!} \quad (n > 0,\ x > 0)$$

$$e^x > \left(1 + \frac{x}{y}\right)^y > \frac{xy}{e^{x+y}} \quad (x > 0,\ y > 0)$$

$$e^{-x} < 1 - \frac{x}{2} \quad (0 < x \le 1.5936)$$

$$\frac{1}{4}|z| < \left|e^z - 1\right| < \frac{7}{4}|z| \quad (0 < |z| < 1)$$

$$\left|e^z - 1\right| \le e^{|z|} - 1 \le |z| e^{|z|} \quad (\text{all } z)$$

$$e^{2a \arctan \frac{1}{2}} = 1 + \frac{2a}{z - a +} \frac{a^2 + 1}{3z +} \frac{a^2 + 4}{5z +} \frac{a^2 + 9}{7z +} \cdots$$

Polynomial approximations

$$0 \le x \le \ln 2 = .693\ldots$$

$$e^{-x} = 1 + a_1 x + a_2 x^2 + \epsilon(x)$$

$$|\epsilon(x)| \le 3 \times 10^{-3}$$

$$a_1 = -.9664$$

$$a_2 = .3536$$

$$0 \le x \le \ln 2$$

$$e^{-x} = 1 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 + \epsilon(x)$$

$$|\epsilon(x)| \le 3 \times 10^{-5}$$

$$a_1 = -.99986\ 84$$

$$a_2 = .49829\ 26$$

$$a_3 = -.15953\ 32$$

$$a_4 = .02936\ 41$$

$$0 \le x \le \ln 2$$

$$e^{-x} = 1 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5 + a_6x^6 + a_7x^7 + \varepsilon(x)$$

$$|\varepsilon(x)| \leq 2 \times 10^{-10}$$

$$a_1 = -.99999\ 99995$$

$$a_2 = .49999\ 99206$$

$$a_3 = -.16666\ 53019$$

$$a_4 = .04165\ 73475$$

$$a_5 = -.00830\ 13598$$

$$a_6 = .00132\ 98820$$

$$a_7 = -.00014\ 13161$$

$$0 \leq x \leq 1$$

$$10^x = (1 + a_1x + a_2x^2 + a_3x^3 + a_4x^4)^2 + \varepsilon(x)$$

$$|\varepsilon(x)| \leq 7 \times 10^{-4}$$

$$a_1 = 1.14991\ 96$$

$$a_2 = .67743\ 23$$

$$a_3 = .20800\ 30$$

$$a_4 = .12680\ 89$$

$$0 \leq x \leq 1$$

$$10^x = (1 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5 + a_6x^6 + a_7x^7)^2 + \varepsilon(x)$$

$$|\varepsilon(x)| \leq 5 \times 10^{-8}$$

$$a_1 = 1.15129\ 277603$$

$$a_2 = .66273\ 088429$$

$$a_3 = .25439\ 357484$$

$$a_4 = .07295\,173666$$

$$a_5 = .01742\,111988$$

$$a_6 = .00255\,491796$$

$$a_7 = .00093\,264267$$

Surface area of cylinder = $2\pi rh + 2\pi r^2$
Volume of cylinder = $\pi r^2 h$
Surface area of a cone = $\pi r^2 + \pi rs$
Volume of a cone = $\dfrac{\pi r^2 h}{3}$
Volume of a pyramid = $\dfrac{Bh}{3}$
(B = area of base)
Slopes
Equation of a straight line: $y - y_1 = m(x - x_1)$

$$\text{where } m = \text{slope} = \frac{\text{rise}}{\text{run}}$$

$$= \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$$

or

$$y = mx + b$$

where m = slope, b = y-intercept.

