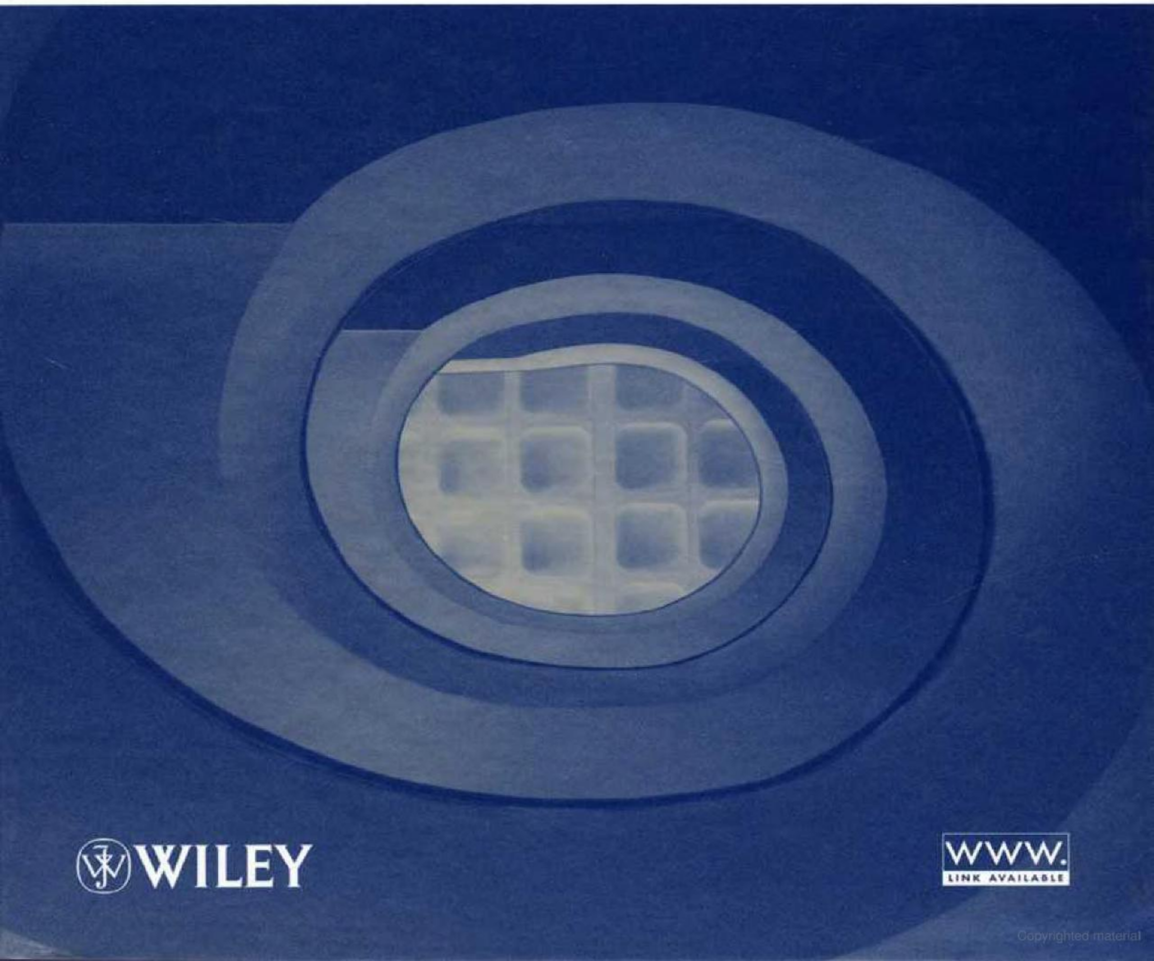
SERVICE SCIENCE

Mark S. Daskin



WILEY

SERVICE SCIENCE

Mark S. Daskin
Department of Industrial and Operations Engineering
University of Michigan
Ann Arbor, MI

WILEY
A JOHN WILEY & SONS, INC., PUBLICATION

All referenced files may be found at http://umich.edu/~msdaskin/servicescience/

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Catologing-in-Publication Data

Daskin, Mark S., 1952–
Service science / Mark S. Daskin.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-52588-3 (cloth)
1. Operations research. 2. Service industries–Management. 3. Service industries–Planning.
I. Title.
T57.6.D3733 2010
658.4′034–dc22

2010010792

Printed in Singapore

10 9 8 7 6 5 4 3 2 1

Appendixes and a full list of References are posted online.

All referenced files may be found at http://umich.edu/~msdaskin/servicescience/

LIST OF FIGURES

LIST OF TABLES

PREFACE

We depend on services and service providers for many of our day-to-day activities, from the news we wake up to on our clock radio to the e-mail we check before breakfast, from the dry cleaner we stop at on our way to work to the express mail delivery service that dropped off our latest holiday gifts, from the cute corner bistro we patronize daily for lunch to the movie theater at which we unwind on the weekends. Many services are implicit in our lives including banking, investments, insurance, police and fire services, and (hopefully) our heath care providers. Without service providers, our lives would simply not be what they are today.

The service sector in the United States is rapidly growing as a percentage of the economy. Sixty years ago, the service sector represented only 20 percent of the gross domestic product; today, services account for over 40 percent of the gross domestic product. The percentage of the GDP accounted for by the production of non-durable goods (e.g., food, clothing, and energy) has seen a commensurate decline. In 1960, slightly less than one out of every two people employed in the United States was employed in the service sector; today, more than two out of every three employees work in some form of service industry.

Given our daily dependence on services and the enormous role that the service sector plays in the economy—not only of the United States, but also of every developed country around the world—it is important that we understand the operation of this sector and that services be provided in an efficient and effective manner. Much of the current debate in the United States over health care reform—and health care is part of the service sector—focuses on ways of increasing access, reducing inequities, and containing costs.

This book will provide students with the tools and background needed to analyze and improve the provision of services in our economy.

Following a brief introduction to the service sector, Part I of the text deals with the methodological background needed to analyze service systems. Two core methodologies are introduced: optimization and queueing modeling. For students who have not had a course on one or both of these topics, these chapters provide the background necessary to master the material in the remainder of the text. In addition, the online Appendix B summarizes probability theory at a level that will allow students who have limited backgrounds to understand the chapter on queueing models.

While many students may have a background in optimization and queueing, topics covered near the end of each chapter are typically not included in introductory courses. Section 2.8 deals with multi-objective optimization. This is critical in the analysis of many services because service providers must often balance conflicting objectives. A local government operating an emergency medical service department (ambulances) must carefully balance the need for rapid response against the demands for fiscal responsibility. Similarly, a cell phone company must balance the demands for expanded and enhanced service area coverage against the need to show a profit at the end of the year. Section 2.9 addresses a number of common mistakes that students (and professionals) make in formulating optimization problems. Section 3.5 summarizes key queueing results that extend beyond those included in many introductory stochastic processes books. Séction 3.6 outlines how to solve queueing models numerically using Excel and section 3.7 discusses queueing problems in which the input or operating conditions change over time. Such problems are critical in the analysis of services. For example, there is typically a three or four to one ratio between the peak and off-peak call rates for emergency medical service. Planning for the average daily arrival rate of calls would lead to serious delays during the peak and excess capacity during the off-peak periods. Many other services experience daily, monthly, or annual spikes in demand. Even students with good backgrounds in optimization and queueing might find these sections useful.

The remainder of the text is devoted to the application of optimization and queueing to the analysis and design of service systems. Chapter 4 deals with strategic decisions regarding the number and location of service facilities. Cell phone service providers must, for example, determine the number and location of their cell phone towers to provide cost-effective coverage to a service region. Fast food restaurants must also determine how many stores to have and where they should be to balance easy access against the possibility of self-cannibalizing the market. Many service providers partition the service region into districts that are then served by individual customer service agents. The chapter concludes with a discussion of districting problems.

Many authors argue that the inability to store services in inventory is a key differentiator between the service sector and the manufacturing sector. A car that is not sold today can be stored in inventory for sale tomorrow or next week. On the other hand, a surgeon who takes an afternoon off from work to watch his son star in a school play cannot place the missed operating room hours in inventory for use later in the week. While the service itself cannot be stored in inventory for future use, many service providers depend critically on the ready availability of inventory to assist in the provision of the services they deliver. The same surgeon relies on the availability of sterile instruments in the operating theater at the beginning of each procedure. The local shoe store must stock shoes in numerous styles and sizes for its customers. Thus, an understanding of inventory problems and decisions is critical for students of the service industries. Furthermore, in contracting for services, individuals and firms must often make commitments before the demand for the services is realized. For example, many of us can place pre-tax funds in special accounts to pay for qualified medical expenses. We must decide in the Fall of one year how much money to set aside during the following year before knowing what our medical expenses will be during the coming year. Any unused funds at the end of a year are lost. Such problems are known as newsvendor problems and are discussed at the end of Chapter 5 on inventory modeling.

At its core, many decisions in the provision of services boil down to resource allocation decisions. A college or university must allocate classroom space to courses. Colleges and universities must also allocate limited dormitory space to students and must also assign students to courses based on the students' preferences and requirements and the availability of space in the courses. Airlines must allocate gates.to aircraft. Chapter 6 deals with resource allocation decisions.

Chapters 7 and 8 address short-term and long-term workforce management decisions and problems. In the short-term, service providers must determine how many staff to employ during each period of the day. For example, a hospital must decide how many full-time nurses to hire during each shift. It must also determine how it will staff each unit in the event that the number of patients on the unit exceeds the expected number. Typically, nurses are asked to perform overtime duty or more expensive temporary nurses are hired to fill in for the permanent staff.

In the long-term, providers must determine how many employees to hire, to promote, to release, and to retrain. A consulting firm, for example, needs to determine how many college seniors to hire each year in each of the specialty areas of the firm. Some of the more senior analysts at the firm may be targeted for management training. When the firm's business base changes, the firm may need to either retrain some of its employees or release the less productive members of its staff to make room for newer, better-trained employees in the growth areas. Chapter 8 addresses these problems.

Chapter 9 extends the discussion of queueing models to three particular topics that arise in many service providers. Not all customers are equal. An elderly woman presenting in an emergency room in active cardiac arrest is likely to be served long before a six-year-old boy who fell of his bicycle and who may have a broken leg or twisted ankle. Frequent customers may be flagged for improved service in many industries. Thus, priority service systems comprise the first part of this chapter. Nearly every major company and government service provider operates a call center to provide service to its customers. When it comes to call centers, bigger really is better. The second section of this chapter addresses the design and operation of call centers. Finally, in many services, customers can be scheduled for service. A dermatologist can schedule most of her patients. Issues in customer scheduling are outlined in the final portion of the chapter.

Finally, many services entail the delivery or pickup of customers or goods. A local public school must provide bussing to its students to pick them up from their homes in the morning and to return them home at the end of the school day. Large white goods (refrigerators, freezers, dish washers, washing machines, and dryers) must be delivered to customer homes in a timely manner following the purchase of the items. The mail must be delivered daily and streets must be cleaned during the summer and plowed during the winter. Chapter 10 introduces vehicle routing problems and models as they arise in the delivery of services.

In addition to the mathematical derivation and formulation of the models outlined in the text, the book includes numerous sections summarizing how to implement the models using Microsoft® Office Excel®. These sections are highlighted in the text, just as this paragraph is highlighted. The example spreadsheets are available from the author's website. Equations or formulae in Excel are enclosed in single quotes such as 'IF(C1<0,1,0)'. This discussion and the spreadsheets should make the models accessible to a broader audience.

All referenced files may be found at http://umich.edu/~msdaskin/servicescience/

The course that I taught at Northwestern University, which operates on a 10-week quarter system, had prerequisites of (a) deterministic optimization, (b) probability, (c) statistics, and (d) stochastic modeling, including an introduction to queueing theory. Thus, students were largely well-prepared in terms of methodological backgrounds. Because of their background in optimization, after a quick summary of the first chapter, I was able to cover only sections 2.8 and 2.9 of Chapter 2. I usually did a one- or two-day review of queueing theory including a quick introduction to time-dependent problems covered in section 3.7. I typically would spend two weeks on location models (Chapter 4). My coverage of inventory theory focused on the newsvendor problem (section 5.6). Resource allocation problems (Chapter 6) were typically introduced during the review of optimization. I would often spend a week each on Chapters 7 and 8 on short-term and long-term workforce management. Topics from Chapter 9 on priority queueing systems, call centers, and customer scheduling typically rounded out the course. Routing and inventory were, with the exception of the newsvendor problem, not covered in the course as there was a separate supply chain management course as well as a production scheduling course that covered routing and inventory. Although students had to take only one of the three courses—supply chain management, production scheduling, or service operations management—many students took two or even all three of the courses. Excessive duplication of material was deemed inappropriate by those of us teaching these three courses.

In short, I encourage faculty and students using the text to pick and choose those topics that are of most interest to them. For students with a strong methodological background in optimization and stochastic modeling, Chapters 4 through 10 should generally stand on their own and can, to a large extent, be covered in any order and in a level of detail that suits the instructor and the class.

I hope you enjoy using the text as much as I enjoyed writing it and teaching the course, which was the genesis of the book.

Mark S. Daskin

ACKNOWLEDGMENTS

The genesis of this book was a course that I created for the undergraduates of the Department of Industrial Engineering and Management Sciences of Northwestern University. In the six quarters that I taught this course, dating back to the Fall of 2003, over 200 students took the course. I am indebted to them for their insightful questions and comments and for their enthusiasm for the course. I also owe a deep debt of gratitude to the six teaching assistants who helped me with this course over these years: Taylan Ilhan, Bilal Gokpinar, Michael Lim, Zigeng Yin, Yao Cheng and Yue Geng. Your help greatly improved both the course and this text. Many thanks.

I also want to thank the many other friends that I made in the IEMS department—students, faculty, and staff—for their friendship and encouragement during my years in the department. I particularly want to thank Barry Nelson, the chair of the department, who afforded me the time to finish writing this text during my last quarter at Northwestern. Without that time, this book would not have been finished on time, and maybe not at all.

From my initial meeting with her, Susanne Steitz-Filler, my editor at John Wiley, expressed strong enthusiasm for this project. She and Stephen Quigley, who filled in for her during her maternity leave, have been with me at every step. Many thanks to both of them.

This book is dedicated to my parents, Walter and Betty; to my two daughters, Tamar and Keren; and to my wife, Babette. My parents always said that if subsequent generations had not surpassed the generation before them, then we would all still be living in caves. I hope that this text enables and motivates the next generation of students of the service industries to surpass me and my generation of researchers. My children, Tamar and Keren, have always been a source of joy and light in my life. Finally, without Babette's support, encouragement, and friendship over the 35 years that we have known each other, I would not be where I am today and this book would never have been written. I love them all and I truly thank them for everything.

This page intentionally left blank

1

WHY STUDY SERVICES?

1.1 WHAT ARE SERVICES?

We use services every day of our lives. When we wake up to the voice of our favorite radio broadcaster on our bedside alarm clock, we are using the services of the radio station as well as those of the electric company providing our home with power. A quick shower entails the use of the water supplied by the water company. The bowl of cereal that we eat at breakfast was purchased from a retail grocery store, or perhaps delivered to our house by a food delivery service, either of which received the box of cereal using trucking services. The pants that we put on before leaving the house were cleaned at the local dry cleaners, yet another service we use implicitly before leaving the house.

After alighting from our bus (a public transportation service), we stop at the local coffee shop for a latte before walking to the office. Chances are we work for a service provider, such as a lawyer, doctor, accountant, or consultant. During the day at the office, we use numerous communication services, including e-mail, phones, faxes, the Internet, and instant messaging. To learn more about the new software our employer installed last week, we may employ the services of an online training program or a consultant at the software vendor's call center.

Service Science, by Mark S. Daskin

Lunch with friends at the nearby burger house demands the services of this establishment, including the waiters, chefs, busboys, and cashier. As we pay with our credit card, we utilize the financial services of our bank. The afternoon call to our dentist to schedule our next checkup brings us in touch with the vast array of medical services at our disposal. After a quiet dinner at home, we drive to a nearby theater to catch the latest movie, taking advantage of the myriad of entertainment services that surrounds us.

We take many services for granted. For example, in driving to the movie theater, we may have forgotten about the public road maintenance services that patched the potholes left by last winter's snows. We surely forgot about the snow plowing services that removed the snow. Unless we recently had our car repaired, we didn't consider the auto maintenance and repair services that keep our cars going. We took for granted the police services that kept our streets·and neighborhoods safe. Most of us are probably unaware of the traffic monitoring services provided by many large municipalities and counties to ensure the smooth flow of traffic on our highways and arterials. If the theater was one we frequent often, we probably did not need the GPS service in our car or on our cell phone. And throughout the day, we implicitly relied on our life, auto, health, homeowner's, and property insurance services.

In short, we rely on and utilize services for many of our daily needs.

So what are services and how can we characterize them? Table 1.1 lists some of the services that we frequently encounter. This should give the reader a sense of the broad spectrum of economic activity that is encompassed by the service sector. In further defining the service sector, it may be easier to define what it is not. Agriculture is not part of the service sector, though agriculture relies on transportation carriers (part of the service sector) to move seed, feed, and fertilizer to farms and to ship produce and animals to producers and markets. The construction industry, which builds roads, airports, manufacturing plants, office buildings, and much more, is not part of the service sector, though it too depends on many services. Manufacturers of durable goods such as automobiles, washers, driers, and refrigerators are not part of the service sector, though they too rely heavily on services for the goods they produce. Manufacturers depend on the financial industry for loans for inventory, raw materials and machinery; on accountants to help manage their finances; on the legal profession to ensure that they comply with government regulations; and consultants for a wide range of professional services. Some durable goods manufacturers are also in the service business. For example, General Motors seems to spend as much time advertising its OnStar emergency and vehicle diagnostic service as it does marketing the vehicles it produces. Often, government activities and employees are separated from the service sector, though many governmental activities are rightly considered services. For example, police, fire, and emergency medical personnel clearly provide important services to the communities in which they are located, as do teachers, many of whom are employed by their local public school districts.

Several attributes characterize most services. First, it is difficult to inventory services. Empty seats at a Sunday matinee of a musical cannot be sold for the

TABLE 1.1. Some of the services we commonly use

Financial	Savings	**Education**	Primary school
	Checking		Secondary school
	ATMs		Community college
	Credit card		Four-year college
	Mortgage services		University
	CDs		Online training
	Stocks		Private school
	Bonds		Parochial school
	Mutual funds		
Insurance	Life	**Entertainment**	Movies
	Health		Live theater
	Dental		Concerts
	Disability		Sports
	Automobile		Television
	Homeowner/property		Radio
	Long-term care		
Retail	Grocery store	**Transportation**	Parking
	Book store		Taxi
	Gasoline station		Bus
	Jewelry store		Train
	Department store		Airplane
	Specialty clothing		Truck
	Electronics warehouse		Rail freight
Health	Social worker	**Personal Services**	Hair stylist
	Psychiatrist		Spa
	Personal physician		Massage
	Specialist		Dry cleaning
	Dentist		Shoe repair
	Clinic		Accountant
	Hospital		Financial consultant
	Emergency room		Lawyer
Communication Services	Post office/U.S. mail	**Public Services**	Library
	Express mail		Police
	Land line telephone		Fire
	Cell phone		Emergency services
	E-mail		Garbage collection
	Text message		Roads
	Internet		Water
			Electricity
Dining	Fast food	**Other**	Call center
	Family dining		Lawn care
	School cafeteria		Snow removal
	Ethnic restaurant		Pet care
	Fine dining		

following Saturday night when demand may be high. Students who miss a class on optimization cannot sit in on a class on French literature and make up for the material on optimization that they missed. A dry cleaner who goes on vacation for two weeks cannot make up for the lost business during that time by working overtime because her customers are likely to have taken their clothing to another cleaner during that time. Airline seats that are flown empty on one flight cannot later be used for a flight that is overbooked. Nevertheless, services often depend critically on the careful management of inventories. Without an adequate inventory of sterile equipment and materials, surgeries cannot be performed. A retail establishment must maintain an adequate inventory of goods to be able to serve its customers. Airlines think of their seats as a perishable inventory that they manage using sophisticated revenue management techniques. Videotaping and recording of lectures, coupled with online material, may allow students to "inventory" a lecture for future reference.

Second, there is an intangible quality associated with services. You cannot touch the interaction between a psychiatrist and his or her patient, just as you cannot hold the education that a student receives in high school. Third, services typically produce some value for customers, or the service provides a solution to a customer need. The value may be in the form of improved health care, entertainment, or education. The value may also provide peace of mind (as in the case of insurance, police, fire, and emergency medical services) or convenience (as provided, for example, by dry cleaners). Fourth, services typically involve an interaction between a service provider and a consumer of the service. For example, doctors interact with patients in providing health care services. Teachers and professors interact with students in the educational process. Lawyers provide advice and counsel to their clients.

