Women in Engineering and Science

Alice E. Smith *Editor*

Women in Industrial and Systems Engineering

Key Advances and Perspectives on Emerging Topics

Springer

Women in Engineering and Science

Series Editor
Jill S. Tietjen
Greenwood Village, Colorado, USA

More information about this series at http://www.springer.com/series/15424

Alice E. Smith
Editor

Women in Industrial and Systems Engineering

Key Advances and Perspectives on Emerging Topics

Springer

Editor
Alice E. Smith
Department of Industrial
and Systems Engineering
Auburn University
Auburn, AL, USA

ISSN 2509-6427 ISSN 2509-6435 (electronic)
Women in Engineering and Science
ISBN 978-3-030-11865-5 ISBN 978-3-030-11866-2 (eBook)
https://doi.org/10.1007/978-3-030-11866-2

© Springer Nature Switzerland AG 2020
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

This volume is dedicated to my parents, Lois Elizabeth Krutsch Chupp and John Paul Chupp, now both deceased. My mother was the first in her family of immigrants from Russia to attend college, and while she was a mathematics student at Purdue University, she had the good fortune to meet Dr. Lillian Gilbreth. My father came from a humble Swiss Amish farming stock, but his father had advanced himself by earning a doctorate at Cornell University and was then a professor there for many years. My father, a PhD in chemistry, was proud when I earned my doctorate, making three generations of PhDs. My parents' examples and encouragement facilitated my career in engineering academia, and I am grateful to them for that, as well as for so much more.

Foreword

Margaret L. Brandeau is the Coleman F. Fung Professor of Engineering and Professor of Medicine (by Courtesy) at Stanford University. My research focuses on the development of applied mathematical and economic models to support health policy decisions. My recent work has examined HIV and drug abuse prevention and treatment programs, programs to control the opioid epidemic, and preparedness plans for public health emergencies.

My undergraduate studies were at MIT, in mathematics. I followed in my father's footsteps to MIT. However, while he studied electrical engineering, I chose math, a subject I have always loved. I was finished with my degree by the end of junior year, but did not want to graduate so soon, so I started taking some interesting applied mathematics and systems analysis classes. Then, I found out that the courses I was taking would fulfill the requirements for a master's degree in operations research—a discipline I had never heard of—so I also earned an MS degree in operations research. After working for 2 years, I moved to Stanford, where I earned a PhD in Engineering-Economic Systems—again taking interesting classes in applied mathematics and systems analysis. Along the way, I published a number of papers about the projects I was working on. I didn't realize it then, but this was great preparation for being a faculty member.

I joined the Stanford faculty in 1985 and have been there ever since, working on interesting problems with amazingly talented students and colleagues. I wouldn't change a single day!

Broadly speaking, industrial engineering focuses on determining how best to organize people, money, and material to produce and distribute goods and services.

Industrial engineering has its roots in the industrial revolution in the mid-18th to early nineteenth century. As production shifted from small enterprises to large-scale factories, and the production of goods became increasingly mechanized and specialized, factory owners realized that improving the efficiency of these new production processes could reduce waste and increase productivity.

One of the first scientific studies of work processes was *The Principles of Scientific Management* by Frederick Taylor (1911). Taylor, who is known as the father of industrial engineering, set forth principles for organizing, planning, and standardizing work. Around this time, a young man named Frank Gilbreth, who had started a job as a bricklayer's helper, began to study the practices of different bricklayers, trying to determine "the one best way" to perform the task. In 1904, he married Lillian Moller, an engineer who also became his work partner in their business and engineering consulting firm.

Lillian Moller Gilbreth was one of the first women engineers to earn a PhD (in psychology). She worked for many years applying industrial engineering concepts such as time and motion studies to improve work processes, first with her husband and then on her own for many years after his death. Her work emphasized a human approach to scientific management. During her career, Lillian Gilbreth published numerous books and papers, some with her husband and some on her own. If Frederick Taylor is the father of industrial engineering, Lillian Gilbreth is surely the mother of industrial engineering.[1]

From this beginning nearly 100 years ago, it is wonderful to see an entire volume of work by women industrial engineers. Since those early days, industrial engineering has of course changed, and this is reflected in this volume. Once focused on factory control, industrial engineering now focuses more broadly on both manufacturing and services. Once focused on techniques such as time and motion studies and Gantt charts, industrial engineering now includes a wide range of modern computational and analytical techniques.

In this volume, 59 women (and 3 male coauthors) present their work in 25 chapters covering such diverse topics as logistics costs in warehousing, container depot operations, multimodal transportation systems, price contracts in manufacturing, crop cultivation, food supply chains, healthcare operations, patient safety, clinical decision-making, disease modeling, and education. Methodologies discussed in these chapters are similarly broad and include human factors engineering, statistics,

[1]Lillian Gilbreth also had 12 children. Her family life was famously immortalized in the book *Cheaper by the Dozen*, written by two of her children, Frank Gilbreth Jr. and Ernestine Gilbreth Carey. Growing up, this was one of my favorite books. In 1994, I had the great pleasure of meeting Ernestine.

decision analysis, graph theory, simulation, optimization, stochastic modeling, and machine learning.

Industrial engineering has come a long way since its beginnings on the shop floors of England. Looking to the future, services are forming an ever-increasing share of economic output, both in the United States and elsewhere. Entire industries are being rapidly transformed via analytics and computation. Digitization and machine learning in the workplace are changing the nature and structure of work and the nature and structure of organizations. Automation and robotics have replaced many jobs once done by people. Increasing numbers of people are employed as "knowledge workers." Digital platforms that allow for spontaneously matching customer needs with available resources are becoming more pervasive. Industrial engineering has evolved and will continue to evolve in the face of these and other changes.

I hope that, as industrial engineering evolves, the numbers and roles of women in industrial engineering will also continue to evolve. The field of industrial engineering has been greatly enriched by the contributions of women. Women bring a diversity of experiences and viewpoints and, often, creative new ways of solving problems. This book showcases the work of 59 such women. I hope that many more amazing women will contribute to solving the important problems of the future—and help us determine how best to organize people, money, and material to produce and distribute goods and services in our changing world.

Contents

Dedication v

Foreword vii

Part I Background

1 Lillian Moller Gilbreth: An Industrial Engineering Pioneer 3
Jill S. Tietjen

Part II Analytics

2 Emergence of Statistical Methodologies with the Rise of BIG Data 27
Nedret Billor and Asuman S. Turkmen

3 Specifying and Validating Probabilistic Inputs for Prescriptive Models of Decision Making over Time 49
Sarah McAllister Ryan

4 Towards a Stable Graph Representation Learning Using Connection Subgraphs 71
Saba A. Al-Sayouri and Sarah S. Lam

5 Parameter Tuning Problem in Metaheuristics: A Self-Adaptive Local Search Algorithm for Combinatorial Problems 93
Cigdem Alabas-Uslu and Berna Dengiz

6 A Partition-Based Optimization Approach for Level Set Approximation: Probabilistic Branch and Bound 113
Zelda B. Zabinsky and Hao Huang

Part III Education

7 Modeling Engineering Student Success Needs 159
Tracee Walker Gilbert, Janis Terpenny, and Tonya Smith-Jackson

8 A Study of Critical Thinking and Cross-Disciplinary Teamwork in Engineering Education 185
Hulya Julie Yazici, Lisa A. Zidek, and Halcyon St. Hill

Part IV Health

9 Healthcare Teams Can Give Quality Patient Care, but at Lower Environmental Impact: Patient-Centered Sustainability 199
Janet Twomey and Michael Overcash

10 Improving Patient Care Transitions at Rural and Urban Hospitals Through Risk Stratification 211
Shan Xie and Yuehwern Yih

11 To Be Healthy, Wealthy, and Wise: Using Decision Modeling to Personalize Policy in Health, Hunger Relief, and Education 233
Julie Simmons Ivy, Muge Capan, Karen Hicklin, Nisha Nataraj, Irem Sengul Orgut, Amy Craig Reamer, and Anita Vila-Parrish

12 Improving Patient Safety in the Patient Journey: Contributions from Human Factors Engineering 275
Pascale Carayon and Abigail R. Wooldridge

13 Advanced Medical Imaging Analytics in Breast Cancer Diagnosis ... 301
Yinlin Fu, Bhavika K. Patel, Teresa Wu, Jing Li, and Fei Gao

14 Decision-Making in Sequential Adaptive Clinical Trials, with Implications for Drug Misclassification and Resource Allocation 321
Alba C. Rojas-Cordova, Ebru K. Bish, and Niyousha Hosseinichimeh

15 Calibration Uncertainty and Model-Based Analyses with Applications to Ovarian Cancer Modeling 347
Jing Voon Chen and Julia L. Higle

Part V Logistics

16 Contributions to Humanitarian and Non-profit Operations: Equity Impacts on Modeling and Solution Approaches 371
Burcu Balcik and Karen Smilowitz

17 Simulation-Based Approach to Evaluate the Effects of Food Supply Chain Mitigation and Compliance Strategies on Consumer Behavior and Risk Communication Methods 391
Jessye Talley and Lauren B. Davis

18 Contributions of Women to Multimodal Transportation Systems 417
Heather Nachtmann

19 Combining Exact Methods to Construct Effective Hybrid Approaches to Vehicle Routing 435
Rym M'Hallah

20 Modeling and Analysis of the Port Logistical Business Processes and Categorization of Main Logistics Costs 457
Carla Vairetti, Rosa G. González-Ramírez, Luisa Fernanda Spaggiari, and Alejandra Gómez Padilla

21 Using Simulation to Improve Container Depot Operations 487
Jimena Pascual and Alice E. Smith

Part VI Production

22 Sustainability and Life Cycle Product Design 517
Deborah Thurston and Sara Behdad

23 Dynamic Price and Lead Time Quotation Strategies to Match Demand and Supply in Make-to-Order Manufacturing Environments 541
Esma S. Gel, Pinar Keskinocak, and Tuba Yilmaz

24 Oyster Mushroom Cultivation as an Economic and Nutritive Alternative for Rural Low-Income Women in Villapinzón (Colombia) 561
Natalia Vargas, Carmen Gutierrez, Silvia Restrepo, and Nubia Velasco

25 Data-Driven Intelligent Predictive Maintenance of Industrial Assets 589
Olga Fink

Index 607

Part I
Background

Chapter 1
Lillian Moller Gilbreth: An Industrial Engineering Pioneer

Jill S. Tietjen

Contents

1.1 Introduction 3
1.2 Early Years 4
1.3 The One Best Marriage 5
1.4 On Her Own 11
1.5 Honors and Awards 19
References 22

1.1 Introduction

Career interest tests. The butter dish, egg tray, and vegetable and meat drawers in your refrigerator. The pump and return water hose on your washing machine. The foot pedal trash can. The design of the kitchen triangle. Accommodations for disabled people. What do these all have in common? They are the legacy of the "First Lady of Engineering" also called "The Mother of Industrial Engineering," "the Mother of Ergonomics," and "the greatest woman engineer in the world," one of the founders of the field of industrial engineering, Lillian Moller Gilbreth. She and her husband, Frank Gilbreth, are considered two of the cornerstones of the field of industrial engineering—a branch of engineering that is concerned with optimizing complex systems, processes, and organizations. Frank's focus was "The One Best Way" to do any task or series of tasks. Lillian's strength was bringing the social sciences to bear in combination with the mathematical and physical sciences.

Popularized in books and movies as the mother of 12 children (*Cheaper by the Dozen* and *Belles on Their Toes*), Gilbreth (see Fig. 1.1) was not only a mother but also a significant force as a pioneering woman industrial psychologist and engineer. Her story is fascinating and too rarely known (Des 2013; Lancaster 2004).

J. S. Tietjen (✉)
Technically Speaking, Inc., Greenwood Village, CO, USA

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_1

Fig. 1.1 Portrait of Lillian Moller Gilbreth. Courtesy of Walter P. Reuther Library, Wayne State University

1.2 Early Years

Growing up in a conventional household of the day, Lillian Moller (Gilbreth) (1878–1972) was the oldest of nine children, expected to conform to what was then deemed proper behavior for women. Very gifted academically, she was able to convince her father to let her attend the University of California while she lived at home and cared for the family. The first woman in the university's history to speak at commencement in 1900, Lillian received her B.A. in literature at the top of her class (although she did not make Phi Beta Kappa due to her gender). After briefly attending Columbia University, she reentered the University of California, earned her master's degree in literature in 1902 and began work on her Ph.D. (Des 2013; Proffitt 1999).

As was also common for women of her social class of her day, Lillian took a trip abroad before delving too deeply into her doctoral work. While in Boston preparing to board her ship, the chaperone for the trip—Miss Minnie Bunker, who was a teacher in the Oakland, California schools—introduced Lillian to her cousin Frank, who owned a construction business. Frank Bunker Gilbreth, who had not attended college and whose passion was finding the "One Best Way" to do any task, and Lillian became enamored with each other. Frank and Lillian decided to marry—embarking on the One Best Marriage—which involved a sharing of home life and work life. After their engagement was announced but before their marriage—Lillian on the West Coast, Frank on the East Coast—Lillian was already editing Frank's manuscripts for publication and critiquing his advertising brochures which he sent to her for this exact purpose. She edited his confidential management booklet "Field

System," reorganized it, fixed the grammar, and added an index. They were married in October 1904 (Des 2013; Lancaster 2004; Proffitt 1999; Yost 1943; Gilbreth Jr 1970).

1.3 The One Best Marriage

After their marriage, it became apparent that Lillian's selected areas of study for her Ph.D.—English and Comparative Literature—were not going to work for the couple's idea of shared work life. Instead, she became Frank's engineering apprentice, learning the types of work he used in his construction business. That education began in earnest on their honeymoon. As their family began to grow, much of that apprenticeship actually occurred at home. And, it was decided that her Ph.D. would be in the field of industrial psychology (Proffitt 1999; Yost 1943, 1949).

As Frank wrote about his original work methods, Lillian served as editor, thus learning the business thoroughly. She also took care of all client calls. In addition, Lillian was the researcher. She also went on site visits and met Kate Gleason, one of the very few, if not the only, woman heading an engineering company at the time, during a visit to Rochester, New York. Lillian located and sifted through the materials that would be incorporated in Frank's speeches at universities and at pitch meetings to clients. Her role of editor and writer was such that she should have been acknowledged as the co-author of Frank's books including *Concrete System* (1908), *Bricklaying System* (1909), and *Motion Study* (1911). *Concrete System* and *Bricklaying System* were two books that Lillian insisted be written to document methods already in practice on Frank's jobs and to expand the Gilbreth system. *Bricklaying System* described what Frank called "Motion Study" to cut product costs and increase efficiency. Frank and Lillian said "Motion Study" should be applied to all industries so that workers and management would share the benefits (Des 2013; Lancaster 2004; Yost 1943; Gilbreth Jr 1970).

Lillian also became convinced that human beings in the industry needed to be approached through psychology. The tragic fire at the Triangle Shirtwaist Factory in 1911 further reinforced her belief that the workers needed to be considered and she worried that much damage had been done through the introduction of efficiency mechanisms without consideration of the cost to human beings (Des 2013; Lancaster 2004; Yost 1943).

Her ideas began to appear in these works. For example, in *Motion Study* (1911), there is mention of a worker's physiology as well as his temperament and contentment in the factory. Further, workers needed adequate tools, an environment that was pleasing, some form of entertainment, and a clear understanding of the reward and punishment system in place. [These same ideas appear in Lillian's first doctoral dissertation.] In *Field System* (1908), employers were encouraged to set up suggestion boxes and to ensure that workers had periodicals that would provide mental stimulation. All workers, including factory hands, office workers,

schoolteachers, homemakers, farmers and store clerks—a much broader range of "worker" than incorporated in the new field of Scientific Management—were included in the Gilbreths' writings (Des 2013).

Their home became their office and laboratory. Their children shared family responsibilities that included investigating the One Best Way. These included the One Best Way for dusting, for setting and clearing the dinner table and for washing dishes, among others. During summers, their efforts were filmed and the children could watch themselves to determine how to do a task more efficiently and in less time. Sometimes the tasks applied to work projects that Frank and Lillian were working on—such as the best way to pack soap in boxes. One time, it involved picking berries—which turned out to be one of the earliest films ever made to show motions in agricultural processes. Another time it involved touch typing. The children tested two theories—one of which involved color coding keys and fingers—and went to school with multi-colored fingernails! (Yost 1943; Gilbreth Jr 1970).

Each person was expected to participate according to his or her aptitudes and abilities. The 3-year-old participated, but only to the extent that worked and made sense. Lillian believed that personal capabilities were a sacred trust that each individual should develop. She helped management and workers understand the benefits of collaboration and to accept the responsibility for working together and not at odds with each other. She became an expert in the areas of worker fatigue and production. Her expertise and insights were of great benefit as these were the years during which scientific management was being developed and just coming into general use (Proffitt 1999; Yost 1943).

Lillian's remarks at the Tuck School of Dartmouth College for the first Conference on Scientific Management in 1911 at which she was probably the only female presenter, where she reported on the key tenets of her first Ph.D. dissertation, offered the perspective that humans were the most important element of engineering tasks and thus psychology needed to be considered by industrial engineers in putting together their programs. She had been introduced for her turn at the podium as "We have all been watching the quiet work of one individual who has been working along lines apparently absolutely different from those being followed by another worker in the scientific management field and I wonder if Lillian Gilbreth would like to say a few words about her work" (Des 2013; Proffitt 1999; Yost 1943, 1949). Lillian remarks included (Lancaster 2004; Graham 1998):

> *I did not expect to speak in this place but I feel as though I must. I feel that the gap between the problems of academic efficiency and industrial efficiency, which is after all only an apparent gap, can be easily closed if only we will consider the importance of the psychology of management. I spent several years examining and studying it and it seems to me that Scientific Management as laid down by Mr. Taylor conforms absolutely with psychology. Principles of vocational guidance may be studied along psychological lines to train the individual so he will know exactly what he does want to do. It is the place of the colleges to train the man so that when he comes into his work there will be no jar. Since the underlying aim is the same and since psychology is the method by which we are all getting there, isn't it merely a question of difference of vocabulary between academic work and scientific work? Why not bridge this gap and all go ahead together?*

The audience was receptive to her comments and the Dartmouth Conference has been referred to as Lillian's "coming out." The audience consisted of manufacturers and businessmen and provided an opportunity for a dialogue with engineers who had applied the principles of scientific management to industrial operations (Des 2013; Proffitt 1999; Yost 1943, 1949).

Lillian was interested in keeping people happy and eliminating antagonistic behavior, as well as such questions as: if a trained pianist makes a faster typist than an untrained pianist, what skills are transferable? In addition, she advocated for having workers be responsible for ideas for greater efficiency as well as training others on the new techniques (today what we call "buy-in"). Her emphasis on the worker's psychology made the Gilbreths different from the other scientific managers and Frank's emphasis on motion instead of time looked more humane than other methods in vogue at the time (Des 2013; Proffitt 1999; Yost 1943).

Their process of evaluating work efforts relied on basic elements or "therbligs" (very close to Gilbreth spelled backwards). These elements (which had associated colors and symbols) are [Graham]:

- Search
- Find
- Select
- Grasp
- Transport Loaded
- Position
- Assemble
- Use
- Disassemble
- Inspect
- Preposition for Next Operation
- Release Load
- Transport Empty
- Rest for Overcoming Fatigue
- Wait (Unavoidable Delay)
- Wait (Avoidable Delay)
- Plan

This framework could be applied to any job whether in the classroom, kitchen, or at an industrial plant. In a breakthrough application, the Gilbreths relied on films from moving picture cameras that recorded movements of workers with a clock in the frames. A film could be run over and over again, run in slow motion, stopped, backed up. Machines could be redesigned to be operated far more safely and with less fatigue to the workers. Chairs could be what we today call "ergonomically" designed at a height to fit the motions of the operator so to keep him/her off his/her feet and to reduce fatigue. They also measured a worker's movement spatially as well as with lights to make time-exposed photographs. Their motion studies included process charts, micromotion photography, and chrono-cyclegraphs. This led to "The One Best Way"—the least taxing method moving the fewest parts of the

body in the least amount of space in the fastest time. Frank considered micromotion study his most important contribution to motion study because it could be used in any field—from surgery to loading a machine gun. The micromotion, chronocyclegraph, and therbligs comprised the Gilbreth system (Des 2013; Gilbreth Jr 1970; Yost 1949; Goff 1946).

The Gilbreths introduced the process (or flow) chart to the field of scientific management. The charts show graphically the arrival of materials at a factory and each step in the "process" as those materials move from piece of equipment to piece of equipment and result in a final product. Such a chart quickly and obviously demonstrates bottlenecks and any backtracking that occurs. Like process charts, the Gilbreths invented the chrono-cyclegraphs—this time to study the motions of speed typists. A small flashing light was attached to a hand, finger, or moving part of a machine and then a time-exposure picture was taken of the entire cycle. The result of the time-exposure photography was a dotted white line on a black background with the path of motion in two dimensions. It was possible to take pictures stereoscopically to create three-dimensional images. Then the Gilbreths were able to determine time and speed as well as acceleration and retardation (Gilbreth Jr 1970).

Lillian became well known as an outstanding psychologist in scientific management circles and was asked to present her views and work in print and at meetings around the world. Her dissertation was completed and much to the Gilbreths' surprise, rejected by the University of California at Berkeley in 1912 as university officials decided that her residency requirement had not been waived (Des 2013; Gilbreth 1990). She instead published the work serially under "L.M. Gilbreth" in the journal *Industrial Engineering* as publishers were not willing to have what would be obviously a woman-authored work published in the field. In 1914, it finally was published in book form (authored by L.M. Gilbreth whom Frank humorously said he was related to "only by marriage") as *The Psychology of Management: The Function of the Mind in Determining, Teaching and Installing Methods of Least Waste* and then reprinted in 1917 and 1918. Its "human" approach to scientific management attracted immediate attention (Des 2013; Lancaster 2004; Proffitt 1999; Yost 1943; Gilbreth Jr 1970; Graham 1998).

A key quote from the book is: *The emphasis in successful management lies in the man, not the work. Efficiency is best secured by . . . modifying the equipment, materials, and methods to make the most of the man. . . with knowledge will come ability to understand the rights of others . . . lead the way to the true Brotherhood which may some day come to be* (Gilbreth Jr 1970).

In 1921, when she became an honorary member of the Society of Industrial Engineers (the first honorary female member and the second in total—Herbert Hoover being the first), it said (Des 2013; Yost 1949; Trescott 1983):

> she was the first to recognize that management is a problem of psychology, and her book, *The Psychology of Management*, was the first to show this fact for both the managers and the psychologists. . . Today it is recognized as authoritative.

Her book was the first time that anyone had brought together the basic elements of management theory including (Trescott 1983):

1. Knowledge of individual behavior
2. Theories of groups
3. Theories of communication
4. Rational bases of decision-making

She dealt with problem-solving, decision-making, planning, communicating, measuring, and evaluating in various work and managerial environments. Her book focused on group behavior and the coordination of activities among work groups. She pioneered in considering the individuality of the work and focused on individual teaching, incentives, and welfare. She even dealt with how workers feel; a topic that had received little attention up to that point in time in scientific management literature. Lillian set the groundwork for further developments in modern management including the field of human relations (Des 2013; Lancaster 2004; Graham 1998).

In 1912, the Gilbreths moved to Providence, Rhode Island to work on a project with the New England Butt Company which manufactured braiding machines for shoelaces and wire insulation—one of the leading firms in this field. Lillian enrolled at Brown University intent on completing her doctorate in the area of applied management—a program of study created especially for her. She would have to do additional coursework and write a new dissertation. When completed, this second dissertation was titled "Some Aspects of Eliminating Waste in Teaching." She received the degree in 1915. Now that she had a Ph.D. after her name, her name as well as Frank's could appear on their professional papers; 50 were produced in the next 9 years. She was the first of the scientific management pioneers to earn a doctorate. And, she was now Phi Beta Kappa, having been elected an alumna of the University of California at Berkeley (Des 2013; Lancaster 2004; Gilbreth Jr 1970).

Her efforts in validating teaching (as the focus of her dissertation), even though it was women's work, was important for the work she would become known for later in her career. At her home, she established a micromotion laboratory for her experiments with women. These initial experiments involved making beds and, similarly to her Ph.D. work with teachers, would become important later in her career (Des 2013).

In 1914, Lillian and Frank started the Summer School of Scientific Management. Here, students learned new ideas about management with an emphasis on the study of motion and psychology. The school filled the need that the Gilbreths saw to teach the teachers—bridging the gap between the academic and the practical. Professors were invited to attend so that they could obtain information on scientific management and then use that material to develop courses at their home college or university. They were exposed to the theories that Lillian had espoused in her first doctoral dissertation: that the psychological element was the most important one in scientific management and that workers needed to be taught properly in order for scientific management to succeed. She addressed overfatigue from a psychological viewpoint as well as insecurity that resulted from work inconsistencies or foremen

who did not value the workers. Frank addressed the developing partnership between his motion study approach and her psychological approach: "I did not know anything about psychology until I was married, and Mrs. Gilbreth told me the courses she had taken. The new animal psychology that has been put out by Professor Thorndike and Professor Colvin has quite revolutionized the whole thing, and I believe we are going to see that the psychology in this management is the big thing." The school operated for 4 years (Des 2013; Lancaster 2004; Proffitt 1999; Yost 1949; Graham 1998; Gilbreth 1998).

She considered herself and was considered by others to be an expert in fatigue study, in the study of skill and its transference among industries and jobs, in precision in measurement, and in standardization of the work of both managers and laborers, as well as in the psychological areas (Trescott 1983). Their book, *Fatigue Study*, which came out in 1916, includes both of their names as authors and was written primarily by Lillian. A second edition was published in 1917, and it sold more copies than their other books. Many steps were recommended to improve productivity by minimizing fatigue. As much as could be individualized to the worker was seen as key—chairs, footrests, armrests, and an adjustable workbench. Additional steps to reduce fatigue included improved lighting, sensible clothes, supplies located close at hand, and regular rest periods. The book concluded: "The good of your life consists of the quantity of 'Happiness Minutes' that you created or caused. Increase your own record by eliminating unnecessary fatigue of the workers!" (Des 2013; Lancaster 2004; Gilbreth Jr 1970).

The years prior to and during World War I called on the strengths of the Gilbreths as industries and the country geared up for national defense. Believing that women would be called on to support the industrial efforts, they wrote papers on how to reorganize work and make it more efficient. Their motion study work was deemed by the press to be the Gilbreths' patriotic contribution to the country (Des 2013).

As part of their joint work, the Gilbreths showed how work could be adapted so that a disabled person could perform jobs that previously had only been considered possible for able-bodied individuals. This was particularly important in the aftermath of World War I and the many returning disabled veterans, especially amputees. It was also important to provide employment opportunities for individuals who had been injured in industrial accidents [Perusek]. The Gilbreths focused on individuals with disabilities and ways to make work environments more accommodating so that such individuals could be productive and efficient. Their work, *Motion Study for the Handicapped*, was published in 1920, after the war and after Frank had recovered from a significant illness (Des 2013).

Frank suffered his illness while he was in the Army at Fort Sill, Oklahoma. Prior to getting sick, his job was to make training films on efficient ways to conduct the business of the artillery such as loading a rifle or caring for a horse. Even in this work, Lillian was his unofficial "Advisor on the Project." She suggested ways to make the films more effective and how to deal more tactfully with officers and soldiers. When she was elected to honorary membership of the Society of Industrial Engineers in 1921, the citation recognized her contribution in this area: "acted as Consulting Psychologist in the field, working under the general Staff, standardizing

the methods for teaching the 4,000,000 officers and men" (Des 2013; Lancaster 2004).

Lillian began lecturing independently on industrial psychology and time and motion studies. Her first solo appearance was at the Massachusetts Institute of Technology in 1918—her talk was titled "The Place of Motion Study and Fatigue Study in Industrial Development." A few months later, she substituted for Frank at a meeting of the American Society of Mechanical Engineers and spoke about the use of motion study films to retrain disabled veterans. During the rest of her career, she would make hundreds of speaking appearances and occasionally broadcast on the radio (Lancaster 2004). She was also invited to lecture at universities around the USA including Stanford, Harvard, and Yale (Proffitt 1999).

During her lectures, she used the family and humorous events to illustrate her speeches:

> *In our family, we make a game out of Motion Study and we all try to see how we can cut down our own motions. This is especially important in the mornings when you have seven or eight children to get ready for school. One of my young sons insists he could improve his efficiency by at least fifty per cent, if we could eliminate baths and replace the back stairs with a fireman's pole. And a young daughter who has the job of setting the breakfast table says the One Best Way to do her job is to have everybody go out into the panty and get his own dishes and silver. Her suggestion, which was rejected by a ten-to-one vote in our Family Council, bears out what our cook and handyman says about us, I fear. He says the Gilbreth System is to get everyone else to do your work for you. (Gilbreth Jr 1970)*

1.4 On Her Own

After Frank Gilbreth dropped dead of a heart attack in 1924, Lillian had the responsibility of educating the 11 surviving children, all under the age of 19, and carrying on the business of Gilbreth, Inc. And, carry on, she did. She decided that she would work for the acceptance of the Gilbreth System and its creation of Happiness Minutes for workers and the disabled (Gilbreth Jr 1970).

She went to Europe in the summer of 1924 as she and Frank had planned and gave a talk on "First Steps in Fatigue Study." She published her biography of Frank *The Quest of the One Best Way*. She continued with the Family Councils (meetings with the children and people who ran the household) and they decided to stay in Montclair, New Jersey. She completed consultancy work for European clients. But, all was not so easy in America (Lancaster 2004).

With Frank gone, Lillian was exposed to sexism, some blatant, some more subtle. Major clients gave notice that they would not be renewing their contracts. They were not willing to have a woman—no matter her level of competence—upsetting their factories. She could not install the Gilbreth System in their plants due to her gender (Gilbreth Jr 1970).

She was paid less than Frank had been when she lectured at universities. She was turned away from venues where she was an invited guest or speaker due to their male-only rules. Although some men whom she had worked with or knew

professionally were willing to help her professionally, in general, engineers (who were almost all men) were not going to hire Lillian to install the Gilbreth system in their plants. She was going to have to use a different approach (Des 2013; Lancaster 2004).

One possibility was suggested to her by a vice president at Johnson and Johnson—she could teach at a school of motion study for Johnson and Johnson managers. Through press releases and letters, she described the course she would be offering: (Des 2013; Proffitt 1999; Gilbreth Jr 1970; Yost 1949; Graham 1998)

> *. . . to prepare a member of an organization, who has adequate training in scientific method and in plant problems, to take charge of Motion Study work in that organization. The advantage of this Course is that his understanding of both plant problems and of plant psychology usually insures cooperation and is a great assistance both in teaching and maintaining the better methods involved by the Motion Study investigation. The cost of the course is $1000. This Course can [as] desired be supplemented by a certain amount of subsequent teaching, inspection or consultation on the Motion Study problems of the organization sending the student. We also furnish reports and recommendations which are in the nature of a Survey, based upon more or less extended investigations of members of our staff. These indicate possible savings and outline methods.*

Her first class in 1925 included managers from Johnson and Johnson, Borden Milk and Barber Asphalt. The next semester students came from as far away as Germany and Japan. Motion study techniques were being taught to "disciples" of sorts and being spread around the world. She continued her classes—a total of seven courses over a period of 6 years—until she saw that engineering schools were now teaching time and motion complete with laboratories outfitted with photographic devices and movement measurement tools (Des 2013; Proffitt 1999; Gilbreth Jr 1970; Yost 1949).

Lillian knew that membership in professional societies was needed for peer recognition and she believed that membership in the American Society of Mechanical Engineers (ASME) was imperative. She wrote to the Membership Committee to determine if her application for membership would be received favorably. Although initial reaction was mixed, she lectured at several ASME technical sessions including in December 1925 when she spoke on "The Present State of Industrial Psychology." She was admitted to full membership in July 1926 (Lancaster 2004; Graham 1998; Perusek 2000). Her description of her work experience in her application for membership in the American Society of Mechanical Engineers reads:

> *I was also engaged in the perfecting of the methods and devices for laying brick by the packet method, and in the design and construction of reinforced concrete work. This work had to do with the management as well as the operating end.*
>
> *In 1914 our company began to specialize in management work. I was placed in charge of the correlation of engineering and management psychology, and became an active member of the staff making visits to the plants systematized in order to lay out the method of attack on the problems, being responsible for getting the necessary material for the installation into shape, working up the data as they accumulated, and drafting the interim and final reports. I was also in charge of research and teaching, and of working up such mechanisms, forms and methods as were needed for our type of installation of scientific management, motion study, fatigue study and skill study. These had to do not only with the handling of men, but with the simplification and standardization of the machinery and tools, for the use*

of both the normal and the handicapped. During Mr. Gilbreth's frequent and prolonged absences, both in this country and abroad, I was in responsible charge of all branches of the work. This was also the case while he was in the service, and while he was recovering from his long illness incurred therein.

Since Mr. Gilbreth's death, June 14, 1924, I have been the head of our organization, which consisted of Consulting Engineers and does work in management, and I have had responsible charge of the research, installation and the teaching, in this country and abroad. (Trescott 1983)

She began to make a name for herself in the field of industrial psychology. The editors of *Industrial Psychology* and *Iron Age* asked her to contribute articles to their magazines. Her expertise on the topic of women industrial workers had become more widespread after her participation in the Woman's Industrial Conference of 1926 and her work for the U.S. Department of Labor's Women's Bureau. Companies came to her with problems related to their women workers as well as serving women customers (Des 2013).

She secured Macy's as a client from 1925 to 1928 at a time when American retailers were desperately trying to figure out how to appeal to the female customer. Lillian's status as a psychologist and a mother led others to believe that she had the right combination of scientific thinking and intuition. Her work involved revamping the physical layout of the aisles in the New York flagship store to make it both more aesthetically pleasing and easier for customers to navigate. Her other efforts included better systems for posting and filing employee records, different light fixtures to reduce eye fatigue, padding walls to reduce noise, and eliminating duplicate recording of sales checks. She introduced procedures to reduce counting errors and to minimize the time that a customer needed to wait for change (Des 2013; Lancaster 2004).

Since management wanted to generate greater profits, an evaluation of the psychology of the female work force was very important. Lillian found that the information male researchers had uncovered in efforts before hers had not gotten to the root of the issues facing the female workers. Very little of the workers' issues related to the physical requirements of the job. Instead, they related to family burdens or social plans after the work shifts. She recommended that managers communicate with the sales clerks and endeavor to understand the wants and needs of each employee on an individual basis. She understood that how individuals related to the larger social group was also important. Although common practice today, these recommendations were unheard of at the time! (Des 2013).

Her work with Johnson and Johnson expanded. She tackled a problem that no male executive at Johnson and Johnson had been able to solve: how to develop and market sanitary napkins. Lillian was the right woman for the job! She hired female market researchers who gathered data from the targeted customers. They found out that women wanted greater comfort, protection, and inconspicuousness with a product that could be discreetly obtained and thrown away. At her home, Lillian put together a consumer testing lab analyzing the products on the market to come up with a design that met customer needs. In the end, the product that

was developed—Modess—had a slogan that was accurate and effective: "Women designed Modess. Johnson and Johnson made it" (Des 2013).

Her work at companies including The Dennison Company and Sears and Roebuck involved understanding the psychology of the female work force and undertaking time and motion studies on female employees. Companies in Belgium wanted her to assist them in understanding and motivating their employees, too. The study of women's work around the USA and worldwide became her bread and butter. By 1926, she was presenting herself as a role model on the compatibility between marriage and a career. She interwove the theme of how scientific management could make this balance possible. Her position proved very popular in the 1920s (Des 2013; Lancaster 2004; Yost 1949).

Lillian began a long-lasting relationship with the Women's Bureau of the U.S. Department of Labor in 1926. She worked with Mary Anderson, the bureau's director, in enacting protective legislation for women workers. She attended the Women in Industry conference and later served on the technical committee whose function was to research the effects of labor legislation on women's work (Lancaster 2004).

By 1926, Lillian had decided that the best way forward was to present herself as an expert on women's work. Her differentiating skills were in her concentration on the minimization of fatigue and the application of psychology—what could be termed "household engineering." Her "coming out" of sorts occurred at a conference she organized and directed at Teachers College, Columbia University in 1927. This conference was the first organized effort to explain scientific management to home economists. The home economists were impressed with the pairing of efficiency and psychology (Lancaster 2004; Graham 1998).