Trigonometric ratios

$$\tan\theta = \frac{\sin\theta}{\cos\theta}$$

$$\sin^2\theta + \cos^2\theta = 1$$

$$1 + \tan^2\theta = \sec^2\theta$$

$$1 + \cot^2\theta = \csc^2\theta$$

$$\cos^2\theta - \sin^2\theta = \cos 2\theta$$

$$\sin 45^\circ = \frac{1}{\sqrt{2}}$$

$$\cos 45^\circ = \frac{1}{\sqrt{2}}$$

$$\tan 45^\circ = 1$$

$$\sin(A+B) = \sin A \cos B + \cos A \sin B$$

$$\sin(A-B) = \sin A \cos B - \cos A \sin B$$

$$\cos(A-B) = \cos A \cos B - \sin A \sin B$$

$$\cos(A-B) = \cos A \cos B + \sin A \sin B$$

$$\tan(A+B) = \frac{\tan A + \tan B}{1 - \tan A \tan B}$$

$$\tan(A-B) = \frac{\tan A - \tan B}{1 + \tan A \tan B}$$

$$\sin\theta = \frac{y}{r}(\text{opposite/hypotenuse}) = 1/\csc\theta$$

$$\cos\theta = \frac{x}{r}(\text{adjacent/hypotenuse}) = 1/\sec\theta$$

$$\tan\theta = \frac{y}{x}(\text{opposite/adjacent}) = 1/\cot\theta$$

$$\sin 30^\circ = \frac{1}{2} \quad \sin 60^\circ = \frac{\sqrt{3}}{2}$$

$$\cos 30^\circ = \frac{\sqrt{3}}{2} \quad \cos 60^\circ = \frac{1}{2}$$

$$\tan 30^\circ = \frac{1}{\sqrt{3}} \quad \tan 60^\circ = \sqrt{3}$$

Sine law

$$\frac{a}{\sin A} = \frac{b}{\sin B} = \frac{c}{\sin C}$$

Cosine law

$$a^2 = b^2 + c^2 - 2bc\cos A$$

$$b^2 = a^2 + c^2 - 2ac\cos B$$

$$c^2 = a^2 + b^2 - 2ab\cos C$$

$$\mathrm{q} = 1\ \mathrm{rad}$$

$$2\mathrm{p}\ \mathrm{rad} = 360^\circ$$

Algebra

Expanding

$$\mathrm{a}(b + c) = \mathrm{a}b + \mathrm{a}c$$

$$(\mathrm{a} + b)^2 = \mathrm{a}^2 + 2\mathrm{a}b + b^2$$

$$(\mathrm{a} - b)^2 = \mathrm{a}^2 - 2\mathrm{a}b + b^2$$

$$(\mathrm{a} + b)(c + d) = \mathrm{a}c + \mathrm{a}d + bc + bd$$

$$(\mathrm{a} + b)^3 = \mathrm{a}^3 + 3\mathrm{a}^2b + 3\mathrm{a}\mathrm{b}^2 + b^3$$

$$(a - b)^3 = \mathrm{a}^3 - 3\mathrm{a}^2b + 3\mathrm{a}\mathrm{b}^2 - b^3$$

Factoring

$$\mathrm{a}^2 - b^2 = (\mathrm{a} + b)(\mathrm{a} - b)$$

$$\mathrm{a}^2 + 2\mathrm{a}\mathrm{b} + b^2 = (\mathrm{a} + b)^2$$

$$\mathrm{a}^3 + \mathrm{b}^3 = (\mathrm{a} + b)(\mathrm{a}^2 - \mathrm{a}b + b^2)$$

$$\mathrm{a}^3b - \mathrm{a}b = \mathrm{a}b(\mathrm{a} + 1)(\mathrm{a} - 1)$$

$$\mathrm{a}^2 - 2\mathrm{a}b + b^2 = (\mathrm{a} - b)^2$$

$$\mathrm{a}^3 - b^3 = (\mathrm{a} - b)(\mathrm{a}^2 + \mathrm{a}b + b^2)$$

Roots of quadratic
The solution for a quadratic equation $ax^2 + bx + c = 0$

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Law of exponents

$$a^r \cdot a^s = a^{r+s}$$

$$\frac{a^p a^q}{a^r} = a^{p+q-r}$$

$$\frac{a^r}{a^s} = a^{r-s}$$

$$(a^r)^s = a^{rs}$$

$$(ab)^r = a^r b^r$$

$$\left(\frac{a}{b}\right)^r = \frac{a^r}{b^r} \ (b \neq 0)$$

$$a^0 = 1 \ (a \neq 0)$$

$$a^{-r} = \frac{1}{a^r} \ (a \neq 0)$$

$$a^{\frac{r}{s}} = \sqrt[s]{a^r} \quad a^{\frac{1}{2}} = \sqrt{a} \quad a^{\frac{1}{3}} = \sqrt[3]{a}$$