Finally, the need for services varies significantly with time. This is one of the key distinguishing features of services as opposed to manufacturing operations where the number of products produced each day is likely to remain roughly constant. An automobile assembly plant can produce approximately one vehicle a minute or about 1,000 cars per day during two eight-hour shifts. The production rate remains the same throughout the day. Emergency room arrivals, on the other hand, vary significantly over the course of the day. For example, in Austin, Texas, we found that the ratio of the peak hourly demand for ambulance services to the demand during the least busy hour exceeded four to one (Eaton et al., 1979). Traffic on most urban highways also exhibits clear daily patterns with low volumes in the late evening and early morning hours and heavy traffic during the morning and evening rush hours. The demand for air travel also exhibits weekly and annual peaks. The Sunday of Thanksgiving weekend is typically one of the busiest times of the year for most airlines. Retail establishments must often sell a large fraction of their goods during the Christmas holiday season if they are to have a successful year. Accountants experience a rush of business in the first three and a half months of the year, as clients try to meet the April 15 tax filing deadline. Some services, such as summer-stock theaters or winter ski resorts, operate only during certain times of the year. Other services, such as insurance sales and library usage, exhibit significantly lower degrees of temporal variation.

Planning for these peaks and valleys in demand is a critical task for a service manager. If possible, one would like to smooth the demand. In some cases, pricing can be used to shift demand from peak to off-peak periods. For some services, this is not possible. The demand for emergency medical services, for example, is not likely to be influenced significantly by any pricing scheme that a hospital or municipality might consider putting in place. Similarly, Thanksgiving holiday travel may decrease overall with significant price increases, but people are not likely to shift that travel to other times of the year since Thanksgiving is when most employees and most students have two (or more) days off from work or classes. Supply can also be adjusted in many cases. Additional flights can be scheduled for busy periods. Larger aircraft can be assigned to those routes that are known to experience the most significant percentage increases in demand during a busy period. Additional temporary accounting staff are hired during the tax filing period. More doctors and nurses can be scheduled to work in an emergency room during the busier hours of the day and week. Reversible traffic lanes are used in many urban areas to accommodate the daily rush hour traffic.

Many services have evolved significantly since the turn of the century. The Internet has spurred much of this evolution. For example, in booking an airline reservation, one used to have to contact the airline or a travel agent directly. Today, many people book reservations online with little to no human interaction. This function of travel agents has been largely replaced by the Internet. Toward the end of the twentieth century, it was nearly impossible to research the housing stock in a community without the aid of a realtor. Today, most sophisticated home buyers conduct a significant amount of research regarding communities and home prices before visiting their first house. Smart and sophisticated realtors encourage and facilitate this sort of research.

To illustrate the evolution of services, consider what has happened to photography during the twenty-first century. In the pre-Internet, pre-digital camera era, photographs were taken on film. The film was developed at a local photo shop and converted into pictures and negatives or slides. Alternatively, some film manufacturers provided customers with mailers to return the film for processing by the company. Today, relatively few people rely on film; most have converted to digital photography. With this conversion has come an explosion in the number of options people have for processing photographs. It is worth noting, that most digital photographs are probably never printed as they can be viewed on the camera or on a computer. Only the winners need be printed and processed. In addition to the local photo shop as a means of printing pictures, consumers can now print their photographs on an inexpensive color printer hooked to their home computer. Many printers now allow users to bypass the computer completely with the camera or memory card linked directly to the printer. Photographs can be uploaded to an online photo processor. Pictures can then be ordered for mail delivery or pickup at a local store. Many drug stores and other establishments have stand-alone kiosks to allow photo enthusiasts to print pictures, again with little or no interaction with the store owners. Finally, digital photos are finding their way into far more than pictures. Online companies allow customers to order mugs, photo books, magnets, calendars, clothing and much more.

Many other services are also being transformed by the Internet. In his seminal book entitled, *The World is Flat,* Thomas Friedman (2005) points out that many accountants are sending simple tax returns to India and other low-labor-cost countries for processing. Accountants are now finding more sophisticated ways of meeting their clients' needs. Medical images taken at a hospital in Boston, Massachusetts, may be read by a radiologist in Boston; they could just as easily, however, be read by a radiologist in Beer Sheva, Israel or Bangalore, India.

1.2 SERVICES AS A PERCENT OF THE ECONOMY

In measuring the economy, the U.S. Bureau of Economic Analysis divides the gross domestic product–-the value of all goods and services produced in the country—into these broad categories:

a. **Durable goods,** including such items as motor vehicles and parts, furniture and household equipment
b. **Non-durable goods,** including food, clothing, shoes, gasoline, oil, fuel, and other energy supplies
c. **Exports,** which is the value of all exported goods and services
d. **Imports,** which is the value of all imported goods and services and which contributes *negatively* to the gross domestic product
e. **Gross private domestic investment,** which includes the value of all inventory and machinery in the country
f. **Government expenditures,** including federal defense and non-defense related expenditures as well as all state and local government expenditures
g. **Services** including expenditures on housing, transportation, household operating expenses (electricity, gas and other operating costs), medical care, recreation, and so on.

Figure 1.1 plots the value of these components of the gross domestic product beginning with the first quarter of 1947 and running through the first quarter of 2008 in current dollars. The explosive growth of the value of the service sector of the economy is readily apparent. In 2000 dollars, the GDP increased from $1.57 trillion to $11.7 trillion ($0.237 to $14.2 trillion in current dollars) during

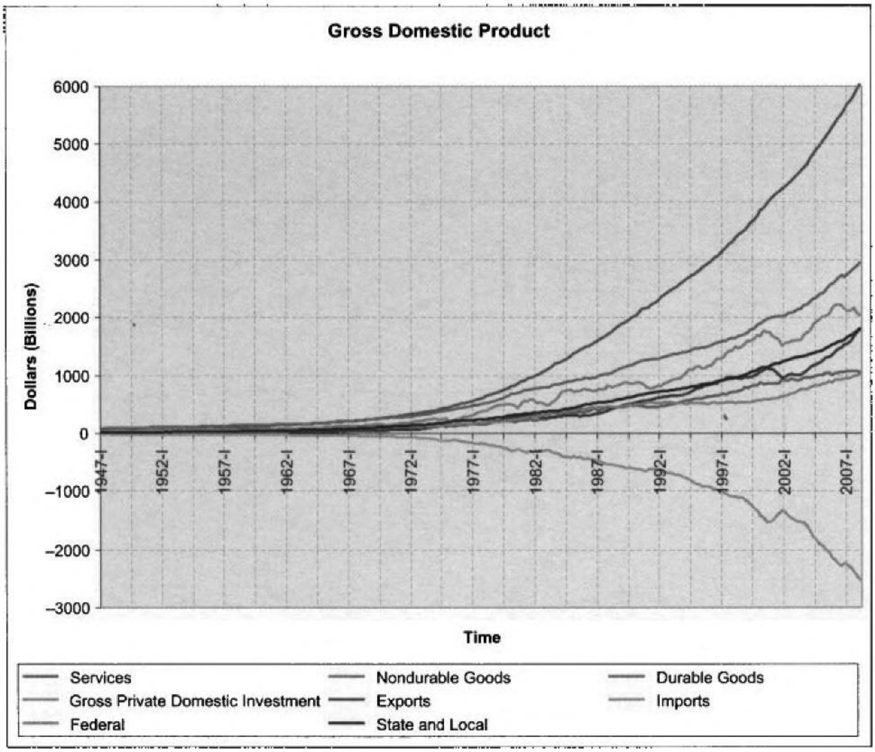


Figure 1.1. Gross domestic product over 60 years. (See color insert)
Based on the U.S. Department of Commerce, Bureau of Economic Analysis, National Income and Product Accounts Table 1.1.5 Gross Domestic Product (Seasonally adjusted at annual rates)

this interval. This is an increase of 745% in real dollars or a compounded growth rate of over 3.3 percent per year. In the 60-year period from 1940 to 2000, the U.S. population grew from 142.2 million to 291.4 million, a compound growth rate of only 1.2 percent.

Figure 1.2 is perhaps more enlightening, as it plots the percentage of the gross domestic product in each of these areas in 10-year increments beginning with the first quarter of 1948 and going through the first quarter of 2008. While the value of durable goods remained roughly constant at about 8 or 9 percent of the GDP, the value of services more than doubled from 20.5 percent in 1948 to 42.5 percent of the GDP in 2008. During this interval, the value of non-durable goods has decreased from 36.5 percent of the GDP to 20.8 percent. Up until the late 1960s, the United States was generally a net exporter; the United States has been a net importer of goods and services during the subsequent four decades (http://www.u-s-history.com/pages/h980.html).

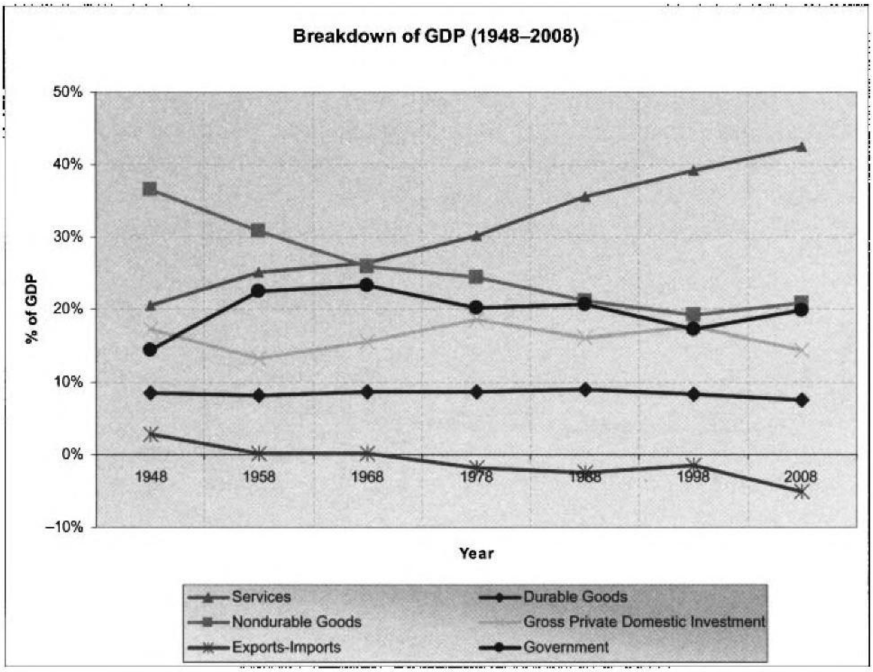


Figure 1.2. Percentage breakdown of the gross domestic product

While services represent roughly 45 percent of the gross domestic product, the number of people employed in the service industry is an even greater percentage of the total employment. Table 1.2 shows the breakdown in non-farm employment in the United States in June 2009. (Note that less than 2 percent of the U.S. workforce is employed on farms, so looking only at non-farm employment is quite reasonable.) Of the 131.4 million employees, less than 15 percent are employed in goods-producing industries including natural resource extraction, construction, and manufacturing. About the same percentage (17.1 percent) are government employees (many of whom can be viewed as being in service-related jobs), while over two out of every three employees (nearly 69 percent) work in the service sector according to the U.S. Bureau of Labor Statistics.

Figure 1.3 shows the breakdown of the U.S. employment in percentage terms for June 2009. Figure 1.4 breaks down the service sector employment into the major categories shown in Table 1.2. Two out of every seven service sector employees work in the trade, transportation, or utility industries, another 19 percent work in professional and business services, while 17 percent work in health care.

TABLE 1.2. Breakdown of U.S. employment in June 2009

Non-farm Jobs in the US—June, 2009					
		Raw #s		% of previous	
Total Non-farm Employment (×10^3)	131,411				
Goods Producing		18,713		14.2%	
Natural Resources			715		3.8%
Construction			6,162		32.9%
Manufacturing			11,836		63.3%
Private Service Providing		90,223		68.7%	
Trade, Transportation, Utilities			25,174		27.9%
Information			2,834		3.1%
Financial Activities			7,737		8.6%
Professional and Business Services			16,624		18.4%
Education			3,072		3.4%
Health care			16,190		17.9%
Leisure and Hospitality			13,177		14.6%
Other			5,415		6.0%
Government		22,475		17.1%	
Federal			2,826		12.6%
State			5,149		22.9%
Local			14,500		64.5%

Based on the U.S. Bureau of Labor Statics, Table B-3. Employees on nonfarm payrolls by major industry sector and selected industry detail, seasonally adjusted (http://www.bls.gov/ces/#tables, ftp://ftp.bls.gov/pub/suppl/empsit.ceseeb3.txt)

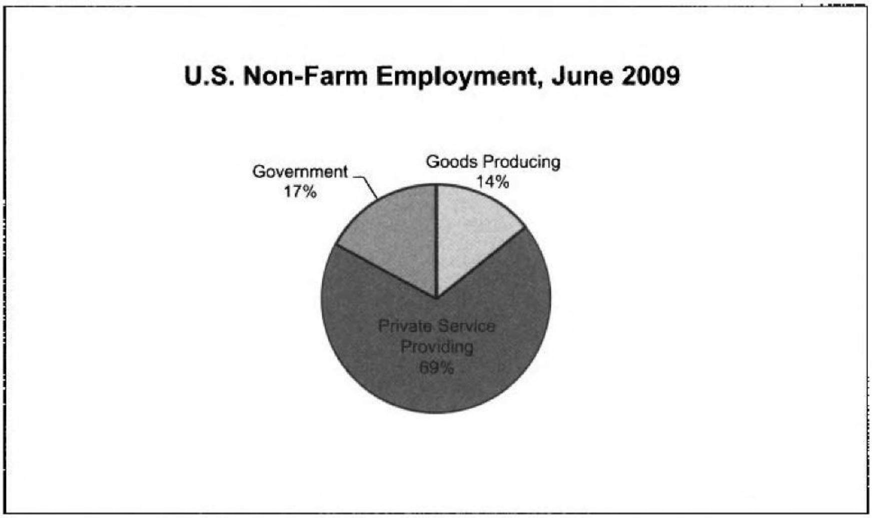


Figure 1.3. U.S. non-farm employment by major sector, June 2009

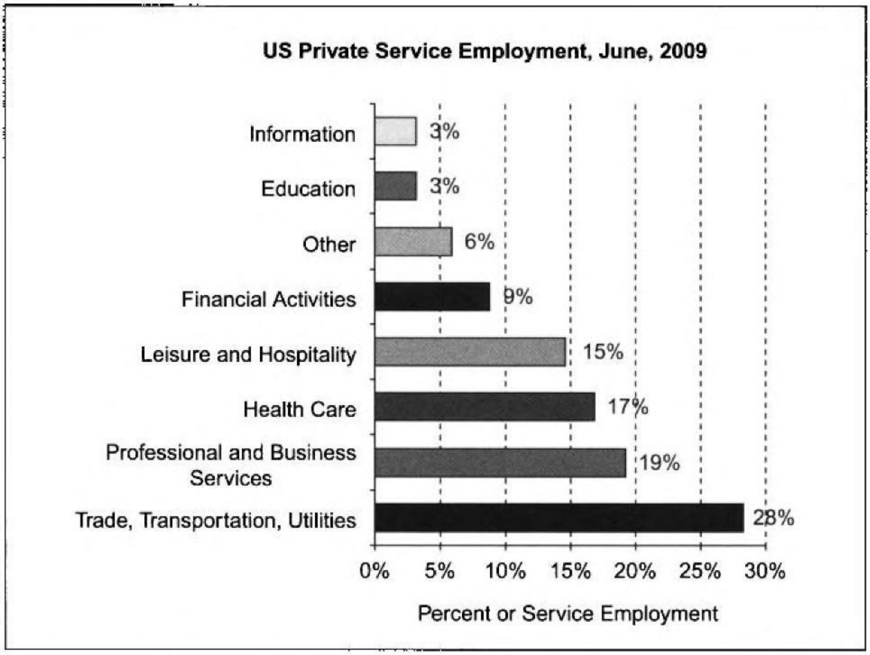


Figure 1.4. U.S. private service employment, June 2009

1.3 PUBLIC VERSUS PRIVATE SERVICE DELIVERY

Services are provided by both the public and private sectors. The public sector refers to governmental agencies, while the private sector encompasses private companies. For the purposes of this discussion, we will consider non-profit agencies, like many food pantries, to be governmental agencies since, in important ways, they behave in a manner that is similar to government agencies.

One of the fundamental differences between services provided by a public agency and those delivered by a private company has to do with the service provider's objective. Private companies, even though they are in the service provision business, often act just like other companies. Their primary objectives have to do with profit maximization or maximizing the return on their shareholders' equity. For example, the primary objective of a company providing cellular phone service is to maximize its profits. In those cases in which their (short-term) revenues are fixed, such firms attempt to minimize the cost of delivering the service. In modeling such firms, it is important to include a constraint that guarantees a minimum level of service. For example, while the long-term objective of a package delivery service company is to maximize profit, its short-term day-to-day objective might be to minimize its costs. Clearly, the firm can minimize its costs by doing nothing! However, it typically is constrained to deliver the packages scheduled for that day. With this constraint, it wants to do so at minimum total cost.

Public agencies and most non-profit agencies operate with different objectives and different constraints. Such agencies typically want to maximize some measure of service, such as the number of clients served, while operating within a fixed, and often very tight, budget. For example, an ambulance service might want to locate its bases to maximize the number of people who live within a given time of the nearest ambulance base, since time is of the essence in medical emergencies. The ambulance service will typically have sufficient funds to build and operate only a limited number of bases. In other cases, an agency may want to maximize some measure of the equity of the service provided to its clients. For example, maximizing the minimum fill rate (the ratio of the delivered food to the requested food) was the objective used in allocating collected food from food donors to food pantries in the Greater Chicago Food Depository program (Lien, Iravani, and Smilowitz, 2007). A vehicle would visit donor companies—typically grocery stores or restaurants—collecting fresh food donations in the morning and would then deliver the food to roughly half a dozen food kitchens over the course of the afternoon. Without some attempt to achieve equity between the food kitchens, the first kitchen visited might receive a large allocation, leaving little if any food for the last kitchens visited on the vehicle's route.

Not only is the profit maximization objective typically replaced by a service-related objective when considering public sector service providers, but also we typically have to capture multiple objectives (Cohon, 1978). For example, in locating ambulances, not only is it important to maximize the number of people served adequately—those for whom an ambulance is located within a specified time standard—but it is also important to ensure equity across different socioeconomic groups within the community. There may well be important tradeoffs between the number of people who can be served adequately and a measure of equity across all groups. Also, a public agency may want to minimize the disruption to the community, thereby introducing a third objective. In the case of ambulance bases, the city providing the service may want to maximize the number of bases that are located on existing city land, or equivalently to minimize the number of pieces of property that need to be seized using eminent domain laws. This will introduce additional tradeoffs and will make the analysis and modeling of the problem that much more complicated and difficult. Johnson and Smilowitz (2007) outline many important issues in community-based operations research, a growing field that is closely related to public sector service modeling.

While many of the models that we outline will be appropriate for both public and private sector service providers, it is important that we bear in mind the differences between these two contexts when determining the appropriate objective(s) and constraints for any model.

1.4 WHY MODEL SERVICES?

From the discussion above, it is clear that services constitute a significant part of the U.S. economy and increasingly of the global economy. Over two out of every three employees work in the service industry and over 45 percent of the $13.8 trillion 2007 GDP is attributable to services. A number of good books have been written outlining the service industry (Davis and Heineke, 2003; Fitzsimmons and Fitzsimmons, 2008; Metters, King-Metters, and Pullman, 2003). While providing an excellent introduction to the service industry and its components in general, these texts do not focus on modeling some of the key decisions that managers of service-oriented firms or agencies need to make.

Many service providers face similar decisions. Much of the focus of this book is on those decisions or classes of problems that are common across many service providers, be they in the public or private sectors. These decisions include:

- Location decisions – how many facilities to have and where should they be located
- Resource allocation decisions – how to allocate scarce resources (such as class capacity) to demands (such as students wishing to take each course)
- Short-term workforce scheduling issues – how many employees to schedule for each shift and when each shift should begin
- Long-term workforce management problems—how many employees to hire, release, and retrain, at which locations, and in which positions
- Inventory problems (arising in retail services as well as a range of other services) – how much safety stock inventory to maintain, when to place orders, and how much inventory to order
- And vehicle routing problems—how many routes are needed to serve a customer base (e.g., clients for Meals on Wheels), how to assign customers to routes, and how to sequence the stops on the route to achieve a service provider's objectives

In many of these cases, the fundamental underlying issue is that of striking the right balance between the demand for services and the supply of those services. Alternatively, this can often be viewed as striking a balance between the level of customer service that is achieved and the cost of providing that level of service. Intuitively we would expect that as more service is provided, the cost of providing the service increases but the quality of service also improves. For example, if a grocery store increases the number of cashiers on duty, its labor costs will go up, but the quality of service, as measured by the time patrons must wait to pay for their groceries, will improve. Similarly, if a popular restaurant increases its seating capacity with a concomitant increase in wait staff, chefs and other personnel, its costs will increase. At the same time, however, the waiting time for a table during the peak dining hours will decrease. As more screeners are added at an airport security checkpoint, costs increase but passenger delay times decrease.