A significant boost to her fortunes came in 1928 when the University of Michigan made her an honorary master of engineering—the first time such a degree had been awarded to a woman by any college. Now, she had an engineering credential in addition to psychology (Gilbreth Jr 1970).

This was particularly fortuitous during and after the Great Depression. Her work in classrooms and department stores would now move to the kitchen—with her emphasis on frugality, efficiency, and psychology so relevant to those difficult economic years. Lillian used her femininity to her—and women's—advantage by bringing efficiency to women's domestic endeavors. She now undertook to systematize women's operations. By framing such innovations in the home as a matter of economic necessity (helping out American families, not solely American women), she was able to gain acceptance where others had been less successful. Out of economic necessity and the reluctance of the scientific management profession to accept her, Lillian reinvented herself as a domestic consultant (Des 2013).

Her timing was excellent—and Lillian had found a good niche. Indoor plumbing and electricity were widely available. Women were beginning to demand labor-saving appliances and efficiently designed kitchens. The days of servants were diminishing which meant the lady of the house needed to do the work herself. Refrigerators were just electric iceboxes without the shelves, drawers, and accessories that we know today—Lillian came to the rescue. What did housewives

use the most? Eggs, milk, and butter. Lillian applied therbligs and recommended putting them at a level where the housewife wouldn't need to stoop. Women did not want to open a valve to drain soapy water bucket by bucket from an electric washing machine—Lillian had the washing machine manufacturers install a pump and wasterwater hose. Voila—more Happiness Minutes (Gilbreth Jr 1970; Graham 1998).

Her decision began to pay off. In 1927 and 1928, she published two books, *The Home-Maker and Her Job* and *Living with Our Children*. In these volumes, she advised on the One Best Way to can baby food and to design a workspace. The One Best Way was communicated through radio addresses and, most successfully, through her kitchen designs. In 1929, she designed the "Kitchen Practical" (which was really the Gilbreth Motion Study Kitchen) for the Brooklyn Gas Company which was unveiled at the national Women's Exposition. Although that kitchen, to meet the needs of her client, was outfitted with gas appliances, the one for Narragansett Light Company had electrical outlets to display the company's light fixtures. She encouraged women to customize the arrangements and appliances to their individual needs (Des 2013; Gilbreth Jr 1970).

Lillian designed not only the kitchens, but also items to go with them. These included the "Door Closet" and the "Management Desk." The closet was a thin cabinet attached to the back of the kitchen door that housed mops, cleansers, and associated items for ease of access. The desk had a clock, adding machine, radio, telephone, children's reference books, and charts to allow the person to organize household chores. Homemakers and corporate men liked her designs—she designed a Management Desk for IBM for the Chicago World's Fair in 1933 (Des 2013).

Lillian saw the need for efficiency in both the home and the workplace. In fact, she (Dr. Gilbreth, internationally famous industrial engineer—as she was referred to in customer booklets) said that if homemakers would employ her recommended methods in the kitchen that they could reduce the distance they traveled in a year from twenty-six miles to nine! She recommended heights of shelves, stoves, sinks, and counters to minimize fatigue. She also advocated for splitting housework fifty-fifty (wife/husband) and her Teamwork Kitchen, with its ability to adjust for height and lengths of workspaces, actually served to accommodate women, children, and men. She designed a foot-pedaled trashcan to minimize kitchen movements. She developed electric stoves, refrigerators, and washing machines. She designed specially rigged kitchens for the American Heart Association that would benefit wheelchair-bound women and women who suffered from heart disease (Des 2013; Lancaster 2004; Gilbreth Jr 1970; Graham 1998).

A detour to politics occurred, however, along the way. An active supporter of Henry Hoover's campaign for President, she had been friends with both the President and his wife Lou Henry Hoover, who were both engineers, since their Stanford days at the turn of the century. She often was invited to events at the White House after he won the office. Lou Henry Hoover asked her to join the Girl Scouts national Board of Directors in 1930, an offer she accepted; she served until 1947. In August of 1930, Hoover put her on a subcommittee of the National Conference on Home Building and Home Ownership (Des 2013; Lancaster 2004; Gilbreth Jr 1970; Yost 1949).

Fig. 1.2 Dr. Lillian Moller Gilbreth with Colonel Arthur Woods, President's Emergency Committee on Employment, 1930. Courtesy, Library of Congress

In October 1930, President Hoover asked her to head the women's division of the President's Emergency Committee on Employment (PECE) which required her to spend much time in Washington, DC (see Fig. 1.2). Her children were quite supportive of her accepting this assignment. Lillian instituted a "Spruce Up Your Home" program where American homes who could afford "handymen" were matched with unemployed workers with the requisite skills. She mobilized nearly three million middle-class women to generate jobs. She developed a "Follow Your Dollar" campaign to encourage women to buy American goods but also to investigate the companies behind the products to ensure that they were working to stabilize employment and make work better for their employees. During this time, Lillian made Ida Tarbell's list of the "Fifty Foremost Women of the United States." She was also one of 22 women featured in a *Good Housekeeping* readers' poll to discover America's 12 greatest living women; she did not make the final cut. Following her service on PECE, she served on the President's Organization on Unemployment Relief. She left government service and returned to her non-political life in 1931 (Des 2013; Lancaster 2004; Yost 1949).

In 1932, in a radio talk, she reported (Yost 1949):

> *. . .It was heartening to find that the best thinkers in the European group agree with ours that what is needed today is not less but more planning. . . .The manufacturer must think back to his raw materials, machines and men, and forward to the distribution and use of his product. . . . The engineer has done a fine job of making things, possibly – it was felt – too good a job. That is what he was asked to do, make things as cheaply and as well as possible. The need to extend the same careful techniques to distribution and consumption should be a challenge and not a warning, and not only to engineers but to industrial and business leaders and to the consumers.*

In 1934, Lillian designed three of the rooms in "America's Little House" for Better Homes of America: the kitchen, a clothery, and a nursery. These rooms were designed to deal with "the food problem, the clothing problem, the care of the child problem, and the problem of keeping the house clean and in order." Columbia Broadcasting System was a co-sponsor of the project and she broadcast from the

house in February 1935. She investigated the correct height for kitchen equipment. A somewhat hostile magazine article said that "Dr. Lillian Moller Gilbreth does a man's job in a woman's sphere—the home" (Lancaster 2004).

After providing many guest lectures at Purdue University starting in 1924 and looking for a source of steady income, Lillian joined the Purdue University faculty in 1935 as a full Professor of Management in the School of Mechanical Engineering, the first woman to be so appointed in an engineering school. After Amelia Earhart's death (leaving vacant the position as advisor on careers for women), Lillian took over her position on the Staff of the Dean for Women. While at Purdue, she did some consulting work in addition to her university duties. At Purdue, she lectured on management engineering in all of the schools of engineering (Mechanical, Civil, Electrical, and Chemical), created a motion study laboratory, and helped set up an honors course where students worked in local industries. She was asked to retire in 1948—at the age of 70! This freed her to consult for the Girl Scouts, serve as one of two women on the Chemical Warfare Board, and serve on the Civil Defense Advisory Council under President Harry S. Truman (Des 2013; Lancaster 2004; Yost 1949).

During World War II, Lillian undertook three types of work: she served as a government advisor, as a role model to other women, and as an ergonomics expert (Lancaster 2004). She sat on the education subcommittee of the War Manpower Commission, on the education advisory committee of the Office of War Information and on the boards of the women's army and navy auxiliaries—WACS and WAVES (Des 2013). She also continued her association with the Women's Bureau. Her fingerprints are visible on a 1942 publication issued by the National Industrial Information Committee titled "Recreation and Housing for Women War Workers." The Code of Ethics for Volunteers included as Appendix B incorporates language that reads "I believe that all work should be carefully analyzed in order that work methods may be standardized. I believe that people should be studied in order to determine what jobs they can do and like to do and that, as far as possible, they should be assigned to jobs they can do well and enjoy." Her influence can also be seen in "Employing Women in Shipyards" published in 1944 that includes in its table of contents: "Select and place women carefully," "schedule rest periods," and "set up an effective women counselor system" (Lancaster 2004).

One specific example of her consulting is illustrative. The Arma plant in Brooklyn, New York had an all-male workforce of several hundred and was getting ready to hire 8000 people, including 3000 women. They were panicked and didn't know what to do about the coming influx of women. They said to Lillian "We've never had women in the shop before. We don't know how to start. We're counting on you to tell us everything we have to do to get ready for them." Her stunning reply: "If that's all my job is, I can finish it with this one sentence: Build separate rest rooms" (Gilbreth Jr 1970).

She was called upon by President Franklin Delano Roosevelt to devise work methods for crippled and female workers. Teaming with author Edna Yost, in 1944, their book titled *Normal Lives for the Disabled* was published, in memory of Frank. Lillian believed that her work for the handicapped had been her most important

career achievement as she said it had done the most good (Perusek 2000). As she lectured around the country as a psychologist and engineer, she said "The mental state of the disabled is all-important. If a person has the normal American outlook, the optimism, the belief in God, man and the future, it is a beginning." She served on committees for other Presidents as well including Eisenhower, Kennedy, and Johnson dealing with topics including civil defense and the problems of aging (Lancaster 2004; Gilbreth Jr 1970).

At the end of the postwar recession, she had numerous work assignments both in the USA and overseas. She also was well known as a scientific researcher in academic circles. Her work style was described thusly: (Lancaster 2004)

> *The pattern is always the same: first, Dr. Gilbreth has a helpful idea; next she inspires someone to start a pilot project to explore the idea. She herself stands by to help if needed. She offers few suggestions but asks many, many penetrating questions. As the pilot project develops she spreads the news, mentions it in her talks, discusses it with people who have something helpful to offer, particularly management people and generally stimulates an exchange of ideas until finally the baby project is "born" into a welcoming climate where it can grow and prosper and expand.*

From 1954 to 1958, she worked with Harold Smalley to apply industrial engineering to hospitals. In his textbook, issued in 1966, he stated "one of the most significant developments in the methods improvement movement occurred in 1945 when Dr. Lillian M. Gilbreth . . . began to urge that hospitals take advantage of the tools and techniques of industrial engineering." With Smalley, Lillian researched nursing, organization of hospital supplies and the best types of hospital beds (Lancaster 2004).

Her postwar work also extended her efforts with the disabled; her audience was now primarily "handicapped homemakers" in lieu of the "crippled soldiers" with whom she was involved after World War I. For almost 10 years, she worked in this area which she regarded as her most important contribution to motion study. She demonstrated how disabled women could perform a variety of tasks around the house including keeping house in a wheelchair, peeling a potato with one hand, and making a bed while on crutches. The Heart Kitchen, which she developed in collaboration with the New York Heart Association, was an outgrowth of the Kitchen Practical where the kitchen was fitted to the height of its occupants. She taught courses at Rutgers where students learned to place items requiring water near the sink and those implements needed for cooking near the stove. She worked with teams comprised of industrial engineers, home economists, rehabilitation experts, psychologists, and architects to build a model kitchen. Many non-disabled people would see the kitchen and wonder if it was possible for them to acquire the Heart Kitchen (Lancaster 2004; Gilbreth Jr 1970).

From 1953 to 1964, Lillian served as a consultant to the University of Connecticut. Originating from a conference she organized on work simplification for the handicapped, Lillian helped the University procure a vocational rehabilitation grant to study work simplification for handicapped homemakers. Part of the grant was the production of a movie *Where There's a Will* and Lillian appeared on camera at the beginning and end of the film. Her efforts in work simplification led to

her 1954 book *Management in the Home: Happier Living Through Saving Time and Energy*. In 1957, she was instrumental in ensuring that a conference of home economists and psychologists was organized to discuss the feasibility and contents of a course on work simplification for working women. This led to a book issued by the U.S. Office of Education titled *Management Problems for Homemakers* (Lancaster 2004).

In 1955, the University of Wisconsin named her the Knapp Visiting Professor (Lancaster 2004). She maintained a torrid pace of consulting, travel and lectures for 20 years after her "retirement"—from Purdue—until 1968 when her doctor forced her to rest (Des 2013; Lancaster 2004).

In 1952, she was described as "The World's Greatest Woman Engineer" because of "her impact on management, her innovations in industrial design, her methodological contributions to time and motion studies, her humanization of management principles, and her role in integrating the principles of science and management. Although we may be unaware today, she influenced the way we work, the way we arrange our houses, and our attitude toward time" (Lancaster 2004).

1.5 Honors and Awards

The recipient of 23 honorary degrees, her first honorary doctorate of engineering degree came from the University of Michigan—the first time a woman was so honored. The institution that had refused to grant her a Ph.D.—the University of California at Berkeley—named her its Outstanding Alumnus in 1954, while praising the work which they had refused to acknowledge earlier. In 1931, she received the first Gilbreth Medal, awarded by The Society of Industrial Engineers, "For distinguished contribution to management." In 1940, she was made an honorary life member of the Engineering Women's Club of New York. That citation read: "For your scientific achievements in the field of industrial psychology, for your pioneer work in applying these principles to the practical problems of the efficiency of human labor, for your intelligent womanhood, and for the esteem in which you are held by your fellow members" (Des 2013; Lancaster 2004; Yost 1949; Goff 1946; Chaffin n.d.).

Lillian and Frank were both honored (Frank, posthumously) with the 1944 Gantt Gold Medal: "To Dr. Lillian Moller Gilbreth, and to Frank B. Gilbreth posthumously . . . the 1944 Gantt Medal, in recognition of their pioneer work in management, their development of the principles and techniques of motion study, their application of those techniques in industry, agriculture and the home, and their work in spreading that knowledge through courses of training and classes at universities." Lillian said receipt of the Gantt Gold Medal was the best news of her life as it meant Motion Study and the Gilbreth System had been acknowledged by the arbiter of professional accomplishment—the American Society of Mechanical Engineers (Des 2013; Lancaster 2004; Gilbreth Jr 1970; Yost 1949; Goff 1946). The Western Society of Engineers presented her with its Washington Award in 1954 "for accomplishments which pre-eminently promote the happiness, comfort

and well-being of humanity" and for her "unselfish devotion to the problems of the handicapped" (Lancaster 2004).

In 1965, Lillian became the first woman elected to the National Academy of Engineering. In 1966, she received the prestigious Hoover Medal, an indication that she was regarded by her peers as having achieved the pinnacle of the engineering profession. This medal is jointly bestowed by four leading engineering organizations (Des 2013). The citation read: *Renowned engineer, internationally respected for contributions to motion study and to recognition of the principle that management engineering and human relations are intertwined; courageous wife and mother; outstanding teacher, author, lecturer and member of professional committees under Herbert Hoover and four successors. Additionally, her unselfish application of energy and creative efforts in modifying industrial and home environments for the handicapped has resulted in full employment of their capabilities and elevation of their self-esteem* (Lancaster 2004). She remained the only woman to have received that medal until 2005 (Giges n.d.).

A strong supporter of the Society of Women Engineers (SWE), Dr. Gilbreth was the first honorary member of the organization (see Fig. 1.3). Her membership

Fig. 1.3 Dr. Lillian Moller Gilbreth on the left of the head table at the 1957 Society of Women Engineers National Convention, Houston, Texas. Courtesy of Walter P. Reuther Library, Wayne State University

Fig. 1.4 Barbara Jean Kinney receiving the Lillian Moller Gilbreth Scholarship of the Society of Women Engineers, 1965 from President Isabelle French. Courtesy of the Walter P. Reuther Library, Wayne State University

number was one. Upon accepting membership, she said "I appreciate the honor and I hope that I will be a useful member of the Society." A scholarship was established in her name in 1958 and is still awarded annually by SWE today (see Fig. 1.4) (Perusek 2000).

Gilbreth has been inducted into the National Women's Hall of Fame and honored on a 1984 US postage stamp in the "Great American" series. Among her many other honors and awards was the Gold Medal of the National Institute of Social Services that included in its citation "distinguished service to humanity" (Gilbreth Jr 1970).

California Monthly (1944) summarizes her accomplishments thusly (Graham 1998):

> *Lillian Moller Gilbreth is a genius in the art of living. Known throughout the world as an outstanding woman engineer who successfully combined her unique engineering career with a delightful home life centering around a beloved husband and twelve well assorted children, Dr. Gilbreth amazes one with the breadth of her interests, the sheer quantity of her activities, the dynamic quality of her daily living and her own unassuming simplicity. One feels conclusively that here is a woman whose history bears inspection . . .*

References

Chaffin D (n.d.) The first 50 years of the Department of Industrial and Operations Engineering at the University of Michigan: 1955–2005. Maize Books, Michigan Publishing. https://quod.lib.umich.edu/m/maize/13855463.0001.001/1:3/%2D%2Dfirst-50-years-of-the-department-of-industrial?rgn=div1;view=fulltext

Des Jardins J (2013) Lillian Gilbreth: redefining domesticity. Westview Press, Philadelphia

Giges N (n.d.) Lillian Moller Gilbreth. https://www.asme.org/career-education/articles/management-professional-practice/lillian-moller-gilbreth

Gilbreth LM (1990) The quest of the one best way: a sketch of the life of frank Bunker Gilbreth. The Society of Women Engineers

Gilbreth LM (1998) As I remember: an autobiography. Engineering & Management Press, Norcross

Gilbreth FB Jr (1970) Time out for happiness. Thomas Y. Crowell Company, New York

Goff A (1946) Women can be engineers. Edward Brothers, Inc., Youngstown

Graham L (1998) Managing on her own: Dr. Lillian Gilbreth and women's work in the interwar era. Engineering & Management Press, Norcross

Lancaster J (2004) Making time: Lillian Moller Gilbreth—a time beyond "Cheaper by the Dozen". Northeastern University Press, Boston

Perusek A (2000) "The First Lady of Engineering", SWE: Magazine of the Society of Women Engineers, January/February 2000. pp 82–92

Proffitt P (ed) (1999) Notable women scientists. Gale Group, Inc., Detroit

Trescott MM (1983) Lillian Moller Gilbreth and the founding of modern industrial thinking. In: Rothschild J (ed) Machina ex dea: feminist perspectives on technology. Pergamon Press, Oxford

Yost E (1943) American women of science. J.B. Lippincott Company, Philadelphia

Yost E (1949) Frank and Lillian Gilbreth: partners for life. Rutgers University Press, New Brunswick

Jill S. Tietjen P.E., entered the University of Virginia in the Fall of 1972 (the third year that women were admitted as undergraduates—under court order) intending to be a mathematics major. But midway through her first semester, she found engineering and made all of the arrangements necessary to transfer. In 1976, she graduated with a B.S. in Applied Mathematics (minor in Electrical Engineering) (Tau Beta Pi, Virginia Alpha) and went to work in the electric utility industry.

Galvanized by the fact that no one, not even her Ph.D. engineer father, had encouraged her to pursue an engineering education and that only after her graduation did she discover that her degree was not ABET-accredited, she joined the Society of Women Engineers (SWE) and for almost 40 years has worked to encourage young women to pursue science, technology, engineering, and mathematics (STEM) careers. In 1982, she became licensed as a professional engineer in Colorado.

Tietjen starting working jigsaw puzzles at age two and has always loved to solve problems. She derives tremendous satisfaction seeing the result of her work—the electricity product that is so reliable that most Americans just take its provision for granted. Flying at night and seeing the lights below, she

knows that she had a hand in this infrastructure miracle. An expert witness, she works to plan new power plants.

Her efforts to nominate women for awards began in SWE and have progressed to her acknowledgement as one of the top nominators of women in the country. Her nominees have received the National Medal of Technology and the Kate Gleason Medal; they have been inducted into the National Women's Hall of Fame (including Lillian Moller Gilbreth in 1995) and state Halls including Colorado, Maryland, and Delaware and have received university and professional society recognition. Tietjen believes that it is imperative to nominate women for awards—for the role modeling and knowledge of women's accomplishments that it provides for the youth of our country.

Tietjen received her MBA from the University of North Carolina at Charlotte. She has been the recipient of many awards including the Distinguished Service Award from SWE (of which she has been named a Fellow and is a National Past President), the Distinguished Alumna Award from the University of Virginia, and she has been inducted into the Colorado Women's Hall of Fame. Tietjen sits on the boards of Georgia Transmission Corporation and Merrick & Company. Her publications include the bestselling and award-winning book *Her Story: A Timeline of the Women Who Changed America* for which she received the Daughters of the American Revolution History Award Medal.

Part II
Analytics

Chapter 2
Emergence of Statistical Methodologies with the Rise of BIG Data

Nedret Billor and Asuman S. Turkmen

Contents

2.1 Introduction 27
2.2 A Review of Statistical/Machine Learning Algorithms 29
2.2.1 Random Forests 30
2.2.2 Support Vector Machine 31
2.2.3 Sparse Modeling 31
2.2.4 Dimension Reduction 33
2.2.5 Deep Learning 34
2.3 Developments in Inferential Analyses of New Algorithms 36
2.3.1 Large-Scale Hypothesis Testing 36
2.3.2 Random Forests 37
2.3.3 Sparse Modeling 38
2.3.4 SVM 39
2.3.5 Dimension Reduction 40
2.3.6 Deep Learning 41
2.4 Discussion 42
References 43

2.1 Introduction

For decades, enormous volumes of datasets have been routinely collected as part of the everyday operation of any manufacturing enterprise, however these potentially valuable knowledge resources have not been fully understood or exploited which led to the "rich data but poor information" problem (Wang and McGreavy 1998). Both volume and complexity of such data have generated a necessity of automated analysis to extract knowledge in a form that can benefit the business. Therefore,

N. Billor (✉)
Department of Mathematics and Statistics, Auburn University, Auburn, AL, USA
e-mail: billone@auburn.edu

A. S. Turkmen
Department of Statistics, The Ohio State University, Columbus, OH, USA
e-mail: turkmen@stat.osu.edu

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_2

data-driven discovery techniques, such as statistical/machine learning algorithms, have begun to be utilized in manufacturing systems toward the end of twentieth century (Lee 1993; Irani et al. 1993; Bertino et al. 1999), and have become extremely important tools in the area. The use of these techniques produces relatively modest reductions in equipment failures, better on-time deliveries, slight improvements in equipment, and meets countless other goals that drive toward better margins for the manufacturers. In recent years, with the 20-20-20 goals (i.e., 20% increase in energy efficiency, 20% reduction of CO2 emissions, and 20% renewables by 2020) proposal by the European Union, countries have announced their development plans in manufacturing. For instance, the Industry 4.0 is the current trend of automation and data exchange in manufacturing technologies initiated by Germany. The advances in computing systems, the rapid maturation of statistical/machine learning algorithms, advanced technologies and strategies like Industry 4.0 have created new opportunities for intercross of statistical/machine learning and manufacturing streams which provide manufacturers to ability to collect, store, and analyze huge amounts of datasets, and consequently to gain predictive insights into their production.

Machine learning (ML) is a type of artificial intelligence utilizing automatic techniques to obtain deep insights, recognize unknown patterns, and create accurate predictive models based on past observations. The field of ML has advanced significantly in recent years, in part due to need for handling large and complex amounts of information. These advances have opened new avenues for strong collaborations between materials scientists, computer scientists, and statisticians. Consequently the past decade has been witnessed accelerated progress in the use of ML algorithms that are designed to learn continually from data and seek optimized outcomes in minutes rather than days or months. The primary categories of ML are supervised, unsupervised, and semi-supervised learning. The supervised learning consists of an outcome variable which is to be predicted from a given set of predictors such as regression, decision tree, and random forest. In the unsupervised learning, there is no outcome variable, and the primary goal is to segment observations in different groups based on provided predictors such as K-means algorithm. Semi-supervised learning falls between unsupervised learning and supervised learning in the sense that some observations have outcome values whereas some do not. Although unsupervised and supervised ML methods have already been mostly used in manufacturing, the relatively recently introduced semi-supervised is expected to be more popular in the near future. Unsupervised learning is generally utilized for clustering, dimensionality reduction, data visualization, and outlier detection in manufacturing applications whereas supervised learning is popular for fault classification, quality improvement, process monitoring, quality prediction, and soft sensor modeling. In addition to these three categories, reinforcement learning is another algorithm that discovers through trial and error. Although it is considered as a special form of supervised learning by some researchers, it differs from unsupervised and supervised learning in the sense that learner has to also discover which actions are needed to generate the best results. Regardless of the category,

these data-driven approaches are able to find highly complex patterns in data, which are then applied for prediction, detection, classification, regression, or forecasting.

Statistical learning (SL) refers to a collection of tools for modeling and understanding complex datasets (Hastie et al. 2009). Even though SL theory was initially based on purely theoretical analysis of the problem from a given data, new types of learning algorithms have been emerged in the late 90s and it has started to blend with developments in ML. Statistical and machine learning, although sometimes used interchangeably, have subtle differences (Breiman 2001b). SL involves making assumptions which are validated after building the models, and it emphasizes the statistical inference (confidence intervals, hypothesis tests, optimal estimators). In contrast, ML requires no prior assumptions about the underlying relationships between the variables and mainly focuses on producing predictions rather than drawing calculations from data. Despite these differences, both approaches share the same goal of learning from data using mathematical techniques to solve problems. ML provides quick and automatic algorithms to produce models that can analyze bigger, more complex data and deliver faster, more accurate results. However, statistical inference has become less important and asymptotics are underrepresented for most of these immense computer-based ML algorithms. Given the importance of common dialogue between these two learning approaches, this chapter provides not only a summary of some promising methods in the overlap of both statistical and machine learning communities, but also provides information on making inferences.

This chapter presents a very general overview of statistical/machine methods and their applications in manufacturing. The remainder of the chapter is organized as follows: Sect. 2.2 reviews the commonly used learning algorithms, including random forests, support vector machine, sparse modeling, dimension reduction, and deep learning. Section 2.3 reviews the developments in inference theory for the algorithms since inferential justifications are very important aspect of a statistical analysis. The concluding remarks are given in Sect. 2.4.

2.2 A Review of Statistical/Machine Learning Algorithms

This section reviews prominent ML tools, and provides an overview for interested readers. For the sake of brevity, we only included a small part of the related and relatively new literatures on ML applications in manufacturing. The growth and popularity of R programming (R Core Team 2016) has been helping data-driven organizations succeed for years. In this paper, we provide citations for the commonly used R packages that implement the machine learning algorithms described in this section.

2.2.1 Random Forests

A decision tree is a type of directed (acyclic) graph in which the nodes represent decisions, and the edges or branches are binary values (yes/no, true/false) representing possible paths from one node to another. Random forest (RF) algorithm is a supervised regression and classification algorithm proposed by Leo Breiman (2001a), and as the name suggests, it creates a forest based on the aggregation of a large number of decision trees. The algorithm constructs an ensemble of trees from a training data to predict the outcome for future observations. Given the training dataset, the decision tree algorithm will come up with some set of rules which can be used to perform the prediction on a test dataset. This approach also ranks variables with respect to their ability to predict the outcome using variable importance measures (VIMs) that are automatically computed for each predictor within the RF algorithm. This is an important asset given that the selection of the most relevant variables is a crucial task for high-dimensional data. The review by Genuer et al. (2008) provides practical guidelines for understanding the method. In the original RF method, each tree is considered as a standard classification or regression tree (CART) which uses a numerical criterion to split. Each tree is constructed from a bootstrap sample drawn with replacement from the original dataset, and the predictions of all trees are finally aggregated through majority voting. The main idea in RF is reducing the variance by averaging and reducing bias using bootstrapping. In general, the higher the number of trees in the forest gives the more accuracy results for the RF classifier while having more trees in the forest does not result in an overfit model. RF algorithm can be used both for classification and the regression of problems (where the system is trained to output a numerical value, rather than "yes/no" classification), and has the ability to handle both missing values and high dimensionality. The R implementations of RF are available in the package "*randomForest*" (Liaw and Wiener 2002).

Applications

RF algorithm has a variety of applications in manufacturing including classification of sensor array data (Pardo and Sberveglieri 2008), machine fault diagnosis (Yang et al. 2008), fault detection (Auret and Aldrich 2010; Puggini et al. 2016). Another important application of RF is the tool wear classification which is one of the important factors for guaranteeing the reliability and stability of manufacturing systems since the excessive wear of cutting tools will result in a sharp increase in cutting forces. Wu et al. (2017) realized tool wear prediction in milling operations by utilizing RFs, and showed that RFs performed better than its other close competitors.

2.2.2 Support Vector Machine

Support vector machines (SVMs) introduced by Vladimir Vapnik (1995) are supervised learning techniques that are intuitive, theoretically well-founded, and have shown to be practically successful for classification and regression analysis. In addition to performing linear classification and regression, SVMs can efficiently perform a non-linear classification and regression using kernel functions by implicitly mapping their inputs into high-dimensional feature spaces. In general, the basic idea for a classification problem is to find a hyperplane which separates the multi-dimensional data perfectly into its two (or more) classes using the notion of a kernel that maps the original data points into a higher dimensional feature space in which they can be separated by a linear classifier. Since the projection of a linear classifier on the feature space is non-linear in the original space, users have the ability to generate non-linear decision boundaries using methods designed for linear classifiers. Compared to other widely used supervised learning methods such as artificial neural networks, SVM may have better generalizations under many cases. An excellent overview of the SVM algorithms can be found in Vapnik (1995), Schölkopf et al. (1999). The "*e1071*" package in R can be used for the implementation of the algorithm (Meyer et al. 2017).

Applications

Due to its good generalization and ability to produce a smooth classifier, SVM has been a popular tool in industrial applications. A major application area of SVM in manufacturing is monitoring, such as machine condition and quality monitoring, fault detection, and tool wear prediction (Cho et al. 2005; Ribeiro 2005; Widodo and Yang 2007; Zhang 2008; Xiao et al. 2014; You et al. 2015; Tian et al. 2015; Benkedjouh et al. 2015). SVMs are also successfully applied to other problems, including—but not limited to—identification of damaged products (Caydas and Ekici 2010), classification of gasoline (and gases) in oil industry (Saybani et al. 2011), control chart pattern recognition (Xanthopoulos and Razzaghi 2013), bearing faults detection (Saidi et al. 2015), soft sensor developments and quality prediction (Jain et al. 2007; Ge and Song 2010; Yu 2012; Zhang et al. 2013).

2.2.3 Sparse Modeling

Given the massive amount of data collected from industrial processes, it is inevitable to have many redundant, irrelevant, and noisy variables in the data leading to unreliable inferences. Although regression models are useful to build models for prediction in manufacturing process, classical methods such as ordinary least squares (OLS) lead to an over-fitted, or more often an under-fitted system of

equations when data consist of correlated and very few quality measurements. Sparse (regularized) modeling aims to achieve best predictability with the most parsimonious model, i.e. with fewest predictors. In other words, regularized models simultaneously perform variable selection and prediction. Ridge regression (Hoerl and Kennard 1970) based on L2 penalty and least absolute shrinkage (LASSO, Tibshirani 1996) based on L1 penalty are well-known examples of penalized regression models. Zou and Hastie (2005) proposed elastic-net penalty which is a linear combination of L1 and L2 penalties, and such method emphasizes a grouping effect, where strongly correlated predictors tend to be in or out of the model together. In general, it has been argued that a good penalty procedure should have oracle property that it performs as well as if the true model were given in advance. The LASSO can perform automatic variable selection since it uses L1 penalty that is singular at the origin. On the other hand, the LASSO penalty may produce inefficient estimation and inconsistent variable selection results in linear regression modeling due to uniformly imposed penalty to all features without considering importance of each feature. It is known that the LASSO penalty may introduce a substantial amount of bias to the estimators of the large coefficients (Fan and Li 2001), thus, the oracle property does not hold for the LASSO estimator. To overcome this problem, Zou (2006) proposed an adaptive version of the LASSO in which data dependent weights are used for penalizing different coefficients in the L1 penalty and showed that the adaptive LASSO possesses the oracle property while having computational easiness of optimizing a convex function. Although there are a variety of regularization methods available, aforementioned methods have been predominantly used for application. The R package "*glmnet*" includes procedures for fitting LASSO or elastic-net regularization path for linear regression, logistic and multinomial regression models (Friedman et al. 2010).

Applications

A complex modern semi-conductor manufacturing process is normally under consistent surveillance via the monitoring of signals/variables collected from sensors and/or process measurement points. However, not all of these signals are equally valuable in a specific monitoring system. The measured signals contain a combination of useful information, irrelevant information as well as noise. It is often the case that useful information is buried in the latter two. Engineers typically have a much larger number of signals than are actually required. If we consider each type of signal as a feature, then feature selection may be applied to identify the most relevant signals. The process engineers may then use these signals to determine key factors contributing to yield excursions downstream in the process. This will enable an increase in process throughput, decreased time to learn and reduce the per unit production costs. In general, since the semiconductor manufacturing process consists of a huge amount of (relevant and irrelevant) predictors that are correlated and relatively very few observations, predictions of quality and/or yield using OLS

are very complicated. To address these issues, penalized regression methods are utilized in the literature for the semiconductor manufacturing process.

Chang and Mastrangelo (2011) underlined the collinearity problem in semiconductor environment, and suggested to address multicollinearity using variable elimination, orthogonal transformation, and adoption of biased estimates such as ridge regression. In semiconductor manufacturing plants, although monitoring physical properties of all wafers is crucial to maintain good yield and high quality standards, such approach is too costly, and in practice, only few wafers in a lot are actually monitored. Virtual metrology (VM) systems allow to partly overcome the lack of physical monitoring. Susto and Beghi (2013) utilized least angle regression to overcome the problem of high dimensionality and model interpretability for VM module. Similarly, Melhem et al. (2016) compared regularized linear regression methods based on dimension reduction and variable selection methods to predict the wafer quality based on the production equipment data. Motivated by the adverse effects of outliers on the modeling, monitoring, and diagnosis of profile monitoring data, Zou et al. (2014) proposed an outlier detection method based on the penalized regression. For semiconductor packaging process, the final step of semiconductor manufacturing, Lim et al. (2017) predicted the final failure of a printed circuit board lot based on observed event sequences in the wire bonding process step using LASSO.

2.2.4 Dimension Reduction

Multi-parameter setting in multi-stage of the modern manufacturing industry brings about the curse of dimensionality, leading to the difficulties for feature extraction, learning, and quality modeling. One imperative way of addressing this issue is to enhance feature extraction capability using a dimension reduction technique that helps to reduce invalid information interference. The main goal of the dimension reduction is to explore the data, and to look for some hidden structure among them, and it is mainly used for information extraction, data visualization, and outlier detection for industry applications. Principal component analysis (PCA) is the most popular statistical procedure aiming to find an orthogonal transformation of data to transform a set of correlated variables into linearly uncorrelated variables that are fewer than the number of original variables (Jolliffe 2002). The original PCA is an unsupervised and linear method, but its different variations that can handle these limitations. There are also other dimension reduction methods used in the literature such as independent component analysis (ICA), partial least squares (PLS), and isometric mapping (Isomap). ICA is an extension of PCA that not only reduces the variables into a set of independent components by decorrelation but also makes them as independent as possible (Hyvorinen et al. 2001). ICA exploits inherently non-Gaussian features of the data and employs higher moments which distinguish it from PCA utilizing the first and second moments of the measured data, hence relying heavily on Gaussian features. PLS is a supervised dimension reduction

technique aimed to find a compromise between the explanation of factors space and the target variable (Wold 1975; de Jong 1993). Isomap is a representation of manifold learning technique (Tenenbaum et al. 2010), and is a non-linear dimension reduction method. R packages "*stats*" (*prcomp* or *princomp* functions, R Core Team 2016), "*FactoMineR*" (*PCA* function, Le et al. 2008) for PCA; "*pls*" (*plsr* function, Mevik et al. 2016) for PLS; "*fastICA*" for ICA (Marchini et al. 2017), and "*vegan*" (*isomap* function, Oksanen et al. 2017) can be used for the implementation of these techniques.

Applications

Regardless of the differences among them, all mentioned dimension reduction techniques are beneficial to explore sophisticated relationships between the multi-parameter manufacturing information and the quality. PCA and some of its variations have been used frequently for process monitoring which is a critical requirement in any manufacturing process as producing quality products within specification reproducibly is required from an economically viable process (Thornhill et al. 2002; Qin 2003; Lee et al. 2004; Zhang et al. 2012; Yao and Wang 2015). A comprehensive review of this literature can be found in de Ketelaere et al. (2015). In addition to that, there are a variety of PCA applications that have been used for dimensionality reduction, data visualization (Dunia et al. 2013), and outlier detection (Chiang et al. 2003). Although PCA is the most commonly utilized dimension reduction technique in literature, there are many applications of ICA, PLS, and Isomap as well. For instance, Kao et al. (2016) used multi-stage ICA to extract independent components from the monitoring process data to distinguish unnatural control chart patterns from the normal patterns. Zhang et al. (2010) used kernel PLS method for the quality prediction in complex processes. Bai et al. (2018) compared promising dimension reduction techniques on two experimental manufacturing data, and demonstrated that Isomap might be a better first option for the multi-parameter manufacturing quality prediction.