Logarithms

Example

$$\mathrm{Log}(xy) = \mathrm{Log}\,x + \mathrm{Log}\,y \quad \mathrm{Log}\left(\frac{x}{y}\right) = \mathrm{Log}\,x - \mathrm{Log}\,y$$

$$\mathrm{Log}\,x^r = r\,\mathrm{Log}\,x$$

$$\mathrm{Log}\,x = n \leftrightarrow x = 10^n \text{ (Common log)} \quad p \simeq 3.14159265$$

$$\log_a x = n \ \leftrightarrow \ x = a^n \text{ (Log to the base a)} \quad e \simeq 2.71828183$$

$$\mathrm{Ln}\,x = n \ \leftrightarrow \ x = e^n \text{ (Natural log)}$$

Six simple machines for materials handling

Material handling design and implementation constitute one of the basic functions in the practice of industrial engineering. Calculations related to the six simple machines are useful for assessing mechanical advantage for material handing purposes. The mechanical advantage is the ratio of the force of resistance to the force of effort:

$$MA = \frac{F_R}{F_E}$$

where

MA is the mechanical advantage
F_R is the force of resistance (N)
F_E is the force of effort (N)

Machine 1: The lever

A lever consists of a rigid bar that is free to turn on a pivot, which is called a fulcrum

The law of simple machines as applied to levers is

$$F_R \cdot L_R = F_E \cdot L_E$$

Machine 2: Wheel and axle

A wheel and axle consist of a large wheel attached to an axle so that both turn together:

$$F_R \cdot r_R \cdot = F_E \cdot r_E$$

where

F_R is the force of resistance (N)
F_E is the force of effort (N)
r_R is the radius of resistance wheel (m)
r_E is the radius of effort wheel (m)

The mechanical advantage is

$$MA_{\text{wheel and axle}} = \frac{r_E}{r_R}$$

Machine 3: The pulley

If a pulley is fastened to a fixed object, it is called a fixed pulley. If the pulley is fastened to the resistance to be moved, it is called a moveable pulley. When one continuous cord is used, the ratio reduces according to the number of strands holding the resistance in the pulley system.

The effort force equals the tension in each supporting stand. The mechanical advantage of the pulley is given by formula

$$MA_{\text{pulley}} = \frac{F_R}{F_E} = \frac{nT}{T} = n$$

where

T is the tension in each supporting strand
N is the number of strands holding the resistance
F_R is the force of resistance (N)
F_E is the force of effort (N)

Machine 4: The inclined plane

An inclined plane is a surface set at an angle from the horizontal and used to raise objects that are too heavy to lift vertically.

The mechanical advantage of an inclined plane is

$$MA_{\text{inclined plane}} = \frac{F_R}{F_E} = \frac{1}{h}$$

where

F_R is the force of resistance (N)
R_E is the force of effort (N)
1 is the length of plane (m)
h is the height of plane (m)

Machine 5: The wedge

The wedge is a modification of the inclined plane. The mechanical advantage of a wedge can be found by dividing the length of either slope by the thickness of the longer end.

As with the inclined plane, the mechanical advantage gained by using a wedge requires a corresponding increase in distance.

The mechanical advantage is

$$MA = \frac{s}{T}$$

where

MA is the mechanical advantage
s is the length of either slope (m)
T is the thickness of the longer end (m)

Machine 6: The screw
A screw is an inclined plane wrapped around a circle. From the law of machines,

$$F_R \cdot h = F_E \cdot U_E$$

However, for advancing a screw with a screwdriver, the mechanical advantage is

$$MA_{\text{screw}} = \frac{F_R}{F_E} = \frac{U_E}{h}$$

where
F_R is the force of resistance (N)
F_E is the effort force (N)
h is the pitch of screw
U_E is the circumference of the handle of the screw

Mechanics: Kinematics

Scalars and vectors
The mathematical quantities that are used to describe the motion of objects can be divided into two categories: scalars and vectors.

a. Scalars: Scalars are quantities that can be fully described by a magnitude alone.
b. Vectors: Vectors are quantities that can be fully described by both a magnitude and direction.