1.5 KEY SERVICE DECISIONS

In the long-run, service providers must make decisions regarding the *number and location of facilities*. For example, a hospital system may need to decide how to allocate services to hospitals and whether additional sites are needed to accommodate a growing demand for medical services. Complicating the planning process is the need to have some services co-located at the same site. Also, any plan must evolve over time as the population ages and changes demographically (Santibáñez, Bekiou, and Yip, 2009).

A coffee company like Starbucks must decide how many stores to open and where they should be. In fact, Starbucks seems to have opened too many stores and, in the summer of 2008, announced that it was closing 600 stores in the United States (de la Merced, 2008). Just as the decision of where to open stores can be modeled, so too can the decision regarding the number of stores to close and the location of those stores. While this is a private sector problem for which we earlier argued that profit maximization should be the long-run objective, news reports, editorials and commentaries (Wong and Rose, 2008) suggest that even such corporate decisions must account for community-based objectives if the firm is to avoid negative publicity.

In some industries, the service region needs to be partitioned or divided between sales representatives. This is a districting problem that generally has multiple objectives. Typically, one wants the districts to be contiguous so that a salesperson does not service one county in New York, for example, and another in southern California. In addition, one often wants to equalize the workload (e.g., the number of potential customers) in such districts. At the same time, one would typically prefer a compact (close to circular) district with small customer-to-base travel distances to one that is highly elongated with longer travel distances. This too is a form of a location problem.

Resource allocation problems are also common in the provision of services. For example, at Northwestern University, half of the freshmen are in the Weinberg College of Arts and Sciences (WCAS). Each of these 1,000 students must take a seminar in the fall quarter and another in the winter or spring quarter. During the fall quarter, there are roughly 70 seminars, each with a capacity of about 15 students. Students must be assigned to seminars so that each student is in exactly one seminar, the seminar capacities are not violated and the assignment maximizes a measure of student satisfaction.

Similarly, each quarter or semester, at every college or university in the country, classes need to be assigned to rooms and times. Clearly one objective is to minimize the number of classes that do not have adequate seating capacity for those wishing to enroll in the course. At the same time, we do not want to assign a small class to a large auditorium as this makes for an uncomfortable teaching and learning experience for all involved. (I know this from personal experience.) In addition, we typically want to minimize both faculty and student walking times. Faculty walking times or distances are generally easier to estimate because we can generally assume that faculty leave for class from their offices, while the location of students immediately before a class is less certain. They could be in their dormitories, at a student cafeteria, in the library, or in another class. Finally, we need to ensure that no faculty member has to teach two classes at the same time. We also would like to respect faculty requests for teaching schedules as much as possible. Some faculty members prefer mornings to afternoons; some prefer longer Tuesday/Thursday classes to shorter Monday/ Wednesday/Friday classes.

Airlines operating at large hubs like Chicago, Atlanta, Los Angeles, the various New York airports, Dallas, Houston, and San Francisco must assign planes to gates. Airlines today operate hub and spoke systems with aircraft converging on the hub at about the same time and then leaving at about the same time roughly an hour or so later. This allows time for passengers to change planes and to make connections. Airlines may try to assign flights to gates to minimize the total walking distance of all connecting passengers. Typically, this problem is constrained by the capacity of specific gates to accommodate different aircraft types.

Almost all services, with the possible exception of Internet-based services, need to *schedule employees*. As noted above, the demand for services varies dramatically over time. In many cases, such as the provision of emergency services and restaurant services, the demand exhibits daily patterns with significant peaks and troughs. Key questions that can be answered with appropriate models include:

- How many employees to have on duty at each time of the day?
- When should the employees begin and end their shifts?
- How many full-time and part-time employees should be utilized?
- How should cross-trained employees (those with multiple skills) be deployed?
- What is the tradeoff between employee costs and customer service?

For some industries, the temporal variability occurs over the time span of months. For example, the demand for accounting services peaks in the first three and a half months of each year in the United States as citizens prepare their personal income tax returns. Community organized or for-profit summer camps, parks, and swimming pools must add staff during the summer months. Retail stores add staff for the pre-holiday sales season. The key decision in these cases revolves around how much staff to add and when to begin the staff in any required training program.

Many service providers must deal with *long-term personnel management*. For example, a management consulting firm must determine how many new hires to bring on board with various skills and how to assign those employees to projects. While this may seem on the surface to be a simple issue of managing an inventory of employees—a problem for which inventory management is, perhaps, well-suited—employees differ significantly from an inventory of drill bits at a home improvement store. First, employees (unlike drill bits) have preferences for where they are deployed and which projects they work on. A failure to account for such preferences can result in disgruntled employees and higher than necessary turnover rates. Second, employees typically come with a bundle of skills whereas a $\frac{1}{4}$-inch drill bit has only a limited number of attributes. Third, and perhaps most importantly, employees acquire and can be taught new skills while untapped and unused knowledge may be forgotten or become out-of-date. A drill bit will not "forget" what it is supposed to be able to do; at the same time, it is unlikely to acquire a new skill and suddenly know how to rip a sheet of $\frac{1}{2}$-inch plywood. In short, while inventory theory may teach us some lessons about personnel management, we need to account for the unique features of people in managing personnel.

While most writers use the presence or absence of inventory as a key distinguishing feature between manufacturing sector and service sector, *inventory decisions* play a key role in the provision of many services. As such, it is appropriate and important that students and managers of the service industries understand fundamental issues of inventory management. For example, retail stores must manage large quantities of inventory if they are to remain in business. When the skies open up in a sudden downpour, customers running into the local drug store for shelter want to be able to buy an umbrella. They are not willing to backorder one, only to pick it up a week later on a sunny day. Auto maintenance and repair facilities must carry an inventory of the commonly used spare parts. Few, if any, drivers would be willing to wait three days while a local service station ordered a few quarts of motor oil for a $20 oil change.

But inventory issues extend far beyond the obvious ones of managing a physical stock of items. Airlines view the seats on their flights as a perishable inventory in the sense that once the flight has departed any unsold seats have "perished" or "spoiled." Yield and profit management systems are used to manage this inventory of seats. Hotels face a similar problem. A room that is not rented on a given night cannot be placed in inventory to be rented during an evening when the hotel is full. At the same time, the hotel does not want to rent a room several months in advance at a low family rate when they could hold the room in "inventory" and rent it to a businessperson willing to pay a far higher rate.

Finally, the newsvendor problem in inventory theory has much to teach us about services as well. In its classical statement, the newsvendor problem is that of a small convenience store determining how many of the daily newspaper to order. If it orders too many, it loses money on unsold copies. If the store orders too few, customers arriving late in the day will be disappointed at not being able to buy a paper, resulting in a loss of goodwill on the part of the store. In the long run, the store may lose such customers as they choose to buy not only newspapers, but other items as well, at a store that does have the paper in stock. Such models are applicable not only in this context but also in the context of seasonal clothing purchases for which there is likely to be a single purchasing window months in advance of the realization of demand. These models can also inform decisions about contracting for such services as snow plowing of one's driveway.

Finally, employees who must make decisions about how many pre-tax dollars to set aside in flexible health care spending accounts are making decisions similar to those made by the newsvendor; the dollars must be committed before the need for those moneys is realized.

Finally, *vehicle routing* is important in the delivery of many services. Express mail companies such as UPS and FedEx are in the business of delivering small packages to home and business locations. For such firms, efficient vehicle routes can mean the difference between profitability and bankruptcy. In other cases, vehicle routing is a necessary part of doing business. For example, a company selling white goods (refrigerators, freezers, washing machines, and driers) must often deliver and install the items purchased by its customers. The firm must decide when to offer delivery to customers in each region that it services and then how to sequence customers on routes on the day of delivery. In a different context, many regions share an inventory of books across multiple libraries serving the area. If a book is not available at your local library, you can often request the book on an inter-library loan system. Such books are sent from one library to another on a dedicated fleet of vehicles operated by a regional library system. A library system may encompass hundreds of primary, middle and high school, university, hospital, and public libraries. Determining the frequency of pickups and dropoffs at each library in the system as well as the route of each of the vehicles in the fleet is a complex vehicle routing problem that significantly impacts the time a library patron must wait for a book to become available at his or her local library.

1.6 PHILOSOPHY ABOUT MODELS

George Box, the noted statistician, said, "Essentially, all models are wrong, but some are useful" (Box and Draper, 1987). A model is an abstraction of reality that may ignore many details of the real world context. This is true of all models. A picture of the heart in a text on biology is a model of the heart. Clearly, the model does not beat. The picture also does not pump any blood. The picture is often cut away in a manner that would be fatal to any patient with such a "wound" in his or her heart. Nevertheless, the picture, coupled with the associated text, is useful in explaining to biology and medical students the various components of the heart, their placement relative to each other, and the functions they perform.

The same is true of models of services. For example, while we know that the demand for emergency services varies by time of day, most facility location models ignore this temporal variability as it is unlikely to impact long-term decisions about where to locate fixed ambulance bases. The model may not explicitly account for the uncertain nature of ambulance demand. The model may also assume that all ambulances are always available for service. Differences between patients in the required on-scene time as well as whether or not transport to a hospital is or is not required are also typically ignored. Nevertheless, location models are often used to assist decision makers in locating ambulances because the models provide valuable *insights* into the problem and the impact of the decisions that need to be made (Eaton et al., 1985).

Similarly, a model of the delays experienced by callers to a call center is likely to make a number of assumptions that may not be supported by empirical studies. For example, we are likely to assume that calls arrive at the center according to a particular stochastic process known as a Poisson process. While this is often supported by the data, we may also assume that the duration of calls follows an exponential distribution. This is rarely validated by empirical results. Despite these flaws, the models are likely to provide valuable insights into the operation of a call center and are likely to help us determine the appropriate number of customer service agents to employ to balance operating costs and customer waiting time.

But the value of modeling a process extends far beyond any insights that the model may provide regarding the operation of the underlying system and the decisions needed to design the service. The process of developing a model often yields benefits in and of itself. This process is shown in Figure 1.5.

Modeling generally begins with a problem. For example, the problem might be that of assigning freshmen at a university to seminars in a more efficient and more equitable manner than the current manual process. The problem might be that of determining how many parking spots to build in a municipal parking garage. The problem might be that of identifying bases for emergency medical service vehicles. It could be identifying improvements in a hospital discharge planning process.

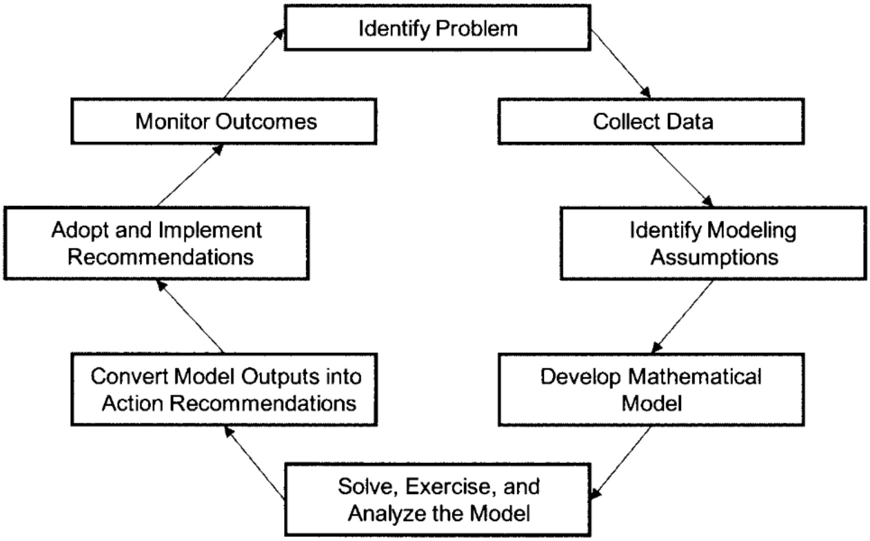

Identify Problem
Collect Data
Identify Modeling Assumptions
Develop Mathematical Model
Solve, Exercise, and Analyze the Model
Convert Model Outputs into Action Recommendations
Adopt and Implement Recommendations
Monitor Outcomes

Figure 1.5. Schematic of the modeling/decision-making process

There are three key facets associated with identifying a problem for modeling and analysis. First, we need to identify the key stakeholders or groups impacted by the process. For example, in the case of hospital discharge planning, key constituents clearly include: (a) the patients and their families, (b) the physicians, (c) the nursing staff, (d) other patient-oriented staff including social workers and discharge planners, (e) the hospital administration, and (f) any institutions such as nursing facilities and rehabilitation institutes to which patients might be sent. Each of these groups may have different subgroups. For example, patients might be stratified by the cause of their hospitalization. The discharge process for a teenager with a broken leg is likely to differ significantly from that of an elderly patient who was hospitalized for a heart attack. Physicians include in-house hospitalists, specialists who treated the patient, and the patient's own internist. Each of these groups might have different perspectives on the issue.

Second, we need to identify the key objectives to be achieved. Two objectives in the case of discharge planning are (a) expediting the process to increase the flow of patients through the hospital and (b) enhancing the communication of discharge orders between the medical personnel (physicians and nursing staff) and the patient.

Third, we need to identify any constraints that may be affecting the system. In the case of discharge planning, constraints may include language barriers between patients and the medical staff. One community hospital in the Chicago area has to deal with over 40 languages for informed consent forms for surgery; a large metropolitan hospital is likely to encounter far more languages. Such language barriers are likely to be present not only at the time a patient is admitted for a surgical or medical procedure but also at the time the patient is discharged. Another significant constraint is that the medical staff has many other tasks to perform each day other than discharging patients.

Once the problem—stakeholders, objectives and key constraints—is identified, the next step involves collecting data on the current process. In the case of discharge planning, this would include data on the number of patients discharged each day, the destinations to which they are discharged, the reasons for their initial hospitalization, and so on. It would also include mapping the current discharge process so that the key steps are clearly identified.

The third key step in modeling a service operation is to identify any key assumptions that the model(s) might make. For example, if we are developing a simulation model of the discharge process, we need to structure the "arrival" process of patients ready to be discharged on any given day. What probabilistic laws does this process follow? Will we model the process as if the same number of patients are discharged each day? This is likely to be a very bad assumption. Which of the various steps in the discharge process will we model explicitly and which will we handle implicitly? How will we represent the delays in this process that are caused by physicians or nurses dealing with other patients who are not being discharged? If the model is an optimization model, explicitly enumerating the assumptions often entails identifying the nature of the relationships between the key decision variables as well as how those variables interact in the objective function(s). For queueing models, we need to be explicit about the probabilistic form of the customer arrival and service processes as well as the queue discipline that specifies the order in which customers are to be served.

Developing the model is the next key step. In the case of a simulation model, this entails writing the computer code in the relevant simulation language to represent the modeled process. In the case of an optimization model, this involves writing down the objective function and constraints in terms of the input data and the decision variables. This mathematical model must then be translated into some form that a computer can understand. For example, many optimization models can be solved within spreadsheet systems like Microsoft Excel. Larger and more complex problems may require the use of specialized optimization languages such as AMPL (Fourer, Gay, and Kernighan, 2002) and IBM's ILOG CPLEX optimizer. Developing the model involves setting up the appropriate spreadsheets and/or writing the code needed to invoke specialized languages. Developing a queueing model may be simple if the model already exists in the literature. If it does not, the underlying equations for the queueing process need to be written down. Sometimes these can be solved in closed form, while in other cases they need to be solved numerically. In the latter case, developing the model entails writing the code or developing the spreadsheet model capable of solving these equations.

Solving, exercising, and analyzing the model is the next step. In many cases, solving the model is actually fairly straightforward once the model has been developed. For simulation models, one simply has to run the model. Similarly, many optimization models can be solved with software built into spreadsheet systems or with off-the-shelf optimization packages. If the results of a queueing model are known, there is nothing to solve! If they are not, numerical procedures in a spreadsheet can often solve a given instance of a queueing model quite adequately.

Exercising the model entails changing the key inputs to see how the model responds. This is necessary (1) to validate the model, (2) to test alternative policies, and (3) to determine the sensitivity of the results to changes in key model inputs. For example, in a simulation model we might increase the rate at which customers arrive at the system. If the performance metrics improve (i.e., if the customer waiting time decreases as the arrival rate increases) we have strong reasons to suspect that the model is not valid. Similarly, if attempts to replicate current conditions (e.g., the current discharge process at a hospital) result in times that differ significantly from those observed in practice, we again have reason to suspect that there is a problem with the model. If the model appears to be valid, we generally want to test alternative policies with the model. For example, if we simulated the discharge planning process, we might want to use the model to understand the impact of moving some steps of the process earlier. We might want to know, for example, how the average time required to discharge a patient—measured from the time a physician says a patient is ready to leave the hospital until the time the patient actually vacates the bed—changes if we begin arranging for a nursing home bed at the time critically ill patients are admitted to the hospital, rather than waiting to do so until they are discharged. In the case of an optimization model for assigning students to seminars, we might want to see how many more students can get their first-choice seminar if we increase the number of students allowed in each seminar by one. For a queueing model, we might want to see how the average waiting time before service is impacted by changes in the average arrival rate of customers, the number of servers on duty, or the mean service time.

The model results must also be analyzed, often statistically. In the case of a simulation model, this is particularly important since the output of any simulation model is typically one realization of a random process. Running the simulation model a second time with different random numbers is likely to result in different output values. Thus, it is important to analyze the results to see if the predicted results of two different policies are indeed statistically significantly different. In the case of an optimization model, there may be multiple objectives and the tradeoff between these objectives may need to be identified as part of analyzing the results. For example, in assigning students to seminars, there is likely to be a tradeoff between maximizing the satisfaction with the assignments averaged over all students and minimizing the number of students who are particularly adversely affected by the assignment (e.g., those who get their fourth choice or worse!).

Once the model has been validated, exercised, and analyzed, we must convert the model results into action recommendations. In the case of discharge planning, this might involve recommending that additional hospitalists be employed to expedite the discharge process. It might mean recommending that physicians spend *additional* time with patients to ensure that the patients fully understand their discharge instructions. Perhaps one of the recommendations would be to begin certain processes, like planning for rehabilitation care or nursing care, as soon as it becomes apparent that the patient will need this additional level of care upon discharge, even if this means beginning the planning significantly earlier in a patient's stay.

It is worth noting that this is often the stage at which the formal modeling by an analyst ends and the work of the decision maker takes over. For example, in assigning students to seminars, it is probably not appropriate for a technical analyst to recommend one assignment plan over another. Rather, the job of the analyst may be to make explicit to the decision maker—the dean in this case—the tradeoffs that must be made in choosing one assignment over another.

The next step involves adopting and implementing the action recommendations. This is often complicated by the many stakeholders involved in the process. For example, a recommendation might be that a patient not be discharged until he or she can tell the nurse the top three discharge instructions. This sounds simple and is intended to ensure that the physician has adequately communicated the discharge instructions to the patient. However, adopting and implementing such a recommendation may be anything but simple. First, it requires that the physician identify the top three instructions and communicate to the nurse that these are, in fact, the key instructions that the patient needs to understand before being discharged from the hospital. Second, the nurse now has added work as he or she must ask the patient what these key instructions are. Third, what should happen if the patient does not understand the instructions or cannot recite them back for the nurse? Should the physician be called back in to discuss the instructions again with the patient? Should the nurse discuss the instructions with the patient? In assigning students to seminars, the seminar assignments must be uploaded to the university or college course scheduling software and conflicts that might arise between the assigned seminars and other courses individual students have enrolled in must be resolved.

The last step involves monitoring the process to assess how well the recommendations have worked. Are discharge times reduced? Is there evidence that patients understand their discharge instructions better than they did before and that they are healthier following the revised discharge process than they had been under the old process? Are students happier with the new assignment process for freshman seminars? Have the number of complaints from parents been reduced as a result of the new assignment process?

Finally, Figure 1.5 shows a feedback link between the step of monitoring the process and the initial step of identifying the problem. In some cases, we find that the problems that initiated the study are still not resolved and the process needs to be repeated. Perhaps we need a more careful understanding of the objectives of the constituents. Maybe the initial process failed to account for some key constraints that resulted in an incomplete or failed implementation of the recommendations. Maybe the data that were collected were biased or erroneous in important ways, which resulted in serious discrepancies between the modeled results and those obtained in the real world. Perhaps the differences between the real and modeled world resulted from unrealistic assumptions that were made in the process of modeling the problem. In any event, it may be necessary to cycle through this modeling/decision making process until the monitored results are satisfactory.