2.2.5 Deep Learning

Artificial neural networks (ANNs) are a family of models inspired by biological neural networks, and are presented as systems of interconnected "neurons" that exchange messages between each other. Numeric weights are assigned to the connections in an adaptive way to inputs that makes the method capable of learning. With advances in hardware and the emergence of big data, we can create neural networks with many layers, which is known as deep learning neural networks. Most ML algorithms mainly focus on function optimization, and the solutions do not necessarily always explain the underlying trends nor give the inferential power aimed by artificial intelligence. Therefore using ML algorithms often becomes a

repetitive trial and error process, in which the choice of algorithm across problems yields different performance results. Deep learning can be considered as the subfield of ML that is devoted to building algorithms that explain and learn data that classical ML algorithms often cannot. In other words, the deep learning algorithms do not only have predictive and classification ability, but they also have the ability to learn different levels of complex patterns in large amounts of data. The single layer perceptron (SLP) model is the simplest form of neural network and the basis for the more advanced models that have been developed in deep learning. Most commonly used deep learning models are convolutional neural network (CNN), restricted Boltzmann machine (RBM), deep belief network (DBN), auto encoder (AE), and recurrent neural network (RNN). CNN is the most frequently used for computer vision and image processing (Lecun et al. 1998). RBM is a two-layer neural network where the connections between units form a directed cycle, and often used for speech and handwriting recognition (Smolensky 1986). DBN is constructed by stacking multiple RBMs to reduce computational complexity (Hinton et al. 2014). Deng et al. (2010) proposed AE by adding more hidden layers to deal with highly non-linear input. The idea behind RNN is to make use of sequential information. RNNs are called recurrent because they perform the same task for every element of a sequence, with the output being depended on the previous computations (Hihi and Bengio 1996). A detailed overview of the deep learning algorithms can be found in Schmidhuber (2015). The "*keras*" package in R (Allaire and Chollet 2018) can be used for the implementation of deep learning algorithms.

Applications

With automatic feature learning and high-volume modelling capabilities, deep learning provides an advanced analytics tool for smart manufacturing in the big data era. Deep learning is a field of study within computer science that aims to equip machines with the capability to think more like human beings. In other words, its main goal is to provide machines with the ability to think and learn such as self-driving cars. From the point view of manufacturing, these algorithms can help companies to develop machines that have capability to determine when a possible defect occurs. Consequently, producers could automatically take corrective action before a defect was likely to occur. Despite these advantages, the use of deep learning is relatively new compared to aforementioned learning methods. Some of the manufacturing applications implementing deep learning are surface integration inspection (Weimer et al. 2016), machinery fault diagnosis (Yu et al. 2015; Janssens et al. 2016; Jia et al. 2016), predictive analytics, and defect prognosis (Malhi et al. 2011; Wang et al. 2017). The paper by Wang et al. (2018) presents a comprehensive survey of commonly used deep learning algorithms, and discusses their advantages over traditional ML and their applications toward making manufacturing "smart." Undoubtedly, deep learning provides advanced analytics tools for processing and analyzing big manufacturing data. We believe that increasing consumer demand for better

products and companies seeking to leverage their resources more efficiently will lead to increasing research effort of deep learning in applications of manufacturing.

2.3 Developments in Inferential Analyses of New Algorithms

There are two aspects of statistical analysis: algorithmic and inferential. We can give very broad definitions for these two aspects. Algorithms are what statisticians do while inference labors to rationalize these algorithms (Efron and Hastie 2016). To differentiate between these two aspects, we can give a very simple example. Assume that we observed defective chips manufactured in 20 production lines in a technology factory, we are interested in summarizing these observed data in a single number, which is the most popular and simplest statistics, the average, and then we wish to know the accuracy of this number. Here averaging is an algorithm which comes generally first in data analysis. Then we wish to know how accurate this number is. Namely, we are concerned about the *algorithm's accuracy*. Therefore, as an accuracy measure, for instance, the standard error is calculated to provide an inference of the algorithm's accuracy.

Due to advancements in computer technology, researchers collect massive datasets which require new innovative computer-based algorithms. Consequently, there has been an upsurge in the developments of algorithmic inventions in computer-age period—from the 1950s to the present—where the traditional bottleneck of statistical applications became faster and easier. All of these algorithms focus on prediction. Prediction is certainly an interesting subject. However, one should be careful in using these algorithms by only focusing on prediction. For instance, estimation and testing are a form of prediction: "In a sample of 25 chips material A (used for manufacturing chip) outperformed material B; would this still be true if we went on to test all possible chips?" The answer for this question does require inferential study.

Inferential aspects of these new algorithms have not been developed in the same pace as in these algorithmic inventions. However, recently there has been great effort among statisticians to contribute to statistical inferential aspects of these new algorithms. Statistical inference (SI) is a very important topic that powers all these algorithmic inventions. In this section, we will give a brief survey on these advances.

2.3.1 Large-Scale Hypothesis Testing

In many complicated manufacturing lines, usually more than one attribute quality variable needs to be monitored simultaneously. For example, delamination is one of the most critical damage modes observed in laminated carbon/epoxy structures in semiconductor packaging lines. It is very important for the manufacturers to monitor the number of delamination of different positions of one product at the same time (Li

and Tsung 2012). There is a need to carry out many simultaneous hypothesis tests, done with the prospect of finding only a few interesting ones among a haystack of null cases.

There is a substantial theory of simultaneous inference which was developed aiming at the frequentist control of family-wise error rates for a small number of hypothesis tests, maybe up to 20 (Miller 1981; Westfall and Young 1993). The breakthrough method which was given in Benjamini and Hochberg's seminal 1995 paper is a false discovery rate (FDR) control approach for multiple hypothesis testing for large-scale datasets involving thousands of simultaneous tests. This method was inspired for microarray applications first, then applied to many areas including manufacturing and engineering. FDR is used to quantify the expected ratio of incorrect rejections to the number of all the rejected hypotheses. Benjamini and Hochberg (1995) theoretically proved that when all the null hypotheses are true, controlling the FDR is equivalent to controlling the type I error rate. Also, when some of the alternative hypotheses are true, controlling the FDR would provide much higher detection power, especially when the number of hypotheses is large. There is an excellent reference on large-scale simultaneous testing method by Efron (2010). All of these methods have great deal of mathematical originality.

The R package, "*locfdr*" on CRAN is an R program that provides FDR estimates for large-scale hypothesis testing procedures (Efron et al. 2015).

2.3.2 Random Forests

Due to rapid developments in computer technology, prediction has become a major task in recent data analysis problems. There are two problems in this context. The first one is how to construct an effective prediction rule which is more algorithmic, the second one is how to estimate the accuracy of its predictions which is more inferential. For the prediction assessment (more inferential), there are nonparametric approaches such as cross-validation (Hastie et al. 2009), and model-based approaches such as Mallows' C_p (Mallows 1973) and the Akaike information criteria (Akaike 1974) that are developed in the 1970s for prediction error assessment.

Many classical prediction algorithms cannot easily handle massive datasets, even if they can, their predictive power is generally either too modest or poor. The method of RF (one of the ensemble methods) has become very popular as learning algorithms that have good predictive performance. Some other ensemble methods, such as boosting and bagging, in addition to RF have also been developed to handle big datasets and improved the predictive performance. They fit models of breathtaking complexity compared with classical linear regression, or even with standard generalized linear modeling. They are routinely used as prediction engines in a wide variety of industrial and scientific applications. We will give some recent studies focused on statistical inferential aspects of these algorithms.

Wager et al. (2014) studied the variability of predictions made by RFs, and showed how to estimate standard errors for these methods. This work builds on variance estimates for bagging that are based on the jackknife and the infinitesimal jackknife (IJ). Moreover, the sampling distributions of the jackknife and IJ variance estimates were studied. Most existing theoretical results about RFs aim to establish the consistency of RF predictions (e.g. Biau et al. 2008; Mentch and Hooker 2014). There has been much less work, however, on understanding the sampling variance of RFs. Wager and Athey (2018) studied an RF model based on subsampling, and showed that RF predictions are asymptotically normal under certain conditions on the sample size. The asymptotic variance can be consistently estimated by using an IJ for bagged ensembles recently proposed by Efron (2014). Mentch and Hooker (2016) have also developed a formal SI procedure for predictions generated by supervised learning ensembles. Although the ensemble methods have improved the predictive accuracy of individual trees, they failed to provide a framework in which distributional results can be derived easily. They showed that the estimator obtained by averaging over trees built on subsamples of the training set has the form of U-statistic. Consequently, predictions for individual feature vectors are asymptotically normal, allowing for confidence intervals to accompany predictions.

In summary, all of these efforts take a step towards making RFs tools for SI instead of just being *black-box* predictive algorithms.

2.3.3 Sparse Modeling

In recent years, there has been a booming interest in ML research community to model the variable selection problem in high dimensional systems ($n \ll p$) using sparse linear models after the seminal research of Tibshirani (1996) proposing LASSO. The focus of these new methods is mostly on building interpretable models for prediction, with little attention paid to inference. Inference is generally difficult for adaptively selected models based on these new regularized regression techniques. Classical regression theory aimed for an unbiased estimate of each predictor variable's effect. High dimensional datasets ($n \ll p$), often with enormous number of predictors p, make that an unattainable goal. These methods, by necessity, use shrinkage methods, biased estimation, and sparsity.

After we do the selection (which is the algorithmic approach), we would be interested in their SIs, such as confidence intervals for the parameters, statistical properties (consistency, oracle property, etc.) of the estimators in the final model, determining the accuracy of the final model. This is called *post-selection inference* (Efron and Hastie 2016). We will briefly attempt to summarize the issues in post-selection inference and the recent studies on the inferential aspects of these methods.

Suppose we have fit a LASSO regression model with a particular value for the tuning parameter, leading to selecting a subset of k of the p ($k < p$) available variables. The question arises as to whether we can assign p-values to these selected variables, and produce confidence intervals for their coefficients. In this context,

one question that arises is that whether we are interested in making inferences about the population regression parameters using the full set of p predictors, or only the subset of k variables. For the first case, it has been proposed that one can view the coefficients of the selected model as an efficient but biased estimate of the full population coefficient vector. The idea is to then *debias* this estimate, allowing inference for the full vector of coefficients. For the second case, the idea is to *condition* on the subset with k variables and then perform conditional inference on the unrestricted (i.e. not LASSO-shrunk) regression coefficients of the response on only the variables in the subset (Efron and Hastie 2016). Most of the activities around post-selection inference were inspired by the work of Berk et al. (2013) which is based on investigating the post-selection inference for linear models where *statistical tests* and *confidence intervals* are pursued after variable selection. Under certain conditions and assumptions, for instance, Gaussian homoscedastic model errors and the design matrix to be rank-deficient, they showed that valid post-selection inference is possible through *simultaneous inference*. For the *debiasing* approach, Zhang and Zhang (2014) proposed methodologies for SI of low-dimensional parameters with high-dimensional data, such as construction of confidence intervals for individual coefficients and linear combinations of several of them in a linear regression model. The theoretical results include sufficient conditions for the asymptotic normality of the proposed estimators along with a consistent estimator for their finite-dimensional covariance matrices. Further, van de Geer et al. (2014) proposed asymptotically *optimal* confidence regions and tests for high-dimensional models. Javanmard and Montanari (2014) proposed an efficient algorithm for constructing confidence intervals with nearly optimal size, and hypothesis testing with nearly optimal power. The *conditional inference* approach was developed first by Lockhart et al. (2014). A series of papers (e.g., Lee et al. 2016) was published on this approach following the work in Lockhart et al. paper. The R package called "*selectiveInference*" implements some of the selective inference methods described in this section (Tibshirani et al. 2016).

2.3.4 SVM

The support vector machine has been successful in a variety of applications. Also on the theoretical front, statistical properties of the support vector machine have been studied quite extensively with a particular attention to its *Bayes risk consistency* under some conditions.

In nonparametric classification and regression problems, regularized kernel methods, in particular support vector machines, attract much attention in theoretical and in applied statistics. In an abstract sense, regularized kernel methods (simply called SVMs here) can be seen as regularized M-estimators for a parameter in a (typically infinite dimensional) reproducing kernel Hilbert space. For smooth loss functions, Hable (2012) showed that the difference between the estimator and the theoretical SVM is asymptotically normal, converges weakly to a Gaussian process

in the reproducing kernel Hilbert space. Several researchers studied the Bayes risk consistency of the SVM (e.g. Lin 2002; Zhang 2004; Steinwart 2005) and its rate of convergence to the Bayes risk (Lin 2000; Blanchard et al. 2004; Scovel and Steinwart 2004; Bartlett et al. 2006).

2.3.5 Dimension Reduction

There have been extensive work on studying the statistical inferential aspects of dimension reduction methods since the 1960s. We will attempt to provide a summary of these activities on this topic.

PCA is the oldest dimension reduction method. Anderson (1963) is the first statistician who studied the inferential aspect of this method. He provided the asymptotic distribution of characteristic roots and vectors of a sample covariance matrix under the assumption of multivariate normality. Dauxois et al. (1982) studied the limiting distribution of the eigenvalues and eigenvectors based on the results of convergence by sampling in linear PCA (of a random function in a separable Hilbert space). Karoui and Purdom (2016) studied the properties of the bootstrap as a tool for inference concerning the eigenvalues of a sample covariance matrix. Through a mix of numerical and theoretical considerations, they showed that the bootstrap performs as it does in finite dimension when the population covariance matrix is well-approximated by a finite rank matrix. There are some further studies based on the inferential aspect of PCA. We can give one or two references within this framework. Critchley (1985) studied the theoretical influence function and various sample versions were developed to provide methods for the detection of influential observations in PCA. Further, Boente (1987) studied the asymptotic distribution of the eigenvalues and eigenvectors of the robust scatter matrix, many activities followed these papers.

The derivation of statistical properties for PLS regression (PLSR) has been a challenging task. The reason is that the construction of latent components from the predictor variables also depends on the response variable. While this typically leads to good performance and interpretable models in practice, it makes the statistical analysis more involved. Not many studies can be found in the SI for PLSR until recently. Kräemer and Sugiyama (2011) studied the intrinsic complexity of PLSR and assessed an unbiased estimate of its degrees of freedom. They have established two equivalent representations that rely on the close connection of PLS to matrix decompositions and Krylov subspace techniques. They have also showed that how this newly defined degrees of freedom estimate can be used for the comparison of different regression methods. A recent development, consistent PLS (PLSc) has been introduced to correct for bias (Dijkstra and Henseler 2015). Aguirre and Rönkkö (2017) employed bootstrap confidence intervals in conjunction with PLSc.

ICA has been widely used for blind source separation in many fields, such as brain imaging analysis, signal processing, telecommunication, and monitoring process. Chen and Bickel (2006) analyzed ICA using semiparametric theories

and proposed an asymptotically efficient estimate under moderate conditions. Wei (2015) studied the asymptotic normality and derived a closed-form analytic expression of the asymptotic covariance matrix of the generalized symmetric FastICA estimator using the method of estimating equation and M-estimator. Sokol et al. (2014) studied the consistent estimation of the mixing matrix in the ICA model when the error distribution is close to (but different from) Gaussian. Miettinen et al. (2015) investigated IC functionals based on the fourth moments in detail, starting with the corresponding optimization problems, deriving the estimating equations and estimation algorithms, and finding asymptotic statistical properties of the estimates.

2.3.6 Deep Learning

Deep learning has attracted tremendous attention from researchers in various fields of information engineering such as artificial intelligence, computer vision, and language processing, and even more traditional sciences such as physics, biology, and manufacturing. Neural networks, image processing tools such as convolutional neural networks, sequence processing models such as recurrent neural networks, and regularization tools such as dropout, are used extensively. However, fields such as physics, biology, and manufacturing are ones in which representing model uncertainty is of crucial importance.

With the recent shift in many of these fields towards the use of Bayesian uncertainty, new needs arise from deep learning. Yarin (2016) developed tools to obtain practical uncertainty estimates in deep learning, casting recent deep learning tools as Bayesian models without changing either the models or the optimization. First, he developed the theory for such tools and tied approximate inference in Bayesian models to dropout and other stochastic regularization techniques, and assess the approximations empirically. He also discussed what determines model uncertainty properties, analyzed the approximate inference analytically in the linear case, and theoretically examined various priors.

Mohamed, senior research scientist at Google DeepMind in London, constructed a view of deep feed-forward networks as a natural extension of generalized linear regression formed by recursive application of the generalized linear form (http://blog.shakirm.com/ml-series/a-statistical-view-of-deep-learning, 2015). Maximum likelihood was shown to be the underlying method for parameter learning. He also showed that auto-encoders address the problem of SI, and provide a powerful mechanism for inference that plays a central role in our search for more powerful unsupervised learning.

It is clear that there are great efforts to develop the SI of these methods or finding connections between these and their SI. We believe all of these statistical inferential advances towards these methods will power these algorithms. Furthermore, they will not be named with *black-box* predictive algorithms.

2.4 Discussion

In this chapter, data-driven methodologies in the manufacturing industry and their statistical inferential aspects are reviewed. The main advantage of ML algorithms in manufacturing is the ability to handle high dimensional data. Especially with the increasing availability of complex data, this feature will intuitively become even more important in the future. In addition, ML algorithms are very advantageous to discover implicit knowledge, and increased availability of source programs implementing the algorithms has allowed easy applications on real datasets. Although ML algorithms generally are very feasible for big data, there are still challenges to remain such as impact of redundant information (due to high-dimensionality) and pre-processing of data on the performance of learning algorithms, selection of an appropriate ML algorithm depending on the question of interest, and correct interpretation of the results. The developments in SI aspects of ML algorithms build strong foundations for the use of these algorithms beyond prediction. Prediction is definitely very important topic, however there are other questions that the researchers would be interested in such as interpretation of the results, estimation, statistical accuracy of ML algorithms, determining causal factors, significance of discovered patterns and so on, which require the use of SI methods.

We would like to end our chapter with an excellent example on how these fast algorithms may provide misleading results which may affect the society in general. In 2008, Google Flu Trends claimed it can tell us whether "the number of influenza cases is increasing in areas around the U.S. earlier than many existing methods." The algorithm, based on counts of internet search terms, outperformed *traditional medical surveys* in terms of speed and predictive accuracy. In 2013, Google Flu Trends was predicting more than double the proportion of doctor visits for flu than the Center of Disease Control (CDC). The algorithm badly failed by overestimating what turned out to be a nonexistent flu epidemic! This is an excellent example of why the predictive accuracy and the speed of the algorithms should not be the main focus. Disregarding inferential justification of any type as massive datasets have become available from the internet can yield quite dangerous results as in the case of Google Flu Trends. It is clear that purely empirical approaches are ultimately unsatisfying without some form of principled justification. The goal of SI is to connect ML algorithms to the central core of well-understood methodology. The connection process has already started. For instance, Efron and Hastie (2016) showed very elegantly how Adaboost (RF algorithm), the original ML algorithm, could be defined in the framework of logistic regression. All of the challenges that ML and SI communities possess demonstrate the astounding opportunities that lie ahead for collaboration in developing new powerful ML techniques equipped by the SI tools. Given the multi-disciplinary aspect of ML applications in manufacturing, it is very important to make efforts to develop connections between SI and ML to allow experts from different cultures to get together and tackle problems in a more efficient and accurate way.

References

Aguirre-Urreta MI, Rönkkö M (2017) Statistical inference with PLSc using bootstrap confidence intervals. MIS Quarterly.

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control, 19(6):716–723

Allaire JJ, Chollet F (2018) Keras: R interface to "Keras". R package version 2.1.3

Anderson TW (1963) Asymptotic theory for principal component analysis. Ann Math Stat 34(1):122–148

Auret L, Aldrich C (2010) Unsupervised process fault detection with random forests. Ind Eng Chem Res 49(19):9184–9194

Bai Y, Sun Z, Zeng B, Long J, Li L, Oliveira JVD, et al. (2018) A comparison of dimension reduction techniques for support vector machine modeling of multi-parameter manufacturing quality prediction. J Intell Manuf (in press). http://dx.doi.org/10.1007/s10845-017-1388-1

Bartlett PL, Jordan MI, McAuliffe JD (2006) Convexity, classification, and risk bounds. J Am Stat Assoc 101:138–156

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Ser B 57(1):289–300

Benkedjouh T, Medjaher K, Zerhouni N, Rechak S (2015) Health assessment and life prediction of cutting tools based on support vector regression. J Intell Manuf 26(2):213–223

Berk R, Brown L, Buja A, Zhang K, Zhao L (2013) Valid post-selection inference. Ann Stat 41(2):802–837

Bertino E, Catania B, Caglio E (1999) Applying data mining techniques to wafer manufacturing. In: Zytkow JM, Rauch J (eds) PKDD'99, LNAI, vol 1704. Springer, Berlin, pp 41–50

Biau G, Devroye L, Lugosi G (2008) Consistency of random forests and other averaging classifiers. J Mach Learn Res 9:2015–2033

Blanchard G, Bousquet O, Massart P (2004) Statistical performance of support vector machines. Technical Report

Boente G (1987) Asymptotic theory for robust principal components. J Multivar Anal 21:67–78

Breiman L (2001a) Random forests. Mach Learn 45:5–32

Breiman L (2001b) Statistical modeling: The two cultures. Stat Sci 16(3):199–231

Caydas U, Ekici S (2010) Support vector machines models for surface roughness prediction in CNC turning of AISI 304 austenitic stainless steel. J Intell Manuf 23:639–650

Chang YC, Mastrangelo C (2011) Addressing multicollinearity in semiconductor manufacturing. Qual Reliab Eng Int 27:843–854

Chen A, Bickel PJ (2006) Efficient independent component analysis. Ann Stat 34(6):2825–2855

Chiang LH, Pell RJ, Seasholtz MB (2003) Exploring process data with the use of robust outlier detection algorithms. J Process Control 13(5):437–449

Cho S, Asfour S, Onar A, Kaundinya N (2005) Tool breakage detection using support vector machine learning in a milling process. Int J Mach Tools Manuf 45(3):241–249

Critchley F (1985) Influence in principal components analysis. Biometrika 72:627–636

Dauxois J, Pousse A, Romain Y (1982) Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. J Multivar Anal 12(1):136–154

de Jong S (1993) SIMPLS: An alternative approach to partial least squares regression. Chemome Intell Lab Syst 18:251–263

de Ketelaere K, Hubert M, Schmitt E (2015) Overview of PCA based statistical process monitoring methods for time-dependent, high dimensional data. J Qual Technol 47:318–335

Deng L, Seltzer M, Yu D, Acero A, Mohamed A, Hinton GE (2010) Binary coding of speech spectrograms using a deep auto-encoder. In" Proceedings of 11th annual conference of the international speech communication association, vol 3, pp 1692–1695

Dijkstra TK, Henseler J (2015) Consistent partial least squares path modeling. MIS Q 39(2):297–316

Dunia R, Edgar TF, Nixon M (2013) Process monitoring using principal components in parallel coordinates. AIChE J 59(2):445–456
Efron B (2010) Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction. Institute of mathematical statistics onographs, Vol 1. Cambridge University Press, Cambridge
Efron B (2014) Estimation and accuracy after model selection (with discussion). J Am Stat Assoc 109(507):991–1007
Efron B, Hastie T (2016) Computer age statistical inference: algorithms, evidence, and data science. Institute of mathematical statistics monographs, 1st edn. Cambridge University Press, Cambridge
Efron B, Turnbull B, Narasimhan B (2015) locfdr: Computes local false discovery rates. R package version 1.1-8
Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96:1348–1360
Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33(1):1–22
Ge Z, Song Z (2010) A comparative study of just-in-time-learning based methods for online soft sensor modeling. Chemom Intell Lab Syst 104(2):306–317
Genuer R, Poggi JM, Tuleau C (2008) Random forests: some methodological insights. Technical report, INRIA
Hable R (2012) Asymptotic normality of support vector machine variants and other regularized kernel methods. J Multivar Anal 106:92–117
Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: prediction, inference and data mining, 2nd edn. Springer
Hihi SE, Bengio Y (1996) Hierarchical recurrent neural networks for long-term dependencies. Adv Neural Inf Process Syst 8:493–499
Hinton GE, Osindero S, Teh YW (2014) A fast learning algorithm for deep belief nets. Neural Comput 18(7):1527–1554
Hoerl A, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55–67
Hyvorinen A, Karhunen J, Oja E (2001) Independent component analysis, 1st edn. Wiley, New York
Irani KB, Cheng J, Fayyad UM, Qian Z (1993) Applying machine learning to semiconductor manufacturing. IEEE Exp 8:41–47
Jain P, Rahman I, Kulkarni BD (2007) Development of a soft sensor for a batch distillation column using support vector regression techniques. Chem Eng Res Des 85(2):283–287
Janssens O, Slavkovikj V, Vervisch B, Stockman K, Loccufier M, Verstockt S, et al. (2016) Convolution neural network based fault detection for rotating machinery. J Sound Vib 377:331–345
Javanmard A, Montanari A (2014) Confidence intervals and hypothesis testing for high-dimensional regression. J Mach Learn Res 15:2869–2909
Jia F, Lei Y, Lin J, Zhou X, Lu N (2016) Deep neural networks: a promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. Mech Syst Signal Process 72–73:303–315
Jolliffe IT (2002) Principal component analysis. Springer series in statistics, 2nd edn. Springer, New York
Kao LJ, Lee TS, Lu CJ (2016) A multi-stage control chart pattern recognition scheme based on independent component analysis and support vector machine. J Intell Manuf 27(3):653–664
Karoui N, Purdom E (2016) The bootstrap, covariance matrices and PCA in moderate and high-dimensions. arXiv:1608.00948
Kräemer N, Sugiyama M (2011) The degrees of freedom of partial least squares regression. J Am Stat Assoc 106(494):697–705
Le S, Josse J, Husson F (2008) FactoMineR: An R package for multivariate analysis. J Stat Softw 25(1):1–18

Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–324
Lee MH (1993) Knowledge based factory. Artif Intell Eng 8:109–125
Lee J, Sun D, Sun Y, Taylor J (2016) Exact post-selection inference, with application to the Lasso. Ann Stat 44(3):907–927
Lee JM, Yoo C, Choi SW, Vanrolleghem PA, Lee IB (2004) Nonlinear process monitoring using kernel principal component analysis. Chem Eng Sci 59(1):223–234
Li Y, Tsung F (2012) Multiple attribute control charts with false discovery rate control, quality and reliability engineering international. Wiley Online Library, vol 28, pp 857–871. https://doi.org/10.1002/qre.1276
Liaw A, Wiener M (2002) Classification and regression by randomforest. R News 2(3):18–22
Lim HK, Kim Y, Kim MK (2017) Failure prediction using sequential pattern mining in the wire bonding process. IEEE Trans Semicond Manuf 30(3):285–292
Lin Y (2000) Some asymptotic properties of the support vector machine. Technical report 1029. Department of Statistics, University of Wisconsin-Madison
Lin Y (2002) A note on margin-based loss functions in classification. Statist Probab Lett 68:73–82
Lockhart R, Taylor J, Tibshirani R, Tibshirani R (2014) A significance test for the Lasso. Ann Stat 42(2):413–468
Malhi A, Yan R, Gao RX (2011) Prognosis of defect propagation based on recurrent neural networks. IEEE Trans Instrum Meas 60(3):703–711
Mallows CL (1973) Some comments on C_P. Technometrics 15(4):661–675
Marchini JL, Heaton C, Ripley BD (2017) fastICA: FastICA algorithms to perform ICA and projection pursuit. R package version 1.2–1
Melhem M, Ananou B, Ouladsine M, Pinaton J (2016) Regression methods for predicting the product's quality in the semiconductor manufacturing process. IFAC-papers online, vol 49, pp 83–88
Mentch L, Hooker G (2016) Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. J Mach Learn Res 17:1–41
Mentch L, Hooker G (2014) Ensemble trees and CLTs: statistical inference for supervised learning. arXiv preprint arXiv:1404.6473
Mevik BH, Wehrens R, Liland KH (2016) pls: Partial least squares and principal component regression. R package version 2.6-0
Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2017) e1071: Misc functions of the Department of Statistics, Probability Theory Group, (Formerly: E1071), TU Wien. R package version 1.6-8
Miettinen J, Taskinen S, Nordhausen K, Oja H (2015) Fourth moments and independent component analysis. Stat Sci 30:372–390
Miller Jr RG (1981) Simultaneous statistical inference. Springer series in statistics, 2nd edn. Springer, New York
Mohamed S (2015) http://blog.shakirm.com/ml-series/a-statistical-view-of-deep-learning
Oksanen J, Blanchet GF, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H (2017) vegan: Community ecology package. R package version 2.4-5
Pardo M, Sberveglieri G (2008) Random forests and nearest Shrunken centroids for the classification of sensor array data. Sens Actuators B Chem 131:93–99
Puggini L, Doyle J, McLoone S (2016) Fault detection using random forest similarity distance. IFAC-Safe Process 49(5):132–137
Qin SJ (2003) Statistical process monitoring: basics and beyond. J Chemom 17:480–502
R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
Ribeiro B (2005) Support vector machines for quality monitoring in a plastic injection molding process. IEEE Trans Syst Man Cybern C (Appl Rev) 35:401–410
Saidi L, Ail JB, Friaiech F (2015) Application of higher order spectral features and support vector machines for bearing faults classification. ISA Trans 54:193–206

Saybani MR, Wah TY, Amini A, Yazdi S, Lahsasna A (2011) Applications of support vector machines in oil refineries: A survey. Int J Phys Sci 6(27):6295–6302
Schmidhuber J (2015) Deep learning in neural networks: an overview. Neural Netw 61:85–117
Schölkopf B, Burges C, Smola A (1999) Advances in kernel methods: support vector learning. MIT Press, Cambridge
Scovel JC, Steinwart I (2004) Fast rates for support vector machines using gaussian kernels. Technical report LA-UR04-8796, Los Alamos National Laboratory
Smolensky PI (1986) Information processing in dynamical systems: foundations of harmony theory, parallel distributed processing: explorations in the micro structure of cognition. MIT Press, Cambridge
Sokol A, Maathuis MH, Falkeborg B (2014) Quantifying identifiability in independent component analysis. Electron J Stat 8:1438–1459
Steinwart I (2005) Consistency of support vector machines and other regularized kernel machines. IEEE Trans Inform Theory 51:128–142
Susto GA, Beghi A (2013) A virtual metrology system based on least angle regression and statistical clustering. Appl Stoch Models Bus Ind 29:362–376
Tenenbaum JB, Silva VD, Langford JC (2010) A global geometric framework for nonlinear dimensionality reduction. Science 290:2319–2323
Tian Y, Fu M, Wu F (2015) Steel plates fault diagnosis on the basis of support vector machines. Neurocomputing 151:296–303
Tibshirani R (1996) Regression shrinkage and selection via the LASSO. J R Stat Soc Ser B 58(1):267–288
Tibshirani R, Taylor J, Loftus J, Reid S (2016) selectiveInference: tools for post-selection inference, R package version 1.1.3
Thornhill NF, Shah SL, Huang B, Vishnubhotla A (2002) Spectral principal component analysis of dynamic process data. Control Eng Pract 10(8):833–846
van de Geer S, Bühlmann P, Ritov Y, Dezeure R (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. Ann Stat 42(3):1166–1202
Vapnik V (1995) The nature of statistical learning theory. Springer
Wager S, Athey A (2018) Estimation and inference of heterogeneous treatment effects using random forests, J Am Stat Assoc 113:1228–1242. http://dx.doi.org/10.1080/01621459.2017.1319839
Wager S, Hastie T, Efron B (2014) Confidence intervals for random forests: The Jackknife and the infinitesimal Jackknife. J Mach Learn Res 15:1625–1651
Wang XZ, McGreavy C (1998) Automatic classification for mining process operational data. Ind Eng Chem Res 37(6):2215–2222
Wang P, Gao RX, Yan R (2017) A deep learning-based approach to material removal rate prediction in polishing. CIRP Ann Manuf Technol 66:429–432
Wang J, Ma Y, Zhang L, Gao RX, Wu D (2018) Deep learning for smart manufacturing: methods and applications. J Manuf Syst 48(Part C):144–156
Wei T (2015) The convergence and asymptotic analysis of the generalized symmetric fast ICA algorithm. IEEE Trans Signal Process 63(24):6445–6458
Weimer D, Scholz-Reiter B, Shpitalni M (2016) Design of deep convolution neural network architectures for automated feature extraction in industrial inspection. CIRP Ann Manuf Technol 65(1):417–420
Westfall P, Young S (1993) Resampling-based multiple testing: examples and methods for p-value adjustment. Wiley series in probability and statistics. Wiley-Interscience
Widodo A, Yang BS (2007) Support vector machine in machine condition monitoring and fault diagnosis. Mech Syst Signal Process 21:2560–2574
Wold H (1975) Path models with latent variables: the NIPALS approach. In: Quantitative sociology international perspectives on mathematical and statistical model building, pp 307–357. Academic Press
Wu D, Jennings C, Terpenny J, Gao RX, Kumara S (2017) a comparative study on machine learning algorithms for smart manufacturing: tool wear prediction using random forests. J Manuf Sci Eng 139:071018–071027

Xanthopoulos P, Razzaghi T (2013) A weighted support vector machine method for control chart pattern recognition. Comput Ind Eng 66:683–695
Xiao Y, Wang H, Zhang L (2014) Two methods of selecting gaussian kernel parameters for one-class SVM and their application to fault detection. Knowl-Based Syst 59:75–84
Yang B, Di X, Han T (2008) Random forests classifier for machine fault diagnosis. J Mech Sci Technol 22:1716–1725
Yao M, Wang H (2015) On-line monitoring of batch processes using generalized additive kernel principal component analysis. J Process Control 103:338–351
Yarin G (2016) Uncertainty in deep learning. Ph.D. thesis, Cambridge University
You D, Gao X, Katayama S (2015) WPD-PCA-based laser welding process monitoring and defects diagnosis by using FNN and SVM. IEEE Trans Ind Electron 62(1):628–636
Yu J (2012) A Bayesian inference based two-stage support vector regression framework for soft sensor development in batch bioprocesses. Comput Chem Eng 41:134–144
Yu H, Khan F, Garaniya V (2015) Nonlinear Gaussian belief network based fault diagnosis for industrial processes. J Process Control 35:178–200
Zhang T (2004) Statistical behavior and consistency of classification methods based on convex risk minimization. Ann Stat 32:56–84
Zhang Y, Teng Y, Zhang Y (2010) Complex process quality prediction using modified kernel partial least squares. Chem Eng Sci 65(6):2153–2158
Zhang Y (2008) Fault detection and diagnosis of nonlinear processes using improved kernel independent component analysis (KICA) and support vector machine (SVM). Ind Eng Chem Res 47(18):6961–6971
Zhang W, He D, Jia R (2013) Online quality prediction for cobalt oxalate synthesis process using least squares support vector regression approach with dual updating. Control Eng Pract 21(10):1267–1276
Zhang Y, Li S, Teng Y (2012) Dynamic processes monitoring using recursive kernel principal component analysis. Chem Eng Sci 72:78–86
Zhang C-H, Zhang S (2014) Confidence intervals for low-dimensional parameters with high-dimensional data. J R Stat Soc Ser B 76(1):217–242
Zou C, Tseng ST, Wang Z (2014) Outlier detection in general profiles using penalized regression method. IIE Trans J Inst Ind Syst Eng 46(2):106–117
Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B 67:301–320
Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101:1418–1429

Nedret Billor has received her Ph.D. in Statistics from University of Sheffield, UK. She is currently serving as a professor of Statistics in the Department of Mathematics and Statistics at Auburn University, Alabama. She served as the Director of Statistics Program from 2008–2010. Her primary interests include robust multivariate data analysis, robust functional data analysis, and outlier detection. Dr. Billor has published numerous technical papers and co-authored a paper which received an honorable mention award in Computers and Biology in Medicine. She is Elected Member of ISI since 2012 and serves as the country representative of the ISI Committee on Women in Statistics (CWIS). Dr. Billor is the associate editor of Communications in Statistics in Theory and Methods and Simulation and Computation. She was awarded to be Fullbright Specialist Scholar in 2016. She has advised numerous Ph.D./M.S. students and served on numerous interdisciplinary Ph.D./M.S. graduate committees. She is the recipient of the

awards for Outstanding Teacher of Graduate Students and Excellence in Teaching. Dr. Billor made the decision of choosing a path in STEM field at a very young age because she had strong belief that it consists of the concepts that relate to solving real-life problems thus, technology and science would have the power to change the world. The developments and innovations in science and technology have been leading the way to find solutions to some of the world's biggest problems and drastically change our society for the better. She is honored to be a member of Mathematics and Statistics research community, continue to contribute to this field and work towards making a difference in the society.