Distance and displacement

a. Distance: Distance is a scalar quantity that refers to how far an object has gone during its motions.
b. Displacement: Displacement is the change in position of the object. It is a vector that includes the magnitude as a distance, such as 5 miles, and a direction, such as north.

Acceleration
Acceleration is the change in velocity per unit of time. Acceleration is a vector quality.

Speed and velocity

a. Speed: The distance traveled per unit of time is called the speed, for example, 35 mph. Speed is a scalar quantity.
b. Velocity: The quantity that combines both the speed of an object and its direction of motion is called velocity. Velocity is a vector quantity.

Frequency
Frequency is the number of complete vibrations per unit time in simple harmonic or sinusoidal motion.

Period
Period is the time required for one full cycle. It is the reciprocal of the frequency.

Angular displacement
Angular displacement is the rotational angle through which any point on a rotating body moves.

Angular velocity
Angular velocity is the ratio of angular displacement to time.

Angular acceleration
Angular acceleration is the ratio of angular velocity with respect to time.

Rotational speed
Rotational speed is the number of revolutions (a revolution is one complete rotation of a body) per unit of time.

Uniform linear motion
A path is a straight time. The total distance traveled corresponds with the rectangular area in the diagram v – t.

a. Distance:

$$S = Vt$$

b. Speed:

$$V = \frac{S}{t}$$

where
s is the distance (m)
v is the speed (m/s)
t is the time (s)

Uniform accelerated linear motion

1. If $\nu_0 > 0$; a > 0, then
   a. Distance:

$$s = n_0 t + \frac{at^2}{2}$$

b. Speed:

$$\mathrm{n} = \mathrm{n}_0 + \mathrm{a}t$$

where
s is the distance (m)
v is the speed (m/s)
t is the time (s)
ν_0 is the initial speed (m/s)
a is the acceleration (m/s^2)

2. If $\nu_0 = 0$; a > 0, then
a. Distance:

$$s = \frac{\mathrm{a}t^2}{2}$$

The shaded areas in diagram v − t represent the distance s traveled during the time period t.
b. Speed:

$$\mathrm{n} = \mathrm{a} \cdot t$$

where
s is the distance (m)
v is the speed (m/s)
ν_0 is the initial speed (m/s)
a is the acceleration (m/s^2)

Rotational motion
Rotational motion occurs when the body itself is spinning. The path is a circle about the axis.

1. Distance:

$$s = \mathrm{I}\mathrm{j}$$

2. Velocity:

$$\mathrm{n} = \mathrm{I}\mathrm{w}$$

3. Tangential acceleration:

$$\mathrm{a}_t = r \cdot \mathrm{a}$$

4. Centripetal acceleration:

$$\mathrm{a}_n = \mathrm{w}^2 r = \frac{\mathrm{n}^2}{r}$$

where
- $\hat{\jmath}$ is the angle determined by s and r (rad)
- ω is the angular velocity (s^{-1})
- α is the angular acceleration ($1/s^2$)
- a_t is the tangential acceleration ($1/s^2$)
- a_n is the centripetal acceleration ($1/s^2$)

Distance s, velocity v, and tangential acceleration a_t are proportional to radius r.

Uniform rotation and a fixed axis

ω_0 = constant; $\alpha = 0$,

a. Angle of rotations:

$$\jmath = \omega \cdot t$$

b. Angular velocity:

$$\omega = \frac{\jmath}{t}$$

where
- φ is the angle of rotation (rad)
- ω is the angular velocity (s^{-1})
- α is the angular acceleration ($1/s^2$)
- ω_0 is the initial angular speed (s^{-1})

The shade area in the diagram $\omega - t$ represents the angle of rotation $\varphi = 2\pi n$ covered during time period t.