It should be clear that there are likely to be many ancillary benefits associated with this modeling process aside from the direct benefits that result from the recommendations. First, the process forces us to identify the key stakeholders, their objectives, and the constraints under which they and the process operate. Identifying these people and sharing the various objectives with all stakeholders is likely to be a valuable exercise as all parties will begin to see the problem from a new perspective. Doctors will be forced to consider the discharge process from the perspective of patients, nurses, and social workers, while administrators will better understand the myriad demands placed on a doctor's limited time.

Similarly, the data collection process is likely to yield important side benefits. In the course of collecting data relevant to a problem, one often finds that important information is missing, thereby making it hard to assess quantitatively how well the current system is performing. What metrics should be used to measure a patient's understanding of the discharge instructions that he or she has been given? In many hospitals, there is limited, if any, data on such issues. Proxies, such as the number of calls back to a physician regarding medication instructions, may be the best information available. Thus the data collection process may suggest that new metrics need to be devised and new data collected. In the course of collecting and using the data, errors in the dataset are likely to be discovered. Correcting these errors for future use is another ancillary benefit of any modeling process.

The process of identifying the key modeling assumptions further facilitates discussion among the stakeholders and between the stakeholders and the technical staff performing the analysis. Agreement must often be reached regarding which aspects of a problem are likely to impact the key decisions and are therefore worth modėling. Is there reason to believe that it is important to model patients being discharged to home separately from those being discharged to a nursing home or rehabilitation facility? Is the age of the patient an important determinant of the discharge time and, if so, should this be included in the model? If we are developing a model for surgical scheduling, is there agreement that uncertainty in the duration of surgical procedures may be contributing to delays in the starting times of subsequent surgeries and that modeling this uncertainty is worthwhile? Finally, by making the assumptions explicit, we are likely to be better able to discuss the results of the model and the recommendations that are derived from the modeling process. Both the technical analysts and the stakeholders will have a better appreciation for whether or not a particular result or recommendation is likely to have been caused by a modeling assumption or is the consequence of underlying properties of the process being analyzed.

In summary, in addition to the insights and action recommendations provided by the modeling process, there are likely to be significant ancillary benefits in the form of more reliable and useful data for future modeling, enhanced communication among stakeholders, and an improved understanding by all of the modeling being done.

1.7 OUTLINE OF THE BOOK

Thomas Davenport and Jeanne Harris argue that "analytics" is the new tool for competitive firms (Davenport and Harris, 2007). Analytics means data-driven decision making.They cite numerous examples of companies and organizations—from professional sports teams to pharmaceutical manufacturers, from banks to book sellers, from casinos to international insurance companies—that are getting ahead faster than their competitors by using quantitative data-based decision tools. They argue that firms that use analytic techniques routinely manage to out-think and out-perform their competitors in a systematic way.

Ian Ayers, in a related book, argues that our lives are increasingly being ruled by "super crunchers" (Ayers, 2007), individuals and organizations capable of managing and analyzing huge quantities of data to extract important trends. Like the Davenport and Harris's firms that compete on analytics, Ayers' super crunchers are found in a broad variety of industries, from medical diagnostics to airline pricing, from studies of Supreme Court decisions to online matchmaking and dating services. Ayers argues that two methodologies—randomization and regression—are the keys to super crunching. Randomization basically means that treatments are applied randomly to elements of a population to see which treatment is most effective. Medical science has used randomization for many years in testing new drugs and procedures. Other industries are taking up the banner of randomization as well. Ayers reports, for example, that it is now possible for a firm like Amazon.com to determine in a matter of hours which of several different home page web designs generates the most viewing time and the most clicks by randomly showing the different designs to different customers over the course of the testing period. Because pages are shown randomly, preference differences that might exist based on age or gender or geographical location or ethnicity or any other individual characteristic are washed out due to the large sample that Amazon.com can amass in a short period of time. For samples that include data on individual characteristics, regression can be used to assess the impact of a treatment as a function of the individual characteristics. For example, in drug testing, individual characteristics that might impact the efficacy of a new drug include (1) how long a patient has had a disease, (2) the age of the patient, (3) the patient's gender, (4) the patient's family history, and (5) any other health conditions (e.g., obesity) that the patient may currently have or may have had in the past. Near the end of the text, Ayers introduces a third methodology, hypothesis testing, which can be used to determine whether or not the differences between two samples are statistically significant. For example, if, in a sample of 200 teenagers, we find that boys have more traffic accidents per mile driven than do their female counterparts, we can test whether the difference is statistically significant or if the difference could have been caused purely by chance.

Just as Ayers focuses on two methodologies, so too this text will focus primarily on two methodologies: optimization modeling and queueing theory. Optimization involves determining the values of key decision variables—those factors associated with a problem about which we need to make decisions—to minimize or maximize some objective. For example, in locating ambulances, we may want to determine the base locations that maximize the number of people who are within 8 minutes of the nearest ambulance base using a limited number of ambulances. Queueing models are typically closed-form mathematical equations that tell us how long a customer will have to wait for service. For example, we can use queueing theory to estimate the average time a person will spend in line and being served at a vehicle emission testing station if we know how many testing bays are operating, some characteristics of the rate at which cars arrive to be tested, as well as key attributes of the time required to test each vehicle. In some important cases, optimization and queueing will merge. In scheduling nurses in an emergency room, for example, we may want to examine two competing objectives: minimizing the cost of the nursing staff that we use and minimizing the total time that patients wait.

Chapter 2 reviews key concepts and issues in optimization modeling. The chapter outlines five key questions that need to be addressed in developing any optimization model. It then reviews linear programming, one of the fundamental tools used in optimization. Network optimization, a special form of linear programming modeling, is then summarized, as is integer programming, another extension of linear programming, in which some variables must take on only integer values. Integer programming is particularly useful when yes/no decisions need to be made. Returning to the example of locating emergency medical service bases, there is likely to be a yes/no variable associated with each candidate location at which a base can be located. As indicated above, many problems are characterized by multiple objectives. Chapter 2 discusses two methods of finding efficient points on a tradeoff curve. Chapter 2 concludes with a list of 10 rules for formulating problems, rules that many students and practitioners seem to violate with disturbing regularity.

Chapter 3 introduces queueing theory. After introducing four key performance metrics of interest in almost any queueing problem and Little's law relating the average time in the system to the average number in the system, the chapter develops a general approach to analyzing a commonly studied class of queues. This general approach is then applied to a range of queueing problems. A more general model is introduced for the case of systems with a single server. Finally, since many service systems are characterized by time-varying conditions (arrivals of customers and service capabilities), the chapter concludes with a discussion of a numerical approach to solving queueing problems in the face of temporally changing conditions.

After reviewing these two key methodologies, the remainder of the text focuses on problems that arise frequently in the design and operation of services. Chapter 4 examines location models, which are typically used to assist in making strategic decisions regarding the number and location of service facilities. The chapter includes a discussion of districting problems that also arise in services.

Inventory theory is closely related to queueing theory and serves as the topic of Chapter 5. While services themselves can rarely be carried in inventory, most service providers require an inventory of some goods or equipment to aid in the service delivery process. As in other chapters, we begin with a simple model, the economic order quantity model, and progressively extend the model in a variety of ways. The chapter concludes with an analysis of the newsvendor problem.

Chapter 6 focuses on decisions related to the allocation of resources to customers or demands. For example, if we know where schools are located, we need to determine which students should attend each of the available schools. This can be thought of as a districting problem or as a resource allocation problem in which we are allocating the scarce resource of classroom space to the students in the district. Similarly, assigning students to seminars is also a resource allocation problem in which the resources are the courses allocated to the students. Assigning gates to aircraft at an airport is also a resource allocation problem. Many such problems can be solved using what is called an *assignment model*, perhaps with minor modifications.

Short-term workforce scheduling decisions and problems are the focus of Chapter 7. Many simple scheduling problems can be modeled as network problems. This allows us to solve large instances of such problems very quickly and easily. The basic model can be extended in important ways to include part-time workers, overtime shifts, and cross-trained employees. More complex problems, including cases in which employees must be scheduled over a 24-hour period or situations in which employees are permitted scheduled breaks, require more complex modeling techniques. The chapter concludes with a discussion of a multiple-objective employee scheduling problem in which employee costs and customer delays are traded off against each other.

Chapter 8 discusses long-term workforce management. This is an evolving area and models of this important topic are still in their nascent stage. The chapter begins with a very simple model and then evolves to increasingly complex problems including a model that accounts (approximately) for employee attrition, hiring and firing in response to anticipated workload requirements. The chapter concludes with a model that examines the tradeoff between employee costs and schedule delays for projects. Such a model could be of particular value in assessing the true cost of new projects in a consulting environment, for example.

Chapter 9 examines a variety of applications of queueing theory in the design and operation of service systems. Priority queueing problems arise in the operation of emergency response systems. Clearly, an elderly patient in active cardiac arrest will be given priority over a young child with a sore throat in an emergency room. Call center design is an important application of queueing theory in service systems. Finally, many services, including non-emergent health care systems, can schedule their customers (or patients). Chapter 9 concludes with a discussion of this problem.

Chapter 10 deals with vehicle routing decisions and problems. As in the case of long-term workforce management, vehicle routing models explicitly devoted to service problems are relatively few and far between. The chapter therefore focuses on traditional models and outlines ways in which such models might be extended to account for the unique issues that arise in the service industries.

Any text on a topic as broad as modeling the service industries can only begin to scratch the surface of the problems that we are likely to encounter. This is certainly true of this book as well. Chapter 11 outlines a number of additional directions for future modeling of the service industries that are not explicitly covered in this book.

1.8 PROBLEMS

1. List the services that you employed:
   a) in the last 24 hours
   b) in the last week
   c) in the last month

Be sure to include services that you implicitly used, such as insurance.

2. Identify two services that you have used recently and briefly discuss ways in which each can be improved.

REFERENCES

Ayers, I., 2007, *Super Crunchers: Why Thinking by Numbers Is the New Way to Be Smart*, Bantam Books, New York.

Box, G. E. P., and N. R. Draper, 1987, *Empirical Model-Building and Response Surfaces*, Wiley, New York, NY.

Cohon, J. L., 1978, *Multiobjective Programming and Planning*, Academic Press, New York, NY.

Davenport, T. H., and J. G. Harris, 2007, Competing on Analytics: The New Science of Winning, Harvard Business Press, Boston.

Davis, M. M., and J. Heineke, 2003, *Managing Services: Using Technology to Create Value*, McGraw Hill/Irwin, Boston, MA.

de la Merced, M. J., 2008, "Starbucks announces it will close 600 stores," *The New York Times*, July 2, 2008, online at http://www.nytimes.com/2008/07/02/business/02sbux.html?_r=1&scp=1&sq=starbucks%20600%20stores&st=cse.

Eaton, D. J. et al., 1979, *Location Techniques for Emergency Medical Service Vehicles: Volume I—An Analytical Framework for Austin, Texas*, Lyndon B. Johnson School of Public Affairs, The University of Texas at Austin, Austin, TX.

Eaton, D., M. S. Daskin, D. Simmons, B. Bulloch, B., and G. Jansma, 1985, "Determining emergency medical service vehicle deployment in Austin, Texas," *Interfaces*, *15*, 1, 96–108.

Fitzsimmons, J. A., and M. J. Fitzsimmons, 2008, *Service Management: Operations, Strategy, Information Technology*, 6th ed., McGraw Hill/Irwin, Boston, MA.

Fourer, R., D. M. Gay, and B. W. Kernighan, 2002, *AMPL: A Modeling Language for Mathematical Programming*, Duxbery Press/Brooks/Cole Publishing Company, South San Francisco, CA.

Friedman, T. L., 2005, *The World Is Flat: A Brief History of the Twenty-first Century*, Farar, Straus and Giroux, New York, NY.

Johnson, M. P., and Smilowitz, K., 2007, "Community-Based Operations Research," in *TutORials in Operations Research*, INFORMS, Institute for Operations Research and the Management Sciences, Chapter 6, 102–123.

Lien, R., Iravani, S., and K. Smilowitz, 2007, *Sequential allocation problems for nonprofit agencies*, Technical report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL.

Metters, R., K. King-Metters, and M. Pullman, 2003, *Successful Service Operations Management*, Thomson/South-Western, Mason, OH.

Santibáñez, P., G. Bekiou, and K. Yip, 2009, "Fraser Health uses mathematical programming to plan its outpatient hospital network," *Interfaces*, *39*, 3, 196–208.

Wong, W., and B. Rose, 2008, "Starbucks' exit leaves some areas wondering," *The Chicago Tribune*, July 19, 2008, online at www.chicagotribune.com/business/chi-sat-starbucksjul19,0,7595492.story.

PART I

METHODOLOGICAL FOUNDATIONS

This page intentionally left blank

2

OPTIMIZATION

Rosemary, an assistant dean in charge of the freshman class, stared at the piles of cards on the table. Most of the piles contained 14 or 15 cards, each card representing an incoming freshman. There was one pile for each of the 68 seminars being offered in the Fall term. One larger pile contained about 40 cards. These students had not yet been assigned to a seminar and were the "problem children" with whom Rosemary had been wrestling for days. The top five choice seminars of each of these students were all at capacity. Every time Rosemary tried to shuffle a few students around to get another student into a seminar, she seemed to end up back where she started, with a large pile of unassigned freshmen.

The good news was that this two-week process would come to an end one way or another in two days, on Friday, when the assignments were due at the registrar's office. The bad news was that each year some students ended up with their seventh or eighth choice (or worse). She knew she would hear from many of the parents of those students. After all, at an expensive private university, parents expected their children to be welcomed and to be able to attend the classes they wanted.

Service Science, by Mark S. Daskin

2.1 INTRODUCTION

"Whatever you do in life, do your best." I still remember that charge from my parents, and I suspect that most parents wish this, among many other things, for their children. I know that I do. Doing your best is really what optimization is all about. While Rosemary was doing her best to assign freshmen to seminars, it seemed to her that there had to be a better way to do so. There is, and optimization is the key to the problem.

Optimization arises in many services. For example, in the late 1970s, Austin, Texas, had approximately a dozen ambulances to serve a population of roughly 350,000 people. At the time, the crews for these ambulances were housed in public apartment buildings scattered throughout the city. Drugs and expensive equipment could not be stored in an unsecured ambulance parked outside an apartment building. As a result, when a crew received a call for service, they had to bring the drugs and equipment from the apartment to the ambulance and then carefully stow them before beginning to drive to the scene of the emergency. When seconds and minutes may determine whether or not a patient survives, the additional delays introduced by this process were deemed unacceptable. Consequently, the city began working with a team of students and faculty at the University of Texas at Austin to identify permanent, secure ambulance sites for their crews and vehicles.

To analyze the problem, the team began with a history of all calls for emergency medical services during a five-month period. Calls for service were geocoded to one of 358 zones that had been established for traffic modeling by one of the city agencies. In addition, the team had access to demographic data from a special census that was conducted in 1976. After significant discussion with city and emergency response personnel, the team focused on a model that maximizes the number of people who are no more than a pre-specified service time from the nearest ambulance base. A number of different service times, and demographic and medical emergency proxies for demand were utilized to develop a large number of locational options for the city (Eaton et al., 1985). Chapter 4 deals with location problems of this sort.

At Northwestern University, the Weinberg College of Arts and Sciences routinely uses an optimization model to assign roughly 1000 freshmen to approximately 70 seminars each fall quarter. During the summer preceding their freshman year, each new student provides the dean's list with their top choice seminars ranked from number 1 (their most preferred seminar) on down. The objective of the optimization model is to minimize the sum of the rankings of all student/seminar assignments. Thus, if a student is assigned to her third choice seminar, she contributes three units to the objective function. Clearly, an objective function value of 1000 would be perfect, as it would mean that each student was assigned to his or her first-choice seminar. Unfortunately, this cannot generally occur. The problem has two key constraints. First, each student must be assigned to exactly one seminar. Second, no seminar can be assigned more students than it can accommodate. It is this second constraint that prevents each student from being assigned to his or her first-choice seminar. Seminars typically accommodate 15 students. One year, over 100 students requested a seminar on forensic pathology as their first choice. Clearly, 85 or more of these students would not be assigned their first-choice seminar. (Equally clearly, many Northwestern students want to be the next *CSI* star!) Chapter 6 focuses on resource assignment problems like this one.

As indicated in Chapter 1, services are almost always provided by people. As a result, service providers are faced with the problem of determining staffing schedules. The schedule determines, in part, how many people will be on duty at each point in time. Although Chapter 3 and part of Chapter 9 deal explicitly with the relationship between staffing levels and the quality of service provided by a service system, it should be intuitively clear that more staff generally will mean better service and shorter waiting times for service. The schedule determines the staffing level at each point in time during a day (or service period) by deciding how many staff should report for duty at each hour. The detailed schedule will determine not only *how many* staff to report for duty at each time period, but also *which* individuals should be on duty at each time period. In determining the number of employees to start work at each time period, we may want to try to minimize the employee costs while ensuring that an adequate number of staff are on duty at each time period. Alternatively, we may want to examine the *tradeoff* between minimizing the employee costs and maximizing a measure of service to the customers. In assigning individual staff to the schedule, we often need to account for work rule regulations that limit the number of consecutive hours that an employee can work, individual preferences for certain times of the day or week, seniority issues, and a host of other requirements. Chapter 7 examines short-term scheduling problems of this sort.

2.2 FIVE KEY ELEMENTS OF OPTIMIZATION

While numerous textbooks have been devoted to optimization modeling and theory (e.g., Dantzig, 1998; Nemhauser and Wolsey, 1988; Winston and Venkataramanan, 2003), the essence of optimization modeling can be boiled down to five key questions. Once we have the answers to these five questions, we are likely to be well on the way to developing a useful model of a service operation or, for that matter, any problem amenable to optimization.

2.2.1 What Do You Know?

Perhaps the first step in studying any problem is to determine what you know. In the ambulance location problem outlined in section 2.1, the team had lots of information on the location, time, and severity of over 4000 calls for emergency services. In addition, they had information on the demographics of 358 zones within the city as well as the travel times between each pair of zones. In assigning students to seminars, the dean's office knows the preferences of each student as well as the number of students each seminar can accommodate. In scheduling employees, we are likely to know either the minimum number of employees that the provider deems necessary to have on duty at each time period or the relationship between the number of employees on duty and a measure of the quality of service that is provided. We are also likely to have information on the preferences of each employee for different times of the day or week as well as knowledge of any work rules that a schedule must obey. Finally, we will have information on the costs of different types of employees (e.g., full-time and part-time employees, employees working the night shift and those beginning during the day, and so on).

What you know before studying the problem constitutes the *inputs* to the problem.

2.2.2 What Do You Need to Decide?

The second aspect of optimization is figuring out what you really need to decide. In some cases, parts of this may be clear. For example, in locating ambulances in the city of Austin, Texas, we clearly want to decide *where* to locate the ambulances. In assigning students to seminars, we want to determine *which* students are allowed to take each seminar.

In other cases, what needs to be decided may be less clear. In scheduling employees, for example, it is not immediately obvious what we need to decide. At the aggregate level, we typically need to decide how many employees should start working at each possible starting time. Note that having fixed, nonoverlapping employee shifts (e.g., 8 A.M. to 4 P.M., 4. P.M. to midnight, and midnight to 8 A.M., for a service provider with 8-hour shifts and 24-hour coverage) is *not* likely to achieve the best possible results. Instead, it is often better to have some employees start at different times of the day and to allow shifts to overlap. Problem 1 illustrates this phenomenon. In addition to making it easier to minimize the number of required employees or the total labor cost, overlapping shifts provide for smoother transitions between shifts in many cases because only a fraction of the employees will be turning over at any point in time. This may be particularly relevant in nurse scheduling, for example, where the handoff between nurses is critical to patient safety and the quality of care received by the patients.

In addition, sometimes we need decision variables to help compute the objective function value. For example, in locating ambulances, in addition to location variables, we need another set of variables to indicate whether or not each zone is within the service standard of the nearest ambulance.

It is critically important that we distinguish between *inputs*, those data items that we know before beginning the modeling process, and *decision variables*, those variables whose values are the *output* of the modeling process and whose values we are trying to determine through exercising the model. All too many students and practitioners fail to distinguish adequately between these two categories and refer to both as "variables." This inevitably leads to confusion and to errors in the model.