Asuman S. Turkmen has received her Ph.D. in Statistics from Auburn University, AL. She is currently employed by The Ohio State University at Newark as an associate professor of Statistics. Her research focuses on multivariate statistical methods that deal with robust estimation, and statistical genetics, specifically the identification of rare variant associations with complex traits. She is the recipient of the Faculty Scholarly Accomplishment Award (2015), and Service Award (2017) at OSU Newark. Math was always my favorite subject in school—and later as a professional, I was enamored by numbers. I was aware of the fact that pursuing a STEM major will not only allow me a wide variety of future opportunities after graduation but also help me to make a positive contribution to society since everyday life is constantly affected by professionals from STEM programs.

Chapter 3
Specifying and Validating Probabilistic Inputs for Prescriptive Models of Decision Making over Time

Sarah McAllister Ryan

Contents

3.1 Introduction 49
3.1.1 Recurring Applications 51
3.2 Stochastic Process Modeling 52
3.3 Discretization 55
3.3.1 Scenario Reduction 57
3.3.2 Comparative Granularity 58
3.4 Solution Methods 59
3.5 Comprehensive Assessment 61
3.5.1 Direct Assessment of Scenario Generation Methods 61
3.5.2 Assessing Solutions by Re-enactment 64
3.6 Conclusions 65
References 66

3.1 Introduction

Each day, a flowergirl must decide on a quantity of fresh flowers to purchase at the wholesale cost before finding out how many she is able to sell at the retail price that day. Unlike her brother, the newsboy, she is able to hold onto some fraction of unsold inventory—flowers that remain fresh enough to sell on the following day. Her daily problem is to choose a purchase quantity to maximize profit over a sequence of days. If she buys too many, she wastes money on flowers that cannot be sold. If she buys too few, then she either incurs an opportunity cost of uncollected revenue (Casimir 1990) or is forced to pay the retail price to make up the difference (Pflug and Pichler 2014). Having operated this business for a while, she has data on past

This chapter is dedicated to the memory of my mother, Janice Crawford McAllister (1929-2017).

S. M. Ryan (✉)
Iowa State University, Ames, IA, USA
e-mail: smryan@iastate.edu

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_3

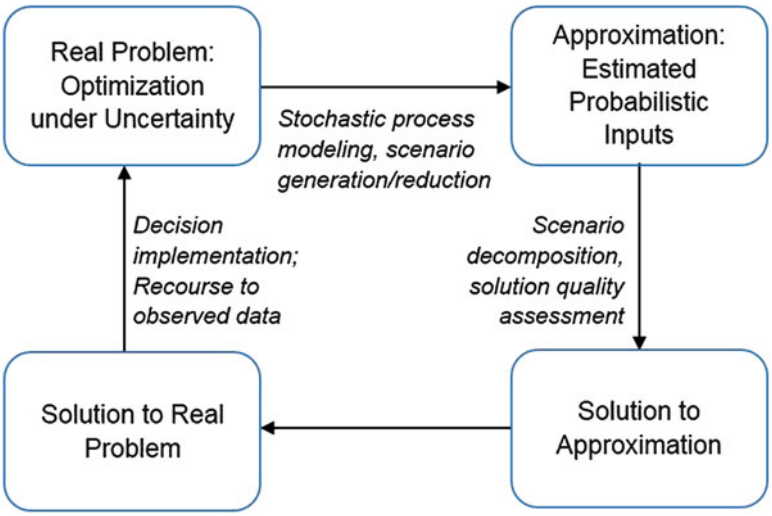


Fig. 3.1 Stochastic modeling and optimization process

demand that she can combine with knowledge about future events such as holidays to estimate a joint probability distribution for demand in the days ahead. The ability to carry inventory means that, unlike the newsboy, she cannot simply compute and purchase a critical fractile of the demand distribution for a single day.[1]

A rolling strategy for solving problems such as the flowergirl's is depicted in Fig. 3.1. At each time, t, given a decision problem with uncertain parameters, we fit a stochastic process model using the data available at that time. Next, we approximate the space of possible realizations, keeping solution procedures in mind, and solve the approximated problem. The solution procedure itself may introduce further levels of approximation. Finally, we implement the decisions that must be taken at time t, roll forward to time $t + 1$, and repeat the process.

Errors can enter the process at any step—in fact, they are deliberately introduced in the form of approximations employed for computational tractability. For the decision maker to accept the sequence of decisions suggested, she must be persuaded that the whole process of modeling, approximation, and solution is sound. Given that, in an uncertain world, more or less sound decisions can be followed by either good or bad consequences, the best decision justification we can offer is that we continually did the best we could with the information we had available. How can we support this claim with data?

[1]Generally I prefer the gender-neutral term "newsvendor," but in this book chapter I wish to emphasize that the professional problem solved by the woman may be more complicated than the one her male counterpart faces!

The following sections describe research on various aspects of this rolling process for decision making under uncertainty conducted over the past three decades. Recurring themes that pervade this work include the use of data-driven methods to specify instances, an emphasis on optimizing the initial one of a sequence of decisions, and a reliance on approximation and solution methods that decompose the stochastic optimization problem according to the possible future realizations of the uncertain parameters. A section is devoted to each step of the rolling process, culminating in some recent efforts to comprehensively assess the whole cycle. Forecasters test the quality of outputs by *backcasting*; i.e., testing how accurately their methods would have predicted values in a historical dataset, while financial traders test trading strategies by *backtesting* them on historical datasets. Similarly, I argue that the process depicted in Fig. 3.1 should be assessed by *re-enactment* over a sequence of historical instances. This process is distinguished from simulation by the use of actual observations rather than randomly generated data. The chapter concludes with research needs.

3.1.1 Recurring Applications

Much of the work described in this chapter has focused on two types of resource management problems, one with a long term orientation and the other having a short term perspective. Planning for the long term inherently involves uncertainty because of the difficulty of forecasting demands and costs in advance. Uncertainty is also present in short-term scheduling of assets when they rely on variable inputs while demand is governed by both physical processes and consumer decisions. Both applications concern the provision of services where the impracticality of either storing inventory or delaying delivery mandates that demand is satisfied when it is experienced.

Capacity Expansion Planning capacity additions to meet a growing demand for service is challenging because the rate of future demand growth is difficult to predict, while the facilities that can be employed to meet the demand typically are not continuously expandable. Capacity comes in chunks because of physical constraints and/or scale economies. Different types of facilities may exist with various combinations of investment levels, operational costs, and lead times for building or installing them. Demand may vary continuously over time and be spatially distributed with physical limits on transportation from the facilities to the demand locations. The time value of money affects the relationship between immediate and future costs. Ryan et al. (2011) summarized challenges of resource planning in the electric power industry.

Unit Commitment In electric power systems, thermal generating units are subject to operating constraints and cost characteristics that limit their flexibility to change

production levels. Unit commitment is the problem of scheduling which units will run at each future time point in view of these considerations, which take the form of minimum up- and downtime constraints; limits on how quickly units can start up or shut down and how fast production can ramp up or down when they are running; and fixed costs for producing at any positive level, which result in minimum economically feasible production levels in addition to upper limits imposed by their capacities. Different types of units vary in the severity of these constraints as well as in their variable production costs, which mostly depend on fuel cost. The deepening penetration of variable renewable generation, such as wind and solar energy, has increased the uncertainty in the net demand for electricity that thermal units are required to supply. At the same time, wholesale market rules impose strict limits on the amount of time available for optimizing the unit commitment schedule. The challenge is to build a credible model of uncertainty and produce a high-quality solution within the time limits.

3.2 Stochastic Process Modeling

Ryan (1988) used an indirect approach to addressing uncertainty in optimization problems, such as production planning or capacity expansion, formulated over indefinite time horizons. This work began with a deterministic formulation of a dynamic optimization problem over an infinite time horizon and used the concepts of a *forecast horizon* and a *solution horizon*. A forecast horizon is a time horizon sufficiently long that any information pertaining to events after it ends has no impact on the decision that is optimal to implement immediately. Under the assumption of a unique optimal initial decision, a forecast horizon was defined in Ryan and Bean (1989) as a time horizon length such that, if the problem is solved over a finite horizon at least that long, the (unique) optimal decision will coincide with the optimal initial decision for the infinite horizon problem. Relaxing the uniqueness assumption, Ryan et al. (1992) defined a (general) solution horizon as a time horizon long enough that, if the problem is solved over a finite horizon at least that long, *any* optimal initial decision identified is guaranteed to be optimal for the infinite horizon problem. The underlying assumption in this work is that events can be forecast with reasonable accuracy over some initial time period. The question was, how long must that initial time period be for a decision maker to confidently implement an initial decision? If future costs are discounted sufficiently, then future uncertainties diminish in importance.

Ryan et al. (1992) continued in this vein by developing a method for breaking ties among alternative optimal solutions so that a solution horizon could be identified. This paper included a numerical study of capacity expansion assuming a finite set of facility types characterized by their capacities and fixed costs, as well as a known function for the cumulative demand for capacity over time, eventually bounded by an exponential function with a rate less than the interest rate used to continuously discount future costs. Because facility costs exhibited economies

of scale, eventually a "turnpike" facility, offering the most cost-efficient capacity growth per unit time, would be adopted. However, finding the initial sequence of optimal installations before this turnpike policy took effect required optimization. The numerical results indicated that solution horizons could be substantially shorter than forecast horizons, so that an optimal initial decision could be identified easily even under almost complete uncertainty about demand growth in the long run.

Experience with forecasting demand growth for use in expansion planning in the utilities division of a large chemical manufacturer prompted me to consider how to model uncertainty in a realistic way that managers would appreciate. This division was responsible for producing steam used for process heat, to run mechanical equipment, and to cogenerate electric power. While electricity generation was supplemented by outside purchases, steam production was entirely internal to the plant. Because having insufficient steam pressure could result in product quality degradation or even force a partial plant shutdown, planning sufficient boiler capacity was critical. The existing planning process was to annually generate a five-year forecast for demand growth. To protect against forecast errors and allow for the lead time required to procure and install equipment, a fixed margin was added to the forecast. The need to expand capacity was signaled if the augmented demand growth was projected to equal or exceed existing capacity within this five-year planning horizon.

Two major features appeared to be important to include in a model for uncertain demand growth for use in such an environment. One was to explicitly represent forecast revisions that occurred in response to demand data observed each year. The other was to replace the fixed margin with a probabilistic envelope around the forecast that would reflect the increase in uncertainty associated with forecasts of more remote time periods. The utility managers were comfortable with statistical confidence intervals and had retained historical records of monthly steam usage. Some questions that arose were how often to update the forecast and what would be a suitable confidence level to use in constructing prediction intervals. These two model parameters are related because less frequent forecast revisions could decrease the accuracy of the prediction limits, while requiring a higher confidence level would magnify the effect of inaccurate point forecasts. Ryan (1998) describes an empirical study in which a time series model was fit to the historical data and the process of repeatedly generating forecasts with the latest available data was *re-enacted*, augmented by either a fixed or a probabilistic margin. Upon each forecast revision, the optimization problem to determine the optimal timing of additional boiler installation was solved using the augmented demand and implementation of any capacity increment to occur before the next roll forward was recorded. The resulting re-enacted capacity expansion policies were compared according to their combinations of total discounted cost and measures of insufficient capacity. Based on an efficient frontier constructed with these two performance measures, the value of more frequent forecasts became apparent while cost-risk profiles of the different capacity margins could be assessed. By assigning a penalty value to capacity shortages, McAllister and Ryan (2000) used first-order stochastic dominance in an

expanded simulation study to select the best combination of forecast frequency and capacity margin.

Although the use of fixed or probabilistic capacity margins was partially motivated by the lead times required to expand capacity, those lead times had not been modeled explicitly in optimization models, going all the way back to classical work by Manne (1961) and Freidenfelds (1981). The model formulated in Ryan (2003) considered them as fixed constants while representing demand according to a time series model with parameters that could be estimated from data. The choice of an integrated moving average model allowed different aspects of the uncertainty in demand; namely, its autocorrelation which results in nonstationary expected growth and its random variation about the expectation, to be isolated. The presence of lead times suggested that expansions should be based on the capacity *position*, similar to the inventory position commonly used in inventory models, that includes not only existing capacity but any capacity in the process of being added. The optimality of an expansion timing policy based on the proximity of uncertainty-inflated demand to this capacity position could then be proved. An approximation for the optimal expansion size was found by adapting a continuous-time optimal expansion policy to the discrete time setting compatible with the demand growth model. Simulation studies revealed the effects of autocorrelation and randomness in the demand growth on the threshold of excess capacity position that would optimally trigger an expansion. The main conclusion was that failing to account for autocorrelation (or nonstationarity) in the demand growth model could lead to overestimating the randomness and expanding capacity too early, resulting in higher than necessary discounted costs.

To explore the effect of expansion lead times analytically, Ryan (2004) modeled demand growth as following a geometric Brownian motion (GBM) process, in line with earlier work by Manne (1961) and Bean et al. (1992), and showed how to optimize expansions to minimize their expected discounted cost subject to a service level constraint. Meanwhile, real options models had been rapidly increasing in popularity as a way to assess the value of the flexibility provided by some investment alternatives to respond to uncertain future events. Motivated by the success of the Black-Scholes formula for the value of a European call option on a stock along with analogies between financial options and operational flexibility, many authors formulated models for investment or even operational decision problems that relied on an explicit or implicit assumption that the value of some "underlying asset" would follow a GBM process. In the engineering economic analysis literature, some examples of such GBM-following variables cited by Marathe and Ryan (2005) included both sales volume and price of a product; internal production, outsourcing and delivery costs of a product; prices of commodities derived from natural resources; present values of cash flows pertaining to equipment operation; and various other physical asset values. Such assumptions did not frequently appear to be based on any analysis of data. Relying heavily on Ross (1999), we proposed simple statistical tests of historical time series data that would verify the fundamental assumptions of the GBM model and allow for reliable estimates of the model's parameters. Applying them to historical data concerning electricity

consumption, airline passenger enplanement, revenue from cell phone subscriptions and number of Internet hosts, we found that the data were consistent with the GBM assumptions in the first two instances but not the last two. It would be interesting to re-examine updated data concerning demand for capacity in the two then-nascent industries for which we found the GBM model to not fit well. Meanwhile, Marathe and Ryan (2009) employed formulas for pricing exotic options to evaluate the potential for shortage during the lead time required to add capacity, assuming GBM demand growth.

The capacity expansion studies described above employed stochastic process models directly in continuous-time dynamic programming problem formulations. The formulations were simple enough that at least some aspects of the form of an optimal policy could be derived analytically and only modest computation was necessary to find optimal solutions. Sample paths of the stochastic process models were generated only for the purpose of simulating or re-enacting the process of estimating parameters, constructing forecasts and probability limits, and computing the corresponding decision sequences. In situations where operational costs vary widely according to the investment decisions chosen, stochastic programming models are more suitable. Efforts to discretize stochastic process models in order to generate scenarios are described more in Sect. 3.3. While the emphasis shifts to finding a relatively small set of scenario paths that well represent the whole space of possible future realizations, it is important to not neglect the identification of an appropriate stochastic process. For example, Jin et al. (2011) applied the statistical tests suggested by Marathe and Ryan (2005) to validate the use of GBM models for both demand for electricity and the price of natural gas before applying a moment-matching procedure to generate scenarios for a two-stage model of electricity generating capacity expansion.

This section concludes with a recent effort to build a stochastic process model that can be used to generate probabilistic scenarios for short-term planning. Feng and Ryan (2016) combined various methods including a functional regression method based on epi-splines (Royset and Wets 2014) to develop a model of demand for electricity based on a day-ahead weather forecast while capturing typical temporal patterns and accounting for seasonal and geographic information. While the accuracy of the forecast can be assessed according to the usual measures of mean squared error and mean absolute percentage error, the shape of the distribution of forecast errors plays an important role in generating probabilistic scenarios based on the model. Our approach resulted in both tighter and less skewed error distributions than commonly used benchmark models.

3.3 Discretization

Once a stochastic process model for the evolution of uncertain parameters is identified, the next step is to find a tractable representation of it for use in optimization. With a few exceptions, such as the infinite horizon generic capacity

expansion models described in Sect. 3.2, a continuous-time and -state stochastic process model cannot be used directly in optimization. The process of formulating a discrete set of future realizations, with associated probabilities, has been investigated under the label of scenario generation. One popular approach is to randomly generate a large collection of sample paths and then apply so-called scenario reduction procedures to identify a representative subset. While some methods for stochastic optimization embed the scenario generation or sampling process within the optimization procedure, I focus attention on methods where the representation of uncertainty is completed before the solution procedure commences.

Study of a medium-term energy planning problem sparked my interest in this issue. Quelhas et al. (2007) had formulated a multiperiod generalized network flow model for bulk energy flows in the US, and Quelhas and McCalley (2007) had validated it against actual utilization of different primary energy sources to meet the demand for electric energy over a year. While coal prices were quite stable, the volatility in the prices of natural gas and crude oil made the assumption of deterministic fuel costs seem unrealistic. In the first few years of the twenty-first century, natural gas generation had grown to account for a significant share of electricity generation in the US because of the relative flexibility and lower emissions of gas-fired generating units compared to coal-fired ones. However, before innovations in shale gas extraction took hold, the price of natural gas was generally increasing with considerable volatility from year to year. Our goal in Wang and Ryan (2010) was to add a representation of the fuel cost uncertainty to the network flow model and investigate the impact of this uncertainty on resource utilization decisions. Because the model of Quelhas et al. (2007) was formulated on a discrete-time basis, a natural approach was to formulate the deterministic equivalent of a two-stage stochastic program where flows for one period composed the first-stage decisions and flows for later periods composed the recourse decisions that could be delayed until after the fuel prices for those periods were realized. We adopted a receding horizon approach to simulate the process of monthly decision making with updates on the fuel price forecasts. We used just three possible values of natural gas price in each month, corresponding to the point forecast and its lower and upper confidence limits according to forecasts published by the US Department of Energy. Even so, the assumption of independence between periods resulted in a large number of scenario time series. When they were combined with the thousands of nodes and arcs in the original deterministic formulation, the extensive form of the deterministic equivalent became prohibitively large.

Several approaches exist for managing the computational difficulties associated with solving with the large-scale deterministic equivalent, and most realistic applications require a combination of them. One is to apply decomposition methods, such as those based on scenario decomposition as described in Sect. 3.4 or Benders decomposition as applied by Wang and Ryan (2010). Another is to limit the number of scenarios used, as described in Sect. 3.3.1. Both of these approaches assume the deterministic formulation is fixed and approximate either the joint probability distribution of the uncertain parameters (in the form of a scenario tree) or the optimal solution of the problem based on a given scenario tree. A third approach,

discussed in Sect. 3.3.2, is to consider more carefully the relative value of detail in the deterministic formulation as opposed to the scenario tree.

3.3.1 Scenario Reduction

Wang (2010) investigated the existing scenario reduction approaches based on probability metrics (Dupačová et al. 2003; Heitsch and Römisch 2003) and deemed them unsatisfactory because they operate entirely within the probability space of the stochastic process realizations without considering the optimization context. In fact, they are motivated by results concerning the stability of the optimal first-stage decisions with respect to the discrete approximations of the continuous probability distributions for the uncertain parameters. However, there are two levels of approximation present in these, by now, "classical" methods of scenario reduction. First, proximity of the solution found using the reduced scenario set to the true optimal solution is expressed in terms of an upper bound on the distance in cost, not the distance itself. Second, the optimization problem to find a reduced scenario set that minimizes this upper bound is only approximately solved using fast heuristics such as fast forward selection (Heitsch and Römisch 2007)—otherwise, the scenario reduction procedure could be less tractable even than the optimization problem it is intended to simplify. To inject some information about the optimization context into the reduction procedure, Wang developed a heuristic approach that employed the forward selection heuristic within clusters of scenarios identified on the basis of their similarity in terms of their optimal first-stage decisions.

Feng and Ryan (2013) elaborated this idea and applied it to the electricity generation expansion planning model of Jin et al. (2011). The moment-matching procedure was simplified to take advantage of the stationarity property of the GBM processes. However, even with only two or three branches in the scenario tree each period, the number of scenario paths was too large to allow for solving the extensive form of the deterministic equivalent over a realistic time horizon. As Wang had proposed, we solved the deterministic "wait-and-see" subproblem for each scenario, characterized the optimal decision in terms of a few summary descriptors, and then clustered the scenarios based on similarity of these descriptors for the resulting optimal decisions. By applying fast forward selection to choose one scenario from within each cluster, we obtained a reduced set of scenarios that performed similarly to a set of the same size found by applying fast forward selection to the whole set of scenarios. However, the time required for our reduction procedure was much lower and, unlike forward selection, remained approximately constant regardless of the desired number of scenarios in the reduced set.

For stochastic unit commitment, limiting the number of scenarios is critical for obtaining high quality commitment schedules in the limited time allowed by market rules on the day before the target day. Feng et al. (2015) combined segmentation of similar days with epi-spline functional regression to develop stochastic process models for the hourly load, incorporating the uncertainty associated with weather

forecasts. Rather than generating randomly sampled paths, we carefully constructed probabilistic scenarios by approximation using conditional expectations. Probabilistic scenarios for wind energy generation were obtained from a commercial vendor based on numerical weather prediction models. The net load scenarios, representing possible time series of load less the wind generation amounts, were formed by crossing the two sets of scenarios. Thus, although the sets of scenarios had been carefully constructed to be small, we still ended up with large sets of net load scenarios. To reduce their number, Feng and Ryan (2016) further developed the approach of Feng and Ryan (2013). In this variant, scenarios were clustered based on the major components of the objective function; namely, the production cost and the positive or negative imbalance between energy produced and the net load in each hour. Compared with the unit commitment schedules found by using fast forward selection, those produced by optimization with our reduced scenario set provided more reliable electricity delivery and were more similar to the schedules produced by using the whole set of scenarios.

3.3.2 Comparative Granularity

Quelhas and McCalley (2007) validated their deterministic model by comparing its optimal network flows with the actual amounts of fuel transported and utilized for electricity generation as well as the electricity transmitted among regions in case studies of two separate past years. They attributed differences between the optimal and actual network flows to the spatial and temporal aggregation necessitated by limitations in the available data and the absence of market interactions in the model, as well as the lack of representation of uncertainty and future expectations by decision makers. Wang and Ryan (2010) attempted to represent uncertainty in fuel costs, as well as changing expectations concerning them, by re-enacting the solution of a stochastic program where the scenarios represented both the forecasts and the associated levels of uncertainty, with forecast updates included in the receding horizon procedure. When this was done, the multiperiod flows comprising the sequence of first-stage decisions that would be implemented in the receding horizon procedure were quite similar to the actual decisions that had been made. As we concluded in the paper, "When model validation is unsatisfactory, analysts frequently strive to include more temporal or spatial detail. Our results suggest that incorporating stochastic variability may be another practical way to improve model fidelity, especially when historical forecasts are available but disaggregated temporal and spatial data are not."

Similar issues arise in infrastructure planning, specifically electricity generation and transmission expansion planning. Practitioners advocate a procedure called scenario planning, where they define a *scenario* as a description of possible future conditions under which the infrastructure would be operated, usually at a single future time point. Electricity system resource planners sometimes use the word "future" instead, where a future could describe global system characteristics

and policy choices such as degree of penetration of renewable energy; and the presence or absence of carbon emission regulations, large-scale energy storage, and demand response mechanisms. A detailed deterministic operational model is used to optimize investments in infrastructure for each future, with the goal of identifying investment decisions that are common across all futures. Muñoz et al. (2014) provide a clear description and critique of this approach, as compared with stochastic programming, in a transmission expansion planning case study for the western US. The weaknesses of scenario planning include the lack of any assessment of the relative likelihood of the futures considered and the possibility that a decision that is optimal for each scenario individually is not optimal when they are considered simultaneously.

However, the intuitive appeal of this approach has led to its widespread adoption and the related assumption that operational models must be sufficiently detailed to accurately assess the value of infrastructure investments. Including a high level of operational detail produces a large scale multiperiod optimization model, with both high-dimensional decision variables, some of which are discrete to represent nonconvexities, and many constraints to capture the details of system operation under temporal variation. As a result, planners are reluctant to consider many different futures or scenarios because simulating operation with each one is so expensive computationally. In such a context, a stochastic program with multiple probabilistic scenarios to be considered simultaneously appears impractical. Jin et al. (2014) formulated a stochastic program for thermal generation expansion planning with probabilistic scenarios representing availability of wind power in a typical year. To control the size of the extensive form, we compared the results of different simplifications. One was to decrease the stochastic granularity by reducing the number of wind energy scenarios considered and the other was to decrease the temporal granularity by dropping the nonconvex unit commitment constraints while retaining the continuous ramping restrictions. In case studies comparing the results of both approximations with the full model, we found that the more granular stochastic representation combined with coarse-grained operational constraints resulted in more accurate solutions and more efficient computation than the coarse-grained stochastic representation combined with highly detailed operational constraints. Accuracy of the solution was judged according to similarity with the solution obtained by solving the full model with high detail in both the stochastic and temporal representations.

3.4 Solution Methods

Both capacity expansion and unit commitment are naturally formulated as stochastic mixed integer programs (SMIPs) because of the discrete character of the primary decisions. In capacity expansion, increments of capacity typically are not available in continuous sizes because of economies of scale and other design considerations for durable equipment or the construction of major facilities. The decision variables

that describe operations may also be discrete because of minimum run-time or production level constraints, discontinuities in marginal cost, or nonlinearities that are approximated as piecewise linear. In unit commitment, binary decision variables are used to express the fundamental on/off decisions as well as nonlinear or nonconvex operational features. In realistically scaled instances, the deterministic subproblem for a single scenario may be challenging to solve in a reasonable amount of time. In both application contexts, considerable research has been devoted to devising reformulations and decomposition methods to solve the deterministic instances efficiently.

Including multiple probabilistic scenarios for parameter values exacerbates the computational challenge and motivates the development of approximate solution methods. Various decomposition methods have been explored including Benders (stage-wise) decomposition and Dantzig-Wolfe decomposition (column generation), as well as Lagrangian relaxation of "complicating constraints." We have focused on scenario decomposition, which can be viewed as relaxation of the nonanticipativity that is expressed either implicitly or explicitly in the formulation of a SMIP. Nonanticipativity is expressed implicitly by formulating the problem in terms of decision stages, where the decision variables in a given stage can depend on realizations of uncertain parameters observed in that stage or earlier, but not on values to be revealed in future stages. In a scenario formulation, all decision variables are scenario-dependent, but explicit nonanticipativity constraints are introduced to force agreement in a given stage for all decision variables corresponding to scenarios that agree up to that stage. When the nonanticipativity constraints are relaxed, the problem decomposes into separate deterministic scenario subproblems that can be solved efficiently using all the solution technology developed for deterministic instances in that application. For example, software for solving unit commitment combines mixed integer programming solvers with specialized constraint management and acceleration techniques such as warm starting.

Scenario decomposition algorithms for solving SMIPs typically produce approximate solutions because exact methods based on branch-and-bound (Carøe and Schultz 1999) converge too slowly to be practical or because guarantees of convergence to optimality that exist in the continuous case (Rockafellar and Wets 1991) fail to hold for nonconvex problems. Focusing without loss of generality on cost-minimization problems, lower bounds on the optimal objective function value are essential, either to employ in branch-and-bound algorithms or to assess the quality of a terminal solution. In the scenario decomposition method known as progressive hedging, Gade et al. (2016) derived a lower bounding approach using the information available in any iteration of the algorithm and demonstrated its practical use in two-stage stochastic server location as well as stochastic unit commitment. Cheung et al. (2015) employed these lower bounds, in stochastic unit commitment instances of the scale typically solved daily by US independent system operators, to demonstrate that parallel progressive hedging could obtain high-quality solutions in a practical length of time. For two-stage SMIPs, Guo et al. (2015) exploited the correspondence between this progressive hedging lower bound and one based on Lagrangian relaxation of the nonanticipativity constraint to speed up convergence of

the exact branch-and-bound algorithm of Carøe and Schultz (1999). Guo and Ryan (2017) extended the progressive hedging lower bound to certain time-consistent formulations of risk-minimizing multi-stage stochastic programs.

3.5 Comprehensive Assessment

Following the sequence of activities discussed in the previous three sections, we have

1. formulated a stochastic process model for uncertain parameters in our optimization model, informed by observational data and allowing parameter estimates to be updated as additional data are collected;
2. carefully discretized the models to produce a modest number of probabilistic scenarios, considering tradeoffs between the amount of detail included in operational considerations and the granularity of the stochastic discretization; and
3. developed a method to assess the quality of an approximate solution to the resulting stochastic mixed integer program.

Steps 2 and 3 have emphasized the role of scenario subproblems. Scenario reduction methods developed for use in Step 2 employed them to characterize and cluster scenarios in terms of the optimal decisions for the associated deterministic subproblems. The lower bound in Step 3 was developed for solution procedures based on scenario decomposition. This section describes approaches to assess the quality of scenario sets and the solutions obtained by optimizing against them. As in the previous work, we employ scenario decomposition and emphasize the influence of different scenarios on the decisions to be implemented at once. In settings where instances of the same problem are solved repeatedly with continually updated parameter values, we argue that *re-enactment* is an appropriate data-driven approach for assessment and develop computationally efficient shortcuts for it. Here we use the term *scenario generation method* (SGM) to denote "any combination of stochastic process modeling, approximation, sampling and reduction techniques that results in a set of probabilistic scenarios based on the information available at the time [when] the [stochastic program] is to be solved" (Sarı Ay and Ryan 2018).

3.5.1 Direct Assessment of Scenario Generation Methods

Before describing methods for assessing scenario generation methods, let us consider some related concepts that have been rigorously defined and tested in the closely related, but not identical, context of probabilistic forecasting. As defined by Gneiting and Katzfuss (2014), "a probabilistic forecast takes the form of a predictive probability distribution over future quantities or events of interest." A

probabilistic forecast is called *calibrated*, or equivalently, *reliable* if the probabilities associated with predicted values correspond closely to their observed frequencies. The goal for a probabilistic forecaster is to produce predictive distributions that are as concentrated, i.e., *sharp* as possible, subject to reliability. The combination of reliability and sharpness is called *skill* (Pinson and Girard 2012). Precise definitions and metrics for these and other desirable characteristics of probabilistic forecasts of scalar quantities have been developed. Various "scoring functions," which measure the distance between a probabilistic forecast and the observed value, are used to compare the predictive performance of competing forecasting methods. Although the observed value could be viewed as a random variable with a degenerate distribution, the probability metrics used for scenario reduction are not mentioned in the probabilistic forecast assessment literature. Moreover, as Gneiting and Katzfuss (2014) note, corresponding metrics and scoring functions for assessing probabilistic forecasts of multidimensional quantities (e.g., scenarios for stochastic programs) are lacking. Many of those that exist were developed in the context of weather forecasting where, typically, equally likely sample paths, called ensemble forecasts, are generated by running multiple replications of numerical weather prediction—simulation—models under different conditions or assumptions. Pinson and Girard (2012) applied some statistical metrics for reliability and skill to evaluate equally likely scenario time series for wind energy production over the short term.

It is important here to note a distinction between the so-called "probability metrics" used in scenario reduction in the stochastic programming literature and the "statistical metrics" used in probabilistic forecast verification. For stability of the optimal solution to a stochastic program, the discretized or reduced scenario set should minimize the distance to the "true" distribution in terms of the Wasserstein distance. Given two cumulative distribution functions (CDFs), F and G for a real-valued random variable, the simplest variant of the Wasserstein distance is (Pflug 2001):

$$d_W(F, G) = \int_{-\infty}^{\infty} |F(u) - G(u)| du, \tag{3.1}$$

that is, the total absolute deviation between the CDFs. This distance measure is often called the mass transportation or earth mover's distance because, for discrete distributions, it can be computed by solving a linear transportation problem to move the probability mass from one distribution to the other with minimal work (defined as mass times distance). On the other hand, in the nonparametric goodness-of-fit testing literature, the distance between empirical distributions is often measured using the energy distance (Székely and Rizzo 2013):

$$d_E(F, G) = \int_{-\infty}^{\infty} (F(u) - G(u))^2 du = 2E|X - Y| - E|X - X'| - E|Y - Y'|, \tag{3.2}$$

where X and X' are independent random variables distributed as F and Y and Y' are independent random variables distributed as G. The name comes from a relation to

Newtonian potential energy within a gravitational space. The energy score used to evaluate probabilistic forecasts is based on the energy distance between the forecast CDF and the observation (Gneiting and Raftery 2007). When probabilistic forecasts and the corresponding observations are available for a collection of historical instances, the skill of the forecasting method can be evaluated in terms of the average energy score over the instances. Both the Wasserstein distance and the energy distance can be computed easily for joint distributions of several discrete variables, such as time series, by solving the corresponding mass transportation problem or evaluating the corresponding expectations as probability-weighted sums.

Another distance-based approach for assessing the reliability of ensemble forecasts of multidimensional quantities, which can be seen as multiple equally likely scenarios, is based on minimum spanning trees (Wilks 2004). Given a collection of historical instances, the idea is to quantitatively assess the degree to which the observation is indistinguishable from an ensemble member. For each instance $d = 1, \ldots, D$, a complete graph is constructed with nodes for each ensemble member, $s = 1, \ldots, S$, as well as the observation where edge lengths are computed according to a suitable distance measure, usually Euclidean distance. Next, a minimum spanning tree (MST) is constructed to connect all the ensemble members and its total edge length is recorded, say as ℓ_0^d. Then, for each ensemble member $s = 1, \ldots, S$, the observation is substituted for member s and the length of the resulting MST over those S nodes, not including member s, is recorded as ℓ_s^d. The $S + 1$ MST lengths for instance d are sorted in increasing order and the rank of ℓ_0 is recorded as r_d. Finally, a histogram with bins for the possible values $1, \ldots, S + 1$ of the ranks $\{r_d, d = 1, \ldots, D\}$ is constructed and evaluated for uniformity. A flat histogram indicates that the observation is equally likely to fall in the middle of the ensemble or its outer reaches. Overpopulation of the lower-valued bins occurs if the ensemble is either underdispersed or biased because the observation tends to be more distant from the ensemble members than they are from each other. A disproportionate number of higher rank values indicates that the ensemble is overdispersed so that the observation falls too often in the middle. Uniformity of the rank distribution can be quantified using a goodness-of-fit statistic but the graphical histogram is appealing because its shape helps diagnose the nature of the errors in ensemble forecasts (or sets of equally likely scenarios).

When a scenario generation method employs approximation rather than generating sample paths of the stochastic process model, or when scenario reduction methods are used, the resulting scenarios generally are not equally likely. To assess the reliability of unequally likely scenarios, Sarı et al. (2016) developed a rank histogram based on the Wasserstein distance. The mass transportation distance (MTD) rank histogram (Sarı and Ryan 2016) is constructed similarly to the MST rank histogram with the following three differences. First, ℓ_0^d is computed as the minimum cost of transporting the probability mass from the scenarios to the observation. Second, when the observation is substituted for scenario s, it is assigned the probability of that scenario and ℓ_s^d is computed as the minimum cost of transporting all the probability mass, including that mass having been re-assigned to the observation, to scenario s. Finally, MTDs are sorted in decreasing order to

find r^d as the rank of ℓ_0^d. In simulation studies, we demonstrated that the MTD rank histogram has a similar shape to the MST histogram under the same conditions of bias, overdispersion or underdispersion. The MTD values can be computed directly (even more efficiently than greedy-algorithm-based MST lengths) as the sum of probability-weighted distances. We applied the MTD rank histogram, as well as energy scores and event-based scores, to assess two different methods for generating wind power scenario time series on the day ahead and found that it could distinguish among scenario sets based on their autocorrelation levels as well as their bias and dispersion.