Uniform accelerated rotation about a fixed axis

1. If $\omega_0 > 0$; $\alpha > 0$, then
   a. Angle of rotation:

$$\jmath = \frac{1}{2}(\omega_0 + \omega) = \omega_0 t + \frac{1}{2}\alpha t^2$$

   b. Angular velocity:

$$\omega = \omega_0 + \alpha t = \sqrt{\omega_0^2 + 2\alpha\jmath}$$

$$\omega_0 = \omega - \alpha t = \sqrt{\omega^2 - 2\alpha\jmath}$$

c. Angular acceleration:

$$\alpha = \frac{\omega - \omega_0}{t} = \frac{\omega^2 - \omega_0^2}{2\theta}$$

d. Time:

$$t = \frac{\omega - \omega_0}{\alpha} = \frac{2\theta}{\omega_0 - \omega}$$

2. If $\omega_0 = 0$; a = constant, then
   a. Angle of rotations:

$$\theta = \frac{\omega \cdot t}{2} = \frac{\alpha \cdot t}{2} = \frac{\omega^2}{2\alpha}$$

   b. Angular velocity:

$$\omega = \sqrt{2\alpha\theta} = \frac{2\theta}{t} = \alpha \cdot t; \quad \omega_0 = 0$$

   c. Angular acceleration:

$$\alpha = \frac{\omega}{t} = \frac{2\theta}{t^2} = \frac{\omega^2}{2\theta}$$

   d. Time:

$$t = \sqrt{\frac{2\theta}{\alpha}} = \frac{\omega}{\alpha} = \frac{2\theta}{\omega}$$

Simple harmonic motion

Simple harmonic motion occurs when an object moves repeatedly over the same path in equal time intervals.

The maximum deflection from the position of rest is called "amplitude."

A mass on a spring is an example of an object in simple harmonic motion. The motion is sinusoidal in time and demonstrates a single frequency.

a. Displacement:

$$s = A\sin(\omega \cdot t + \theta_0)$$

b. Velocity:

$$\nu = A\omega\cos(\omega \cdot t + \theta_0)$$

c. Angular acceleration:

$$a = -A\omega^2 \sin(\omega \cdot t + \varphi_0)$$

where
s is the displacement
A is the amplitude
φ_0 is the angular position at time t = 0
φ is the angular position at time t
T is the period

Pendulum
A pendulum consists of an object suspended so that it swings freely back and forth about a pivot.

a. Period:

$$T = 2\pi\sqrt{\frac{I}{g}}$$

where
T is the period (s)
I is the length of pendulum (m)
g = 9.81 (m/s^2) or 32.2 (ft/s^2)

Free fall
A free-falling object is an object that is falling due to the sole influence of gravity.

a. Initial speed:

$$v_0 = 0$$

b. Distance:

$$h = -\frac{gt^2}{2} = -\frac{vt}{2} = -\frac{v^2}{2g}$$

c. Speed:

$$v = +gt = -\frac{2h}{t} = \sqrt{-2gh}$$

d. Time:

$$t = +\frac{v}{g} = -\frac{2h}{v} = \sqrt{-\frac{2h}{g}}$$

Vertical project

a. Initial speed:

$$n_0 > 0, \text{(upwards)}; \quad n_0 < 0, \text{(downwards)}$$

b. Distance:

$$h = n_0 t - \frac{gt^2}{2} = (n_0 + n)\frac{t}{2}; \quad h_{\max} = \frac{{n_0}^2}{2g}$$

c. Time:

$$t = \frac{n_0 - n}{g} = \frac{2h}{n_0 + n}; \quad t_{h\max} = \frac{n_0}{g}$$

where
v is the velocity (m/s)
h is the distance (m)
g is the acceleration due to gravity (m/s^2)

Angled projections

$$\text{Upwards } (a < 0); \quad \text{downwards } (a < 0)$$

a. Distance:

$$s = n_0 \cdot t \cos a$$

b. Altitude:

$$h = n_0 t \sin a - \frac{g \cdot t^2}{2} = s \tan a - \frac{g \cdot s^2}{2{n_0}^2 \cos a}$$

$$h_{\max} = \frac{{n_0}^2 \sin^2 a}{2g}$$

c. Velocity:

$$n = \sqrt{{n_0}^2 - 2gh} = \sqrt{{n_0}^2 + g^2 t^2 - 2g n_0 t \sin a} \tag{11.1}$$

d. Time:

$$t_{h\max} = \frac{n_0 \sin a}{g}; \quad t_{s1} = \frac{2n_0 \sin a}{g}$$

Horizontal projection: ($\alpha = 0$)

a. Distance:

$$s = \mathrm{n}_0 t = \mathrm{n}_0 \sqrt{\frac{2h}{g}}$$

b. Altitude:

$$h = -\frac{gt^2}{2}$$

c. Trajectory velocity:

$$\mathrm{n} = \sqrt{\mathrm{n}_0^2 + g^2 t^2}$$

where
ν_0 is the initial velocity (m/s)
v is the trajectory velocity (m/s)
s is the distance (m)
h is the height (m)