2.2.3 What Are You Trying to Achieve?

The third key component of any optimization problem is determining what you want to achieve. By this we mean, what do you want to optimize or what do you want to maximize or minimize? In the ambulance location problem, the team tried to maximize the number of demands, measured by one of many different surrogates for demand, that were within a specified time limit of the nearest ambulance. Such demands are said to be covered, while those demands that are further away are not covered. Another objective might have been to minimize the cost of the ambulance fleet needed to cover all demands. Yet a third objective might have been to minimize the average response time of the ambulances. Finally, a fourth possible objective might have been to minimize the maximum of worst-case response time. In assigning students to seminars, the dean's office tries to minimize the sum of all assigned preference rankings. This is equivalent to minimizing the average of all assigned preference rankings. Because low preference rankings correspond to more preferred choices, this is equivalent to maximizing the average assigned seminar preference over the freshman class. Other objectives are also possible. For example, they could also minimize the worst assignment over all students. In determining staffing schedules, we might want to minimize the number of employees needed or, more generally, the cost of satisfying all of the requirements in each period.

Determining the appropriate *objective function*(s) for a problem is the third key component of developing an optimization model.

2.2.4 What Inhibits Your Ability to Do So?

In formulating optimization models, either mathematically or conceptually, it is important to ask what inhibits our ability to achieve the desired objective. In locating ambulances, there is likely to be a budget that limits the number of ambulances that can be deployed. In assigning students to seminars, there are two classes of constraints. The first requires each student to be assigned to exactly one seminar while the second limits the number of students assigned to any particular seminar to the capacity of that seminar. In staff scheduling, there is likely to be a constraint that specifies the minimum number of staff that have to be on duty at any time.

Other constraints are less obvious. For example, in locating ambulances, there must be a constraint that links the two classes of variables: the location variables specifying where ambulance bases are to be located and the coverage variables indicating which nodes can be served adequately by the bases that are located by the model. This constraint simply says that a demand node cannot be counted as covered unless we locate one or more bases sufficiently close to the demand node. There must be at least one base located within the service standard of the demand node for the node to be counted as covered.

Constraints inhibit our ability to achieve the objective function. They also link key classes of decision variables to each other.

2.2.5 What Can You Learn?

While answers to the first four questions are adequate for developing an optimization model, we must also ask what we want to learn from the model. Except for the simplest of real-time decisions, the results of an optimization model must be analyzed in a broader context that often includes non-modeled components of the problem. In addition, some inputs, such as the budget for new ambulance bases, the capacity of individual seminars, or the number of staff to have on duty at any time of the day, may not be known perfectly. Instead, we may want to perform some sort of sensitivity analysis on these unknown or uncertain parameters. Thus, we might like to know how much the percentage of the population that is covered by an ambulance within a given service standard increases if we can afford one more ambulance base. By how much would the coverage decrease if we could build fewer bases? How would the average assigned seminar preference change if a particular seminar's capacity increased by 20 percent? If the desired number of staff on duty during some period (e.g., midnight to 4 A.M.) increased by 10 percent, how many more employees would we need and when should they start work? How would the number of employees change if each staff member was on duty for 8 hours but had a 1-hour break during the fifth hour on duty?

In short, the output of optimization models is rarely implemented directly and without additional analysis. The optimization models need to be exercised for a range of parameter values. In setting up an optimization model, either in Excel or in some other modeling environment, we should keep in mind the need to do this sort of analysis.

2.3 TAXONOMY OF OPTIMIZATION MODELS

While there are many ways to categorize optimization models, Figure 2.1 provides one taxonomy of such models. We can first distinguish between models that have a single objective, such as minimizing the cost of providing a service, and those that have multiple objectives. In this book, we will typically limit our attention within the class of multi-objective problems to cases with only two objectives. For example, we may want to minimize the *average* distance between ambulances and population centers and we may also want to minimize the *maximum* distance between ambulance bases and the population being served. Multi-objective problems with two objectives are easier to visualize and easier to understand than are problems with more than two objectives.

Within both the single and multiple objective classes of optimization problems, some models are linear while others are non-linear. By a linear problem, we mean a problem in which the constraints and the objective function can be written as linear functions of the decision variables. This means that we are not using trigonometric functions, we are not raising constants to a power that depends on a decision variable, we are not multiplying decision variables together,

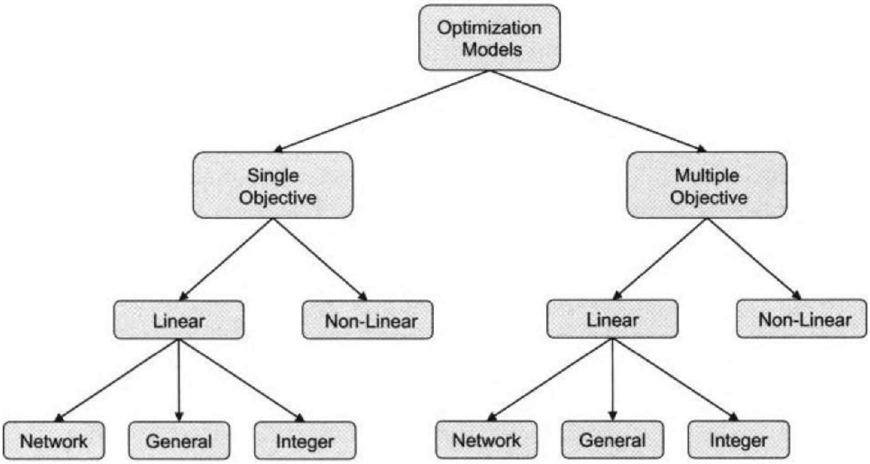


Figure 2.1. Taxonomy of optimization models

and so on. Simply put, the objective function and all constraints can be written as the sum of terms and each term is a constant multiplied by a single decision variable. In Excel, what this also means is that we are not using IF statements to represent logical conditions. For example, in locating ambulances to maximize the number of covered demands, one might be tempted to model the problem with one set of decision variables representing where we locate facilities. We might then determine whether or not a particular demand node was covered within the service standard using a set of IF statements that depend on the location variables. This would make the model non-linear and this approach should be avoided at all costs. Instead, we should use a separate set of decision variables to determine whether or not each demand node is covered. We will see how to do this in Chapter 3 when we discuss location problems explicitly.

It is useful to distinguish between linear and non-linear problems for a number of reasons. First, for linear problems (with the exception of integer linear problems discussed below), when we get a solution, we can be sure that we have obtained an optimal solution. For non-linear problems, we need to be much more careful before asserting with certainty that the solution is truly the best possible solution. Either we need to make fairly restrictive assumptions about the nature of the constraints and the objective function or we need to take extreme care in designing a solution algorithm for non-linear problems to ensure that we have the true optimal solution when the algorithm terminates. Second, linear programming algorithms are pervasive while algorithms for non-linear problems are less readily available, although this is changing rapidly over time. Third, linear problems tend to be easier to understand than are their non-linear counterparts. Finally, the size of the problem (as measured, for example by the number of decision variables or the number of constraints) that we can solve for linear problems is much larger than is the size we can solve for non-linear problems.

Despite the inherent advantages associated with linear optimization problems, there are times when a non-linear formulation is essential. For example, as we will see in Chapter 5, inventory costs typically are non-linear functions of the annual demand, the holding cost per item, and the fixed cost of placing an order. Trying to approximate such inherently non-linear functions in a linear manner will introduce errors. In many such cases, it is better to simply live with the non-linearity inherent in the problem.

Nevertheless, for most of the problems in this text, we will be able to use linear modeling techniques. A good rule is to try to linearize problems as much as possible. The discussion in section 2.7 shows some ways of using integer variables to represent logical conditions while still remaining in the linear family of models.

Within the linear family of models, in addition to the generic linear optimization problems, two classes of problems deserve special mention. Many problems that we will be interested in can be represented as network optimization problems. This is useful for a number of reasons. First, we can solve *very* large network optimization problems very quickly. To get some sense of this, consider the problem faced by Mapquest or Google Maps. When you ask for driving directions from your home in Dallas, Texas, to your university (in Ann Arbor, Michigan, for example), the computer behind the web page must solve a network optimization problem known as a shortest path problem (see section 2.6). Despite the fact that the problem may have thousands of variables, the directions seem to come back almost instantly. Furthermore, the web page is simultaneously determining not only the directions you requested, but also the directions for many other users all of whom are logged on to the site simultaneously. The fact that the underlying problem is a network problem enables this fast response. (In addition, the algorithms underlying such web pages use sophisticated heuristics or rules to eliminate potential solutions that are clearly going to be sub-optimal. For example, such algorithms may stop searching for shortest paths once they get more than a certain number of miles south of your home in Dallas if your destination is north of Dallas in Ann Arbor, Michigan.)

The second reason to focus attention on network optimization problems is that they can be represented graphically as a *network*. This means that we can often visualize problems with thousands or even millions of variables. The U.S. map is one visualization of the inputs to a shortest path problem, for example. As we will see below, the problem of assigning students to seminars in the Weinberg College of Arts and Sciences at Northwestern University is a network problem with over 10,000 variables. Nevertheless, we can readily draw a network that allows us to visualize what the problem looks like.

The third reason for focusing on network problems is that the structure of such problems is always the same, at least for our purposes. The objective will be to minimize (or sometimes to maximize) a linear function of the decision variables. The constraints will fall into two categories. First, for every node in the problem—think about a highway intersection as a node—we will require that the flow into the node equals the flow out of the node. Flow must be conserved at every node. This means, for example, that if we go into St. Louis, Missouri, on our way from Dallas, Texas, to Ann Arbor, Michigan, we must also leave St. Louis. The second set of constraints restricts the amount of flow on each link to a range of values. In other words, for some problems the flow on a link must be greater than or equal to some quantity and less than or equal to some other quantity. In the problem of assigning students to seminars, each student will be represented by a node and each seminar will be represented by a node. There will be a link into each student node and the flow on each of those links must be greater than or equal to one, meaning that each student must be assigned to a seminar. Similarly, there will be a link out of each seminar node and the flow on those links must be less than or equal to the capacity of the associated seminar.

The fourth reason for focusing on network problems is that if a problem can be structured as a network optimization problem, then we can be assured that the decision variables will take on integer values, as opposed to fractional values. This is a very useful property, as we will see.

Integer programming problems represent the other specialized form of linear optimization models that we need to consider. In general linear programming problems, there is no guarantee that the optimal value of a decision variable will be integer-valued. In fact, in many cases, the optimal solution will involve many non-integer-valued decision variables. For some contexts, this is fine. In other cases, however, we need to specify that we want only integer valued solutions. For example, if a variable represents whether or not we should locate an ambulance in the 2700 block of North Halsted in Chicago, a value of $\frac{1}{3}$ would not be useful. The value should either be 1, meaning we should locate an ambulance there, or 0 meaning we should not. Logical relationships between variables are often expressed using integer variables. For example, the condition that we cannot assign demands at node *j* to a facility at node *k* unless we have decided to locate at node *k* can be represented using integer variables. Unfortunately, if the problem structure is such that we are not guaranteed an all-integer solution, then solving the problem becomes more difficult and time-consuming. The good news is that asking Excel or any of the standard Excel add-ins to ensure that certain variables take on only integer values is easy to do; it just may take longer to get the solution.

2.4 YOU PROBABLY HAVE SEEN ONE ALREADY

Finally, before discussing linear programming in detail, it is worth noting that you have probably already seen and used optimization; you just may not have recognized what you were doing as solving an optimization problem.

Students of statistics are familiar with *least squares regression*. We should recognize this as an optimization problem by its very name: *least* squares regression. The problem is that of fitting a straight line to a set of observations.

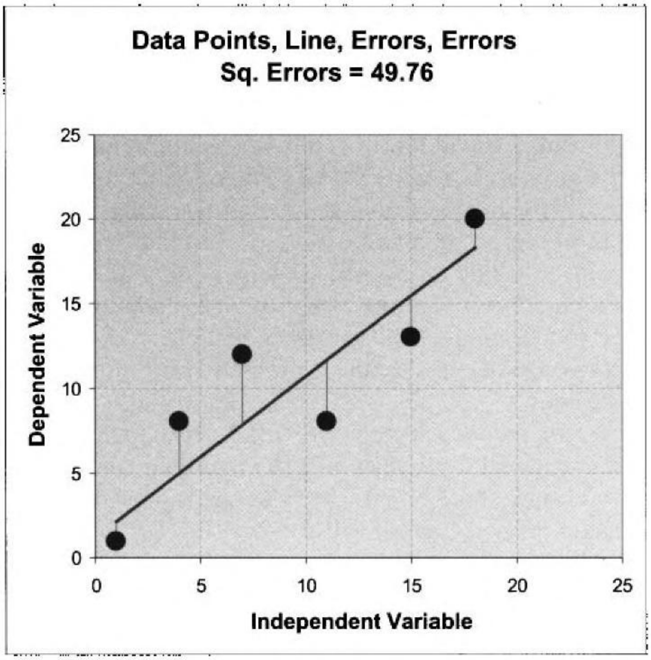


Figure 2.2. Sample data points, one possible line through the data, and the associated errors

The line is defined by its slope and intercept, which are the two decision variables.

Figure 2.2 shows six observations of an independent variable and the corresponding six values of the dependent variable. These are also shown in Table 2.1. Figure 2.2 also shows one possible line through the data. The line is $Y = 1.2 + 0.95X$, corresponding to a line with an intercept of 1.2 and a slope of 0.95. Figure 2.3 shows the "squared errors" in a manner similar to that used by Erkut and Ingolfsson (2000). On the surface, this line looks fairly good. There are three points above the line and three below the line. The errors below the line look about the same as the errors above the line.

See Squared Errors.xls in the online appendix.

However, we can do better using optimization or, as it is called in this case, least squares regression. The error associated with observation i, e_i, is given by

$$e_i = Y_i - (a + bX_i)$$

since $a + bX_i$ is the value on the line corresponding to observation i. If we now try to minimize the sum of the squared errors, we will be minimizing

In short, for those familiar with linear regression, we have already seen a non-linear optimization problem. We will see other non-linear models in Chapter 5 on inventory models. In the remainder of this chapter, however, we focus on linear models.

2.5 LINEAR PROGRAMMING

In this section, we more formally introduce linear programming. We will begin with a review of what we mean by linear in section 2.5.1. In the following section, we will illustrate linear programming with a simple two-variable problem that can readily be graphed. In section 2.5.3, we will present the canonical linear programming problem using algebraic notation. Section 2.5.4 discusses the dual of a linear program and its relationship to the main or primal problem we are trying to solve. Section 2.5.5 formulates a number of key problems as linear programming models. Finally, we show how to structure and solve some of these problems in Microsoft Excel.

2.5.1 What Is a Linear Program?

What exactly is a linear program? Linear programming does not have anything to do, necessarily, with computer programming, though almost every linear program of practical interest today is solved using a computer. Doing so sometimes requires some programming expertise in either a high-level programming language like C or C++ or in a specialized optimization language like AMPL (Fourer et al., 2002). In this text, however, we will explore smaller linear programming problems which can be solved readily in Excel using either the built-in Solver or one of a number of commercial add-in solvers for Excel such as What's Best (2009) by Lindo, Co.

Linear programming means that we are trying to minimize or maximize a linear function of some decision variables subject to constraints that are themselves linear functions of those, and perhaps other, decision variables. By linear we mean that we are not multiplying decision variables by one another, we are not squaring decision variables or taking the square root of decision variables, or more generally, raising a decision variable to any power (other than 1). We are also not raising a constant to a power that depends on a decision variable and so we are not using functions such as e^X, where X is a decision variable. Similarly, we are not dealing with the logarithm of decision variables and we are not dealing with trigonometric functions of decision variables. All we are doing is multiplying a decision variable by a known constant and then adding or subtracting similar such terms in both the objective function and in the constraints. Fortunately, many important problems can be structured as linear programming problems and others that cannot immediately be represented that way can be transformed into linear models or can be approximated by linear models.

2.5.2 Graphical Representation

To illustrate a simple linear programming problem, consider an abstraction of the problem faced by many small towns and cities in deciding how to allocate a budget for municipal services between police and fire protection. In this *very* simplified model, each police patrol costs $200,000 per year and each fire truck costs $1,000,000 per year including the cost of the fire station. The city has only $5,350,000 to allocate to the combined police and fire budgets. In addition, contracts with the unions representing the two city services stipulate that there must be at least 1.5 times as many police patrol units as there are fire trucks and that there cannot be more than 7.5 times as many police units as there are fire units.

The goal of the city is to allocate funds to maximize the number of lives saved over the course of a year. An outside consultant has told them that they can expect 0.2 lives saved per year per police patrol unit and 0.65 lives saved per fire truck.

Figure 2.5 shows the constraints on the problem. The heavy line defines the budget. All solutions below that are feasible in the sense that they do not require the city to spend more than its available budget of $5,350,000, though some of the solutions below this line violate one of the other two conditions. All solutions to the right of the uppermost line that slopes upward satisfy the condition that there must be at least 1.5 police units per fire truck, while all solutions above the

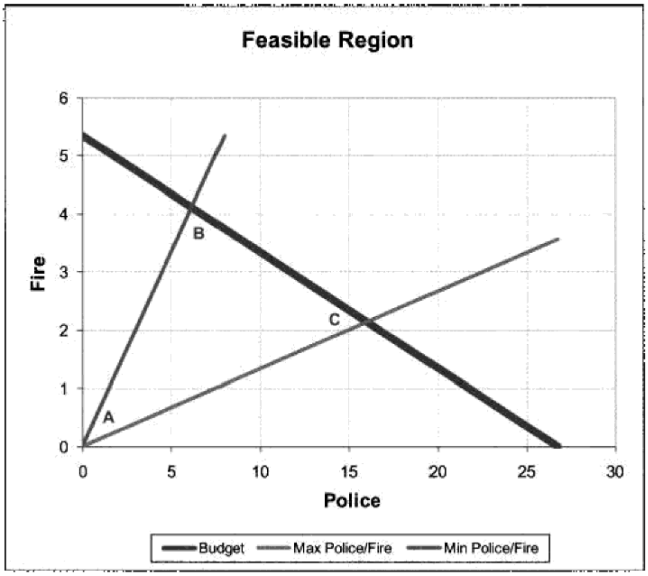


Figure 2.5. Feasible region for a simple linear programming problem

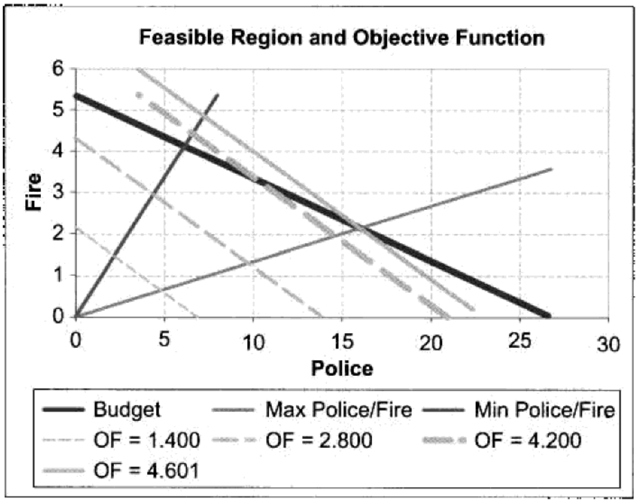


Figure 2.6. Feasible region and objective function contours

lower upward sloping line satisfy the condition that there are 7.5 or fewer police units per fire truck. Thus, the region enclosed by these three lines—the triangle defined by the points A, B, and C—represents the set of all combinations of fire and police units that satisfy all three constraints. This triangular region is the intersection of all of the linear constraints and is called the *feasible* region of the linear programming problem. Any solution to the problem must be inside or on the border of this region.

Figure 2.6 shows the feasible region with four parallel lines representing different values of the objective function. The smallest value of the objective function shown is 1.4. Clearly, there are better values of the objective function—maximizing lives saved—that we can attain and still be within the feasible region. If we double this value, we get to 2.8 and arrive at the second objective function line shown in the figure. Again, we can do better. If we move out to a value of 4.2—another 50-percent increase in the number of lives saved—we still have some solutions that are within the feasible region. Finally, when we move the line representing the objective function value a bit further to the right to a value of 4.601, we find that the line representing the objective function—the solid objective function line in this case—intersects the feasible region at only one point. This is the best we can do and we have now solved the linear programming problem graphically. The optimal objective function value is 4.601 and we should fund 16.05 police units (at a cost of $3,210,000) and 2.14 fire trucks at a cost of $2,140,000, completely expending the budget for the two city services. Furthermore, at this solution, there are exactly 7.5 police units per fire truck and so the constraint representing the maximum number of police per fire truck is also limiting our ability to improve the solution. These two constraints are said to be *binding* constraints. The third constraint stipulating that there be at least 1.5 police units per fire truck is not binding and does not constrain the optimal solution.

See Simple LP Example.xls in the online appendix.