3.5.2 Assessing Solutions by Re-enactment

While reliability of scenario sets may be seen as a necessary condition for obtaining good solutions to stochastic programs, it may not be sufficient. In fact, there seem to be few studies that have "closed the loop" and examined how well the solution to a stochastic programming performs in the target context. The stochastic process modeling step can be assessed by comparing sample paths generated by the model to observed realizations, but studies of this type are rarely reported. Scenario reduction procedures, operating entirely in the realm of probability models, aim to approximate a continuous or highly granular discrete model with a coarse-grained discrete one. We return to the idea of re-enactment as a data-driven approach for assessing the quality of solutions obtained by the whole process of formulating a stochastic program, generating scenarios and obtaining approximate solutions.

The term re-enactment has been used recently, to describe a procedure to assess prediction intervals for wind energy generation, as "a walk forward through date-times in the past, computing prediction intervals using only data available prior to that date-time. In doing so, we compute prediction intervals using only relevant historical information, and are able to assess prediction interval quality using actual observations not used in the computation of those prediction intervals" (Nitsche et al. 2017). Staid et al. (2017) used a similar procedure to evaluate scenarios for wind power time series in terms of energy scores, MST rank histograms, and other metrics. In the context of stochastic unit commitment, Sarı and Ryan (2017) extended this idea to re-enact the process of not only generating scenarios but also solving the extensive forms of the stochastic programs. For each historical day d, we generated scenarios by competing methods, including some variants, using the data available through day $d-1$, then solved the stochastic program to obtain an optimal commitment schedule, and finally simulated dispatching the committed units to meet the observed net load on day d. We found that the variant of the scenario generation method that would be selected according to energy score, MTD rank histogram and some event-based scores produced the lowest average cost over the set of historical days.

Encouraged by these empirical results but cognizant of the computational burden of repeatedly solving stochastic programs to conduct this type of re-enactment, Sarı Ay and Ryan (2018) proposed solution assessment methods for two-stage stochastic programs (SPs) based on MTD rank histograms of the costs of solutions to scenario subproblems. As described in that paper, "for each [historical] instance, a single-scenario version of the SP is solved to find a candidate first-stage solution. Then, for each scenario as well as the observation, the second-stage solution is optimized assuming the candidate solution has been implemented, and the total cost for the scenario is computed. Reliability assessment is then applied to these costs. Variants of this approach differ according to whether the expected value (EV) scenario, perfect information (PI, i.e., the observation), or a randomly selected (RS) scenario is used to find the candidate solution." The use of an RS scenario is consistent with the notion that members of a reliable scenario set are statistically indistinguishable from the corresponding observation. We simulated this process using synthetic data for stochastic server location as well as stochastic unit commitment instances and then applied it to a case study of stochastic unit commitment with uncertain wind energy production. We concluded, "Simulation studies demonstrate that reliability of SGMs can be assessed accurately by the EV-based method. The stochastic unit commitment case study indicated that the PI- and RS-based methods can be used to distinguish between higher and lower quality SGMs, as have been identified by re-enactment" (Sarı Ay and Ryan 2018).

3.6 Conclusions

My current interest in re-enactment as a data-driven strategy for evaluating the entire modeling and solution process depicted in Fig. 3.1 arose while conducting a project on stochastic unit commitment for the Advanced Research Projects Agency-Energy (ARPA-E) of the US Department of Energy. Because of the funding source, the project emphasized engagement with end users to enable transfer of the technology developed. Our team, which included personnel from two universities, a software developer, and a national laboratory in partnership with an independent system operator, readily identified two major barriers to adoption of stochastic optimization by electricity system operators. One was mistrust in the scenario generation process and the other was doubt that high quality solutions could be found within realistic time limits. Some of the research described in Sect. 3.3.1 was aimed at overcoming the former barrier while the work outlined in Sect. 3.4 addressed the latter. The pair of papers by Feng et al. (2015) and Cheung et al. (2015) summarize this project's major accomplishments. However, the real test of our project came when we were asked to demonstrate the cost savings that the system operator might enjoy by replacing their current deterministic optimization with our proposed stochastic programming procedure. To estimate them, our team conducted a detailed and careful re-enactment of daily unit commitment over a year's time. For the stochastic programming model, this process included stochastic process modeling and scenario

generation using the data available up to the target day, followed by dispatch of the committed units to satisfy the observed net load on the target day. The results demonstrating savings of a few percent have been presented at conferences but, unfortunately, not documented in a published paper.

While writing this chapter I was surprised to recall that one of the first papers I independently conceived and wrote (Ryan 1998) involved a similar process of re-enactment with actual data to test different ways of modeling uncertainty in a capacity expansion problem. About a decade later, after spending some time on queuing models of manufacturing systems, I began working on stochastic mathematical programming and, again, used re-enactment to explore the impact of uncertainty on an optimization model intended to simulate actual decision making (Wang and Ryan 2010).

For this form of validation of the stochastic modeling and optimization process to be widely accepted and used, it must developed rigorously. For this development, we need an underlying probability model for the observed data that first inform the stochastic modeling process and later are used to evaluate solutions. The detailed re-enactment procedure is predicated on the idea that a higher quality scenario generation method should result in lower re-enacted costs. Because the cumulative cost over the re-enactment period is a random variable that depends on the observed data collected, comparisons of the costs incurred by different modeling and solution approaches can only be claimed in probabilistic terms. Sarı Ay and Ryan (2018) presented evidence that our faster approach for scenario and solution assessment is itself reliable, but formal proofs of this claim await completion.

In the long run I envision open-source software tools that could streamline the conduct of re-enactment studies. Our R package to compute the MTD rank histogram (Sarı and Ryan 2016) is a tiny step in this direction. The PySP package in Pyomo (Hart et al. 2017) has helped to structure the way I think about a stochastic program, as a deterministic model accompanied by a scenario tree. This structure facilitates the repeated re-formulation of problem instances that differ only according to the scenarios included, which is also necessary for re-enactment. The easy parallelization and inclusion in Pyomo of an extension that computes the progressive hedging lower bound both facilitate scenario decomposition as a fast and effective strategy for repeatedly solving re-enacted optimization models. Transparent validation methods, made easier by software tools, could expand the use of stochastic optimization and result in better decisions in a world of uncertainty.

References

Bean JC, Higle J, Smith RL (1992) Capacity expansion under stochastic demands. Oper Res 40:S210–S216

Carøe CC, Schultz R (1999) Dual decomposition in stochastic integer programming. Oper Res Lett 24:37–45

Casimir RJ (1990) The newsboy and the flower-girl. OMEGA Int J Manag Sci 18(4):395–398

Cheung K, Gade D, Ryan S, Silva-Monroy C, Watson JP, Wets R, Woodruff D (2015) Toward scalable stochastic unit commitment - part 2: Assessing solver performance. Energy Syst 6(3):417–438. https://doi.org/10.1007/s12667-015-0148-6

Dupačová J, Gröwe-Kuska N, Römisch W (2003) Scenario reduction in stochastic programming. Math Program 95(3):493–511. https://doi.org/10.1007/s10107-002-0331-0

Feng Y, Ryan SM (2013) Scenario construction and reduction applied to stochastic power generation expansion planning. Comput Oper Res 40(1):9–23

Feng Y, Ryan SM (2016) Day-ahead hourly electricity load modeling by functional regression. Appl Energy 170:455–465. https://doi.org/10.1016/j.apenergy.2016.02.118

Feng Y, Rios I, Ryan SM, Spurkel K, Watson JP, Wets RJB, Woodruff DL (2015) Toward scalable stochastic unit commitment - part 1: Load scenario generation. Energy Syst. https://doi.org/10.1007/s12667-015-0146-8

Freidenfelds J (1981) Capacity expansion: Analysis of simple models with applications. North-Holland, New York

Gade D, Hackebeil G, Ryan SM, Watson JP, Wets RJB, Woodruff DL (2016) Obtaining lower bounds from the progressive hedging algorithm for stochastic mixed-integer programs. Math Program Ser B 157(1):47–67. https://doi.org/10.1007/s10107-016-1000-z

Gneiting T, Katzfuss M (2014) Probabilistic forecasting. Annu Rev Stat Appl 1:125–151. https://doi.org/10.1146/annurev-statistics-062713-085831

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102(477):359–378. https://doi.org/10.1198/016214506000001437

Guo G, Hackebeil G, Ryan SM, Watson JP, Woodruff DL (2015) Integration of progressive hedging and dual decomposition in stochastic integer programs. Oper Res Lett 43(3):311–316. https://doi.org/10.1016/j.orl.2015.03.008

Guo GC, Ryan SM (2017) Progressive hedging lower bounds for time consistent risk-averse multistage stochastic mixed-integer programs. URL https://works.bepress.com/sarah_m_ryan/93/

Hart WE, Laird CD, Watson JP, Woodruff DL, Hackebeil GA, Nicholson BL, Siirola JD (2017) Pyomo – optimization modeling in Python, 2nd edn. Springer

Heitsch H, Römisch W (2003) Scenario reduction algorithms in stochastic programming. Comput Optim Appl 24:187–206

Heitsch H, Römisch W (2007) A note on scenario reduction for two-stage stochastic programs. Oper Res Lett 35(6):731–738. https://doi.org/10.1016/j.orl.2006.12.008

Jin S, Ryan S, Watson JP, Woodruff D (2011) Modeling and solving a large-scale generation expansion planning problem under uncertainty. Energy Syst 2:209–242. https://doi.org/10.1007/s12667-011-0042-9. URL http://dx.doi.org/10.1007/s12667-011-0042-9

Jin S, Botterud A, Ryan SM (2014) Temporal vs. stochastic granularity in thermal generation capacity planning with wind power. IEEE Trans Power Syst 29(5):2033–2041. https://doi.org/10.1109/TPWRS.2014.2299760

Manne AS (1961) Capacity expansion and probabilistic growth. Econometrica 29(4):632–649

Marathe R, Ryan SM (2005) On the validity of the geometric Brownian motion assumption. Eng Econ 50:159–192. https://doi.org/10.1080/00137910590949904

Marathe RR, Ryan SM (2009) Capacity expansion under a service level constraint for uncertain demand with lead times. Nav Res Logist 56(3):250–263

McAllister C, Ryan SM (2000) Relative risk characteristics of rolling horizon hedging heuristics for capacity expansion. Eng Econ 45(2):115–128

Muñoz FD, Hobbs BF, Ho JL, Kasina S (2014) An engineering-economic approach to transmission planning under market and regulatory uncertainties: WECC case study. IEEE Trans Power Syst 29(1):307–317

Nitsche S, Silva-Monroy C, Staid A, Watson JP, Winner S, Woodruff D (2017) Improving wind power prediction intervals using vendor-supplied probabilistic forecast information. In: IEEE Power & Energy Society General Meeting

Pflug GC (2001) Scenario tree generation for multiperiod financial optimization by optimal discretization. Math Program Ser B 89:251–271. https://doi.org/10.1007/s101070000202

Pflug GC, Pichler A (2014) Multistage stochastic optimization. Springer
Pinson P, Girard R (2012) Evaluating the quality of scenarios of short-term wind power generation. Appl Energy 96:12–20, https://doi.org/10.1016/j.apenergy.2011.11.004
Quelhas A, McCalley JD (2007) A multiperiod generalized network flow model of the U.S. integrated energy system: Part II simulation results. IEEE Trans Power Syst 22(2):837–844
Quelhas A, Gil E, McCalley JD, Ryan SM (2007) A multiperiod generalized network flow model of the U.S. integrated energy system: Part I – model description. IEEE Trans Power Syst 22(2):829–836
Rockafellar RT, Wets RJB (1991) Scenarios and policy aggregation in optimization under uncertainty. Math Oper Res 16(1):119–147
Ross S (1999) An introduction to mathematical finance. Cambridge University Press, Cambridge, UK
Royset JO, Wets RJB (2014) From data to assessments and decisions: Epi-spline technology. INFORMS Tutor Oper Res, 27–53. https://doi.org/10.1287/educ.2014.0126
Ryan SM (1988) Degeneracy in discrete infinite horizon optimization. Ph.D. dissertation, The University of Michigan
Ryan SM (1998) Forecast frequency in rolling horizon hedging heuristics for capacity expansion. Eur J Oper Res 109(3):550–558
Ryan SM (2003) Capacity expansion with lead times and correlated random demand. Nav Res Logist 50(2):167–183. https://doi.org/10.1002/nav.10055
Ryan SM (2004) Capacity expansion for random exponential demand growth with lead times. Manag Sci 50(6):740–748. https://doi.org/10.1287/mnsc.1030.0187
Ryan SM, Bean JC (1989) Degeneracy in infinite horizon optimization. Math Program 43:305–316
Ryan SM, Bean JC, Smith RL (1992) A tie-breaking rule for discrete infinite horizon optimization. Oper Res 40(Supplement 1):S117–S126
Ryan SM, McCalley JD, Woodruff D (2011) Long term resource planning for electric power systems under uncertainty. In: Eto JH, Thomas RJ (eds) Computational needs for the next generation electric grid, U.S. Department of Energy, pp 6–1–41. URL http://energy.gov/sites/prod/files/FINAL_CompNeeds_Proceedings2011.pdf
Sarı D, Ryan SM (2016). MTDrh: Mass transportation distance rank histogram. URL https://cran.r-project.org/package=MTDrh
Sarı D, Ryan SM (2017) Statistical reliability of wind power scenarios and stochastic unit commitment cost. Energy Syst https://doi.org/10.1007/s12667-017-0255-7
Sarı D, Lee Y, Ryan S, Woodruff D (2016) Statistical metrics for assessing the quality of wind power scenarios for stochastic unit commitment. Wind Energy 19:873–893. https://doi.org/10.1002/we.1872
Sarı Ay D, Ryan SM (2018) Observational data-based quality assessment of scenario generation for stochastic programs. Comput Manag Sci https://works.bepress.com/sarah_m_ryan/94/
Staid A, Watson JP, Wets RJB, Woodruff DL (2017) Generating short-term probabilistic wind power scenarios via nonparametric forecast error density estimators. Wind Energy 20(12):1911–1925. https://doi.org/10.1002/we.2129
Székely GJ, Rizzo ML (2013) Energy statistics: A class of statistics based on distances. J Statist Plann Inference 143(8):1249–1272. https://doi.org/10.1016/j.jspi.2013.03.018
Wang Y (2010) Scenario reduction heuristics for a rolling stochastic programming simulation of bulk energy flows with uncertain fuel costs. Ph.D. dissertation, Iowa State University, URL http://search.proquest.com/docview/848503163
Wang Y, Ryan SM (2010) Effects of uncertain fuel costs on optimal energy flows in the U.S. Energy Syst 1:209–243
Wilks DS (2004) The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts. Mon Weather Rev 132:1329–1340

Sarah McAllister Ryan I am the youngest of three daughters born to an accountant who retrained to become a systems analyst and a chemistry major who never worked outside the home after she married. In the 1970s my parents noted the rising divorce rate in the US, along with the high grades my sisters and I achieved in math and science, and concluded that engineering could be a profession that would support us in case our future spouses did not. I followed their urging and my sisters' examples and entered an engineering program. I chose The University of Virginia for its strength in a diversity of fields in case I wanted to switch to a different field. I selected systems engineering in part because of its requirements to take courses in psychology and economics. Then I discovered the field of operations research, where I could combine mathematical analysis with real world considerations. I was fortunate to have both Chip White, who was then department chair, and Carl Harris as mentors who encouraged me to pursue graduate study. I completed my graduate degrees in industrial and operations engineering at The University of Michigan, and have enjoyed faculty positions in industrial engineering at the University of Pittsburgh, the University of Nebraska-Lincoln, and Iowa State University. The research described here has been supported by the National Science Foundation, including a CAREER award, the US Department of Energy, and industry consortia among other sources. I am now a Fellow of the Institute of Industrial & Systems Engineers, the Joseph Walkup Professor of Industrial and Manufacturing Systems Engineering at Iowa State University, and the 2017 recipient of its College of Engineering's D. R. Boylan Eminent Faculty Award for Research. I am deeply grateful to my husband, Steve Ryan, for embracing the role of primary caregiver to our children for many years.

Some say we choose the professional specialization that helps us most personally. I do view life as a series of decisions made under uncertainty. My research and experience have taught me to distinguish what I can control from what I cannot, content myself with having made the best choice I knew with the information available at the time, and eschew regret as a criterion for evaluating past decisions.

Chapter 4
Towards a Stable Graph Representation Learning Using Connection Subgraphs

Saba A. Al-Sayouri and Sarah S. Lam

Contents

4.1 Introduction 71
4.1.1 Representation Learning 71
4.1.2 Need and Challenges for Representation Learning 72
4.1.3 Evolution of Network Representation Learning 73
4.1.4 Taxonomy of Network Representation Learning Algorithms 73
4.2 Related Work 75
4.3 Proposed Method: GRCS 77
4.3.1 *GRCS*—Step 1: Neighborhood Definition 77
4.3.2 *GRCS*—Step 2: Node Representation Vector Update 83
4.4 Experiments 84
4.4.1 Q1: Multi-Label Classification 85
4.4.2 Q2: Representation Learning Stability 87
4.5 Conclusions 88
References 90

4.1 Introduction

4.1.1 Representation Learning

Representation learning has recently become a research trend in graph mining domain, which can be attributed to multiple reasons. First, the notion of representation learning helps to design a variety of deep architectures. Second, representation learning algorithms share statistical information across various tasks, that is, information learned from either unsupervised or semi-supervised tasks can effectively be exploited to perform supervised tasks. Third, they efficiently handle scarce data, where very few labeled examples exist for training. Thus, the model learns from labeled and unlabeled data, which in turn avoids overfitting. Fourth, they help to

S. A. Al-Sayouri · S. S. Lam (✉)
Binghamton University, Binghamton, NY, USA
e-mail: ssyouri1@binghamton.edu; sarahlam@binghamton.edu

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_4

resolve the issue of initial poor representations. One typical example of this is the use of word embeddings, where each word is represented using one-hot vector. Such representations do not convey useful information, as any two different vectors have the same distance from each other. It is important to note that representation learning incorporates two different aspects: (1) The choice of initial parameters of deep architecture can substantially enhance model optimization; and (2) Understanding the input distribution can help to learn about the mapping function from the input space to the output space, which in turn allows generalization over various tasks.

4.1.2 Need and Challenges for Representation Learning

Real-world information networks (a.k.a. graphs), such as social networks, biological networks, co-authorship networks, and language networks are ubiquitous. Further, the large size of networks—millions of nodes and billions of edges—and the massive amount of information they convey (Easley et al. 2010) have led to a serious need for efficient and effective network mining techniques. Modeling the relationships and interactions among network entities enables researchers to better understand a wide variety of networks in a systematic manner. Generally speaking, any model we employ for graph analysis operates on the graph corresponding adjacency matrix or on the learned vector space. Recently, because methods that exploit learned representations can generalize over a broad variety of tasks, they have become widely popular. Basically, the learned representations are ultimately used as an input to the model and parameters are learned using training data. However, it is worth mentioning that learning graph representations has never been an easy task. This can be rooted in multiple challenges (Goyal and Ferrara 2017):

1. **Choice of property**. A good vector representation of a node well-preserves local and global network structure. Therefore, given a plethora of properties defined for graphs, the large number of applications, and the wide variety of tasks, the challenge comes when we need to choose proper properties that embeddings must preserve.
2. **Robustness**. Because of the randomness attached to the ways of generating node representations, it has become more challenging to generate representations that are or near robust. As a result, such embeddings suit single-graph related tasks, while they are inappropriate for multi-graph tasks.
3. **Embedding dimensionality**. Choosing the optimal number of dimensions for the learned representation is not trivial. Intuitively, having a higher number of dimensions may, on the one hand, improve the reconstruction precision, while it would increase time and space complexity, on the other. In addition, what makes the process even more difficult is that the choice of the number of dimensions may be task-dependent. That is, if the number of dimensions is sufficient for one downstream process performance, it can be unsatisfactory for another.

4. **Scalability**. Because the majority of real-world networks involve millions of nodes and billions of edges, representation learning algorithms should be scalable to operate on such large-scale graphs. It is important to point out that defining a scalable model is more complicated when the aim is to preserve network global structure.

4.1.3 Evolution of Network Representation Learning

Under the umbrella of dimensionality reduction, since the early 2000s, researchers have developed different graph embedding methods (Goyal and Ferrara 2017). The essence of graph embedding is to map the nodes from a higher dimensional space D to a lower dimensional space d, where $d \ll D$. The embedding functions aim to embed similar nodes close to one another in the lower dimensional space based on the constructed similarity graph using generated node neighborhoods. As similar nodes can be connected through observed or unobserved connections, distinct connectivity patterns that emerge in real-world networks—i.e., homophily and structural equivalence (Grover and Leskovec 2016), should be preserved in similarity graph. Under the homophily assumption, nodes that are highly interconnected and belong to the same community should be embedded closely together. On the other hand, under the structural equivalence assumption, nodes that play the same structural roles—i.e., serve as hops to their corresponding communities should be mapped close to one another in the lower-dimensional space. The large size of today's real-world graphs make scalability a major issue for the earliest generated embedding algorithms, such as Laplacian eigenmaps (Belkin and Niyogi 2002) and locally linear embedding (Roweis and Saul 2000). Further, inspired by the recent advancements in the domain of natural language processing (Le and Mikolov 2014; Mikolov et al. 2013a,b), where two word2vec models (Mikolov et al. 2013a) have been proposed, namely continuous bag of words (CBOW) and Skipgram, and the analogy in context, various algorithms have been developed to learn graph representations, such as node2vec (Grover and Leskovec 2016), DeepWalk (Perozzi et al. 2014), LINE (Tang et al. 2015), Walklets (Perozzi et al. 2016), and many others.

4.1.4 Taxonomy of Network Representation Learning Algorithms

Representation learning algorithms can be broadly divided into four different categories (Goyal and Ferrara 2017):

1. **Factorization-based Algorithms**. In such algorithms, a matrix is initially utilized to represent the connection among nodes—i.e., adjacency matrix, Laplacian matrix, and others. A factorization technique is employed afterwards that may

vary depending upon the properties of the matrix to obtain the embeddings, such as eigendecomposition and gradient descent. Recently, locally linear embedding (Roweis and Saul 2000), Laplacian eigenmaps (Belkin and Niyogi 2002), graph factorization (Ahmed et al. 2013), among others have been proposed to obtain graph embeddings using factorization.

2. **Random Walk-based Algorithms**. Random walks are known to be very beneficial in cases like handling large-scale graphs or when observing the complete graph is impossible. Therefore, random walks have been widely employed to approximate node centrality (Newman 2005) and node similarity (Fouss et al. 2007) in graphs. DeepWalk (Perozzi et al. 2014) and node2vec (Grover and Leskovec 2016) are among the most recent random walk-based proposed algorithms.
3. **Deep Learning-based Algorithms**. Due to the positive impact deep learning research has played, especially in modeling non-linear relationships, deep learning-based algorithms have been largely used to mining graphs (Wang et al. 2016; Cao et al. 2016). In particular, Bengio et al. (2013) has harnessed deep autoencoders for dimensionality reduction. Recently, SDNE (Wang et al. 2016) and DNGR (Cao et al. 2016) have utilized deep autoencoder architectures to generate graph embeddings.
4. **Other Algorithms**. LINE (Tang et al. 2015) is a representation learning algorithm, which does not fall under any of the previously listed categories. It learns the representations in two separate phases, where each phase learns $d/2$ dimensions, where ultimately the learned representations are concatenated to have $d-$dimensional representation. The first phase preserves local structure using *first order proximity* measure, while the second phase preserves global structure using *second-order proximity* measure, by sampling nodes that are merely 2-hops away from each node.

Inspired by the recent advances in natural language processing (NLP), and the analogy in context, various algorithms have been developed to learn graph representations (Perozzi et al. 2014; Tang et al. 2015; Grover and Leskovec 2016). However, some recently proposed algorithms fail to clearly define and optimize an objective that is tailored for graph nature (Perozzi et al. 2014). Further, they employ completely random walks (Perozzi et al. 2014, 2016) or biased (Grover and Leskovec 2016) random walks to obtain node neighborhoods, which remain unsatisfactory in performing downstream processes. In addition, state-of-the-art algorithms share a major stability issue that makes them less useful and applicable, especially for multiple graph problems. In other words, it seems that while baseline representation learning algorithms strive to preserve similarities among nodes to generate and learn node representations, they fail to maintain similarities across runs of any of the algorithms, even when using the same data set. Therefore, they are not beneficial for canonical multi-graph tasks, such as graph similarity (Koutra et al. 2013).

It is worth mentioning that the quality of the learned representations is heavily influenced by the preserved local and global structure. Consequently, we need

to properly and neatly identify node neighborhoods. In this study, we develop a robust graph embedding algorithm that follows the fourth category of the embedding algorithms we list above, and can preserve connectivity patterns unique to undirected and (un)weighted graphs using connection subgraphs (Faloutsos et al. 2004). Connection subgraphs are proved to be useful in different real-world applications. In social networks, connection subgraphs can help us identify the people that have been infected with a specific disease. In addition, connection subgraphs can indicate the presence of any suspicious relationships that link an individual with a group of people in a terrorism network. For biological networks, connection subgraphs can assist to identify the connection exists between two different proteins or genes, in a protein–protein interaction network or a regulatory network. Further, in a world wide web network, using the hyper-link graph, connection subgraphs help to summarize the connection between two web sites.

Connection subgraphs avail the analogy with electrical circuits, where a node is assumed to serve as a voltage source and an edge is assumed to be a resistor, where its conductance is considered as the weight of the edge. When forming the connection subgraph, we concurrently capture the *node local and global connections*, and account for the node degree imbalances by downweighing the importance of paths through high-degree nodes (hops) and by accounting for both low- and high-weight edges. Further, using connection subgraphs allows to account for *meta-data* that is not well-exploited by existing embedding algorithms. It is important to note that our goal in forming connection subgraphs is to maximize the flow between pairs of non-adjacent nodes along with avoiding long paths, where generally information is lost, therefore, the formation process is distance and flow-driven. Our contributions are as follows:

1. **Novel Flow-based Formulation.** We propose a graph embedding approach that robustly preserves network local and global structure using GRCS algorithm to learn graph representations using the notion of *network* flow to produce approximate but high-quality connection subgraphs between pairs of non-adjacent nodes in undirected and (un)weighted large-scale graphs.
2. **Stable Representations.** Contrary to all state-of-the-art methods, which involve randomness that is reflected on the embeddings and their quality, we propose a deterministic algorithm that produces consistent embeddings across independent runs. We experimentally demonstrate the benefits of stability.

4.2 Related Work

Representation Learning Recent work in network representation learning has been largely motivated by the new progress in NLP domain (Mikolov et al. 2013a,b; Le and Mikolov 2014), because of the existing analogy among the two fields, where a network is represented as a document. One of the NLP leading advancements

is attributed to the SkipGram model, due to its efficiency in scaling to large-scale networks. However, merely adopting the SkipGram model for graph learning representations is insufficient, because of the sophisticated connectivity patterns that emerge in networks but not in corpora. For that, we propose to preserve linear and non-linear proximities while generating neighborhoods and before even being learned by the SkipGram model. Recently proposed algorithms that adopted the SkipGram model, despite their better performance than other methods, are still incapable to satisfactorily and properly capture node neighborhoods in a network (Perozzi et al. 2014; Tang et al. 2015; Grover and Leskovec 2016). Specifically, DeepWalk (Perozzi et al. 2014), for instance, employs random walks to attain the node neighborhoods in a graph, which ultimately introduces noise to the search process. Furthermore, LINE (Tang et al. 2015) captures the network local and global structure using first- and second-order proximities, respectively, along with an edge-sampling algorithm to overcome the limitations of the optimization using stochastic gradient descent. However, the sampling process ignores the strength of weak ties, where crucial information generally flows (Easley et al. 2010). A more recent approach, node2vec (Grover and Leskovec 2016), preserves graph unique connectivity patterns, homophily and structural equivalence, by using biased random walks, which is a more flexible search strategy to identify node neighborhoods. However, relying on such biased walks to identify node neighborhoods introduces some randomness that compromises task performance.

Connection Subgraphs The work on connection subgraphs (Faloutsos et al. 2004), which captures proximity among any two non-adjacent nodes in arbitrary undirected and (un)weighted graphs, is the most relevant to ours. In a nutshell, Faloutsos et al. (2004) includes two prime phases: *candidate generation*, and *display generation*. In the *candidate generation* phase, a distance-driven extraction of a much smaller subgraph is performed to generate *candidate* subgraph. At a high level, *candidate* subgraph is formed by gradually and neatly 'expanding' the neighborhoods of any two non-adjacent nodes until they 'significantly' overlap. Therefore, *candidate* subgraph contains the most prominent paths connecting a pair of non-adjacent nodes in the original undirected and (un)weighted graph. The generated *candidate* subgraph serves as an input to the next phase, i.e., the *display generation*. The *display generation* phase removes any remaining spurious regions in the *candidate* subgraph. The removal process is current-oriented; it aims to add an end-to-end path at a time between the two selected non-adjacent nodes that maximizes the delivered current (network flow) over all paths of its length. Typically, for a large-scale graph, the *display* subgraph is expected to have 20–30 nodes. Connection subgraphs have also been employed for graph visualization (Faloutsos et al. 2004). Our work is the first to leverage connection subgraphs to define appropriate neighborhoods for representation learning.

4.3 Proposed Method: GRCS

In this section, we describe our proposed method, GRCS, a deterministic algorithm that is capable of preserving local and global—beyond two hops—connectivity patterns. It consists of two main steps: (1) Neighborhood definition via connection subgraphs, and (2) Node representation vector update. We discuss the two steps in Sects. 4.3.1 and 4.3.2, respectively. We note that GRCS is deterministic, and thus can be applied to multi-graph problems, unlike previous works (Perozzi et al. 2014, 2016; Grover and Leskovec 2016) that employ random processes, such as random walks.

Our method operates on an (un)weighted and undirected graph $G(\mathcal{V}, \mathcal{E})$, with $|\mathcal{V}| = n$ nodes and $|\mathcal{E}| = m$ edges. For a given node u, we define its 1-hop neighborhood as $\mathcal{N}(u)$ (i.e., set of nodes that are directly connected to u).

4.3.1 GRCS—*Step 1: Neighborhood Definition*

The heart of learning node representations is to obtain representative node neighborhoods, which preserve local and global connections simultaneously. Inspired by Faloutsos et al. (2004), we propose to define node neighborhoods by leveraging the analogy between graphs and electrical circuits, and adapting the connection subgraph algorithm (discussed in Sect. 4.2) to our setting. In Table 4.1, we give a qualitative comparison of GRCS and the connection subgraph algorithm (Faloutsos et al. 2004), highlighting our major contributions.

The notion of connection subgraphs is beneficial in our setting, since they allow us to: (1) Better control the search space; (2) Benefit from the actual flow, metadata, that is being neglected by state-of-the-art algorithms; (3) Exploit the strength of weak ties; (4) Avoid introducing randomness caused by random/biased walks; (5) Integrate two extreme search strategies, breadth-first search (BFS) and depth-first search (DFS) (Zhou and Hansen 2006); (6) Address the issue of high-degree

Table 4.1 Qualitative comparison of the connection subgraph algorithm (Faloutsos et al. 2004) vs. GRCS

	Connection subgraph	GRCS
Purpose	Node proximity (for only 2 nodes)	Neighborhood definition (for the whole graph)
Step 1	Candidate generation (distance-driven)	Neighborhood expansion (distance-driven)
Step 2	Display generation (delivered current-driven)	Neighborhood refinement (current-driven)
Efficiency	Inefficient (for the whole graph)	More efficient (for the whole graph)
Source	u_i	$\forall\, u \in \mathcal{V}$
Target	u_j	Universal sink node z

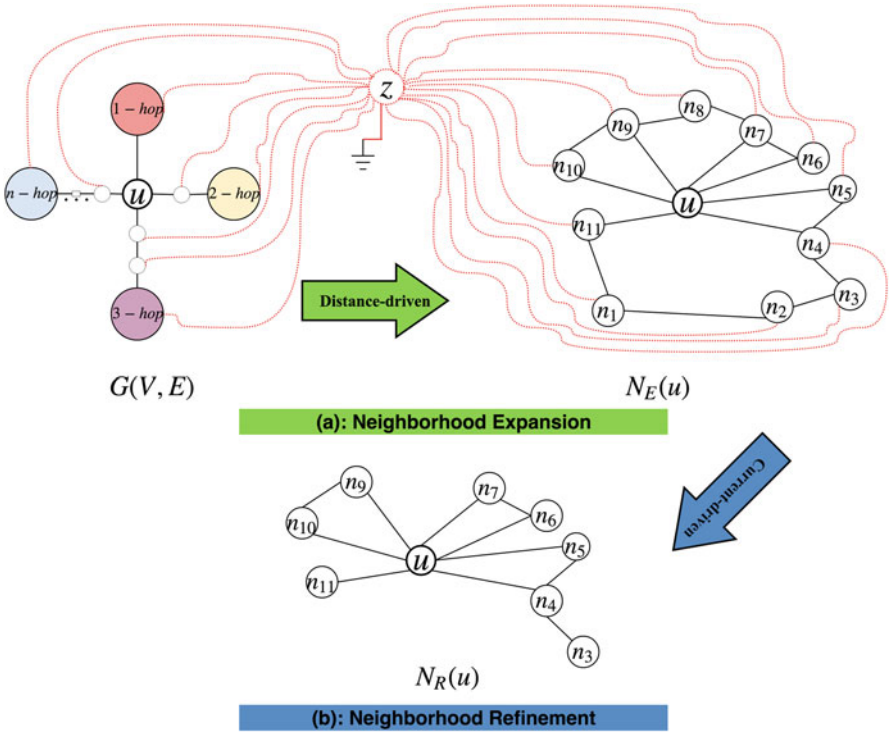


Fig. 4.1 A description of GRCS algorithm neighborhood definition step main phases: **(a)** Neighborhood expansion of node u through $n-hop$ neighbors to generate $N_E(u)$ on distance basis. Node z indicates the grounded universal sink node. **(b)** Neighborhood refinement of $N_E(u)$ to generate $N_R(u)$ on current basis

nodes; and (7) Better handle non-adjacent nodes that are ubiquitous in real-world large-scale graphs.

The neighborhood definition step consists of two phases: (A) *Neighborhood expansion*, and (B) *Neighborhood refinement*. We provide an overview of each phase next, and an illustration in Fig. 4.1. The overall computational complexity of GRCS is $O(\mathcal{V}^2)$.

- **Phase A: Neighborhood Expansion**—$N_E(u)$. Given a node u, we propose to gradually expand its neighborhood on a distance basis. Specifically, we employ the analogy with electrical circuits in order to capture the distances between u and the other nodes in the network, and then leverage these distances to guide its neighborhood expansion.

Graph Construction We first construct a modified network G' from G by introducing a universal sink node z (grounded, with voltage $V_z = 0$), and connect all the nodes (except from u) to that, as shown in Fig. 4.1a. The newly added edges in G' for every node $v \in \{\mathcal{V} \setminus u\}$ are weighted appropriately by the following weight

or conductance (based on the circuit analogy):

$$C(v, z) = \alpha \sum_{w \in \mathcal{N}(u) \setminus z} C(v, w), \tag{4.1}$$

where $C(v, w)$ is the weight or conductance of the edge connecting nodes v and w, $\mathcal{N}(u)$ is the set of 1-hop neighbors of u, and $\alpha > 0$ is a scalar (set to 1 for unweighted graphs).