Sliding motion on an inclined plane

1. If excluding friction ($\mu = 0$), then
   a. Velocity:

$$\mathrm{n} = \mathrm{a}t = \frac{2s}{t} = \sqrt{2\mathrm{a}s}$$

   b. Distance:

$$s = \frac{at^2}{2} = \frac{\mathrm{n}t}{2} = \frac{\mathrm{n}^2}{2\mathrm{a}}$$

   c. Acceleration:

$$\mathrm{a} = g \sin \mathrm{a}$$

2. If including friction ($\mu > 0$), then
   a. Velocity:

$$\mathrm{n} = \mathrm{a}t = \frac{2s}{t} = \sqrt{2\mathrm{a}s}$$

   b. Distance:

$$s = \frac{\mathrm{a}t^2}{2} = \frac{\mathrm{n}t}{2} = \frac{\mathrm{n}^2}{2\mathrm{a}}$$

c. Accelerations:

$$s = \frac{at^2}{2} = \frac{nt}{2} = \frac{n^2}{2a}$$

where

μ is the coefficient of sliding friction
g is the acceleration due to gravity
$g = 9.81$ (m/s^2)
ν_0 is the initial velocity (m/s)
ν is the trajectory velocity (m/s)
s is the distance (m)
a is the acceleration (m/s^2)
α is the inclined angle

Rolling motion on an inclined plane

1. If excluding friction (f = 0), then
   a. Velocity:

$$n = at = \frac{2s}{t} = \sqrt{2as}$$

   b. Acceleration:

$$a = \frac{gr^2}{I^2 + k^2} \sin a$$

   c. Distance:

$$s = \frac{at^2}{2} = \frac{nt}{2} = \frac{n^2}{2a}$$

   d. Tilting angle:

$$\tan a = m_0 \frac{r^2 + k^2}{k^2}$$

2. If including friction (f > 0), then
   a. Distance:

$$s = \frac{at^2}{2} = \frac{nt}{2} = \frac{n^2}{2a}$$

   b. Velocity:

$$n = at = \frac{2s}{t} = \sqrt{2as}$$

c. Accelerations:

$$a = gr^2 \frac{\sin a - (f/r)\cos a}{I^2 + k^2}$$

d. Tilting angle:

$$\tan a_{min} = \frac{f}{r}; \quad \tan a_{max} = m_0 \frac{r^2 + k^2 - fr}{k^2}$$

The value of k can be the calculated by the following formulas:

Ball	Solid cylinder	Pipe with low wall thickness
$k^2 = \frac{2r^2}{5}$	$k^2 = \frac{r^2}{2}$	$k^2 = \frac{r_i^2 + r_0^2}{2} \approx r^2$

where

s is the distance (m)
ν is the velocity (m/s)
a is the acceleration (m/s^2)
α is the tilting angle (°)
f is the lever arm of rolling resistance (m)
k is the radius of gyration (m)
μ_0 is the coefficient of static friction
g is the acceleration due to gravity (m/s^2)

Mechanics: Dynamics

Newton's first law of motion

Newton's first law is called the Law of Inertia. An object that is in motion continues in motion with the same velocity at constant speed and in a straight line, and an object at rest continues at rest unless an unbalanced (outside) force acts upon it.

Newton's second law

The second law of motion is called the Law of Accelerations. The total force acting on an object equals the mass of the object times its acceleration.

In equation form, this law is

$$F = ma$$

where

F is the total force (N)
m is the mass (kg)
a is the acceleration (m/s^2)

Newton's third law

The third law of motion, called the Law of Action and Reaction, can be stated as follows:

For every force applied by object A to object B (action), there is a force exerted by object B on object A (the reaction) which has the same magnitude but is opposite in direction.