Mathematically, we can write the problem as follows:

$$\begin{array}{llllll} Max & 0.2\cdot Police & +0.65\cdot Fire & & & (2.2) \\ s.t. & 200\cdot Police & +1000\cdot Fire & \le & 5350 & (2.3) \\ & -1.0\cdot Police & +1.5\cdot Fire & \le & 0 & (2.4) \\ & 1.0\cdot Police & -7.5\cdot Fire & \le & 0 & (2.5) \\ & Police & & \ge & 0 & (2.6) \\ & & Fire & \ge & 0 & (2.7) \end{array}$$

The objective function (2.2) maximizes the number of lives saved. The first constraint (2.3) is the budget constraint expressed in thousands of dollars. We have simply divided the cost of a police unit, the cost of a fire truck, and the budget by \$1000. The second constraint (2.4) says that we must have at least 1.5 police units per fire truck while the third constraint (2.5) ensures that we have no more than 7.5 police units per fire truck. Constraints (2.6) and (2.7) are *non-negativity* constraints and stipulate that we cannot have a negative number of either police or fire units.

For now, let us not be overly concerned about the fact that the solution tells us to hire a fractional number of police and fire units. We will return to this issue in section 2.7 below when we discuss *integer linear programming problems*.

The graphical view of the linear programming problem allows us to consider the impact of adding an additional constraint to the problem. Suppose the city did not want to hire more than 11 police officers. We would then have to add a constraint of the form

$$Police \le 11 \qquad (2.8)$$

Graphically, this changes the problem as shown in Figure 2.7. The old feasible region defined by points A, B, and C is no longer the correct feasible region. The triangle represented by C, D, and E has been eliminated from the feasible region and the new feasible region is the four-sided figure given by points A, B, D, and E. The previously optimal solution at C is no longer optimal. Instead, the optimal solution is found at point D, at which we employ 11 police officers and 3.15 fire units, resulting in an objective function value of 4.2475, which is less than the value at point C, which was 4.601.

Alternatively, suppose there was a restriction limiting the number of fire units to 3. The new feasible solution is shown in Figure 2.8. Again, the original

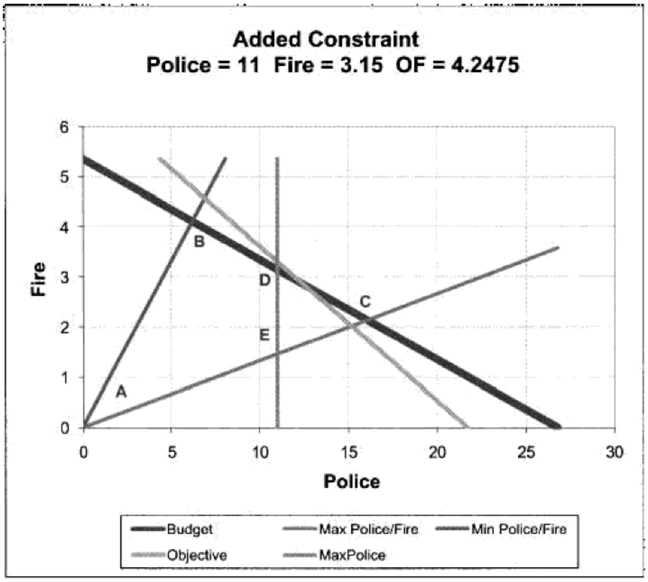


Figure 2.7. Adding a constraint that degrades the solution

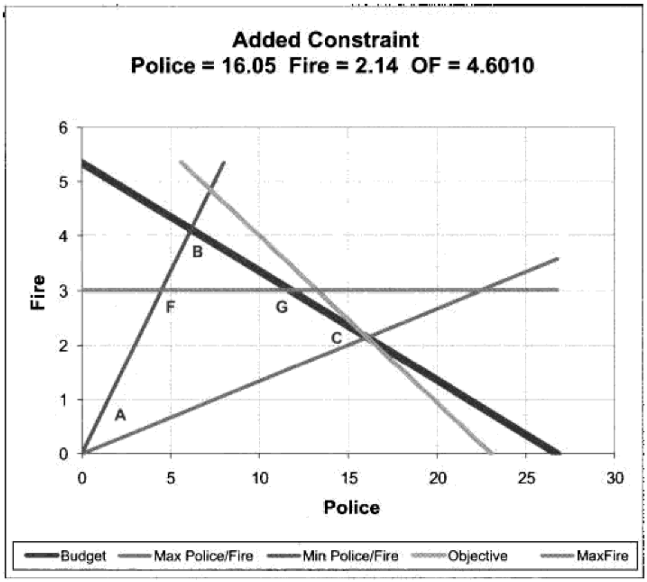


Figure 2.8. Adding a constraint that does not change the solution

Associated with each of the constraints in (2.10) is a dual variable, W_i. As in the primal problem, this variable must be non-negative as stipulated by constraint (2.18). The right-hand side constants in (2.10) become the coefficients of the dual decision variables in the objective function (2.16), which is now to be minimized. Similarly, the coefficients in the primal objective function (2.9) become the right-hand side values in the dual constraints (2.17). Finally, note that while the primal linear programming problem in canonical form was written with less than or equal to constraints, the canonical dual linear programming problem has greater than or equal to constraints. Note also that the coefficient matrix of the constraints (the a_{ij} terms) has been transposed. In our summation notation, this is indicated by the fact that we are now summing over the set of products, I, in the dual problem rather than over the set of inputs, J, as in the primal problem.

The dual problem also has an interpretation. The dual variable, W_i, gives the value to the objective function of having one more unit of resource i. It tells us how much the objective function would increase if we had one more unit of this resource. (See Rubin and Wagner [1990] for a discussion of the limitations of interpreting the dual variables too literally.) Another way of thinking of this is that W_i gives the value of resource i, or how much we should be willing to pay for one more unit of this resource, at the margin. This is why the dual variable is also referred to as the shadow price. The dual objective minimizes the value of all resources, or what we should pay at the margin for these resources. Constraint (2.17) states that the value of all resources that go into making one unit of product j must be at least as much as the profit associated with product j. Constraint (2.18) simply states that the dual variables or shadow prices cannot be negative.

Just as we could write the primal linear programming problem in standard form, so too can we write the dual linear programming problem in standard form by introducing surplus variables, T_j, into constraints (2.17). The dual problem in standard form is then:

$$Min \quad \sum_{i \in I} b_i W_i \tag{2.19}$$

$$s.t. \quad \sum_{i \in I} a_{ij} W_i - T_j = c_j \qquad \forall j \in J \tag{2.20}$$

$$W_i \geq 0 \qquad \forall i \in I \tag{2.21}$$

$$T_j \geq 0 \qquad \forall j \in J \tag{2.22}$$

The primal and dual problems are intimately related.

1. If the primal problem is infeasible—meaning that the intersection of all of the regions formed by constraints (2.10) and (2.11) is non-existent—then either
   a. the dual problem is also infeasible or
   b. the dual problem is unbounded, meaning that the dual objective function, (2.16), can be made arbitrarily small.

2. If the primal is unbounded, then the dual is infeasible.

   Generally speaking, however, we are interested in the cases in which both the primal problem and the dual problem have feasible solutions. In these cases, there are several other properties of interest to us.

3. If both linear programming problems are feasible, then for any feasible set of primal decision variables, X_j, and any feasible set of dual decision variables, W_i, we know that $\sum_{i \in I} b_i W_i \geq \sum_{j \in I} c_j X_j$. This is known as the *weak duality condition*. It states that for any feasible solution to the primal problem and any feasible solution to the dual problem, the value of the primal objective function—the one we are maximizing—will always be less than or equal to the value of the dual objective function—the one we are minimizing.

4. For the optimal solutions to the primal and dual problems, which we will denote by X_j^* and S_i^* for the primal and W_i^* and T_j^* for the dual, $\sum_{i \in I} b_i W_i^* = \sum_{j \in J} c_j X_j^*$. In other words, at the optimal solution to the primal and dual problems, the objective function values are exactly equal.

5. Furthermore, at the optimal solution, $X_j^* T_j^* = 0$ for all products *j*, and $W_i^* S_i^* = 0$, for all inputs or resources *i*. These conditions are known as the *complementary slackness* conditions. These conditions have an important interpretation. Consider the conditions $W_i^* S_i^* = 0$ on the resources. This can also be written as $W_i^* \left(b_i - \sum_{j \in J} a_{ij} X_j^* \right) = 0$ for every resource *i*. This says that if we do not use up all of resource *i*, meaning that $S_i^* > 0$ or $b_i - \sum_{j \in J} a_{ij} X_j^* > 0$, then the value of an additional unit of this resource must be 0 or $W_i^* = 0$. Conversely, if we are willing to pay a positive amount for an additional unit of this resource, meaning $W_i^* > 0$, then we must be using all of that resource that we currently have, or $S_i^* = 0$, or equivalently $b_i - \sum_{j \in J} a_{ij} X_j^* = 0$. The condition $X_j^* T_j^* = 0$ has a similar interpretation in terms of which products we produce. In short, if the value of the inputs required to produce product or service *j* exceeds the profit from the product or service, meaning that $T_j^* > 0$, then we do not produce any of that product or service $X_j^* = 0$.

6. At least one optimal solution occurs at a corner point of the feasible region, meaning at a point at which some set of constraints intersect.

7. The dual of the dual is the primal. This says that if we find the dual problem to the dual problem, we get back to the primal problem.

To illustrate the dual problem, consider again the simple problem of determining the number of police and fire units to deploy that we discussed in section 2.5.2. For ease of use, we repeat that problem here:

$$\begin{array}{lllcll}
Max & 0.2\cdot Police & +0.65\cdot Fire & & & \qquad (2.2)\\
s.t. & 200\cdot Police & +1000\cdot Fire & \le & 5350 \quad Total & \qquad (2.3)\\
 & -1.0\cdot Police & +1.5\cdot Fire & \le & 0 \quad MinPolice & \qquad (2.4)\\
 & 1.0\cdot Police & -7.5\cdot Fire & \le & 0 \quad MaxPolice & \qquad (2.5)\\
 & Police & & \ge & 0 & \qquad (2.6)\\
 & & Fire & \ge & 0 & \qquad (2.7)
\end{array}$$

Associated with constraints (2.3), (2.4), and (2.5) are dual variables *Total*, *MinPolice*, and *MaxPolice*, which are shown above next to each constraint. The dual problem associated with this linear programming problem is

$$\begin{array}{llllll}
Min & 5350\cdot Total & & & & \qquad (2.23)\\
s.t. & 200\cdot Total & -1.0\cdot MinPolice & +1.0\cdot MaxPolice & \ge 0.2 & \qquad (2.24)\\
 & 1000\cdot Total & +1.5\cdot MinPolice & -7.5\cdot MaxPolice & \ge 0.65 & \qquad (2.25)\\
 & Total & & & \ge 0 & \qquad (2.26)\\
 & & MinPolice & & \ge 0 & \qquad (2.27)\\
 & & & MaxPolice & \ge 0 & \qquad (2.28)
\end{array}$$

To solve this problem, we can use the complementary slackness conditions. We know that at the optimal primal solution, constraint (2.4) is not *binding* and so the associated dual variable, *MinPolice*, must be 0. The other two constraints are binding and so their associated dual variables can be greater than 0. Since we know that the dual of the dual is the primal, the dual variables associated with constraints (2.24) and (2.25) are the *Police* and *Fire* variables respectively. Since these variables are both positive at the optimal solution, we know that (2.24) and (2.25) must be satisfied by strict equality (with no slack). Therefore, we have to solve the following three equations to find the optimal solution.

$$\begin{array}{lllll}
200\cdot Total & -1.0\cdot MinPolice & +1.0\cdot MaxPolice & =0.2 & \qquad (2.29)\\
1000\cdot Total & +1.5\cdot MinPolice & -7.5\cdot MaxPolice & =0.65 & \qquad (2.30)\\
 & MinPolice & & =0 & \qquad (2.31)
\end{array}$$

The solution to this problem is $Total=\left[0.65+7.5\left(\frac{0.35}{12.5}\right)\right]\Big/1000=0.00086$, $MinPolice=0$, and $MaxPolice=\frac{0.35}{12.5}=0.028$. The objective function value for the dual problem is 4.601, which is exactly the value obtained in section 2.5.2 for the primal problem. The dual variable associated with the budget, *Total*, says that for every extra \$1000 in the budget, we will save an extra 0.00086 lives. (Recall that the budget constraint was written in terms of \$1000s.)

2.5.5 Example Problems

In this section, we show how two problems can be formulated as linear programming problems. The first is the problem of assigning students to seminars. The second is the problem of finding the shortest path between two cities.

2.5.5.1 Assigning Students to Seminars. We begin by showing how to formulate the problem of assigning students to seminars. For every student, we know the ranking that student gave to each seminar that the student ranked. In particular, let $STUDENTS$ be the set of students and let SEM be the set of seminars. Each student ranks a subset, SEM_j, of the seminars. For each seminar $k \in SEM_j$, we know the ranking, $rank_{jk}$ that student j gave to seminar k. The smaller the ranking, the more preferred the seminar is for that student. We will assign students only to seminars that they rank. We know the capacity, cap_k, of each seminar $k \in SEM$.

We define $ASSIGN_{jk}$ to be a decision variable that will equal 1 if student j is assigned to seminar k, and 0 otherwise.

The problem can now be formulated as follows:

$$Min \quad \sum_{j \in STUDENTS} \sum_{k \in SEM_j} rank_{jk} ASSIGN_{jk} \tag{2.32}$$

$$s.t. \quad \sum_{k \in SEM_j} ASSIGN_{jk} = 1 \qquad \forall j \in STUDENTS \tag{2.33}$$

$$\sum_{j \in STUDENTS} ASSIGN_{jk} \le cap_k \qquad \forall k \in SEM \tag{2.34}$$

$$ASSIGN_{jk} \ge 0 \qquad \forall j \in STUDENTS; \forall k \in SEM_j \tag{2.35}$$

The objective function (2.32) minimizes the sum of all assigned ranks over all students. The first constraint, (2.33), states that each student must be assigned to exactly one of the seminars that he or she ranks. The second constraint, (2.34), states that for each seminar, the number of students assigned to the seminar cannot exceed the seminar's capacity. Finally, constraints (2.35) state that the assignment variables cannot be negative.

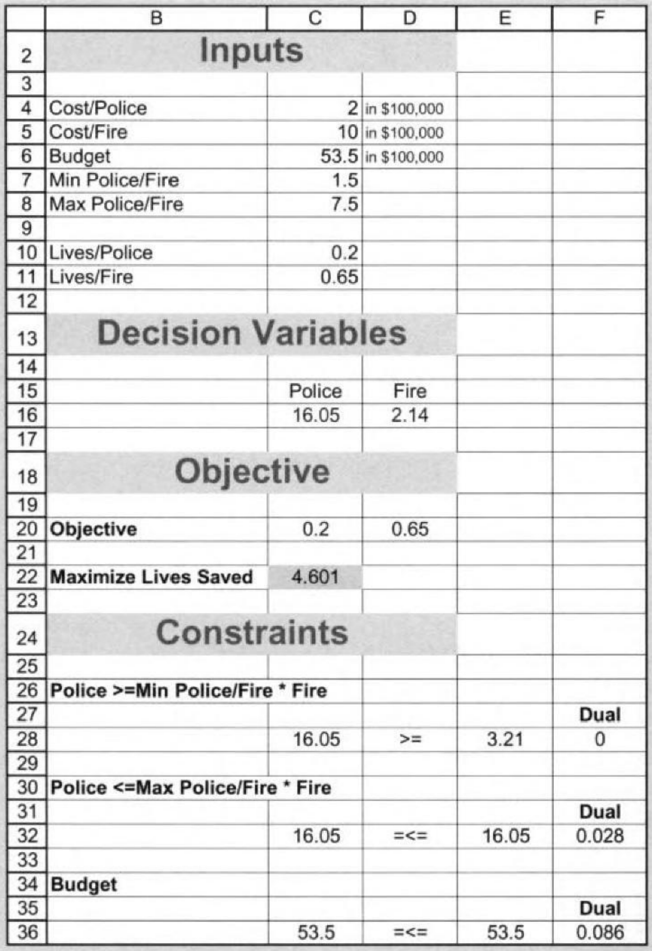
Note that while we defined the assignment variables as binary—either 0 or 1—in our verbal description of the problem, in the mathematical description of the problem we simply defined them as real, non-negative variables. We can do that **in this case** because this problem is a network linear programming problem, as discussed in section 2.6. For network problems, there will always be an all-integer solution and so we do not need to worry about forcing the assignment variables to be either 0 or 1. Furthermore, because each student must be assigned to exactly one seminar, we know that the assignment variable will not take on a value greater than 1.

Finally, we note that the problem may not have a feasible solution for two reasons. First, the selections made by the students may not admit a feasible solution. To illustrate this, suppose we have five students and three seminars, each with a capacity of two. Each student ranks two of the three seminars, indicating implicitly that he or she does not want the third seminar at all. If one of the seminars is not ranked by any of the five students, then we are left with trying to assign all five students to the remaining two seminars, which have a combined capacity of four students. Clearly, we cannot assign all students to only seminars that they rank in this case. Second, if the total number of students to be assigned exceeds the sum of the seminar capacities, then there can be no feasible solution. This situation can arise if each professor teaching a seminar can set the capacity of his/her seminar.

To handle both of these eventualities, we can create a new "dummy" seminar whose capacity is equal to the total number of students to be assigned. Students assigned to this seminar represent students who cannot be assigned to a "real" seminar. For each student, we add the dummy seminar to the set of ranked seminars by the student. The "cost" of assigning student j to the dummy seminar, $rank_{j,dummy}$, should be set to a very large number. For example, this value might be 10 times the total number of students times the number of seminars ranked by each student. In the example above, there are five students and each student ranked two seminars. Therefore, if each student gets his/her second choice, the objective function value would be 10. If we assign a value of 100 to each of the non-ranked seminars, we will quickly be able to tell whether or not any student was assigned to the dummy seminar. If the objective function value exceeds 100, we know this happened. Similarly, if there is not enough capacity in the real seminars, some students will have to be assigned to the dummy seminar. Again, the objective function value will exceed 100 in our simple example if there is not enough capacity (e.g., if there were seven students and only three seminars each with a capacity of two students).

This illustrates an important point about linear programming in particular and optimization in general. **We need to ensure that the model can deal with infeasibility caused by the inputs.** In addition, **we may need to protect against infeasibility caused by an excess of constraints.** For example, we might have additional constraints that stipulate that the ratio of men to women be between 0.5 and 2.0, meaning that there cannot be more than twice as many women as there are men or more than twice as many men as there are women in each seminar. If we add these constraints, along with other diversity-oriented constraints (perhaps regarding mixing of students by intended majors or by geographic background), we can rapidly generate problems without any feasible solution.

2.5.5.2 Shortest Path Problems. The shortest path problem is also a network linear programming problem, as we might expect. In this case, N defines a set of



	B	C	D	E	F
2	Inputs				
3					
4	Cost/Police	2	in $100,000		
5	Cost/Fire	10	in $100,000		
6	Budget	53.5	in $100,000		
7	Min Police/Fire	1.5			
8	Max Police/Fire	7.5			
9					
10	Lives/Police	0.2			
11	Lives/Fire	0.65			
12					
13	Decision Variables				
14					
15		Police	Fire		
16		16.05	2.14		
17					
18	Objective				
19					
20	**Objective**	0.2	0.65		
21					
22	**Maximize Lives Saved**	4.601			
23					
24	Constraints				
25					
26	**Police >=Min Police/Fire * Fire**				
27					**Dual**
28		16.05	>=	3.21	0
29					
30	**Police <=Max Police/Fire * Fire**				
31					**Dual**
32		16.05	=<=	16.05	0.028
33					
34	**Budget**				
35					**Dual**
36		53.5	=<=	53.5	0.086

Figure 2.9. Example budget allocation problem in Excel

Constraints—corresponding to the four key components of an optimization model.

The Inputs section contains all of the known data about the problem. In some problems, it is also useful to include information that can be computed directly from the known inputs in this section. Alternatively, we can add a separate section with a title "Computed Inputs" for such constants.

In the budget allocation example, the first six inputs deal with the budget, including the cost per police unit, the cost per fire unit, and the total budget. In this case, the cost inputs are in terms of $100,000. In addition, we have the key ratios for the constraints.

In addition to these inputs, this section of the worksheet also includes the coefficients for the objective function, which in this case are the numbers of lives saved per police unit and per fire unit.