In the modified network G', the distance, or proximity, between the given node u and every other node is defined as:

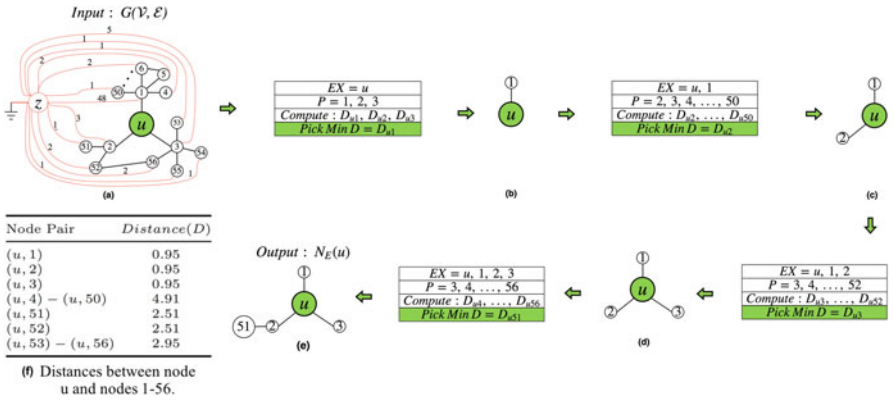
$$D(u, v) = \begin{cases} \log \frac{deg^2(u)}{C^2(u,v)}, & \text{for } v \in \mathcal{N}(u). \\ \log D(u, c) + D(c, v), & \text{for } v \notin \mathcal{N}(u), \text{ and } u, v \in \mathcal{N}(c). \end{cases} \tag{4.2}$$

where $deg(u)$ is the weighted degree of u (i.e., the sum of the weights of its incident edges), and the distance for non-neighboring nodes u and v is defined as the distance from each one to their nearest common neighbor $c \in \mathcal{V}$. This distance computation addresses the issue of high-degree nodes (which could make 'unrelated' nodes seem 'close') by significantly penalizing their effects in the numerator.

Distance-Based Expansion After constructing the circuit-based graph, we can leverage it to expand u's neighborhood. Let EX be the set of expanded nodes that will form the expansion graph $N_E(u)$ (initialized to $\{u\}$), and P be the set of *pending* nodes, initialized to u's neighbors, $\mathcal{N}(u)$. During the expansion process, we choose the closest node to u (except for z), as defined by the distance function in Eq. (4.2). Intuitively, the closer the expanded node v to the source node u, the less information flow we lose. Once a node v is added to the expansion subgraph, we add its immediate neighbors to P, and we repeat the process until we have $|EX| = e$ nodes, where e is a constant that represents the desired size of expanded subgraph. We show the *neighborhood expansion* pseudocode in Algorithm 1a. The procedure of computing the $N_E(u)$ takes $O(\mathcal{V})$ time.

Example 1 Figure 4.2 shows one example of generating $N_E(u)$ for an undirected, unweighted graph G, in which the original edges have conductance (weight) equal to 1, and the size of the expanded neighborhood is set to $e = 5$. The conductances for the new edges in G' (red-dotted lines), computed via Eq. (4.1), are shown in Fig. 4.2a. Based on the distances between u and every other node, which are defined by Eq. (4.2) and shown in Fig. 4.2f, the neighborhood of u is expanded on a distance basis.

- **Phase B: Neighborhood Refinement**—$N_R(u)$. As shown in Fig. 4.1b, the *neighborhood refinement* phase takes an *expanded* subgraph as an input and returns a *refined neighborhood* subgraph as an output, which is free of spurious graph regions. Unlike the previous phase that is based on distances, the *refined* subgraph is generated on a network flow (current) basis.



Node Pair	$Distance(D)$
$(u, 1)$	0.95
$(u, 2)$	0.95
$(u, 3)$	0.95
$(u, 4) - (u, 50)$	4.91
$(u, 51)$	2.51
$(u, 52)$	2.51
$(u, 53) - (u, 56)$	2.95

(f) Distances between node u and nodes 1-56.

Fig. 4.2 Neighborhood expansion example

In a nutshell, in this phase, we first link the nodes of the *expansion* subgraph $N_E(u)$ (except for node u) to the previously introduced grounded node z. Then, we create the *refined neighborhood* subgraph by adding end-to-end paths from node z to node u one at a time, in decreasing order of total current. The underlying intuition of the *refinement* phase is to maximize the current reaches to node z from the source node u. By maximizing the current, we maximize the information flow between the source node u and node z, which ultimately serves our goal of including proximate nodes to the source node u in its $N_R(u)$. The process stops when the maximum predetermined *refined* subgraph size, $|N_R(u)|$, is reached. Each time a path is added to the *refined* subgraph, only the nodes that are not already included in the subgraph are added. We use dynamic programming to implement our *refinement* process, which is like a depth first search (DFS) approach with a slight modification.

To that end, we need to calculate the current I flows between any pair of neighbors in the *expanded* subgraph. In our context, I indicates the meta-data or network flow that we aim to avail. We compute the current I flow from source node s to target node t using Ohm's law:

$$I(s,t) = C(s,t) \cdot [V(s) - V(t)] \tag{4.3}$$

where the $V(s) > V(t)$ are the voltages of s and t, satisfying the *downhill constraint* (otherwise, there would be current flows in the opposite direction). In order to guarantee this satisfaction, we need to sort subgraph's nodes in a descending order, based on their calculated voltage values, before we start current computations. The voltage of a node $s \in \mathcal{V}$ is defined as:

$$V(s) = \begin{cases} \frac{\sum_{v \in (s)} V(v) \cdot C(s,v)}{\sum_v C(s,v)}, & \forall \text{ nodes } s \neq u, z. \\ 1, & s = u. \\ 0, & s = z. \end{cases} \tag{4.4}$$

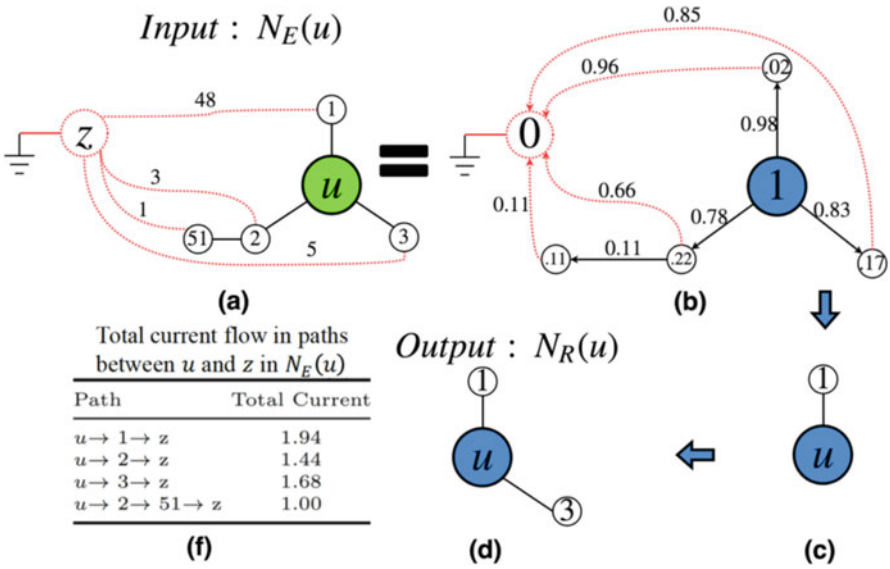


Fig. 4.3 Neighborhood refinement example

where $C(s, v)$ is the conductance or weight of the edge between nodes s and v, as defined in Eq. (4.1).

Example 2 Given the expanded neighborhood $N_E(u)$ in Example 2, the second phase of GRCS gradually refines it on a current basis, as shown in Fig. 4.3. We first compute the voltages by solving the linear system in Eq. (4.4), and include them in the nodes of (b). Then, the current flow of each edge connecting nodes in the expanded neighborhood $N_E(u)$ is computed using Eq. (4.3) such that the 'downhill constraint' is satisfied (current flowing from high to low voltage), as shown over the red-dotted edges in (b). Given the current values, we enumerate all possible paths between nodes u and z, and give their total current flows in (f). The paths are then added in descending order of total current values into $N_R(u)$ until the stopping criterion is satisfied. In (c), we show the first path we add, where the highest current flows (as mentioned in b). Assuming that the size of the refined neighborhood, $|N_R(u)| = r = 3$, the final neighborhood is given in (d).

*Remark 1 (*GRCS *Neighborhood vs. Context in Baseline Methods)* Unlike existing representation learning methods:

(1) We preserve the local and global structure of network by accounting for the immediate neighbors and neighbors at increasing distances of the source node u to identify its neighborhood; (2) We generate neighborhoods on distance and network flow bases; (3) We address the issue of high-node degree distribution; (4) We concurrently identify neighborhoods while maximizing proximity among non-

Algorithm 1: GRCS Algorithm—Step 1

a: Neighborhood Expansion

Input : Graph $G(\mathcal{V}, \mathcal{E})$
u: node to expand
e $(default : 1200)$: max size of $N_E(u)$
Output: $N_E(u)$: expanded neighborhood

Add grounded node z to $G(\mathcal{V}, \mathcal{E})$
Connect all nodes $u \in \mathcal{V}$ (except u and z) to z
Initialize $EX = \{u\}$
Initialize $P = \mathcal{N}(u) = \{v_1, v_2, ..., v_n\}$
while $|EX| < e$ **do**
  $minDist = \infty$
  for $p \in P$ **do**
    $newDist$ = D(u, p) = Eq. (4.2)
    if $minDist > newDist$ **then**
      $minDist = newDist$
    end
  end
  Add node($minDist$) to EX
  Remove node($minDist$) from P
  Add neighbors of node($minDist$) to P
end
Remove node z from $G(\mathcal{V}, \mathcal{E})$
return $N_R(u)$: subgraph of G induced on EX

b: Neighborhood Refinement

Input : $N_E(u)$
r $(default : 800)$: max size of $N_R(u)$
u : node to refine
Output: $N_R(u)$: refined neighborhood

Add node z to graph $N_E(u)$
Connect all nodes in $N_E(u)$ (excl. u, z) to z
Initialize voltages $V(u) = 1$ and $V(z) = 0$
Initialize N_R = { }
Calculate voltage & current for each $u \in N_E(u)$
while $|N_R(u)| < r$ **do**
  Add all the nodes along the path that
  maximizes the current $I(u, z)$ to N_R
end
return $N_R(u)$

adjacent nodes, which are abundant in real-world networks; and (5) We design our algorithm such that it yields consistent stable representations that suit single and multi-graph problems.

*Remark 2 (*GRCS *vs. Connection Subgraph Algorithm (Faloutsos et al. 2004))* It is important to note that the computations of 'current' (in GRCS) and 'delivered current' (in Faloutsos et al. (2004)) are different. The computation of current is not as informative as delivered current, but is more efficient. The use of delivered current was not a major struggle in Faloutsos et al. (2004), because that algorithm only processes one subgraph. However, we find that it is problematic for generating multiple neighborhoods due to: (1) The large size of the *expanded* subgraph, $|N_E(u)|$; (2) The large size of *refined* subgraph, $|N_R(u)|$ (order of 800), compared to the *display generation* subgraph size capped at 30 nodes; and (3) The extremely large number of subgraphs (equal to the number of nodes $|\mathcal{V}| = n$) that need to be processed, to ultimately generate node neighborhoods.

4.3.2 GRCS—*Step 2: Node Representation Vector Update*

After identifying node neighborhoods in a graph, we aim to learn node representations via the standard SkipGram model (Mikolov et al. 2013a). However, since GRCS yields completely deterministic representations, we avoid the randomness implied by the SkipGram model by using the same random seed every time we employ it. The Skipgram objective maximizes the log-probability of observing the neighborhood generated during the neighborhood definition step, given each node's feature representation:

$$\max_{f} \sum_{u \in V} \log(Pr(N_R(u) \mid f(u)) \tag{4.5}$$

where $N_R(u)$ is the refined neighborhood of node u, and $f(u)$ is its feature representation. Following common practice, we make the maximum likelihood optimization tractable by making two assumptions:

Assumption 1 (Conditional Independence) We assume that the likelihood of observing node u's neighborhood is independent of observing any other neighborhood, given its feature representation $f(u)$:

$$Pr(N_R(u) \mid f(u)) = \prod_{w \in N_R(u)} Pr(w \mid f(u)) \tag{4.6}$$

where w represents any node that belongs to node u's refined neighborhood.

Assumption 2 (Symmetry in Feature Space) The source node u and any node w in its refined neighborhood $N_R(u)$ have a symmetrical impact on each other in the continuous feature space. Therefore, the conditional probability, $Pr(w \mid u)$, is modeled using the softmax function:

$$Pr(w \mid f(u)) = \frac{\exp(f(w) \cdot f(u))}{\sum_{v \in V} \exp(f(v) \cdot f(u))} \tag{4.7}$$

Based on the above two assumptions, we can simplify the objective in Eq. (4.5) as follows:

$$\max_{f} \sum_{u \in V} \left[-\log \sum_{v \in V} \exp(f(v) \cdot f(u)) + \sum_{w \in N_R(u)} f(w) \cdot f(u) \right] \tag{4.8}$$

It is important to note that performing such calculations for each node in large-scale graphs is computationally expensive. Therefore, we approximate the function using negative sampling (Mikolov et al. 2013b). We optimize the objective shown in Eq. (4.8) using stochastic gradient decent.

4.4 Experiments

In this section, we aim to answer the following questions: (**Q1**) How does GRCS perform in multi-label classification compared to baseline representation learning approaches? (**Q2**) How stable are the representations that GRCS and baseline methods learn? Before we answer these questions, we provide an overview of the datasets, and the baseline representation learning algorithms that we use in our evaluation.

Datasets To showcase the generalization capability of GRCS over distinct domains, we use a variety of datasets, which we briefly describe in Table 4.2.

Baseline Algorithms We compare GRCS with three state-of-the-art baselines: **DeepWalk** (Perozzi et al. 2014), **node2vec** (Grover and Leskovec 2016), and **Walklets** (Perozzi et al. 2016). The reason why we choose these state-of-the-art methods is the way they adopt for neighborhood definition using random walks. On the contrary, in GRCS, we follow a completely deterministic manner, which makes our method applicable for single and multi-graph problems. Table 4.3 lists the parameters settings of GRCS, DeepWalk, node2vec, and Walklets, respectively. For GRCS, we set the expansion neighborhood subgraph size,

Table 4.2 A brief description of evaluation datasets

Dataset	# Vertices	# Edges	# Labels	Network type
PPI (Breitkreutz et al. 2007)	3,890	76,584	50	Biological
Wikipedia (Mahoney 2011)	4,777	184,812	40	Language
BlogCatalog (Tang et al. 2012)	10,312	333,983	39	Social
CiteSeer (Sen et al. 2008)	3,312	4,660	6	Citation
Flickr (Tang et al. 2012)	80,513	5,899,882	195	Social

Table 4.3 Parameter settings used for GRCS, DeepWalk, node2vec, and Walklets, respectively

Algorithm	$\|N_E(u)\|$	$\|N_R(u)\|$	d	w	l	s	p	q	k
GRCS	1,200	800	128						
DeepWalk			128	80	10	10			
node2vec			128	80	10	10	1	1	
Walklets			128	80	10				2

$|N_E(u)|$, the refinement neighborhood subgraph size, $|N_R(u)|$, and the number of dimensions of the feature representation, d. For DeepWalk, we set the number of walks per node, w, walk length, l, neighborhood size, s, and d. For node2vec, we set w, l, s, and d. In addition, we set the return parameter, p, the in-out parameter, q, in order to capture the homophily, and the structural equivalence connectivity patterns, respectively. With respect to Walklets, we set w, l, d, and the feature representation scale, k, which captures the relationships captured at scale 2.

Experimental Setup For GRCS parameter settings, we set the expansion neighborhood subgraph size $|N_E(u)| = 1,200$. In order to compare with the baseline methods, we set the refinement neighborhood subgraph size, $|N_R(u)| = 800$, and the number of dimensions of the feature representation, $d = 128$, in line with the values used for DeepWalk, node2vec, and Walklets.

4.4.1 Q1: Multi-Label Classification

Setup Multi-label classification is a single-graph canonical task, where each node in a graph is assigned a single or multiple labels from a finite set $\mathcal{L}$. We input the learned node representations to a one-vs-rest logistic regression classifier with L2 regularization. We perform tenfolds cross validation and report the mean Micro-F1 score results. We omit the results of other evaluation metrics—i.e., Macro-F1 score, because they follow the exact same trend. It is worth mentioning that multi-label classification is a challenging task, especially when the finite set of labels $\mathcal{L}$ is large, or the fraction of labeled vertices is small (Rossi et al. 2017).

Results In Table 4.4, we demonstrate the performance of GRCS algorithm and compare it to the three representation learning state-of-the-art methods. Our results are statistically significant with a p-value < 0.02. Overall, GRCS outperforms or is competitive with the baseline methods, while also having the benefit of generalizing to the multi-network problems that the other methods fail to address. Below we discuss the experimental results by dataset.

PPI: It is remarkable that using various percentages of labeled nodes, GRCS outperforms all the baselines. For instance, GRCS is more effective than

Table 4.4 Micro-F1 scores for multi-label classification on various datasets

	PPI			Wikipedia			BlogCatalog			CiteSeer			Flickr		
Algorithm	10%	50%	90%	10%	50%	90%	10%	50%	90%	10%	50%	90%	10%	50%	90%
DeepWalk	12.35	18.23	20.39	42.33	44.57	46.19	30.12	34.28	34.83	46.56	52.01	53.32	37.70	39.62	42.36
node2vec	16.19	20.64	21.75	44.38	**48.37**	48.85	**34.53**	**36.94**	**37.99**	**50.92**	52.49	56.72	38.90	41.39	43.91
Walklets	16.07	21.44	22.10	43.69	44.68	45.17	26.90	29.09	30.41	47.89	52.73	54.83	38.32	40.58	42.62
GRCS	**16.91**	**21.71**	**23.97**	**45.68**	48.10	**49.90**	31.02	34.85	36.42	48.80	**53.36**	**57.12**	**38.98**	**42.31**	**44.26**
G.O. DWalk	36.85	19.08	17.55	7.90	7.91	8.03	3.00	1.63	4.55	4.80	2.59	7.13	3.40	6.79	4.49
G.O. N2vec	4.41	5.16	10.19	2.92	–	2.14	–	–	–	–	1.63	0.70	0.21	2.22	0.80
G.O. Walk	5.19	1.23	8.47	4.53	7.64	10.48	15.27	19.80	19.75	1.87	1.18	4.17	1.72	4.26	3.85

Numbers where GRCS outperforms other baselines are bolded. By "G.O." we denote "gain over"

DeepWalk by 36.85% when the labeled nodes are sparse (10%), 19.08% for 50% of labeled nodes, and 17.55% when the percentage of labeled nodes is 90%.

Wikipedia: We observe that GRCS outperforms the three baseline algorithms by up to 10.48% when using 90% of labeled nodes. In the only case where GRCS does not beat node2vec, it is ranked second.

BlogCatalog: We observe that GRCS has a comparable or better performance than DeepWalk and Walklets for various percentages of labeled nodes. Specifically, it outperforms DeepWalk by up to 4.55% and Walklets by up to 19.75%, when the percentage of labeled nodes is 90%. For more labeled nodes, GRCS achieves similar performance to node2vec.

CiteSeer: Similar to Wikipedia, GRCS outperforms the state-of-the-art algorithms, and achieves a maximum gain of 7.13% with 90% of labeled nodes.

Flickr: We perceive that GRCS outperforms the other three baselines by up to 6.79%, when using 50% of labeled nodes.

Discussion From the results, it is evident that GRCS mostly outperforms the baseline techniques on PPI, Wikipedia, CiteSeer, and Flickr networks, with exceptions, where GRCS was very close to the best method. This can be rooted in the fact that GRCS is more capable in preserving the global structure in such networks. On the other hand, although GRCS has a very comparable performance with node2vec on BlogCatalog dataset, it might be that the 2nd order biased random walks of node2vec are slightly more capable in preserving the homophily, and the structural equivalence connectivity patterns in social networks.

4.4.2 Q2: Representation Learning Stability

Setup Surveying the existing node representation learning methods, we perceive that the tasks for which such algorithms are being evaluated on are limited to single graph-related tasks—i.e., prediction, recommendation, node classification, and visualization. Since many tasks involve multiple networks (e.g., graph similarity (Koutra et al. 2013), graph alignment (Bayati et al. 2009), temporal graph anomaly detection (Koutra et al. 2013), brain network analysis for a group of subjects (Fallani et al. 2014)), we seek to examine the similarity of representations learning approaches to multi-network settings. Heimann and Koutra (2017) states that existing embedding algorithms are inappropriate for multi-graph problems, and attribute this to the fact that different runs of any algorithm yield different representations every time the algorithm is run even if the same dataset is used. To that end, GRCS is fully deterministic, with the goal of achieving stable and robust outcomes. We evaluate its representation stability by verifying the similarity of the learned vectors across different independent runs of the algorithms. Ideally, a robust embedding should satisfy such a criteria.

Results Figure 4.4 shows the embeddings of two different runs of each approach against each other for a randomly selected set of nodes. For $d = 128$, we visualize the results for three randomly selected dimensions of node2vec, DeepWalk, and Walklets. For GRCS, we intentionally choose the same three dimensions randomly selected for each of the baseline methods. In the interest of space, we only show the visualization results of GRCS using the same three dimensions (39, 55, 111) used for Walklets dataset. The results are equivalent for all the dimensions. If all points fall on (or close to) the diagonal, this indicates stability, which is a desirable attribute of a robust graph embedding. Figure 4.4a, b and c shows that, as expected node2vec, DeepWalk, and Walklets suffer from significant variation across runs. To the contrary, Fig. 4.4d shows that GRCS obtain perfectly consistent embeddings across runs, and thus it is robust.

4.5 Conclusions

We propose a novel and stable representation learning algorithm, GRCS, using connection subgraphs. GRCS includes two steps: (1) Neighborhood identification using connection subgraphs, which represents our contribution, and (2) Representation vector update using the established SkipGram model. In essence, our contribution lies in the first step, where we generate node neighborhoods using connection subgraphs. For that, we divide the neighborhood generation process into two phases: (1) Neighborhood expansion, where we extract a much smaller subgraph from original graph to capture proximity between any two non-adjacent nodes in the graph on distance basis, and (2) Neighborhood refinement, where we perform extra pruning to the expanded subgraph in the previous phase using information flow (current) basis. It is worth mentioning that even though connection subgraphs are designated to attain proximity among two non-adjacent nodes, they are still capable of addressing local connectivity between a pair of such nodes. This can be attributed to the nature of the neighborhood generation process.

In contrast to existing representation learning baseline algorithms, GRCS generates entirely deterministic representations, which makes it more appealing for single and multi-graph problems. We empirically demonstrate GRCS efficacy and stability over state-of-the-art algorithms, showing that GRCS is more or as effective as baseline algorithms and is completely stable using multi-label classification problems.

In addition to improving the quality of neighborhoods generated, which ultimately affect the quality of embeddings, we also focus on remedying the issue of randomness of generated embeddings. However, the way we generate the neighborhoods imposes a scalability limitation, where GRCS has higher runtime comparing to baseline methods, which makes it less appealing to scale to large-scale networks. Further, for classification problems, the accuracy results remain unsatisfactory. Therefore, in the future work, we will work to enhance scalability and accuracy. Further, we will account for the interpretability that has never been

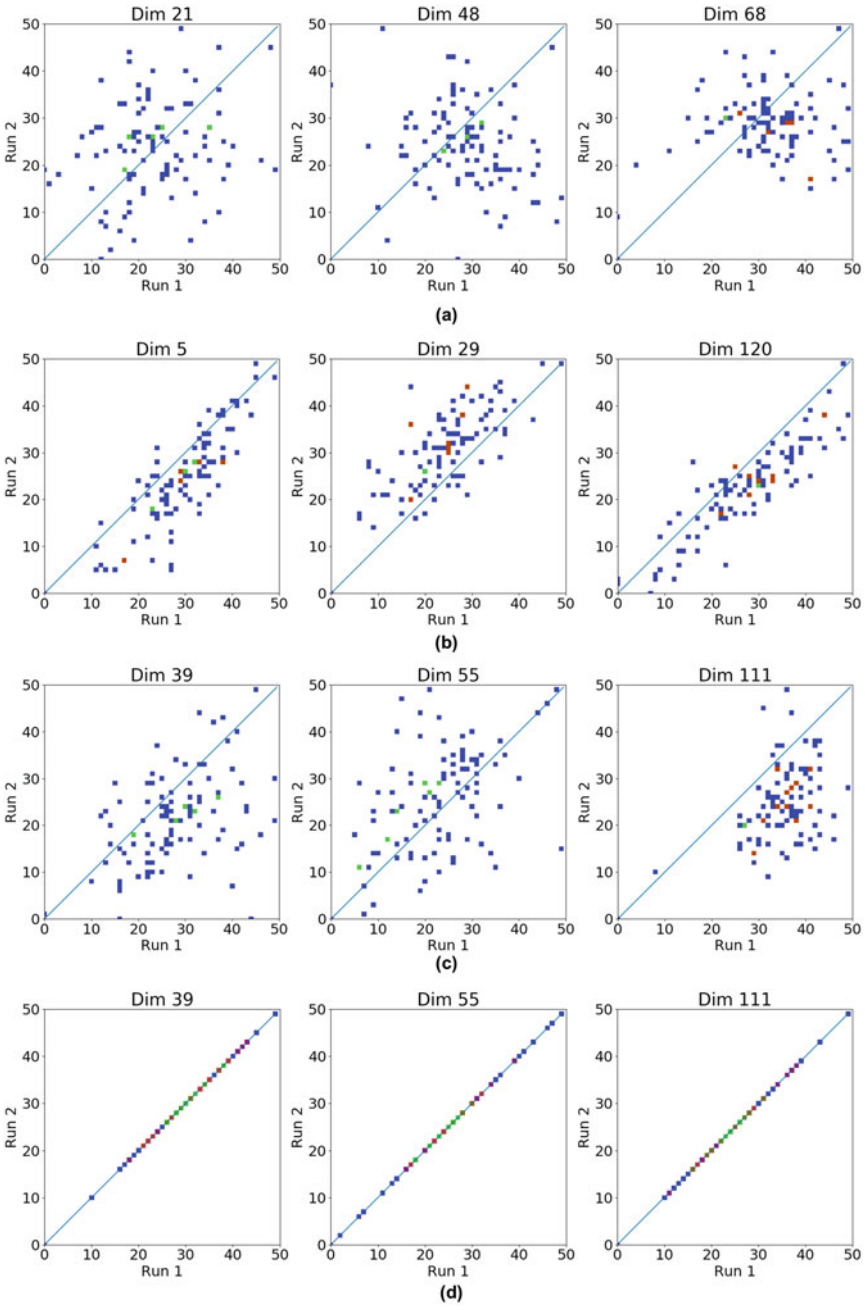


Fig. 4.4 PPI data: Comparison of embeddings per dimension for a random sample of 100 nodes. node2vec, DeepWalk, Walklets, and GRCS are run two times. The x-axis represents first run representations values, and the y-axis represents second run representations values. Three dimensions are selected randomly for each algorithm. The GRCS-based representations are robust across runs (perfectly fall on a straight line $y = x$), which is not the case for node2vec, DeepWalk, and Walklets. The results are consistent for all the datasets. **(a)** Node2vec. Dimension from left: 21, 48, 68. **(b)** DeepWalk. Dimension from left: 39, 55, 111. **(c)** Walklets. Dimension from left: 39, 55, 111. **(d)** GRCS. Dimension from left: 39, 55, 111

addressed yet. We will also address the issue of embedding update, especially for a recently joined nodes that have no evident connections. This problem is very related to the "cold-start" problem in recommendation systems, where a new user joins the system, and we seek external information for them, in order to properly compute their profile. Similarly, we propose to explore different forms of external context and meta-data for the recently joined nodes, which can help us address connection sparsity. Further, we will examine how vertices connectivity would impact the selection of expansion and refinement subgraphs sizes.

References

Ahmed A, Shervashidze N, Narayanamurthy S, Josifovski V, Smola AJ (2013) Distributed large-scale natural graph factorization. In: Proceedings of the 22nd international conference on world wide web. ACM, New York, pp. 37–48

Bayati M, Gerritsen M, Gleich DF, Saberi A, Wang Y (2009) Algorithms for large, sparse network alignment problems. In: Data mining, 2009. ICDM'09. Ninth IEEE international conference on. IEEE, Washington, pp. 705–710

Belkin M, Niyogi P (2002) Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Advances in neural information processing systems. MIT Press, Cambridge, pp. 585–591

Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828

Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bähler J, Wood V et al (2007) The biogrid interaction database: 2008 update. Nucleic Acids Res 36(suppl 1):D637–D640

Cao S, Lu W, Xu Q (2016) Deep neural networks for learning graph representations. In: Association for the advancement of artificial intelligence. AAAI Press, California, pp. 1145–1152

Easley D, Kleinberg J (2010) Networks, crowds, and markets: reasoning about a highly connected world. Cambridge University Press, Cambridge

Fallani FDV, Richiardi J, Chavez M, Achard S (2014) Graph analysis of functional brain networks: practical issues in translational neuroscience. Phil Trans R Soc B 369(1653):20130521

Faloutsos C, McCurley KS, Tomkins A (2004) Fast discovery of connection subgraphs. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp. 118–127

Fouss F, Pirotte A, Renders JM, Saerens M (2007) Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. IEEE Trans Knowl Data Eng 19(3):355–369

Goyal P, Ferrara E (2017) Graph embedding techniques, applications, and performance: a survey. arXiv preprint arXiv:1705.02801

Grover A, Leskovec J (2016) node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp. 855–864

Heimann M, Koutra D (2017) On generalizing neural node embedding methods to multi-network problems. ACM, New York

Koutra D, Vogelstein JT, Faloutsos C (2013) Deltacon: a principled massive-graph similarity function. In: Proceedings of the 2013 SIAM international conference on data mining. SIAM, Philadelphia, pp. 162–170

Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: Proceedings of the 31st international conference on machine learning (ICML-14). Beijing, pp. 1188–1196
Mahoney M (2011) Large text compression benchmark
Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781
Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. Curran Associates Inc., New York, pp. 3111–3119
Newman ME (2005) A measure of betweenness centrality based on random walks. Soc Networks 27(1):39–54
Perozzi B, Al-Rfou R, Skiena S (2014) Deepwalk: online learning of social representations. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp. 701–710
Perozzi B, Kulkarni V, Skiena S (2016) Walklets: multiscale graph embeddings for interpretable network classification. arXiv preprint arXiv:1605.02115
Rossi RA, Zhou R, Ahmed NK (2017) Deep feature learning for graphs. arXiv preprint arXiv:1704.08829
Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500):2323–2326
Sen P, Namata G, Bilgic M, Getoor L, Galligher B, Eliassi-Rad T (2008) Collective classification in network data. Artif. Intell. Mag. 29(3):93
Tang J, Qu M, Wang M, Zhang M, Yan J, Mei Q (2015) Line: large-scale information network embedding. In: Proceedings of the 24th international conference on world wide web. ACM, New York, pp. 1067–1077
Tang L, Wang X, Liu H (2012) Scalable learning of collective behavior. IEEE Trans Knowl Data Eng 24(6):1080–1091
Wang D, Cui P, Zhu W (2016) Structural deep network embedding. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp. 1225–1234
Zhou R, Hansen EA (2006) Breadth-first heuristic search. Artif Intell 170(4–5):385–408

Saba A. Al-Sayouri is a third-year PhD Candidate in the Systems Science and Industrial Engineering Department at Binghamton University. Her desire to join STEM started since she was a kid. Based on her high school GPA, she was admitted to the pharmacy school, which is more competitive than the engineering school to get in. However, since joining STEM, the engineering field in specific has been her ambition, she decided to eventually join the Industrial Engineering Department. She earned her B.S. degree in Industrial Engineering from Jordan University of Science and Technology in 2011. The strong passion she used to and still has towards STEM motivated her to pursue her graduate studies in the U.S. She earned her M.S. degree in Industrial and Systems Engineering from Binghamton University in 2014, where she employed data mining frameworks for healthcare-related applications. Her love for research has been her big motivation to pursue her Ph.D. study. She is currently working under the supervision of Professor Sarah Lam. Her research interest includes mining large-scale graphs by developing algorithmic frameworks that embed large-scale graphs from high-dimensional vector space to low-dimensional vector space.

Dr. Sarah S. Lam is the Associate Dean of the Graduate School and Professor of the Systems Science and Industrial Engineering Department at Binghamton University. She received an M.S. degree in operations research from University of Delaware and a Ph.D. degree in industrial engineering from University of Pittsburgh. A passion for mathematics and problem-solving led her to the field of industrial and systems engineering.

As an Assistant Director of Systems Analysis and Modeling of the Watson Institute for Systems Excellence (WISE) at Binghamton University, she has successfully administrated and led research projects in the areas of system optimization, data analytics, capacity planning and resource allocation, facility layout designs, supply chains, modeling and simulation of enterprise and healthcare systems. As a role model, she has inspired female students entering the STEM fields.

Chapter 5
Parameter Tuning Problem in Metaheuristics: A Self-Adaptive Local Search Algorithm for Combinatorial Problems

Cigdem Alabas-Uslu and Berna Dengiz

Contents

5.1 Combinatorial Optimization Problems 93
5.2 Metaheuristics and the Parameter Tuning Problem 94
5.3 Description of SALS for Single Objective COPs 96
5.4 Description of SALS for Multi-objective COPs 99
5.5 Application of SALS to VRP 100
5.6 Computational Study 102
5.7 Conclusions 107
References 107

5.1 Combinatorial Optimization Problems

The word combinatorial is derived from the word combinatorics which is a branch of discrete mathematics. On the other hand, combinatorial optimization is related with the combinatorics field and operations research. In combinatorial optimization problems (COPs), the desired result is found in the set of possible search space which is a finite or possibly countably infinite set. This is an integer number, a subset, a permutation, or a graph structure (Sait and Youssef 1999). COP, as an abstract representation, can be defined as finding a minimum weight feasible subset: $\text{Min}_{S\subseteq N}\left\{\sum_{j\in S} c_j : S \in \mathcal{F}\right\}$, where $N = \{1, \ldots, n\}$ is a finite set, c_j is weight for each $j \in N$, and $\mathcal{F}$ is the set of all feasible subsets of N.

C. Alabas-Uslu
Department of Industrial Engineering, Marmara University, Istanbul, Turkey
e-mail: cigdem.uslu@marmara.edu.tr

B. Dengiz (✉)
Department of Industrial Engineering, Başkent University, Ankara, Turkey
e-mail: bdengiz@baskent.edu.tr

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_5

Developing a suitable algorithm to find the "best" configuration of a set of decision variables with discrete values to achieve one or more than one goal for a COP is an important process and depends on the problem structure. Due to the many applications of COP in the real world, many algorithms have been developed. These algorithms are either complete or heuristic algorithms. While complete algorithms find an optimal solution, heuristic algorithms are not guaranteed to find an optimal but near optimal one. If COP is an NP-hard, no polynomial time algorithm exists. Therefore the field of metaheuristics for the application to COPs has been a rapidly growing field of operations research during the last 40 years.

5.2 Metaheuristics and the Parameter Tuning Problem

Since there is no polynomial time algorithm for COP that is NP-hard, exact optimization methods with exponential computation time are not practical to solve the large scale real-world COPs. Even though heuristic methods cannot guarantee the optimality, they are effective in finding approximate solutions within a reasonable computer time. For the last 40 years, researchers have been striving for the development of better heuristics to get highly qualified solutions in shorter computer times. The term metaheuristic was first introduced by Glover (1986). Metaheuristics provide a dynamic balance between diversification through the search space and intensification through the search process using different methods and meta-level strategies. Sörensen et al. (2017) give a comprehensive historical survey on metaheuristics including the recent innovations in this field.