In equation form this law is

$$F_B = -F_A$$

where

F_B is the force of action (N)
F_A is the force of reaction (N)

Momentum of force

The momentum can be defined as mass in motion. Momentum is a vector quantity; in other words, the direction is important:

$$p = m\mathrm{n}$$

Impulse of force

The impulse of a force is equal to the change in momentum that the force causes in an object:

$$I = Ft$$

where

p is the momentum (N s)
m is the mass of object (kg)
v is the velocity of object (m/s)
I is the impulse of force (N s)
F is the force (N)
t is the time (s)

Law of conservation of momentum

One of the most powerful laws in physics is the law of momentum conservation, which can be stated as follows:

In the absence of external forces, the total momentum of the system is constant.

If two objects of mass m_1 and mass m_2, having velocity ν_1 and ν_2, collide and then separate with velocity ν_1' and ν_2', the equation for the conservation of momentum is

$$m_1\mathrm{n}_1 + m_2\mathrm{n}_2 = m_1\mathrm{n}_1 + m_2\mathrm{n}_2$$

Friction

Friction is a force that always acts parallel to the surface in contact and opposite to the direction of motion. Starting friction is greater than moving friction. Friction increases as the force between the surfaces increases.

The characteristics of friction can be described by the following equation:

$$F_f = \mu F_n$$

where

F_f is the frictional force (N)
F_n is the normal force (N)
μ is the coefficient of friction ($\mu = \tan \alpha$)

General law of gravity

Gravity is a force that attracts bodies of matter toward each other. Gravity is the attraction between any two objects that have mass.

The general formula for gravity is

$$F = \Gamma \frac{m_A m_B}{r^2}$$

where

m_A, m_B are the mass of objects A and B (kg)
F is the magnitude of attractive force between objects A and B (N)
r is the distance between object A and B (m)
Γ is the gravitational constant (N m^2/kg^2)
$\Gamma = 6.67 \times 10^{-11}$ N m^2/kg^2

Gravitational force

The force of gravity is given by the equation

$$F_G = g \frac{R_e^2 m}{(R_e + h)^2}$$

On the earth surface, h = 0; so

$$F_G = mg$$

where

F_G is the force of gravity (N)
R_e is the radius of the Earth ($R_e = 6.37 \times 10^6$ m)
m is the mass (kg)
g is the acceleration due to gravity (m/s^2)
$g = 9.81$ (m/s^2) or $g = 32.2$ (ft/s^2)

The acceleration of a falling body is independent of the mass of the object.

The weight F_w on an object is actually the force of gravity on that object:

$$F_w = mg$$

Centrifugal force

Centrifugal force is the apparent force drawing a rotating body away from the center of rotation, and it is caused by the inertia of the body. Centrifugal force can be calculated by the formula:

$$F_c = \frac{mn^2}{r} = mw^2 r$$

Centripetal force

Centripetal force is defined as the force acting on a body in curvilinear motion that is directed toward the center of curvature or axis of rotation.

Centripetal force is equal in magnitude to centrifugal force but in the opposite direction.

$$F_{cp} = -F_c = \frac{mn^2}{r}$$

where

F_c is the centrifugal force (N)
F_{cp} is the centripetal force (N)
m is the mass of the body (kg)
ν is the velocity of the body (m/s)
r is the radius of curvature of the path of the body (m)
ω is the angular velocity (s^{-1})

Torque

Torque is the ability of a force to cause a body to rotate about a particular axis. Torque can have either a clockwise or a counterclockwise direction. To distinguish between the two possible directions of rotation, we adopt the convention that a counterclockwise torque is positive and that a clockwise torque is negative. One way to quantify a torque is

$$T = F \cdot I$$

where

T is the torque (N m or 1b ft)
F is the applied force (N or 1b)
1 is the length of torque arm (m or ft)

Work
Work is the product of a force in the direction of the motion and the displacement.