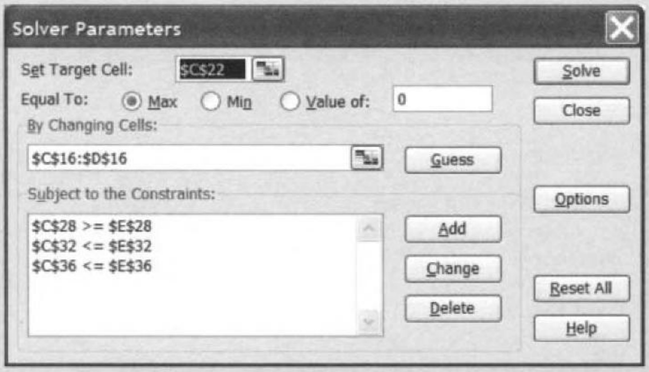
Just as it is good to name the sections of the spreadsheet, it is also good practice to name as many of the cells as possible. For example, cell C6 is named Budget and cell C4 is named Cost_Police.[1] Each of the cells in the inputs is named in the example shown in Figure 2.9.

The next section includes the Decision Variables. In this example, there are only two variables: the number of police and the number of fire units. Again, each is labeled (in this case with a name on top of the cell).

The next section is the Objective function. In this section, we compute the objective as a function of the inputs (lives saved per police and fire unit) and the decision variables (the number of police and fire units, respectively). Cells C20 and D20 repeat the objective function coefficients. Note that they refer to cells C10 and C11 respectively. The equation in cell C20, for example is '=Lives_Police', while that in cell D20 is '=Lives_Fire'. Thus when cells C10 or C11 change so too do cells C20 and D20.

The objective function uses the SUMPRODUCT function in Excel. The equation in cell C22 is '=SUMPRODUCT(C20:D20,C16:D16)'. Note that we needed to put the objective function coefficients into row 20 so that we could use this function as the SUMPRODUCT function requires that the number of rows and columns in the first block of cells be identical to the number in the second block. The SUMPRODUCT function multiples each element in the first block by the corresponding element in the second block and then adds the products. In our case, it multiplies 0.2 times Police and 0.6 times Fire and then adds the two products to give the objective function value.

[1] You can automatically name a group of cells or you can name each cell manually. To automatically name a group of cells, identify what is in each cell to the left (for example) of each cell. In the example of Figure 2.9, the labels are in cells B4 through B11. The cells identify what is to be placed in cells C4 through C11, respectively. Then, select cells B4 through C11, click Insert → Name → Create ... This will bring up a Create Names box with the Left column check box already checked in this case. Simply click OK and each cell in column C will be named with the label next to it in column B. Excel converts some characters (e.g., the / in cell B4) to an underscore (_). In Excel 2007, use Formulas → Create From Selection. To label cells manually, simply select the cell to be labeled and click Insert → Name → Define ... (Formulas → Define name or Formulas → Name Manager in Excel 2007.) This will bring up a Define Name box. In the area labeled Names in Workbook, type the name you would like to assign to the cell and then click OK. You can label groups of cells in this way. For example, you could label cells B10:D20 as Assignments if these correspond to assignment variables in some other worksheet. In all equations using named cells, the name of the cell will appear instead of the row and column reference. **Note that named cells behave like cells with locked references in equations.** In other words, cell C6 would be labeled Budget in Figure 2.9 and any equation using cell C6 would use Budget instead of C6. If you now drag that equation down a column, the reference will not change to C7 and then C8 in the next two rows; rather, it remains as Budget pointing to cell C6. In other words, it operates as though the equation had $C$6 instead of C6. In general, this is a good feature, as it avoids many mistakes. However, users should be aware of this when using named cells.



Set Target Cell: $C$22

Equal To: ● Max ○ Min ○ Value of: 0

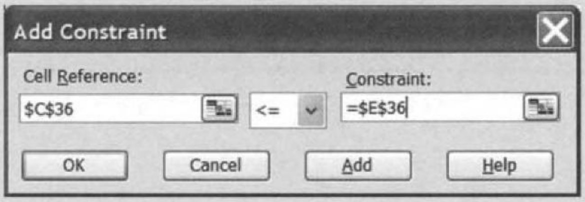
By Changing Cells:

$C$16:$D$16

Subject to the Constraints:

$C$28 >= $E$28
$C$32 <= $E$32
$C$36 <= $E$36

Figure 2.10. Solver Parameters dialog box

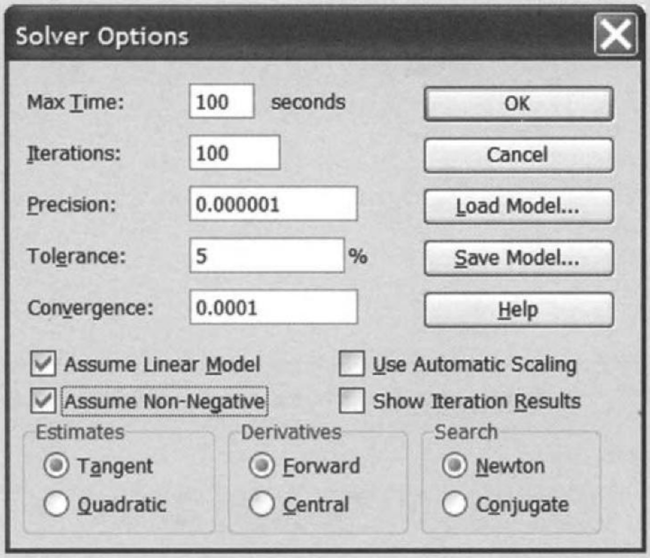


Cell Reference: $C$36 <= Constraint: =$E$36

Figure 2.11. Add Constraint dialog box in Excel Solver

The last section includes the constraints on the problem. Again, it is a good idea to label each constraint so that the user of the spreadsheet knows what each constraint is doing. The first constraint says that the number of police must be greater than or equal to the minimum number of police per fire unit (1.5) times the number of fire units. Cell C28 is simply equal to '=Police' while cell E28 is '=Fire*Min_Police_Fire'. Cells C32 and E32 are similar. Finally, cell C36 is '=Cost_Police*Police+Cost_Fire*Fire' while cell E36 is '=Budget'.

Built into Excel is a Solver. This allows the user to solve linear and integer-linear programming problems. To use the Solver, click Tools → Solver ... (Use Data → Solver ... in Excel 2007.) This will bring up a menu that looks like Figure 2.10. In the Set Target Cell block, you should identify the cell that contains the formula for the objective function. Below this, be sure to specify whether the objective is to be maximized or minimized. The By Changing Cells block contains the decision variables. If the decision variables are not in one contiguous block within the spreadsheet, the different blocks can be separated by commas. Finally, you need to add constraints to the model by clicking the Add button. This brings up a menu similar to Figure 2.11. This menu allows you to identify the left-hand side and right-hand side of each constraint (or of a group of constraints) and whether or not the left hand side should be greater than or equal to (>=) less than or equal to (<=) or simply equal to (=) the right-hand side value. Also, if



Max Time: 100 seconds

Iterations: 100

Precision: 0.000001

Tolerance: 5 %

Convergence: 0.0001

☑ Assume Linear Model ☐ Use Automatic Scaling

☑ Assume Non-Negative ☐ Show Iteration Results

Estimates: ● Tangent ○ Quadratic

Derivatives: ● Forward ○ Central

Search: ● Newton ○ Conjugate

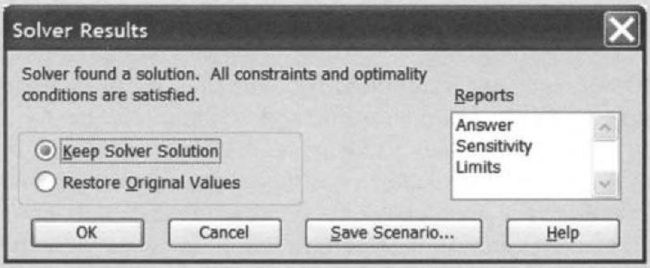
Figure 2.12. Solver Options dialog box

you give the menu a reference to only a left-hand side cell, which is a decision variable, you can specify that the variable must be integer-valued (int) or binary (bin). We will need this later in section 2.7.

After the constraints are fully specified, click OK on the Add Constraint menu to return to the Solver Parameters menu. Now click Options to bring up the Solver Options menu shown in Figure 2.12. Be sure that the boxes for Assume Linear Model and Assume Non-Negative are checked to indicate that the problem is a linear programming problem and that the decision variables are to be non-negative. When you click OK on this menu you are returned to the Solver Parameters menu of Figure 2.10. At this point, click Solve and the Solver will find the optimal values for the decision variables.

When the Solver has finished solving the problem, a menu similar to that shown in Figure 2.13 will be displayed. By highlighting one or more of the reports, you can ask Excel to show additional information, including the Sensitivity report, which gives the values of the dual variables that Excel refers to as the Shadow Prices. After clicking OK, the spreadsheet will update with the optimal values of the linear programming problem.

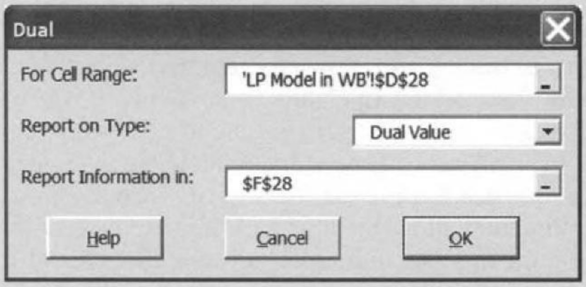
An alternate way of solving linear programming problems in Excel is to use an add-in such as What's Best (2010). One of the key advantages of What's Best over the Solver is that most of the model is shown directly on the spreadsheet and not in menus that are hidden until pulled up. The spreadsheet in Figure 2.9 illustrates this. Once the model is set up as described above, the user need only highlight the decision variables—cells C16 and D16 in Figure 2.9—and either


Solver Results

Solver found a solution. All constraints and optimality conditions are satisfied.

● Keep Solver Solution
○ Restore Original Values

Figure 2.13. Solver Results dialog box


Dual

For Cell Range: 'LP Model in WB'!$D$28

Report on Type: Dual Value

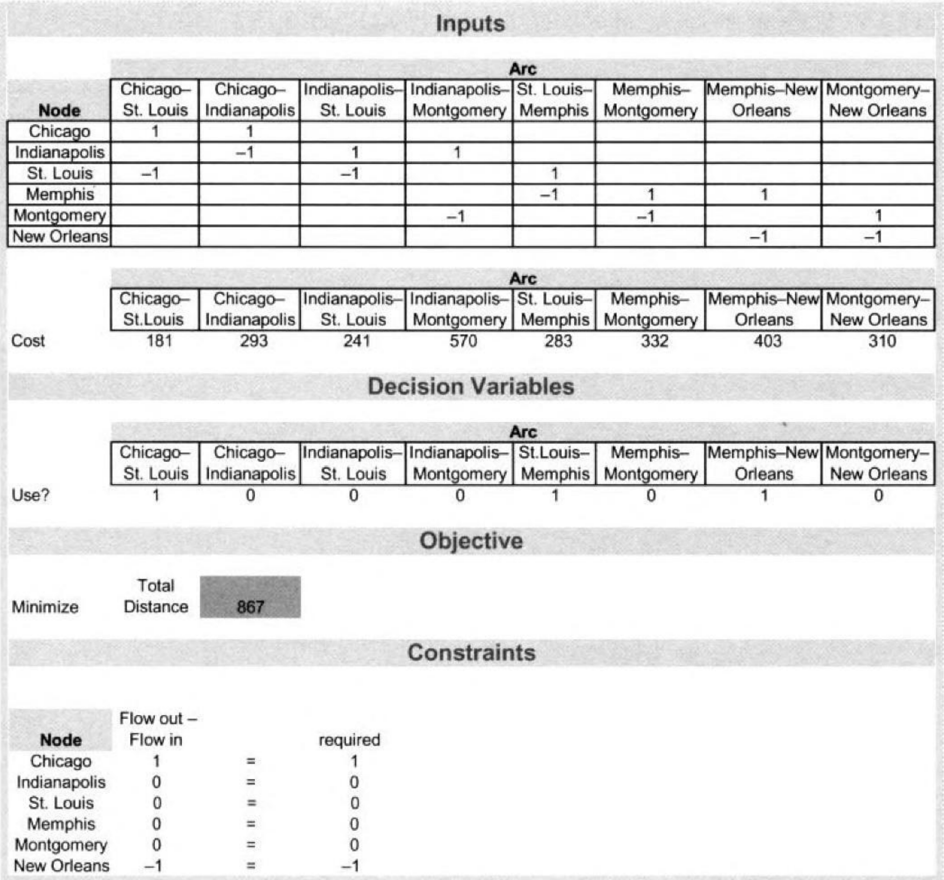
Report Information in: $F$28

Figure 2.14. Dual dialog box in What's Best

click on the What's Best k→x toolbar button or click on WB! → Adjustable ... to identify the cells as decision variables. What's Best then changes the color of the cells to blue. The user immediately can recognize these as decision variables. Next, you highlight the objective function (Cell C22 in Figure 2.9) and, using the What's Best up arrow toolbar button or the WB! → Best ... option, identify the cell as the cell to be maximized. Next, you identify the constraints. What's Best requires a space (a column or a row) between the left-hand side and the right-hand side of each constraint. Click on this cell and use either the appropriate What's Best toolbar button or WB! → Constraints ...

Finally, you can ask What's Best to compute the dual variables directly on the worksheet. To do so, click on the cell immediately to the right of the right-hand side of a constraint (e.g., cell F28 in Figure 2.9). Click on WB! → Advanced ... → Dual ... to bring up the Dual menu shown in Figure 2.14. The dual variable will be reported in cell F28. It refers to the constraint in D28. *Note that the For Cell Range should refer to the cell that has the inequality or equality symbol in it and not to the right-hand side of the constraint.*

Once the model is set up, click on the What's Best toolbar bull's-eye or click WB! → Solve to solve the model. The program will find the optimal solution and will produce a report in a separate worksheet.



Inputs

Node	Arc Chicago–St. Louis	Chicago–Indianapolis	Indianapolis–St. Louis	Indianapolis–Montgomery	St. Louis–Memphis	Memphis–Montgomery	Memphis–New Orleans	Montgomery–New Orleans
Chicago	1	1						
Indianapolis		–1	1	1				
St. Louis	–1		–1		1			
Memphis					–1	1	1	
Montgomery				–1		–1		1
New Orleans							–1	–1

	Arc Chicago–St.Louis	Chicago–Indianapolis	Indianapolis–St. Louis	Indianapolis–Montgomery	St. Louis–Memphis	Memphis–Montgomery	Memphis–New Orleans	Montgomery–New Orleans
Cost	181	293	241	570	283	332	403	310

Decision Variables

	Arc Chicago–St. Louis	Chicago–Indianapolis	Indianapolis–St. Louis	Indianapolis–Montgomery	St.Louis–Memphis	Memphis–Montgomery	Memphis–New Orleans	Montgomery–New Orleans
Use?	1	0	0	0	1	0	1	0

Objective

Minimize	Total Distance	867

Constraints

Node	Flow out – Flow in		required
Chicago	1	=	1
Indianapolis	0	=	0
St. Louis	0	=	0
Memphis	0	=	0
Montgomery	0	=	0
New Orleans	–1	=	–1

Figure 2.16. Solution to shortest path problem of Figure 2.15

nodes. There is one node for each student and one node for each seminar that is offered. If a student ranks a seminar then we connect the node representing that student to the node representing the seminar. In Figure 2.17 each student has ranked three seminars. A student's first-choice seminar is shown with a heavy black arc, his/her second choice seminar with a dashed arc and the student's third choice with a dotted arc. If a student does not rank a seminar, there is no arc between the student node and the seminar node. Note that of the four students shown in Figure 2.17, three ranked Seminar M as their first choice. Also, two of the four ranked Seminar 1 as their last choice and two ranked Seminar 2 as their last choice. The lower bound for each student/seminar arc would be 0 and the upper bound would be 1, meaning that a given student does not necessarily need to be assigned to any particular seminar, but a student can be assigned only once to any seminar. If we want to minimize the average ranking of the assigned seminars, where low rankings correspond to high preferences, we could assign a unit

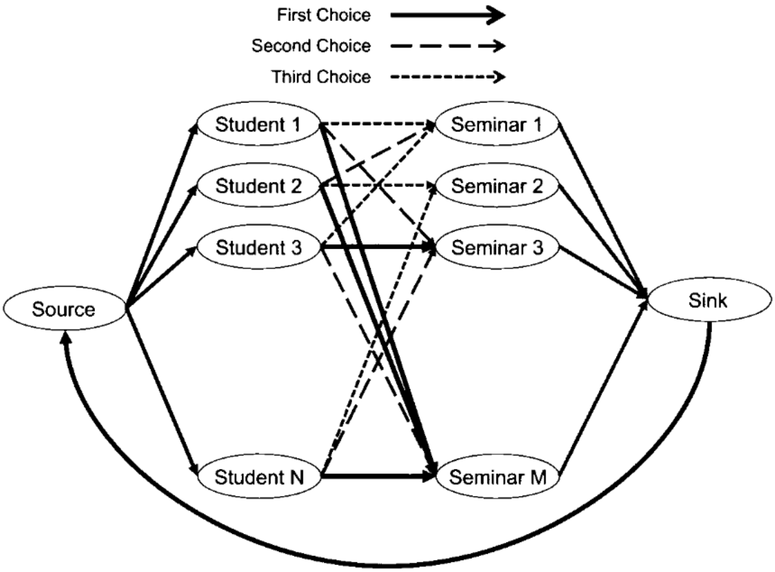


Figure 2.17. Network for student-seminar assignment problem

cost of 1 to the arc which corresponds to each student's first choice, 2 to the arc corresponding to the student's second choice, and so on.

Three other classes of arcs are shown in Figure 2.17 as well as two special nodes, the SOURCE and SINK nodes. The arcs connecting the SOURCE to each student node are where we enforce the constraint that each student is assigned to exactly one seminar. We do so by imposing a lower and upper bound of 1 on the flow on each such arc. This requires that we have a flow of 1 going into each student node. Since we will require the flow into each node to equal the flow out of each node, having a flow of 1 into each student node means that we have to have a flow of 1 out of each student node. In other words, we have to pick one of the arcs out of each student node, meaning we have to assign each student to exactly one seminar. The unit cost of each such arc is 0. The arcs between the seminars and the SINK node enable us to enforce the seminar capacity constraints. For the arc between Seminar *j* and the SINK, we impose a lower bound of 0 (we do not need to assign any students to the seminar), an upper bound equal to the capacity of the seminar, and a unit cost of 0. Finally, the return arc between the SINK and SOURCE nodes simply allows us to have a flow into each node equal to the flow out of each node. The lower bound on this arc can be 0, the upper bound can be equal to the number of students, and the unit cost should equal 0.

If we minimize the total cost in this network—the sum over all arcs of the flow on the arc times the unit arc cost–subject to the flows being between the lower and upper bounds for each arc and the condition that the flow into every node equals the flow out of every node, we will minimize the average rank of the student assignments. In other words, the solution will assign students to their most preferred seminars to the extent possible. From Figure 2.17, it should be clear that not every student is likely to be assigned to his or her first-choice seminar. In this simple example, if the capacity of Seminar *M* is 2 or less, we cannot simultaneously assign students 1, 2, and N to their first-choice seminar. Section 6.4 discusses the freshman-seminar assignment problem in greater detail.

Network linear programming problems are important for four reasons. First, many important problems can be structured as network problems. Second, being able to represent the problem as a network means that we can readily create a mental image of the problem and we can draw a figure that quickly captures the essence of the problem. Third, if we can represent the problem as a network flow problem with integer values for the lower and upper bounds for each arc and integer-values for the difference between the flow into and out of each node, then the optimal flows will be integer-valued when we solve the resulting linear programming problem. This means that we cannot get a solution like the one we found above in section 2.5.2, in which the optimal solution was to employ 16.05 police units and 2.14 fire trucks; these values are clearly not integer values. In other words, we do not need to worry about the solution to the student-seminar assignment problem of Figure 2.17 telling us that we should assign $\frac{1}{2}$ of Student 1 to Seminar 1 and $\frac{1}{2}$ of Student 1 to Seminar 3. Similarly, the solution to the shortest path problem of Figure 2.15 will always tell us to either use an arc, with a flow of 1, or to not use the arc, with a flow of 0. We will never get a "maybe" solution with a flow of $\frac{1}{2}$. Finally, there are exceptionally efficient and fast algorithms for solving network flow problems. MENU-OKF (available from the text's website) solves network flow problems using the out-of-kilter flow algorithm, one such specialized algorithm for solving network flow problems.

As a final example of a network flow problem, consider the problem of maximizing the flow of goods between an origin and a destination. For example, suppose that we wanted to maximize the number of volunteers we could send from Chicago to New Orleans for disaster relief. Perhaps each leg of the trip is limited by the number of seats in vans that we can hire. The number of seats available to us between each of the major city pairs is given in Figure 2.18. While this figure looks similar to Figure 2.15, in this case, the numbers beside each arc represent the maximum number of volunteers who can travel between the two cities per day.