As asserted in the no free lunch theorem (Wolpert and Macready 1997), there doesn't exist a single heuristic method that is superior to all others for all possible decision problems. Therefore, selection of an appropriate metaheuristic for a specific problem as well as setting the parameters of selected metaheuristic is crucial to attain the best performance for the desired solution. In the literature, determination of the best parameter set for a certain heuristic algorithm is called, interchangeably, parameter setting, parameter tuning, parameter calibration, parameter configuration or parameter optimization since this task, itself, is an optimization problem. If P is the performance metric that depends on parameter vector $\mathbf{x}$, selected from all possible combinations of parameters $\mathbf{X}$, parameter optimization problem can be described as $\max\{P(\mathbf{x}): \mathbf{x} \in \mathbf{X}\}$. Mutual influences and also interrelated influences of parameters which may vary on different instances of the decision problem make parameter optimization a complex and time-consuming task. To deal with the parameter optimization problem, the parameters are either fine tuned before the execution of the underlying algorithm by offline techniques, or they are allowed to change dynamically or adaptively throughout the execution by online techniques. However, online techniques may still exhibit parameters which must be determined a priori. In the current literature of offline techniques, automatic algorithm configuration gets an increasing interest since it automates the tuning process. Automatic algorithm configuration methods involve experimental

design and statistical modeling techniques (Coy et al. 2000; Adenso-Diaz and Laguna 2006; Bartz-Beielstein 2006; Dobslaw 2010; Arin et al. 2011), racing algorithms (Barbosa et al. 2015; Birattari et al. 2002; Balaprakash et al. 2007; López-Ibáñez et al. 2016), and meta-optimization approaches which tune the parameters using any other heuristic (Eberhart and Kennedy 1995; Nannen and Eiben 2006; Meissner et al. 2006; Hutter et al. 2007, 2009; Smith and Eiben 2009; Neumüller et al. 2011). Offline techniques have two main drawbacks. Firstly, the best values of parameters depend on the problem and even its various instances and therefore, should be instance specific (Ries et al. 2012). Secondly, the effectiveness of a pre-defined parameter set may change during the progress of search and different parameter values should be selected at different moments of the search (Eiben et al. 2007). Online techniques, also known as parameter control techniques, aim to tune the parameters during the execution of the algorithm to overcome these drawbacks. These techniques, mainly, consists of dynamic update and adaptive update approaches. A dynamic update changes the parameter value using a rule which does not take into account the progress of search while an adaptive approach updates the parameter value using feedback from the current search state. In the literature of genetic algorithms, if the parameters are incorporated into the solution representation, the approach is called self-adaptive parameter control. Thus, the parameter values are allowed to evolve as the population evolves. Robert et al. (1996), Eiben et al. (1999), Smith (2003), Krasnogor and Gustafson (2004) are some examples of self-adaptive parameter control in genetic algorithms. Readers are also referred to representative studies of Eiben et al. (1999), De Jong (2007), Battiti and Brunato (2010), Battiti and Tecchiolli (1994), Hansen and Ostermeier (2001), and Battiti et al. (2008) for the online techniques. Although these techniques aim to provide less parameter configuration, they still have parameters that need to be initialized. Hence, parameter-less metaheuristics have been one of the recent interests in the related literature. "Parameter-less" mainly emphasizes the algorithms which are completely free from parameter tuning and therefore, they are sufficiently simple to use in practice. Most of the efforts on parameter-less algorithms are based on genetic algorithms (Harik and Lobo 1999; Lobo and Goldberg 2004; Lima and Lobo 2004; Nadi and Khader 2011). Despite the amount of works on parameter tuning, parameter-less heuristics would be preferred by many practitioners even a parameterized heuristic could give a better performance.

As metaheuristic algorithms get complicated, in general, the number of parameters increases. Therefore, the parameter optimization problem of these algorithms also becomes more complex. If two algorithms show similar performance but one is significantly simpler, the simpler one will be, obviously, superior to the other in terms of simplicity (Silberholz and Golden 2010). Based on this fact, Alabas (2004) proposed a simple local search algorithm with a single parameter. Moreover, this single parameter, called acceptance parameter, is adaptively tuned during the progress of search without any need of initialization. Some useful feedback from the search process such as the number of improvement solutions and improvement rate of the initial cost is used to update the acceptance parameter. Therefore, the proposed algorithm, named self-adaptive local search (SALS) is a

simple and parameter-free heuristic. Although the acceptance parameter is tuned using the feedback from search process, it is called self-adaptive to emphasize its independence of parameter tuning. An application of SALS to the classical permutation flow shop scheduling problem was presented in Dengiz et al. (2009). Alabas-Uslu and Dengiz (2014) showed that SALS successively solves problems of quadratic assignment, classical vehicle routing and topological optimization of backbone networks. Dengiz and Alabas-Uslu (2015) also gave a detailed application of SALS to the topological optimization problem. Furthermore, the multi-objective application of SALS was presented by Alabas-Uslu (2008) for a bi-objective vehicle routing problem.

In this study, SALS algorithm, firstly, is explained for single objective and multi-objective COPs. Also, the application of SALS to a single objective and bi-objective vehicle routing problem is investigated further to show that it is able to perform better or at least similarly compared to several sophisticated metaheuristics proposed in the related literature.

5.3 Description of SALS for Single Objective COPs

SALS algorithm is based on local search. It starts at a given initial solution and explores the solution space iteratively through neighborhoods of the current solution. Let $\mathbf{X} = [x_1, x_2, \ldots, x_n]$ be the current solution which corresponds to a vector of decision variables in a COP and $\Theta > 1$ be the adaptive parameter of SALS. At any iteration, a neighbor solution $\mathbf{X}'$ is randomly selected from the neighborhood of the current solution, $\mathrm{N}(\mathbf{X})$. If neighbor $\mathbf{X}'$ satisfies the acceptance condition of $f(\mathbf{X}') \leq \Theta f(\mathbf{X})$, then it replaces the current solution $\mathbf{X}$, where $f(.)$ denotes cost of the given solution. It is assumed that the cost cannot be negative. In the case of $f(\mathbf{X}') > \Theta f(\mathbf{X})$, the neighbor solution is rejected. Searching in the neighborhood of $\mathbf{X}$ is continued until a trial neighbor solution satisfies the acceptance condition. Whenever the current solution is replaced, the algorithm progresses to the next iteration.

The essential property of SALS is the adaptive calculation of Θ. For this purpose two indicators are used: Improvement rate of the initial solution (α_1) and rate of the number of accepted neighbors to the current iteration number (α_2). Indicators α_1 and α_2 are computed as given by Eqs. (5.1) and (5.2), where $\mathbf{X}_b^{(i)}$ is the best solution observed until iteration i and $\mathbf{X}_z$ is the initial solution. $C(L^{(i)})$ is a simple counter of the number of improvements in the best solution encountered until iteration i.

$$\alpha_1 = \frac{f\left(\mathbf{X}_b^{(i)}\right)}{f\left(\mathbf{X}_z\right)} \tag{5.1}$$

$$\alpha_2 = \frac{C\left(L^{(i)}\right)}{i} \tag{5.2}$$

Decreasing values of α_1 represent that the solution quality of the best solution is improved when compared to the initial solution. On the other hand, constantly decreasing values of α_2 may indicate flat regions of the solution space, while its fluctuating values may indicate many local minima. At each iteration, α_1 and α_2 are updated and parameter Θ is adaptively calculated as given in Eq. (5.3) using the updated values of α_1 and α_2. Additionally, if the number of rejected neighbors reaches the size of $N(\mathbf{X})$ at any iteration, the acceptance condition is relaxed by increasing Θ by $\alpha_1 \cdot \alpha_2$ only for that iteration. Parameter Θ determines the border of the search region surrounding the current solution $\mathbf{X}$ in terms of the objective function value. In the first iteration, $i = 1$, initial value of Θ is 2 by definition, since $f\left(\mathbf{X}_b^{(i)}\right) = f(\mathbf{X}_z)$ and $C(L^{(i)}) = 1$. Θ tends to take smaller values as the iteration number increases. Obviously, as Θ approaches 1, the acceptance condition forces the search to find better solutions.

$$\Theta = 1 + \alpha_1 \cdot \alpha_2 \tag{5.3}$$

The pseudo-code of SALS is given in Figs. 5.1 and 5.2 also helps to better figure out the searching behavior of the algorithm. Our pre-experiments on each test instance of the classical vehicle routing problem (VRP) showed that as Θ approaches 1, the search also approaches highly qualified solutions. As an example,

---

$i \leftarrow 1$, $\mathrm{C}(\mathrm{L}^{(i)}) \leftarrow 1$;
Randomly create initial solution $\mathbf{X}_z$;
$\mathbf{X} \leftarrow \mathbf{X}_z$; $\mathbf{X}_b^{(i)} \leftarrow \mathbf{X}_z$;
Repeat
$\alpha_1 \leftarrow \dfrac{f(\mathbf{X}_b^{(i)})}{f(\mathbf{X}_z)}$; $\alpha_2 \leftarrow \dfrac{\mathrm{C}(\mathrm{L}^{(i)})}{i}$;
$\Theta \leftarrow 1 + \alpha_1 \cdot \alpha_2$;
i $\leftarrow i + 1$; $r \leftarrow 0$;
Repeat
Select a neighbor solution **X'** randomly from the N′(**X**);
r $\leftarrow$ r + 1;
If $r = |\mathrm{N}(\mathbf{X})|$ *then* $\Theta \leftarrow \Theta + \alpha_1\alpha_2$;
Until $f(\mathbf{X}') \leq \Theta f(\mathbf{X})$
If $f(\mathbf{X}') < f(\mathbf{X}_b^{(i)})$ *then* $\mathrm{C}(\mathrm{L}^{(i)}) \leftarrow \mathrm{C}(\mathrm{L}^{(i)}) + 1$, $\mathbf{X}_b^{(i)} \leftarrow \mathbf{X}'$;
$\mathbf{X} \leftarrow \mathbf{X}'$;
Until ($\Theta \rightarrow 1$)or(Another termination condition is met)

Fig. 5.1 Pseudo-code of SALS for single objective COPs

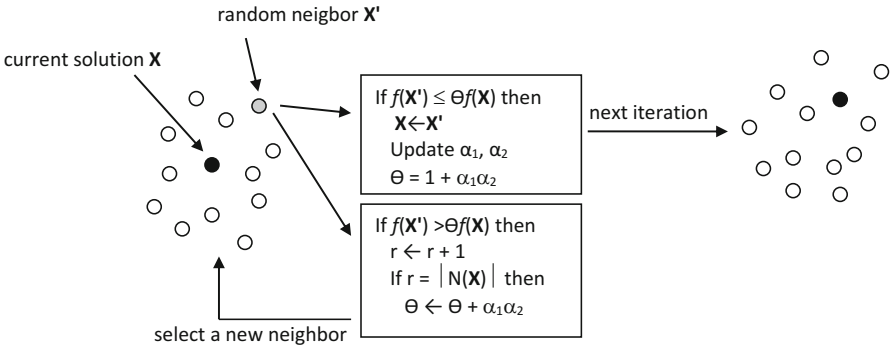


Fig. 5.2 Searching behavior of SALS algorithm

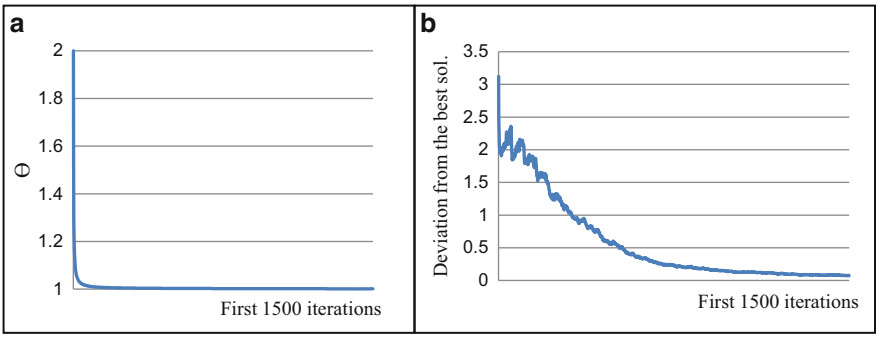


Fig. 5.3 (**a**) Convergence of Θ in the first 1500 iterations of SALS. (**b**) Convergence of the deviation from the best solution in the first 1500 iterations of SALS

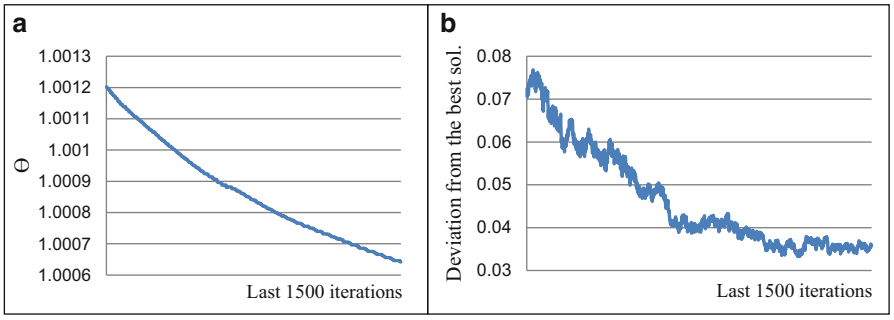


Fig. 5.4 (**a**) Convergence of Θ in the last 1500 iterations of SALS. (**b**) Convergence of the deviation from the best solution in the last 1500 iterations of SALS

Fig. 5.3a shows the values of Θ in the first 1500 iterations of SALS when applied to a moderate sized VRP instance and correspondingly deviations from the best solution obtained throughout these iterations are given in Fig. 5.3b. Similar figures are also shown in Fig. 5.4a, b for the last 1500 iterations to see the convergence of both

Θ and deviation from the best solution more clearly. Even though, neighborhood structures embedded in SALS algorithm have an impact on the solution quality as in all of the local search based heuristics, SALS is able to yield qualified solutions without any tuning effort. Therefore, the success of the algorithm mainly stems from the adaptive parameter Θ.

5.4 Description of SALS for Multi-objective COPs

SALS algorithm needs neither a composite function of the objectives nor preemptively solving the problem according to priorities of the objectives in the case of multi-objective COPs. Instead, the objective functions are minimized simultaneously using the acceptance condition which is adapted from implementation of SALS in single objective problems explained in Sect. 5.3. The adapted acceptance condition is as follows where $\mathbf{X}'$ is a neighbor solution of the current solution $\mathbf{X}$ and k is the number of objectives:

If $(f_1(\mathbf{X}') \leq \Theta_1 f_1(\mathbf{X})) \wedge (f_2(\mathbf{X}') \leq \Theta_2 f_2(\mathbf{X})) \wedge \ \ldots \ \wedge (f_k(\mathbf{X}') \leq \Theta_k f_k(\mathbf{X}))$ then $\mathbf{X} \leftarrow \mathbf{X}'$

Function $f_j(.)$ is the jth objective function assuming that all functions are nonnegative cost functions. Θ_j, for $j = 1, \ldots, k$, are the adaptive parameters of SALS. Exploration in the neighborhood of current solution $\mathbf{X}$ is maintained until a neighbor solution satisfies the acceptance condition. Once a neighbor solution is accepted, the algorithm replaces the current solution with the neighbor and progresses to the next iteration. If the total number of sampled neighbors reaches the neighborhood size, $|N(X)|$, value of parameter Θ_j (for all j) is increased by $\alpha_{j1} \cdot \alpha_{j2}$ only for the exploration in the current neighborhood. Here, α_{j1} and α_{j2} are the indicators used to adaptively compute parameter Θ_j as given in Eqs. (5.4) and (5.5). Θ_j is re-computed using Eq. (5.6) in each iteration for all j.

$$\alpha_{j1} = \frac{f_j\left(\mathbf{X}_{bj}^{(i)}\right)}{f_j\left(\mathbf{X}_z\right)} \quad \text{for } j = 1, \ldots, k \tag{5.4}$$

$$\alpha_{j2} = \frac{C\left(L_j^{(i)}\right)}{i} \quad \text{for } j = 1, \ldots, k \tag{5.5}$$

$$\Theta_j = 1 + \alpha_{j1}\alpha_{j2} \quad \text{for } j = 1, \ldots, k \tag{5.6}$$

In Eq. (5.4), $\mathbf{X}_{bj}^{(i)}$ $(j = 1, \ldots, k)$ is the best solution observed until iteration i according to the jth objective function and $\mathbf{X}_z$ is the initial solution. In Eq. (5.5),

$i \leftarrow 1$, $C(L_j^{(i)}) \leftarrow 1$ for $j = 1,\ldots k$;
Randomly create initial solution $\mathbf{X}_z$;
$\mathbf{X} \leftarrow \mathbf{X}_z$; $\mathbf{X}_{bj}^{(i)} \leftarrow \mathbf{X}_z$ for $j = 1,\ldots k$;
Repeat
  For $j = 1,\ldots k$ *do* $\alpha_{j1} \leftarrow \dfrac{f_j(\mathbf{X}_{bj}^{(i)})}{f_j(\mathbf{X}_z)}$, $\alpha_{j2} \leftarrow \dfrac{C(L_j^{(i)})}{i}$, $\Theta_j \leftarrow 1 + \alpha_{j1} \cdot \alpha_{j2}$;
  $i \leftarrow i + 1$; $r \leftarrow 0$;
  Repeat
    Select a neighbor solution **X'** randomly from the $N'(\mathbf{X})$;
    $r \leftarrow r + 1$
    If $r = |N(\mathbf{X})|$ *then For* $j = 1,\ldots k$ *do* $\Theta_j \leftarrow \Theta_j + \alpha_{j1} \cdot \alpha_{j2}$;
  Until $(f_1(\mathbf{X}') \le \Theta_1 f_1(\mathbf{X})) \wedge (f_2(\mathbf{X}') \le \Theta_2 f_2(\mathbf{X}) \wedge \ldots \wedge (f_k(\mathbf{X}') \le \Theta_k f_k(\mathbf{X}))$
  For $j = 1,\ldots k$ *do if* $f_j(\mathbf{X}') < f_j(\mathbf{X}_b^{(i)})$ *then* $C(L_j^{(i)}) \leftarrow C(L_j^{(i)}) + 1$, $\mathbf{X}_{bj}^{(i)} \leftarrow \mathbf{X}'$;
  $\mathbf{X} \leftarrow \mathbf{X}'$
Until $(\Theta_1 \rightarrow 1) \wedge \ldots \wedge (\Theta_k \rightarrow 1)$or(Another termination criterion is met)

Fig. 5.5 Pseudo-code of SALS algorithm for multi-objective COPs

$C\left(L_j^{(i)}\right)$ gives the number of improvements in the best solution of jth objective function until iteration i. Similarly with the single objective implementation of SALS, parameters Θ_j determine the border of the search region surrounding the current solution **X** respecting the solution quality (α_{j1}) and the ratio of improvement solutions to the iteration number (α_{j2}).

Pseudo-code of SALS for multi-objective COPs is given in Fig. 5.5. The pre-experiments conducted for the problem of minimizing both the number of vehicles and the maximum length of routes in VRP showed that as Θ_j ($j = 1,\ldots, k$) approaches to 1, SALS forces the search to find higher qualified solutions. The experimental results obtained from multi-objective implementation of SALS are consistent with that of single objective implementation.

5.5 Application of SALS to VRP

In this section, the classical VRP and a bi-objective VRP were briefly explained and used to represent the application of SALS.

The Classical VRP The problem is designed for optimal delivery routes from one depot to a number of customers under the limitations of side constraints for minimization of total traveling cost. According to its graph theoretic definition,

$G = (V, A)$ is a complete graph in which $V = \{1, \ldots, n + 1\}$ is the set of vertices and A is the set of arcs. Vertices $i = 2, \ldots, n + 1$ correspond to the customers, whereas vertex 1 corresponds to the depot. A nonnegative cost, c_{ij}, associated with each arc $(i, j) \in$ A represents the travel cost between vertices i and j. Each customer i has a known nonnegative demand, d_i, to be delivered. Side constraints of the problem are given as follows:

- The total demand assigned to any route cannot exceed the vehicle capacity, Q.
- A fleet of m identical vehicles is located at the depot.
- Each vehicle must follow a single route such that it starts and terminates at the depot.
- Each customer must be served exactly once by exactly one vehicle.

The Bi-objective VRP In the literature of multi-objective VRP, different types of objective functions are taken into consideration such as number of vehicles, total traveling distance/cost, the longest route (from/to depot), and total waiting time. In this study, an application of VRP to school bus routing is considered. The school bus routing problem can be represented by $G = (V, A)$ where vertices $i = 2, \ldots, n + 1 \in V$ correspond to bus stops and vertex $i = 1$ represents the school. Each arc $(i, j) \in A$ is associated with a travel time between bus stops i and j. Each bus stop $(i = 2, \ldots, n + 1)$ is associated with a number of students, d_i, to be picked up at bus stop i. Identical m buses with capacity of Q students are located at the school. In addition to the side constraints of the classical VRP, solving this real-life problem involves minimization of the number of buses and also minimization of the maximum time spent by a given student in the route (that is, minimization of the longest route). The first objective is related to the transportation cost while the other ensures to provide higher quality service. Obviously, these objectives are in conflict.

Moving Mechanisms The only design step of SALS algorithm to apply any COP is to define moving mechanisms. A moving mechanism is a perturbation scheme which converts the current solution into its neighbors. Different neighborhoods can be created by using different moving mechanisms. In the application of SALS algorithm to VRP for both single and bi-objective cases, permutation solution representation is used to code a solution point: A solution point $\mathbf{X} = [x_1, x_2, \ldots, x_D]$ corresponds to a sequence of locations (depot and customers, or school and bus stops), where $D = n + m + 1$. As an example, $\mathbf{X} =$ [1 2 4 6 5 8 1 3 7 9 5 8 10 1] means that first route (i.e., first vehicle) starts from depot 1, visits locations 2, 4, 6, 5, 8 in the given order and goes back to depot 1; second vehicle departs from depot 1, visits customers 3, 7, 9, 5, 8, 10, successively, and returns depot 1. SALS benefits from various five moving mechanisms to create a diversification effect on the search process: adjacent swap (M_{AS}), general swap (M_{GS}), single insertion (M_{SI}), block insertion (M_{BI}), and reverse location (M_{RL}). Definitions of the move types are given in Table 5.1.

Initial Solution and Preventing Infeasibility SALS algorithm is initialized with a feasible initial solution. Initial feasible solution is constructed by assigning one vehicle to one location (customer or bus stop) assuming that the number of vehicles

Table 5.1 Definition of moving types

Type	Definition
M_{AS}	Nodes x_i and x_{i+1} are interchanged with probability 0.5 else nodes x_i and x_{i-1} are interchanged, for $i = 2, \ldots, D-1$
M_{GS}	Nodes x_i and x_j are interchanged, for $i, j = 2, \ldots, D-1$ and abs$(i-j) > 1$
M_{SI}	Node x_i is inserted between nodes x_j and x_{j+1}, for $i = 2, \ldots, D-2$ and $j = i+2, \ldots, D-1$
M_{BI}	A subsequence of nodes from x_i to x_{i+b} is inserted between nodes x_j and x_{j+1}, for $i = 2, \ldots, D-2-b, j = i+b+1, \ldots, D-1$ and $b = 1, \ldots, D-4$
M_{RL}	A subsequence of nodes from x_i to x_j is sequenced in the reverse order for $i, j = 2, \ldots, D-1$ and abs$(i-j) > 1$

Table 5.2 Performance of SALS on the classical VRP instances of Christofides and Elion (1969)

n (# of customers)	Reference cost	BD%[a]	AD%[b]	CV[c]	ART[d]
50	**524.61**	0	0.0008	0.0125	1.1483
75	**835.26**	0.02	1.2	0.0081	3.1150
100	**826.14**	0.5	0.8	0.0029	3.9508
100	**819.56**	0	0	0	4.1258
120	**1042.11**	0	0.4	0.0019	5.5717
150	**1028.42**	1.1	1.6	0.0057	8.7982
199	**1291.29**	2.9	3.98	0.0095	16.3203
		0.65	*1.14*	*0.0058*	*6.1472*

[a]The best percentage deviation from the reference cost out of 10 experiments
[b]Average percentage deviation from the reference cost over 10 experiments
[c]Coefficient of variation over 10 experiments
[d]Average run time over 10 experiments in minutes

is not restricted. A trial neighbor solution is found in two steps: First, selecting one of the five move types defined above randomly and then, perturbing the current solution according to that type of move randomly. Additionally, feasibility of each neighbor solution is always maintained by avoiding perturbations which result in infeasibility. According to the permutation solution representation adapted, all neighbor solutions satisfy the side constraints of VRP, except the capacity constraint, unless the first and the last elements (x_1 and x_D) in current solution $\mathbf{X}$ are replaced with other locations (customers/bus stops). Therefore, the feasibility is preserved for these constraints. On the other hand, the neighbors that violate capacity constraints are just rejected.

5.6 Computational Study

Experimental Results on the Classical VRP Performance of SALS on the classical VRP instances, firstly, is represented on the seven well-known instances of Christofides and Elion (1969). As given in Table 5.2, SALS finds the reference

cost in at least one of the 10 experiments for three instances out of 7. For the remaining instances, the best percentage deviation from the reference cost (BD) shows that the best costs generated by SALS are very close to the reference costs. Average deviation from the reference cost (AD) is also very small for each instance. Additionally, the coefficient of variation (CV) of the 10 experiments of SALS is substantially small for each instance and it reveals that SALS yields robust results. Average run time (ART) requirement of SALS also indicates that it can be used in solving of real-world problems.

Twenty large-sized instances of the classical VRP by Golden et al. (1998) are used to compare the SALS with the metaheuristics listed in Table 5.3, after the performance of SALS has been demonstrated on the relatively small instances of Christofides and Elion (1969). Table 5.3 also includes the number of parameters of these metaheuristics and their abbreviations. As outlined in the table, all listed algorithms require the parameter tuning process for a number of parameters changing from 1 to 20. All the heuristics in Table 5.3 are stochastic algorithms which need to be replicated to provide a certain confidence level. Since the number of replications and the complete set of solution costs are not available for each heuristic and each instance, BD is used as a common comparison measure. BD results of all the heuristics are shown in Table 5.4. As seen from the table, SALS has higher solution quality, on average, than five of the metaheuristics out of thirteen. Though MB-AGES algorithm gives quite good results for each instance, GGW-PRRT$_{IP}$ algorithm, which is a parallel implementation of record-to-record algorithm and integer programming, outperforms all algorithms. However, large parameter

Table 5.3 Some successful algorithms for the classical VRP and their parameters

Study	Algorithm	Algorithm abbreviation	Number of parameters
Xu and Kelly (1996)	Tabu Search	XK-TS	20
Golden et al. (1998)	Record-to-Record Travel	GWKC-RRT	3
Tarantilis and Kiranoudis (2002)	Adaptive Memory Programming	TK-AMP	7
Tarantilis et al. (2002a)	Threshold Accepting	TKV-TA1	7
Tarantilis et al. (2002b)	Threshold Accepting	TKV-TA2	1
Toth and Vigo (2003)	Tabu Search	TV-TS	7
Prins (2004)	Evolutionary Algorithm	P-EA	7
Reimann et al. (2004)	Ant Colony	RDH-AC	9
Tarantilis (2005)	Adaptive Memory Programming	T-AMP	8
Li et al. (2005)	Record-to-Record Travel	LGW-RRT	5
Mester and Braysy (2007)	Active Guided Evolution Strategy	MB-AGES	11
Groër et al. (2011)	Parallel Algorithm Combining Record-to-record travel with Integer Programming	GGW-PRRTIP	13

Table 5.4 BD results of some sophisticated metaheuristics and SALS on the classical VRP instances of Golden et al. (1998)

Pr	n	Reference cost	XK-TS	GWKC-RRT	TKV-TA1	TKV-TA2	TK-AMP	TV-TS	P-EA	RDH-AC	T-AMP	LGW-RRT	MB-AGES	GGW-PRRT$_{IP}$	SALS
1	241	5623.47	–	3.754	1.070	1.008	0.951	2.004	0.412	0.365	0.951	0.283	0.072	**0.000**	0.399
2	321	8404.61	–	7.111	1.478	1.285	1.285	1.766	0.515	0.530	0.658	–	0.515	**0.362**	0.764
3	401	11036.22	–	7.645	1.481	1.397	1.481	3.321	**0.000**	**0.000**	**0.000**	–	**0.000**	**0.000**	0.701
4	481	13592.88	–	7.698	0.502	0.838	0.328	9.694	**0.233**	0.782	0.328	–	**0.233**	**0.233**	1.179
5	201	6460.98	–	3.742	0.088	**0.000**	**0.000**	3.661	**0.000**	**0.000**	**0.000**	–	**0.000**	**0.000**	**0.000**
6	281	8400.33	–	7.340	0.345	0.326	0.345	6.702	0.148	0.150	0.166	–	**0.148**	0.150	0.326
7	361	10101.7	–	11.004	1.936	1.708	1.136	4.413	**0.929**	0.929	1.136	–	0.930	**0.929**	1.954
8	441	11635.3	–	7.554	2.738	2.867	2.586	3.446	1.663	1.663	2.586	0.526	0.243	**0.125**	1.980
9	256	579.71	1.620	1.273	2.969	2.698	–	2.353	2.041	1.235	0.987	1.006	0.635	**0.000**	1.070
10	324	736.26	1.399	1.751	3.908	3.887	–	2.092	2.058	1.971	1.399	0.939	0.720	**−0.541**	1.373
11	400	912.84	2.173	2.354	3.545	3.533	–	2.542	2.213	1.581	1.132	1.125	0.615	**0.056**	1.361
12	484	1102.69	3.449	3.128	3.691	3.723	–	4.031	2.820	3.462	2.513	1.227	0.408	**0.006**	1.293
13	253	857.19	2.786	2.782	1.805	1.724	–	1.354	2.096	0.919	0.912	0.605	0.224	**0.000**	2.374
14	321	1080.55	3.474	2.142	2.022	2.046	–	1.446	0.527	1.223	0.511	0.287	0.070	**0.000**	2.071
15	397	1337.92	2.980	1.966	3.447	3.563	–	2.356	2.201	1.517	1.195	1.004	0.546	**0.006**	1.824
16	481	1612.5	2.739	2.817	4.155	4.016	–	2.469	2.384	1.405	1.379	1.238	0.632	**0.072**	2.346
17	241	707.76	5.578	1.792	1.469	1.362	–	0.468	0.376	0.141	0.138	0.123	0.004	**0.000**	0.787
18	301	995.13	7.179	3.425	3.558	3.712	–	2.181	1.977	0.372	1.183	1.279	0.362	**0.000**	2.008
19	361	1365.6	5.146	2.742	3.150	2.993	–	2.589	0.797	0.117	0.396	0.923	0.092	**0.000**	0.880
20	421	1818.25	6.420	3.130	2.969	3.004	–	5.367	1.556	0.258	1.068	1.233	0.101	**0.000**	1.713
Average performance			*3.745*	*4.258*	*2.316*	*2.284*	*1.014*	*3.213*	*1.247*	*0.931*	*0.932*	*0.822*	*0.328*	*0.070*	*1.320*

Table 5.5 The results of SALS for the bi-objective VRP instances of Christofides and Elion (1969)

n (# of customers)	Bestt_{max}/m	Averaget_{max}/m	CV of t_{max}	CV of m	ART (min)
50	101.43/6	104.58/5.9	0.074	0.054	1.597
	126.31/5				
75	91.04/11	91.65/11.0	0.011	0.000	3.705
100	110.60/9	120.75/8.1	0.060	0.039	5.030
	115.82/8				
100	120.73/10	126.03/10.0	0.021	0.000	2.260
120	199.93/8	202.66/8.0	0.010	0.000	7.335
150	103.36/13	111.75/12.7	0.105	0.038	9.110
	111.84/12				
199	102.68/18	109.03/17.6	0.079	0.029	13.280
	104.17/17				

sets of MB-AGES and GGW-PRRT$_{IP}$ make these algorithms complicated to apply different instances. Also it should be noted that the best cost results reported by the researchers, except T-AMP, are all obtained using well-tuned parameter sets. Only T-AMP uses a common parameter set for all the instances. The most important advantage of SALS, on the other hand, is its parameter tuning free structure which let it be simple to apply different instances.

Experimental Results on the Bi-objective VRP Performance of SALS in solving of the bi-objective VRP is tested in the instances of Christofides and Elion (1969) by taking t_{max} (the longest route) and m (the number of vehicles) into account as the two objectives to be minimized. Table 5.5 shows the results of 10 experiments of SALS for the problem instances. SALS generates two alternative solutions for the four instances while it generates a single solution for the remaining problems. The average t_{max} and m values are close to the best values with small CVs and it is an indication of obtaining a good quality solution by this algorithm with reasonable time even for the biggest problem size.

Once the capability of SALS is tested on the instances of Christofides and Elion (1969) considering the bi-objectives, the algorithm is compared with the scatter search algorithm of Corberán et al. (2002), CFLM-SS, and the tabu search algorithm of Pacheco and Marti (2006), PM-TS, on the school bus routing problem. A total of 16 instances of school bus routing problem was taken from Corberán et al. (2002). Corberán et al. (2002) propose constructive methods for the school bus routing problem. The solutions obtained from the constructive methods are also improved and combined within the framework of scatter search. On the other hand, Pacheco and Marti (2006) develop different constructive methods from Corberán et al. (2002) and improve the constructed solutions using a tabu search algorithm. The considered tabu search algorithm includes an intensification phase based on the path relinking. Because the maximum number of buses in a solution is equal to the number of pickup points, assigning one bus to each pickup point. The solution approach used in both CFLM-SS and PM-TS is to minimize t_{max} for possible values

Table 5.6 Comparison of SALS with CFLM-SS and PM-TS for the bi-objective school bus routing problem instances of Corberán et al. (2002)

	Current	CFLM-SS					PM-TS					SALS				
n (# of bus stops)	t_{max}/m	$m-1$	m	$m+1$	$m+2$	$m+3$	$m-1$	m	$m+1$	$m+2$	$m+3$	$m-1$	m	$m+1$	$m+2$	$m+3$
57	70/12	57	56	53	51	48	52	51	**48**	**47**	48	**51**	**48**	**48**	**47**	**47**
24	45/5	**46**	**36**	**32**	**29**		**46**	**36**	**32**	**29**		**46**	**36**	**32**	**29**	
24	60/6	54	46	43	**39**		52	45	43	**39**		**50**	**44**	**39**	**39**	
19	70/3		**52**	42	37	**33**		**52**	**41**	37	**33**		**52**	**41**	**36**	**33**
22	60/4	**55**	44	**39**			**55**	45	**39**			**55**	**43**	**39**		
32	80/4	81	62	53	47	44	**79**	62	52	**45**	43	**79**	**61**	**51**	**45**	**42**
36	60/6	**51**	45	38	**36**		**51**	45	37	**36**		**51**	**43**	**36**	**36**	
53	75/9	61	59	50	47	44	59	53	48	45	42	**54**	**50**	**46**	**43**	**40**
39	90/5	82	65	53	50	48	90	65	53	49	46	**79**	**62**	**52**	**48**	**44**
22	60/6	44	**41**	**40**			**43**	**41**	**40**			**43**	**41**	**40**		
13	60/4	67	51	45	**39**		**42**	**40**	**40**	40		67	51	45	**39**	
4	25/2		**15**	**14**	**9**			**15**	**14**	**9**			**15**	**14**	**9**	
28	45/6	**39**	34	32	**29**		40	34	**29**	**29**		**39**	**33**	**29**	**29**	
23	60/5	**53**	46	**37**			**53**	**45**	**37**			**53**	**45**	**37**		
31	50/7	50	44	42	40		49	**42**	41	40		**47**	**42**	**40**	**39**	
9	60/2	**84**	**51**	**38**	**35**		**84**	**51**	**38**	**35**		**84**	**51**	**38**	**35**	

of m. SALS is also run for the possible values of m by minimizing of $t_{\max}$ for the comparison of two algorithms on the same base line. Comparative results are given in Table 5.6. Column 2 represents $t_{\max}/m$ values. Column segments (3–7), (8–12), (13–17) contain $t_{\max}$ values generated by CFLM-SS, PM-TS, and SALS, respectively, for a particular value of m ($m - 1$ to $m + 3$). Note that m in columns 3–17 corresponds to the current value of m given in column 2. As seen from the table, SALS provides either the same or better $t_{\max}$ than the other algorithms for all instances except the one. It is clearly shown that the performance of SALS algorithm is superior to the two algorithms in terms of the solution quality. Another advantage of SALS is that its implementation is simpler than both CFLM-SS with 7-parameter and PM-TS with 4-parameter in terms of the point of parameter tuning.

5.7 Conclusions

Combinatorial optimization is an important mathematical topic that is related with combinatorics to find an optimal solution from a finite set of search space. Many problems we face in real life are modeled as combinatorial optimization. There is an increasing interest among researchers to develop heuristic algorithms for combinatorial optimization problems, because enumeration based search is not feasible for them. So, obtaining global optimum solutions for these problems, within a reasonable time, is extremely time consuming by exact algorithms. Particularly in recent years, high-level metaheuristics have been developed for combinatorial optimization problems. On the other hand, it is known that metaheuristic or heuristic algorithms are controlled by a set of parameters so the best parameter set reveals better performance in terms of solution quality and computation times.