a. Work done by a constant force:

$$W = F_s \cdot s = F \cdot s \cdot \cos\alpha$$

where
W is the work (N m = J)
Fs is the component of force along the direction of movement (N)
s is the distance the system is displaced (m)

b. Work done by a variable force: If the force is not constant along the path of the object, we need to calculate the force over very tiny intervals and then add them up. This is exactly what the integration over differential small intervals of a line can accomplish:

$$W = \int_{si}^{sf} F_s(s) \cdot ds = \int_{si}^{sf} F(s)\cos\alpha \cdot ds$$

where
$F_s(s)$ is the component of the force function along the direction of movement (N)
$F(s)$ is the function of the magnitude of the force vector along the displacement curve (N)
s_i is the initial location of the body (m)
s_f is the final location of the body (m)
α is the angle between the displacement and the force

Energy
Energy is defined as the ability to do work. The quantitative relationship between work and mechanical energy is expressed by the equation:

$$TME_i + W_{ext} = TME_f$$

where
TME_i is the initial amount of total mechanical energy (J)
W_{ext} is the work done by external forces (J)
TME_f is the final amount of total mechanical energy (J)

There are two kinds of mechanical energy: kinetic and potential.

a. Kinetic energy: Kinetic energy is the energy of motion. The following equation is used to represent the kinetic energy of an object:

$$E_k = \frac{1}{2} m\text{n}^2$$

where
m is the mass of moving object (kg)
ν is the velocity of moving object (m/s)

b. Potential energy: Potential energy is the stored energy of a body and is due to its internal characteristics or its position. Gravitational potential energy is defined by the formula

$$E_{pg} = m \cdot g \cdot h$$

where
E_{pg} is the gravitational potential energy (J)
m is the mass of object (kg)
h is the height above reference level (m)
g is the acceleration due to gravity (m/s^2)

Conservation of energy
In any isolated system, energy can be transformed from one kind to another, but the total amount of energy is constant (conserved):

$$E = E_k + E_p + E_e + \cdots = \text{constant}$$

Conservation of mechanical energy is given by

$$E_k + E_p = \text{constant}$$

Power
Power is the rate at which work is done, or the rate at which energy is transformed from one form to another. Mathematically, it is computed using the following equation:

$$P = \frac{W}{t}$$

where
P is the power (W)
W is the work (J)
t is the time (s)

The standard metric unit of power is the watt (W). As is implied by the equation for power, a unit of power is equivalent to a unit of work divided by a unit of time. Thus, a watt is equivalent to Joule/second (J/s). Since the expression for work is

$$W = F \cdot s,$$

the expression for power can be rewritten as

$$P = F \cdot \mathrm{n}$$

where

s is the displacement (m)
ν is the speed (m/s)

Industrial Engineering

INDUSTRIAL CONTROL SYSTEMS

Mathematical and Statistical Models and Techniques

Adedeji B. Badiru • Oye Ibidapo-Obe • Babatunde J. Ayeni

Issues such as logistics, the coordination of different teams, and automatic control of machinery become more difficult when dealing with large, complex projects. Yet all these activities have common elements and can be represented by mathematics. Linking theory to practice, *Industrial Control Systems: Mathematical and Statistical Models and Techniques* presents the mathematical foundation for building and implementing industrial control systems.

Examining system fundamentals and advanced topics, the book contains mathematically rigorous models and techniques generally applicable to control systems with specific orientation toward industrial systems. It provides a generic project scheduling tool that incorporates resource characteristics and performance interdependencies among different resource groups and then proceeds to map the most adequate resource units to each newly scheduled project activity. Clearly detailing concepts and step-by-step procedures, the book covers:

- Scientific uncertainty, fuzziness, and application of stochastic–fuzzy models in urban transit, water resources, energy planning, and education
- The methodical development of the optimization problem in relation to applications in applied sciences, from classical techniques through stochastic approaches to contemporary and recent methods of intelligent metaheuristics
- The use of statistical control techniques for quality improvement in the manufacturing environment
- Additional statistical tools for quality and process improvement, including factorial designs, response surface methodology, central composite designs, and response surface optimization
- How to use a Bayesian technique for parameter estimation

An amalgamation of theoretical developments, applied formulations, implementation processes, and statistical control, the book includes examples that demonstrate how to use the statistical designs to develop feedback controllers and minimum variance controller designs for industrial applications. It matches mathematics with practical applications, giving you the tools to achieve system control goals.

CRC Press
Taylor & Francis Group
an informa business

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

75586

ISBN: 978-1-4200-7558-8
90000
9 781420 075588

www.crcpress.com