Figure 2.18 includes a return arc from New Orleans to Chicago. If the unit cost on this arc is –1, there will be an incentive to move flow along this arc. This in turn will create an incentive to move flow through the primary arcs of the network from Chicago down to New Orleans. The lower bound on the return arc should be 0 and the upper bound should be a big number. It should be bigger than the maximum possible flow (which clearly is less than the sum of all of the arc capacities, or 2615 in this case). The lower bound on every other arc should be 0 and the upper bound should be the arc capacity shown in Figure 2.18. The unit cost on every other link should be 0. If we now minimize the total cost in this network while ensuring that the flow into every node equals the flow out of

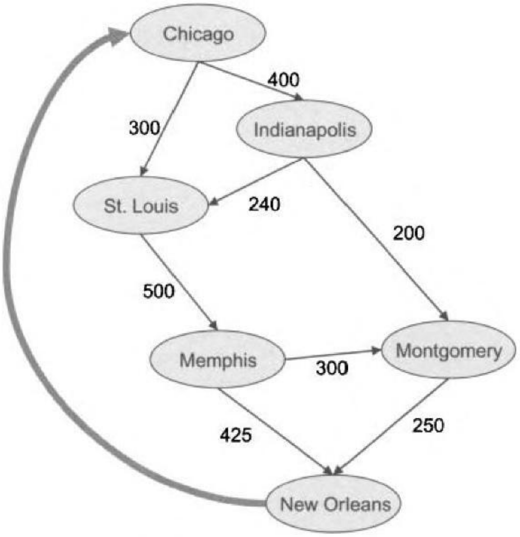


Figure 2.18. Figure for maximum flow problem

every node and that the arc flows are between their minimum and maximum values, we will be maximizing the flow from Chicago to New Orleans.

Maximum flow problems are another important category of network flow problems. Note that in this case, some arc costs are negative.

2.7 INTEGER PROBLEMS

In many of the problems that are of interest to us in the design and operation of services, fractional solutions from optimization problems are not acceptable. We have already encountered one such example. The simple linear programming problem of section 2.5.2 indicated that we should employ 16.05 police units and 2.14 fire units. Clearly this is not possible. It is tempting to simply round these two values down to the next smaller integer values. In other words, why don't we hire 16 police and 2 fire units? This would clearly satisfy the budget. After all, if we can afford 16.05 police and 2.14 fire units we can afford less of each. This highlights one of the problems, however, with rounding. If one of the two values to be rounded needed to be rounded *up*, there would be no guarantee that the solution would still satisfy the budget constraint.

In our case, rounding produces other problems. Recall that the number of police units must be less than or equal to 7.5 times the number of fire units. Hiring 16 police and only 2 fire units does not satisfy this condition, as the ratio is now 8:1 rather than 7.5:1, which is the maximum allowable ratio. Clearly, **rounding a linear programming solution to get the nearest integer solution may not produce**

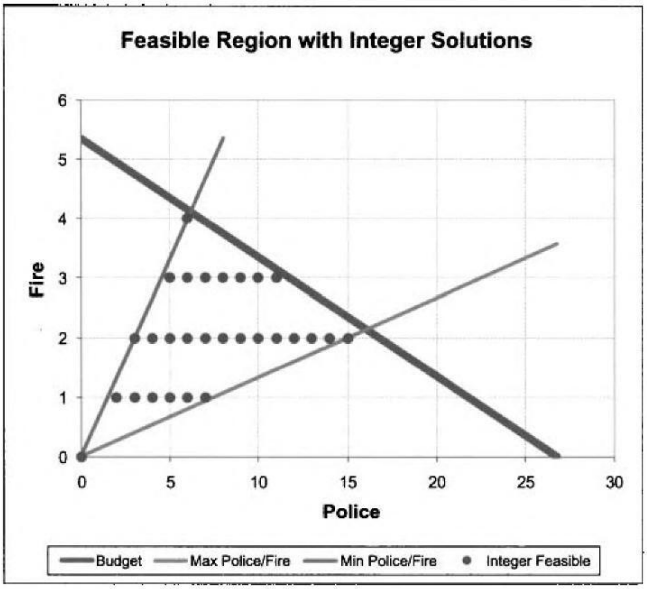


Figure 2.19. Feasible region with integer solutions highlighted

a feasible solution and is not likely to produce the best integer solution. Thus, we need to use some other technique.

The model that we should be solving in the police/fire example is the following.

$$
\begin{array}{llllll}
Max & 0.2\cdot Police & +0.65\cdot Fire & & & (2.2)\\
s.t. & 200\cdot Police & +1000\cdot Fire & \leq & 5350 & (2.3)\\
 & -1.0\cdot Police & +1.5\cdot Fire & \leq & 0 & (2.4)\\
 & 1.0\cdot Police & -7.5\cdot Fire & \leq & 0 & (2.5)\\
 & Police & & \geq & 0 \text{ and integer} & (2.39)\\
 & & Fire & \geq & 0 \text{ and integer} & (2.40)
\end{array}
$$

Note that constraints (2.6) and (2.7) have been changed to (2.39) and (2.40), in which we explicitly state that the decision variables must be integer-valued. In terms of the solution space, the feasible region is now shown in Figure 2.19 in which the feasible integer solutions are shown as black dots.

To inform the Excel Solver that the two decision variables need to be integer-valued, we need to add a constraint stating that these decision variables must take on only integer values. Looking back at Figure 2.9, we see that cells C16 and D16 contain the decision variables for the number of police and fire units to employ, respectively. Figure 2.20 shows how we can tell the Excel Solver that these two values must be integer valued. In What's Best, we highlight the decision increases from 4.3 to 4.35. However, this solution necessitates eliminating 3 of the 15 police units that we employed when the budget was $5,350,000. Eliminating civil service positions is exceptionally difficult. This illustrates one of the problems with implementing some integer programming solutions directly: The optimal solutions can be very sensitive to small changes in the constraint values. Note that the budget would have to increase from $5,350,000 to $6,000,000, more than a 12 percent increase, for the optimal solution to be to employ the same 15 police units plus an additional fire unit. This suggests that multi-year planning is essential in such situations so that we avoid being "painted into a corner." Such problems exist in other contexts as well. The manager of one plant once told me that 100 years of rational decision making had led to a disastrous plant layout. The firm had made good or optimal decisions for 100 years about where to place new equipment conditional on where everything else was at the time. A failure to plan ahead had resulted in a layout that was terrible even though the myopic decisions had all been well-founded. Such problems clearly also arise in service industries.

2.7.1 Uses of Integer Variables

We have just seen one of the important uses of integer variables in optimization. Sometimes a non-integer solution simply does not make sense. We cannot hire $\frac{1}{4}$ of a person. We also showed that rounding the solution may not give a feasible solution. When doing so does result in a feasible solution, there is no guarantee that it is truly the optimal solution for the integer programming problem. Nevertheless, in many practical cases, rounding a fractional solution to get a nearby integer solution often works well, at least as a starting point for discussions. As indicated in Figure 1.5, developing a mathematical model and exercising, solving, and analyzing the model results are only two of many phases in the modeling/decision making process.

The most common form of an integer variable is a *binary* integer variable that can take on a value of either 0 or 1 only. Such variables are typically used for two purposes:

- To represent decisions to either do something or not to do something
- To capture the relationships between key decisions.

In many problems, we have to decide whether or not to take certain actions. For example, in the student/seminar assignment problem, students may have been able to express a preference for any one of 70 different seminars, but we may have rooms, faculty, and resources for only 50 seminars. Thus, we must simultaneously determine *which* 50 seminars to offer (an either/or decision for each seminar) and which students to assign to the seminars we do offer. Note that we clearly would not want to enumerate every possible way of offering 50 seminars out of a total of 70 since there are over 10^{17} ways of doing so. (This is 100,000,000,000,000,000 possible ways!) If we could evaluate 10^9 or 1 billion combinations every second, it would still take us over 5 years to evaluate every possible way of offering 50 out of 70 seminars. The entire freshman class would have graduated long before we could decide which seminars to offer if we had to resort to a total enumeration approach like this. Fortunately, we can formulate and usually solve such problems quite well using integer programming models. For each seminar, we will have a *binary* (0 or 1) variable, which will be 0 if the seminar is not to be offered and 1 if it is. The model will automatically determine which 50 of these variables should take on a value of 1.

Similarly, we may have hundreds of candidate sites at which we can locate automatic meter readers for household electricity use (Gavirneni, Clark, and Pataki, 2004). For each location, we will have a binary variable, which will take on a value of 0 if we do not use that candidate site for a meter reader and 1 if we do use the site. The problem addressed in Gavirneni, Clark, and Pataki (2004) is to select the minimum number of meter reader locations needed to serve all homes subject to distance limitations between the reader and the home meter and a capacity limitation on the number of homes that can be assigned to any meter.

Binary variables are used in many other contexts. In routing vehicles, we may have a binary variable that is equal to 1 if customer *k* follows customer *j* immediately on a route and is 0 if not. In classroom scheduling, we might define room/time combinations representing a specific room and time of day. For example, the Monday, Wednesday, Friday time slot from 9 A.M. to 10 A.M. in lecture room 5 might represent one room/time combination. We could then define a variable that would be 1 if course *m* is assigned to room/time combination *n* and 0 if not; for example, the variable would be 1 if Introduction to Probability is to be taught in lecture room 5 from 9 to 10 on Monday, Wednesday, and Friday mornings.

In addition to making yes/no decisions, binary variables can be used to represent logical conditions that must exist between decisions. In the seminar assignment problem, for example, we cannot pick more than 50 seminars. In the problem of assigning classes to rooms and times, we must pick exactly one room/time combination for each class and we cannot assign more than one class to any particular room/time combination. In the vehicle routing problem, we would impose a condition that states that we go *from* every node to exactly one other node and that we go *into* every node exactly once. These are all examples of relationships between variables within the same class of decision variables. In other words, they represent logical conditions that are imposed on only one type of decision variable as opposed to decisions that cut across two or more types of decision variables.

Such cross-cutting conditions are also important. In the student/seminar assignment problem, in addition to a variable representing whether or not we offer each seminar, there will also be a variable for each student/seminar combination indicating whether or not a particular student is assigned to the particular seminar. One constraint will stipulate that we have to assign each student to exactly one seminar. The cross-cutting condition or linkage condition will ensure that we do not assign a student to a seminar unless we have decided to offer that seminar.

To illustrate how we can use binary (0 or 1) variables in this way, let us define the following two types of decision variables. The first will be a variable that allows us to decide whether to offer a seminar or not. Specifically, let us define X_j to be 1 if seminar j is to be offered and 0 if not, in the student/seminar problem. We also let Y_{ij} be 1 if student i is assigned to seminar j and 0 if not. We will now have constraints of the following form.

$$\sum_{j \in J} Y_{ij} = 1 \quad \forall i \in I \tag{2.41}$$

$$Y_{ij} \le X_j \quad \forall i \in I; j \in J \tag{2.42}$$

$$\sum_{j \in J} X_j \le p \tag{2.43}$$

In these constraints, I is the set of students and J is the set of candidate seminars that might be offered. Constraint (2.41) states that each student must be assigned to exactly one seminar. Note that this constraint applies only to the assignment or Y_{ij} variables. Constraint (2.42) states that the assignments can only be made to seminars that we offer. These are the linkage constraints that link the assignment decision variables, Y_{ij}, to the X_j decision variables indicating whether or not a seminar is to be offered. Finally, constraint (2.43) states that we can only offer p seminars because of budget or room availability constraints. Again, these constraints pertain to only one class of decision variables, the X_j variables. It is worth noting that in many problems, we will not have to require the assignment variables, Y_{ij}, to be binary; in many cases, they will naturally be either 0 or 1. We will generally have to require the decision variables regarding which seminars to offer, the X_j variables, to be binary.

In structuring models in Excel, it is generally a very *bad* idea to use IF statements to implement the constraints imposed by (2.42), despite the temptation to do so. The reason for this is that Excel treats IF statements as non-linear functions, which they potentially can be. Thus, what could be an integer-linear programming problem becomes a non-linear-integer programming problem, which is much harder to solve if we structure the problem using IF statements. Instead, you should set up the two classes of decision variables and you should link them explicitly using constraints of the form shown in (2.42).

In essence, constraints (2.42) represent a form of binary constraints that say that you can only do A if you also have done B, or equivalently, if you want to do A, you must also do B. In the seminar assignment case, these constraints say that you cannot assign student i to seminar j unless you decide to offer seminar j; that is, unless $X_j = 1$. Another way of saying this is that if you want to assign student i to seminar j, you must elect to offer seminar j.

Binary variables can also be used to represent problems in which we must do either A or B, but not both. If we let X_A equal 1 if we choose to do A and 0 otherwise, and if we define X_B similarly in terms of action B, then $X_A + X_B = 1$

requires that we take exactly one of the two actions but not both. This is an either/or constraint. For example, we may decide to offer pizza at our restaurant or to offer hamburgers, but not both since we do not have the kitchen space for both types of food. We must offer one or the other, however, or we will not have a restaurant (at least one that we want to consider operating). Similarly, $X_A + X_B \leq 1$ says that we can take *at most* one of the two decisions, A or B, but not both. We may also decide to take no action. For example, we might decide to assign a faculty member to the department's strategic planning committee (decision A) or to the faculty recruiting committee (decision B). The faculty member may, however, not be assigned to either committee. A single faculty member could not serve on both committees if this constraint is imposed.

Note that constraint (2.43) is a generalization of this form of condition or constraint. This constraint states that we can select at most p seminars. Thus, instead of saying that we can select either seminar A or seminar B, it states that we can select at most p of the seminars that were listed as candidate seminars to be offered.

2.7.2 How to Solve Integer Programming Problems

2.7.2.1 Branch and Bound. For the most part, we will not need to concern ourselves with the details of how the Excel Solver or What's Best solve integer programming problems. Once we state which variable(s) must be either integer or binary as discussed above, the software takes over and ensures that the requisite variables are not fractional in the solution that is returned. However, it is important to have some understanding of what the software is doing so that we can appreciate why integer programming problems may be much more difficult to solve than their linear programming cousins and why they may take much longer to solve.

Let us again consider the budget allocation problem shown in Figure 2.6. The optimal solution was to employ 16.05 police units and 2.14 fire units with an objective function value of 4.601. Clearly, this is not an integer-valued solution and it would be impossible to hire this many units of either type. Because the solution indicates that we should hire 16.05 police units, the optimal integer-valued number must be either less than or equal to 16 or greater than or equal to 17. Therefore, we create two new linear programming problems. In the first problem, we add a constraint that states that the number of police units must be less than or equal to 16 and in the second we add a constraint saying that the number of police units must be 17 or more. Figure 2.24 shows the results of adding these two problems. This is the beginning of what is called a branch and bound tree. We have branched on the Police variable, creating two new problems. In the problem on the left, we require the Police variable to be less than or equal to 16. As a result of adding this constraint, the number of fire units has increased to 2.15 and the objective function value has decreased slightly to 4.5975. Recall that adding a constraint to a maximization problem cannot improve the objective function, and in this case, it degraded the objective function value. The value of

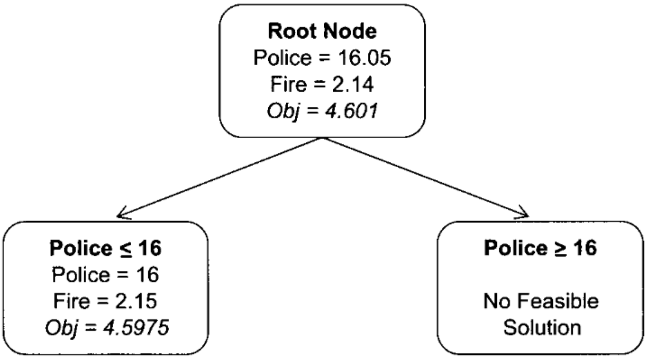


Figure 2.24. First branching in the branch and bound tree

4.5975 is an *upper bound* on the value of the objective function that we would find if we continue processing the left-hand node of the tree, which we are going to have to do since the number of fire units is still not integer-valued.

On the right-hand side, adding the constraint that the number of police units be greater than or equal to 17 creates an infeasible problem. This should also be evident from Figure 2.6 which shows that there is no solution with 17 or more police units. We can *fathom* this node of the tree as any additional constraints added to this problem will not allow the problem to be feasible. In short, we cannot have 17 or more police units.

We now return to the left-hand node. The number of fire units is 2.15 and so we now will branch on the Fire variable, adding two more new problems, one in which we use two or fewer fire units and one in which we use three or more fire units. The continuation of the branch and bound tree is shown in Figure 2.25. Note that both of the new problems inherit the constraints from the node (and all nodes) above. In particular, in solving the two new problems, we continue to impose the constraint that the number of police units must be less than or equal to 16.

When we limit the number of fire units to 2 or less, the optimal solution is to employ 15 police units and 2 fire units with an objective function value of 4.3, as shown in Figure 2.25. This is an all-integer solution and so we know that 4.3 is the worst that we can do. In other words, we now have a *lower bound* on the integer programming objective function.

When we require that the number of fire units be 3 or more, we employ 11.75 police units and 3 fire units, resulting in an objective function value of 4.3 (again) as shown in Figure 2.25. This is still not an integer-valued solution, and so we might be tempted to continue branching from this node. However, in this case, the value of 4.3 is an *upper bound* on the value of any solution that we might find if we continue adding constraints to problems below this node. No solution below this node can do better than an objective function value of 4.3. Since we already have an all-integer solution with a value of 4.3, we know that the solution

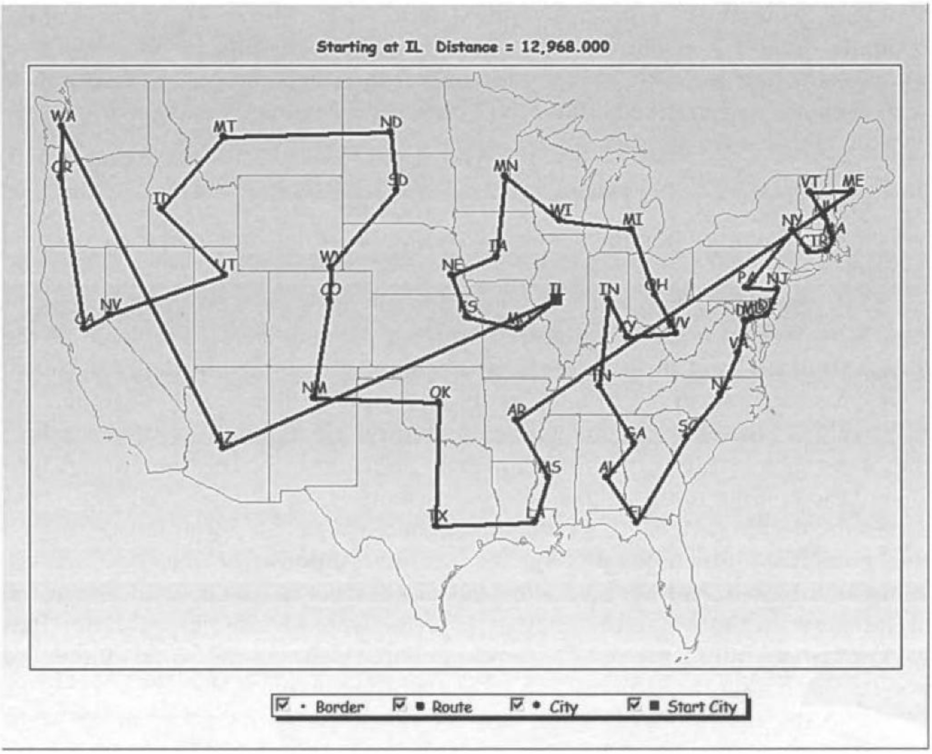


Figure 2.26. Nearest neighbor solution starting at Illinois

states plus Washington, DC. We begin and end the tour in Illinois. We begin by going from Illinois (IL) to Missouri (MO) and from there to Kansas (KS) and so on. The last location visited is in Arizona (AZ) and we have to return from there to IL to complete the tour, incurring a distance of 1311 miles, over 10 percent of the total tour length. Also note that the tour frequently crosses itself. For example, in going from Maine (ME) to Arkansas (AR), the tour crosses itself. It does so again on the final two legs of the tour.

In fact, if we start the algorithm from each of the 49 locations, the tour always crosses itself. The average length of the tours resulting from these 49 applications of the nearest neighbor algorithm is 13,240 miles. The best solution results from starting in Arizona (AZ); this tour has a length of 12,434 miles. On average, the last link in each of these tours, the link connecting the last visited node to the original starting node, is 1147 miles, or 8.63 percent of the tour length. This is roughly four times as long as the average link length. (Because there are 49 links in each of these tours, each link will be about 2 percent of the total tour length.)

See Animated GIF WithOut2Opt.gif in the online appendix.