In this chapter while a survey is carried out about parameter tuning approaches in metaheuristics/heuristics, the performance of SALS algorithm which is a novel algorithm, is investigated on the vehicle routing problem considering both the single and multi-objectives on a large scale suit of test problem. The main focus of the algorithm is to reduce the effort of the parameter optimization/tuning to be able to find an optimal or near optimal solution. State-of-the art techniques are introduced and then it is shown that the SALS algorithm performs better or at least similarly according to several sophisticated heuristics which exist in related literature.

References

Adenso-Diaz B, Laguna M (2006) Fine-tuning of algorithms using fractional experimental design and local search. Oper Res 54(1):99–114

Alabas C (2004) Self-controlled local search heuristic for combinatorial optimization problems. PhD theses. Gazi University, Ankara

Alabas-Uslu C (2008) A self-tuning heuristic for a multi-objective vehicle routing problem. J Oper Res Soc 59(7):988–996

Alabas-Uslu C, Dengiz B (2014) A self-adaptive heuristic algorithm for combinatorial optimization problems. Int J Comput Int Sys 7(5):827–852

Arin A, Rabadi G, Unal R (2011) Comparative studies on design of experiments for tuning parameters in a genetic algorithm for a scheduling problem. Int J Exp Design Process Optim 2(2):103–124

Balaprakash P, Birattari M, Stutzle T (2007) Improvement strategies for the F-race algorithm: sampling design and iterative refinement. In: BartzBeielstein T, Blesa M, Blum C, Naujoks B, Roli A, Rudolph G, Sampels M (eds) 4th International Workshop on Hybrid Metaheuristics, Proceedings, HM 2007. Lecture Notes in Computer Science, vol 4771. Springer, Berlin, pp 108–122

Barbosa EBM, Senne ELF, Silva MB (2015) Improving the performance of metaheuristics: an approach combining response surface methodology and racing algorithms. Int J Eng Math 2015:9. https://doi.org/10.1155/2015/167031

Bartz-Beielstein T (2006) Experimental research in evolutionary computation: the new experimentalism, Natural Computing Series. Springer Verlag, Berlin

Battiti R, Brunato M (2010) Reactive search optimization: learning while optimizing. Chap. In: Gendreau M, Potvin JY (eds) Handbook of metaheuristics, 2nd edn. Springer, Berlin

Battiti R, Tecchiolli G (1994) The reactive tabu search. INFORMS J Comput 6(2):126–140

Battiti R, Brunato M, Mascia F (2008) Reactive search and intelligent optimization, Operations research/Computer Science Interfaces, vol 45. Springer, Berlin

Birattari M, Stutzle T, Paquete L and Varrentrapp K (2002). A racing algorithm for configuring metaheuristics. Proceedings of the Genetic and Evolutionary Computation Conference, 11–18, GECCO'02

Christofides N, Elion S (1969) An algorithm for the vehicle dispatching problem. Oper Res Quart 20:309–318

Corberán A, Fernández E, Laguna M, Martí R (2002) Heuristic solutions to the problem of routing school buses with multiple objectives. J Oper Res Soc 53(4):427–435

Coy SP, Golden BL, Runger GC, Wasil EA (2000) Using experimental design to find effective parameter settings for heuristics. J Heuristics 7:77–97

De Jong K (2007) Parameter settings in EAs: a 30 year perspective. In: Lobo FG, Lima CF, Michalewicz Z (eds) Parameter setting in evolutionary algorithms, Studies in computational intelligence. Springer, Berlin/Heidelberg, pp 1–18

Dengiz B, Alabas-Uslu C (2015) A self-tuning heuristic for design of communication networks. J Oper Res Soc 66(7):1101–1114

Dengiz B, Alabas-Uslu C, Sabuncuoğlu I (2009) A local search heuristic with self-tuning parameter for permutation flow-shop scheduling problem. In: IEEE Symposium on Computational Intelligence in Scheduling. CI-Sched'09, April 2–March 30, Nashville, TN, pp 62–67

Dobslaw F (2010) A parameter tuning framework for metaheuristics based on design of experiments and artificial neural networks. In: Proceedings of the International Conference on Computer Mathematics and Natural Computing, pp 1–4

Eberhart R, Kennedy J (1995) A new optimizer using particle swarm theory. In: Proceedings of the Sixth International Symposium on Micro Machine and Human Science, pp 39–43

Eiben AE, Hinterding R, Michalewicz Z (1999) Parameter control in evolutionary algorithms. IEEE Trans Evol Comput 3(2):124–141

Eiben AE, Michalewicz Z, Schoenauer M, Smith JE (2007) Parameter control in evolutionary algorithms. In: Lobo FG, Lima CF, Michalewicz Z (eds) Parameter setting in evolutionary algorithms. Studies in computational intelligence. Springer, Berlin/Heidelberg, pp 19–46

Glover F (1986) Future paths for integer programming and links to artificial intelligence. Comput Oper Res 13:533–549

Golden BL, Wasil EA, Kelly JP, Chao I-M (1998) The impact of metaheuristics on solving the vehicle routing problem: algorithms, problem sets, and computational results. In: Crainic TG, Laporte G (eds) Fleet management and logistics. Kluwer, Boston

Groër C, Golden B, Wasil E (2011) A parallel algorithm for the vehicle routing problems. INFORMS J Comput 23:315–330

Hansen N, Ostermeier A (2001) Completely derandomized self-adaptation in evolution strategies. Evol Comput 9(2):159–195

Harik GR, Lobo FG (1999) A parameter-less genetic algorithm. In: Banzhaf W, Daida J, Eiben AE, Garzon MH, Honavar V, Jakiela M, Smith RE (eds) GECCO-99: proceedings of the genetic and evolutionary computation conference. Morgan Kaufmann, SanFrancisco, pp 258–267

Hutter F, Hoos H, Stutzle T (2007) Automatic algorithm configuration based on local search. In: Proceedings of the twenty-second conference on Artificial intelligence (AAAI'07), pp 1152–1157

Hutter F, Hoos HH, Leyton-Brown K, Stützle T (2009) ParamILS: an automatic algorithm configuration framework. J Artif Intell Res 36:267–306

Krasnogor N, Gustafson S (2004) A study on the use of "self-generation" in memetic algorithms. Nat Comput Int J 3(1):53–76

Li F, Golden B, Wasil E (2005) Vey large-scale vehicle routing: new test problems, algorithms, and results. Comput Oper Res 32:1165–1179

Lima CF, Lobo FG (2004) Parameter-less optimization with the extended compact genetic algorithm and iterated local search. In: Proceedings of the genetic and evolutionary computation conference GECCO-2004, part I, LNCS 3102, Springer, pp 1328–1339

Lobo FG, Goldberg DE (2004) Parameter-less genetic algorithm in practice. Inf Sci 167(217):232

López-Ibáñez M, Dubois-Lacoste J, Cáceres LP, Birattari M, Stützle T (2016) The irace package: iterated racing for automatic algorithm configuration. Oper Res Perspect 3:43–58

Meissner M, Schmuker M, Schneider G (2006) Optimized particle swarm optimization (OPSO) and its application to artificial neural network training. BMC Bioinform 7:125. https://doi.org/10.1186/1471-2105-7-125

Mester D, Braysy O (2007) Active-guided evolution strategies for large-scale capacitated vehicle routing problems. Comput Oper Res 34:2964–2975

Nadi F, Khader AT (2011) A parameter-less genetic algorithm with customized crossover and mutation operators. In: Proceedings of the 13th annual conference on genetic and evolutionary computation, pp 901–908

Nannen V, Eiben AE (2006) A method for parameter calibration and relevance estimation in evolutionary algorithms. In: Proceedings of genetic and evolutionary computation conference, GECCO 2006, ACM, pp 183–190

Neumüller C, Wagner S, Kronberger G, Affenzeller M (2011) Parameter meta-optimization of metaheuristic optimization algorithms. In: Proceedings of the 13th international conference on Computer Aided Systems Theory, EUROCAST'11, Part I, Las Palmas de Gran Canaria, Spain, pp 367–374

Pacheco J, Marti R (2006) Tabu search for a multi-objective routing problem. J Oper Res Soc 57:29–37

Prins C (2004) A simple and effective evolutionary algorithm for the vehicle routing problem. Comput Oper Res 31:1985–2002

Reimann M, Doerner K, Hartl RF (2004) D-ants: saving based ants divide and conquer the vehicle routing problem. Comput Oper Res 31(4):563–591

Ries J, Beullens P, Salt D (2012) Instance-specific multi-objective parameter tuning based on fuzzy logic. Eur J Oper Res 218:305–315

Robert H, Zbigniew M, Thomas CP (1996) Self-adaptive genetic algorithm for numeric functions. In: Proceedings of the 4th international conference on parallel problem solving from nature, Springer-Verlag

Sait SM, Youssef H (1999) Iterative computer algorithms with applications in engineering. In: IEEE computer society, Los–Alamitos

Silberholz J, Golden B (2010) Comparison of metaheuristics. In: Gendreau M, Potvin JY (eds) Handbook of metaheuristics, 2nd edn. Springer, Berlin

Smith JE (2003) Co-evolving memetic algorithms: a learning approach to robust scalable optimisation. In: Proceedings of the 2003 congress on evolutionary computation, pp 498–505

Smith SK, Eiben AE (2009) Comparing parameter tuning methods for evo-lutionary algorithms. In: IEEE Congress on Evolutionary Computation, pp 399–406

Sörensen K, Sevausx M, Glover F (2017) A history of metaheuristics. arXiv preprint arXiv:1704.00853
Tarantilis CD (2005) Solving the vehicle routing problem with adaptive memory programming methodology. Comput Oper Res 32(9):2309–2327
Tarantilis CD, Kiranoudis CT (2002) BoneRoute: an adaptive memory-based method for effective fleet management. Ann Oper Res 115(1):227–241
Tarantilis CD, Kiranoudis CT, Vassiliadis VS (2002a) A backtracking adaptive threshold accepting metaheuristic method for the vehicle routing problem. Syst Anal Model Simul 42(5):631–644
Tarantilis CD, Kiranoudis CT, Vassiliadis VS (2002b) A list based threshold accepting algorithm for the capacitated vehicle routing problem. J Comput Math 79(5):537–553
Toth P, Vigo D (2003) The granular tabu search (and its application to the vehicle routing problem). INFORMS J Comput 15(4):333–348
Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. IEEE Trans Evol Comput 1(1):67–82
Xu J, Kelly JP (1996) A network flow-based tabu search heuristic for the vehicle routing problem. Transp Sci 30(4):379–393

Cigdem Alabas-Uslu earned her bachelor, master, and doctorate degrees from Gazi University in the Department of Industrial Engineering. She served as an assistant professor in Maltepe University and Okan University between 2004 and 2011. Currently, she is a full professor in the Industrial Engineering Department at Marmara University. Dr. Alabas-Uslu conducts research in the optimization of complex problems such as topological design of backbone networks, flow shop scheduling, vehicle routing, and the design of wind farm layouts using mathematical programming techniques and heuristic optimization algorithms. Simulation optimization is another of her research areas. She published her research findings in the international scientific journals. Her motive for being an engineer was her abilities in design and mathematics since her early years in education. She was mentored by Dr. Dengiz throughout her engineering studies.

Berna Dengiz is a full professor in the Industrial Engineering Department at Başkent University. She has been serving as the Dean of Engineering Faculty of Baskent University since 2008. Prior to becoming a professor at Baskent University, she was a professor in Industrial Engineering department at Gazi University, where she also served as Vice Dean (1996–2000) and Acting Dean (2000–2001) of the Engineering and Architecture Faculty of Gazi University. Dr. Dengiz conducts research in the fields of topology optimization of telecommunication and computer systems using heuristic optimization and the modeling of large scale complex industrial systems by simulation or simulation optimization.

Her research in the abovementioned fields has been funded by the Turkish Scientific and Technical Research Council (TÜBİTAK), and the Government Planning Organization of Turkey (DPT). Her international research collaborations were funded by NSF (USA) and NATO programs. She published her research results in distinguished journals such as IEEE

Transactions on Reliability, IEEE Transactions on Evolutionary Computation, International Journal of Production Economics (IJPE), Simulation Practice and Theory, Journal of the Operational Research Society (JORS), European Journal of Operational Research (EJOR), OMEGA, Computers and Industrial Engineering, IIE Transactions, among others which have garnered over 500 citations (in ISI Web of Science). Dr. Dengiz has also been a visiting professor at Auburn University, and the University of Pittsburgh.

Dr. Dengiz had been loving mathematics and science lessons ever since her young age and always wanted to be an engineer. She achieved this by becoming an engineer and serving as a role model for young girls. Just like her role models, the first female engineers of the Republic of Turkey. She has trained and mentored women engineers and women academics since 1987.

Dr. Dengiz received the "WORMS (Women in Operations Research and Management Science) 2011" award in the USA and the "Leader Women in Science 2012" award by TAUW in Turkey. Dr. Dengiz is a member of the Institute of Electrical and Electronics Engineers (IEEE) and the Institute for Operations Research and Management Science (INFORMS).

Chapter 6
A Partition-Based Optimization Approach for Level Set Approximation: Probabilistic Branch and Bound

Zelda B. Zabinsky and Hao Huang

Contents

6.1 Introduction 113
6.2 Framework of Probabilistic Branch and Bound (PBnB) for Level Set Approximation ... 116
6.3 Algorithm Details of PBnB for Level Set Approximation 120
6.4 Implementation of Probabilistic Branch and Bound 124
6.4.1 Implementation of PBnB for Level Set Approximation 124
6.4.2 Implementation of PBnB for Global Optimization 126
6.5 Performance Analysis 126
6.5.1 Analysis on a Function with Exact Evaluation (No Noise) 127
6.5.2 Analysis on a Noisy Function 134
6.6 Numerical Results 137
6.6.1 Sphere Function with Normal Noise 138
6.6.2 Continuous Test Functions with No Noise 138
6.6.3 Continuous Test Functions with Normal Noise 142
6.6.4 Integer Test Functions 142
6.7 Conclusions 144
Appendix 145
Proof of Theorem 1 145
Proof of Theorem 2 148
References 152

6.1 Introduction

Real world problems are often extremely complicated and involve uncertainty, making decision-making a challenge. For these complex problems, a closed form performance function may not exist. Simulation is considered to be a useful

Z. B. Zabinsky (✉)
University of Washington, Seattle, WA, USA
e-mail: zelda@u.washington.edu

H. Huang
Yuan Ze University, Taoyuan, Taiwan
e-mail: haohuang@saturn.yzu.edu.tw

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_6

approach to estimate the performance of complex systems. A computer simulation may be deterministic, as in a finite element analysis, or stochastic, such as a discrete-event simulation of a complex queueing network.

In order to optimize a complex system, many approaches using simulation optimization have been developed (e.g., see Fu 2015; Fu et al. 2008). Most of the simulation optimization algorithms focus on approximating a local or global optimal solution. Instead of just providing a single solution, our approach is to provide a set of solutions that achieves a target quantile (e.g., best 10%), allowing decision makers to make trade-offs between simulated performance and other issues. Others have also looked at methods to relax the objective of finding a single optimal solution. Ordinal optimization (Ho et al. 2000, 2007) introduced the concept of "goal softening" which finds at least one solution out of the best m solutions. Also, instead of finding the best design, indifference-zone procedures focus on obtaining a design with a "good" performance (Kim and Nelson 2001; Nelson et al. 2001; Rinott 1978; Shi and Ólafsson 2009).

Our approach, called probabilistic branch and bound (PBnB), has been developed over several years and involved several doctoral students (Huang 2016; Prasetio 2005; Wang 2011). It was originally motivated by research in air traffic flow management where weather had a large impact on delay. The research team wanted a method, apart from classical design of experiments, to optimize a detailed discrete-event simulation of air traffic with delay propagation due to weather events. The motivation was to provide a set of solutions that would give the decision-makers insight into the sensitivity of good policy decisions.

The PBnB algorithm was developed in Prasetio (2005), Wang (2011), Zabinsky et al. (2011) for both continuous and discrete optimization problems with black-box noisy function evaluations. This work used partitioning and sampling to prune subregions and concentrate subsequent sampling on promising subregions. Order statistics were used to determine the quality of each subregion for pruning and branching decisions. In Huang (2016), Huang and Zabinsky (2013), the concept of "maintaining" a subregion with statistical confidence was introduced. This chapter presents PBnB with maintaining and pruning for level set approximation, with a few words on variations of PBnB for global optimization.

PBnB is a random search algorithm that uses sampling and partitioning of the solution space, similar to a nested partitioning optimization framework (see Ólafsson 2004; Shi and Ólafsson 2000, 2009). The algorithm iteratively updates its confidence interval on the value of the target quantile, and seeks the target level set, that is, the set of solutions within the target quantile (e.g., all solutions in the best 10%). As the algorithm proceeds, each subregion is statistically identified as: (1) contained within the target level set, called maintained; (2) no intersection with the target level set, pruned; or (3) there is not sufficient evidence to decide. Probability bounds are derived on the maximum volume of incorrectly determined regions at each iteration, providing useful information to the user on the quality of solutions provided.

The benefit of estimating a level set in simulation optimization in contrast to a single solution is that decision makers can consider information that is not evaluated

in the simulation model. For instance, flexible threshold design problems (Pintér 1990) can use a level set solution to ensure that a threshold tolerance is met. Csendes and Pintér (1993) also considered a level set as an approach to sensitivity analysis. When there is noise in the performance measure (as is typically the case in discrete-event simulation), decision makers may be indifferent to small differences in the estimate of the objective function. By providing a set of solutions, PBnB enables decision makers to explore other aspects of the solutions that are in the target level set.

PBnB has been used in several applications to provide insight to decision-makers. In Prasetio (2005), an early version of PBnB was applied to weather forecasting and its impact on air traffic flow management. Zabinsky et al. (2011) applied PBnB to a parameter estimation problem, formulated as a maximum likelihood optimization problem that was nonlinear. In Huang et al. (2016), PBnB was used to perform a sensitivity analysis on a hepatitis C screening and treatment budget allocation problem, where the shape of the level set indicates which time period for the budget allocation has stronger impact on the health utility gain. Huang and Zabinsky (2014), developed a version of PBnB adapted to multiple objective functions. This version of PBnB approximates the Pareto optimal set of solutions to support decision makers in investigating trade-offs between objective functions. PBnB for multiple objective functions was used to study allocating portable ultrasound machines and MRI usage for orthopedic care in a hospital system (Huang et al. 2015). Two objectives of cost and health utility loss were evaluated via a discrete-event simulation, and a Pareto optimal set was approximated. More recently, PBnB was applied to a water distribution network using a simulator to identify the set of feasible solutions for decision makers (Tsai et al. 2018).

A partition based approach is used here for approximating a level set since it can provide a collection of subregions forming the approximation. Partition based algorithms have been used for finding global optimal solutions. The nested partition algorithm (Ólafsson 2004; Shi and Ólafsson 2000, 2009) partitions solution sets into multiple subregions and finds a promising subregion with a high likelihood of containing the global optimal solution. In contrast to nested partition, our algorithm does not narrow in on the global optimum, but creates a set of subregions considered "maintained" because there is statistical confidence at every iteration that the maintained region is contained in the target level set.

In addition, our algorithm considers subregions when determining whether implementing further samples is computationally warranted. A similar concept is used in optimal computing budget allocation (OCBA) algorithms, where a fixed number of simulation runs is given, and the algorithm evaluates the objective function at different solutions (Chen et al. 2008, 2010; Chen and Lee 2011). Also, ranking and selection algorithms group solutions and perform ranking and selection procedures for a maximum sample size (Xu et al. 2016). Another partition based algorithm, named empirical stochastic branch and bound (ESBnB), calculates empirical bounds for each subregion on future sampled performance (Xu and Nelson 2013).

Simulation optimization algorithms often provide asymptotic convergence results to a local or global solution, and maximize the probability of correct selection or minimize a computation budget. Also, stochastic approximation, nested partitioning, and model-based algorithms ensure convergence asymptotically (Fu et al. 2008; Hu et al. 2007; Ólafsson 2004). However, in practice, decision makers may want to run the algorithm for a while, and decide to continue or terminate. In contrast to asymptotic results, we derive bounds on the performance of the algorithm at any iteration. Finite time results help decision makers interpret the quality of the reported solutions, and decide when to terminate the algorithm.

An overview of PBnB for level set approximation is introduced in Sects. 6.2, and 6.3 describes the details of the algorithm. Section 6.4 provides guidelines and discussions of implementing PBnB. Section 6.5 presents several theorems that describe the quality of the level set approximation provided by the algorithm with probability bounds. Numerical results are presented in Sect. 6.6, and conclusions are in Sect. 6.7.

6.2 Framework of Probabilistic Branch and Bound (PBnB) for Level Set Approximation

The proposed algorithm aims to approximate a level set with respect to a performance function $f(x)$ for a black-box simulation model, where the optimization problem is

$$(\mathscr{P}) \min_{x \in S} f(x) \tag{6.1}$$

and $f(x) = E_{\Xi}[g(x, \xi_x)]$ is the expected value of a black-box simulation or a noisy function $g(x, \xi_x)$, and ξ_x is the random variable representing the noise. Typically, ξ_x is assumed to be normally distributed since the sample mean is often used to estimate $f(x)$, and the central limit theorem can be applied (Chen and He 2005). The design variable x is a vector in n dimensions, and the values may be integer or real-valued. In this chapter, the feasible set S is defined by upper and lower bounds on x, which makes partitioning easy and the volumes of the subregions (hyper-rectangles) easily computed.

We are interested in the δ quantile associated with $f(x)$, denoted $y(\delta, S)$, defined as

$$y(\delta, S) = \arg\min_{y} \{P(f(X) \leq y) \geq \delta\}, \text{ for } 0 < \delta < 1, \tag{6.2}$$

where X is a random variable uniformly distributed on the domain S. We recognize that $y(\delta, S)$ can also be expressed in terms of probability as

$$P(f(X) \leq y(\delta, S)) = \frac{v(\{x : f(x) \leq y(\delta, S), x \in S\})}{v(S)} \geq \delta \tag{6.3}$$

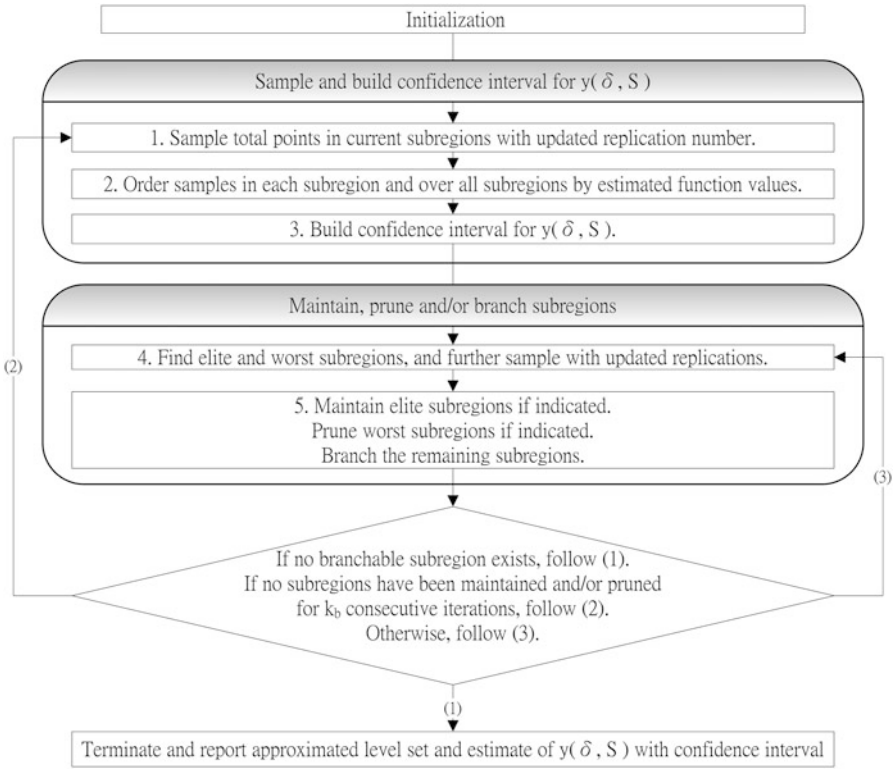


Fig. 6.1 Procedure of PBnB for level set approximation

$$P(f(X) < y(\delta, S)) = \frac{v(\{x : f(x) < y(\delta, S), x \in S\})}{v(S)} \leq \delta, \tag{6.4}$$

where $v(S)$ denotes the n-dimensional volume of set S. Also, we let $L(\delta, S)$ be our desired set of best δ-quantile solutions, where

$$L(\delta, S) = \{x \in S : f(x) \leq y(\delta, S)\}, \text{ for } 0 < \delta < 1. \tag{6.5}$$

The PBnB algorithm for level set approximation has two primary components, as illustrated in Fig. 6.1. The first component has three steps to develop a confidence interval for the target quantile $y(\delta, S)$. In Step 1, the algorithm samples over all current subregions. The number of sample points and replications depends on the desired confidence level and the size of the current subregion. In order to rank the samples correctly under noise, we apply a two-stage procedure based on Bechhofer et al. (1954) to update the replication number, in Step 2. We estimate the target quantile with confidence intervals in Step 3.

The second component focuses on processing subregions of the solution space, by maintaining, pruning, or branching. Based on the confidence interval, Step 4 finds the elite and worst subregions which have a potential to be maintained or pruned, and implements further sampling in the elite and worst subregions to statistically confirm the maintaining and pruning. In Step 5, the algorithm updates the maintained subregions, the pruned subregions and branches the rest of the subregions to create the current set of subregions.

PBnB for level set approximation has a straightforward stopping condition, route (1) in Fig. 6.1. It proceeds until all subregions are either maintained, pruned, or reach a user-defined minimum size that we term "unbranchable." In this chapter, a hyper-rectangle is "unbranchable" when the length of its longest side is less than a percentage (e.g., 4%) of the length of the shortest side of S. The size of an "unbranchable" hyper-rectangle impacts the precision of the level set approximation as well as the computation. In high dimensional problems, it could be difficult to achieve a tight approximation, and the user could terminate the algorithm by other methods, such as the volume of maintained subregions and/or a number of iterations.

If the algorithm does not terminate, it decides whether to update the quantile estimation, route (2) in Fig. 6.1, or to continue branching and sampling with the current quantile estimation, route (3) in Fig. 6.1, based on the number of consecutive iterations with unsuccessful maintaining and pruning. This balances the computation between reducing the number of current subregions (by maintaining or pruning), and narrowing the confidence interval of the quantile estimation for future maintaining and pruning.

Consider a two-dimensional sphere function as an example. The function details are described in Sect. 6.5 in Eq. (6.40). Figure 6.2 illustrates the progress of the algorithm with maintained, pruned, and branched subregions, overlaid on the contours of the sphere objective function, at iterations 6, 8, and 10. The target quantile is set to 10%, and the associated contour is bolded in Fig. 6.2. The white rectangles indicate the pruned subregions, and the numbers in the rectangles indicate the iteration at which they were pruned. The light gray (green) rectangles represent the maintained subregions that are contained in the target level set, and the dark gray (blue) rectangles are the current undecided subregions. As the algorithm progresses, the undecided subregions form the boundary of the target level set. In the example, at the sixth iteration, PBnB already pruned the outer part of the solution space. By iteration 8, four subregions in the center of the level set are maintained, and at iteration 10, several smaller subregions are maintained, and the current undecided subregions trace the level set boundary. Figure 6.3a shows the shrinking quantile interval estimation (upper and lower bounds on $y(\delta, S)$), and the bold line indicates the true quantile value $y(\delta, S)$ of the sphere function. Figure 6.3b shows the fractional volume of the maintained, pruned, and undecided subregions, for iterations 1 to 10 of the algorithm on the sphere function. Notice the volume of the undecided region tends to decrease while the volume of the maintained and pruned regions increases with iteration.

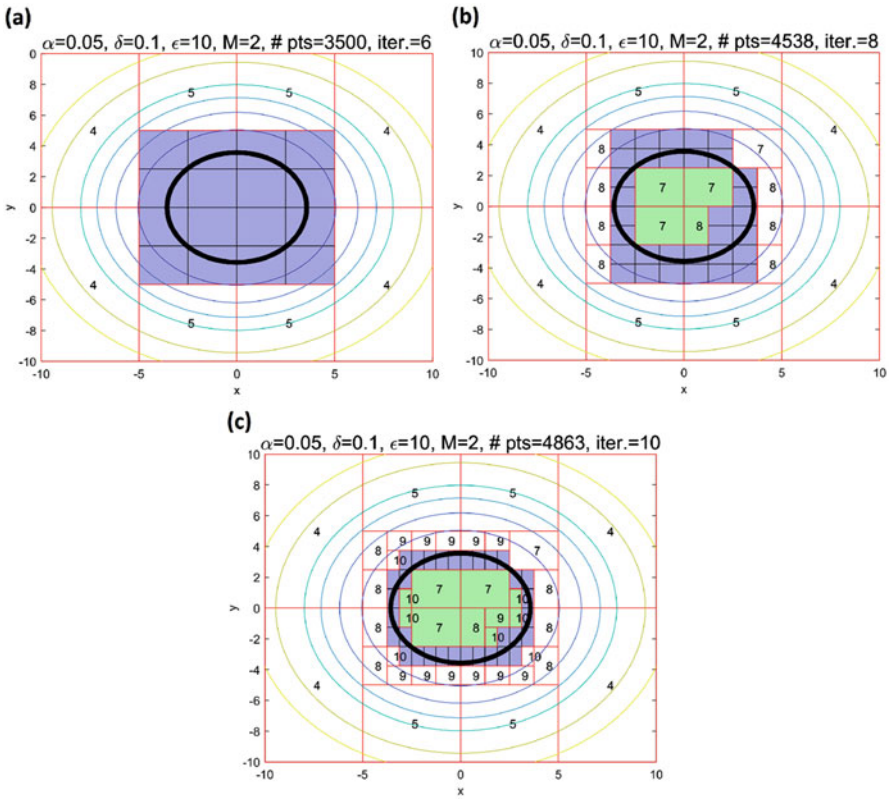


Fig. 6.2 PBnB for level set approximation on a two-dimensional sphere function, at iterations 6, 8 and 10

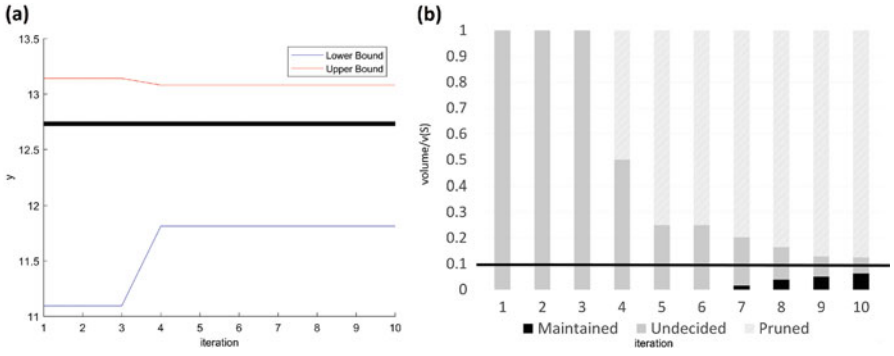


Fig. 6.3 (**a**) Quantile interval estimation (upper and lower bounds) on $y(\delta, S)$ for iterations $1, 2, \ldots, 10$. The bold line in (**a**) indicates the true value of $y(\delta, S)$. (**b**) The ratios of volumes of the maintained, pruned, and undecided subregions to the volume of S. The bold line in (**b**) indicates the target 10% on the two-dimensional sphere function

6.3 Algorithm Details of PBnB for Level Set Approximation

The input parameters to PBnB defined by the user include: δ, α, ϵ, k_b, B, c, and R^0. The parameter δ, $0 < \delta < 1$, is used to define a δ-quantile for the target level set. For example, the user may be interested in the set of solutions in the best 10%, in which case $\delta = 0.1$. The user does not have to specify a range on the objective function. If a specific value is known a priori, the algorithm can be modified to accommodate the information.

The following two parameters, α and ϵ, are used to determine the quality of the level set approximation. The approximation can be wrong in two ways: it could prune a portion of the level set or it could maintain some area that is not in the level set. The parameter α, $0 < \alpha < 1$, is used in the confidence level of the estimation of $y(\delta, S)$ and in the probabilities of incorrectly pruning or maintaining. The choice of α will influence the sample size and number of replications. As the confidence level $(1 - \alpha)$ increases, a larger sample size and more replications are needed to achieve the confidence level. The parameter $\epsilon > 0$ is the volume of solutions that can be tolerated to be categorized incorrectly. We also expect that a high confidence level (low α) will have fewer incorrectly pruned or maintained subregions, since the probability of the incorrect volume exceeding ϵ decreases.

The main results of the analysis are stated in Theorems 4, 5, and 6 giving probability bounds on the quantile estimation, on incorrectly pruning a volume of size ϵ, and on incorrectly maintaining a volume of size ϵ, respectively, for functions $f(x)$ that can be evaluated exactly (with no noise). Corollaries 1, 2 and 3 provide the same probability bounds for noisy functions where $f(x)$ is estimated by $\hat{f}(x)$.

The parameter k_b is the maximum number of consecutive iterations without maintaining or pruning before further overall sampling for quantile estimation. The branching scheme is defined by B, $B \geq 2$, which is the number of evenly sized subregions to create when a subregion is branched, c is the incremental sample size for each estimation of quantile when route (2) in Fig. 6.1 is taken, and R^o is the initial replication number at iteration 1.

The algorithm provides two results, a δ-quantile estimation and an approximation of the target level set. The δ-quantile is estimated by a confidence interval $[CI_l, CI_u]$, and the interval bounds are estimated by $CI_l = \hat{f}(z_{(r)})$ and $CI_u = \hat{f}(z_{(s)})$, where $\hat{f}(z_{(r)})$ and $\hat{f}(z_{(s)})$ are the rth and sth ordered samples at the last iteration. The level set approximation consists of the maintained region, denoted $\tilde{\Sigma}_k^M$, and illustrated as the light gray (green) rectangles in Fig. 6.2. In the analysis, we provide statistical confidence that $\tilde{\Sigma}_k^M$ is contained in the target level set up to a maximum error volume of ϵ. In addition to the maintained region, the algorithm also provides the undecided region $\tilde{\Sigma}_k^C$ shown as the dark gray (blue) rectangles in Fig. 6.2, and the pruned region $\tilde{\Sigma}_k^P$ shown as white rectangles in Fig. 6.2. The analysis provides the opposite information for the pruned region; statistical confidence that it is not in the target level set up to an error of ϵ.

Probabilistic Branch and Bound (PBnB) for Level Set Approximation

Step 0. Initialization: Input user-defined parameters, δ, α, ϵ, k_b, B, c, and R^0. Also, initialize the maintain, prune, and current subregion collections and iterative counters as $\Sigma_1 = \{S\}$, $\widetilde{\Sigma}_1^C = S$, $\widetilde{\Sigma}_1^M = \emptyset$, $\widetilde{\Sigma}_1^P = \emptyset$, $\delta_1 = \delta$, $\alpha_1 = \frac{\alpha}{B}$, $\epsilon_1 = \frac{\epsilon}{B}$, $R_0 = R^o$, and $k = 1, k_c = k_b, c_1 = c$. The set Σ_1 provides a way to keep track of the subregions that are not pruned or maintained, and $\tilde{\Sigma}_1^C$ indicates the union of those subregions. Also, $\tilde{\Sigma}_1^M$ and $\tilde{\Sigma}_1^P$ indicate the union of all maintained and pruned subregions, respectively, and both are empty at iteration 1.

Step 1. Sample c_k points in current subregions with updated replication number: For the current undecided subregion $\widetilde{\Sigma}_k^C$, uniformly sample additional points over the entire $\widetilde{\Sigma}_k^C$ such that the total number of points in $\widetilde{\Sigma}_k^C$ is c_k. For each subregion $\sigma_i \in \Sigma_k$, denote the sample points as $x_{i,j} \in \sigma_i$, for $j = 1, \ldots, N_k^i$ and $i = 1, \ldots, ||\Sigma_k||$, where $||\Sigma_k||$ is the number of subregions in set Σ_k. Note that $\sum_{i=1}^{||\Sigma_k||} N_k^i = c_k$. For notational convenience, let $N_k = c_k$. If $f(x)$ is noisy, PBnB evaluates $g(x, \xi)$ with R_{k-1} replications. Specifically, for each $x_{i,j} \in \sigma_i$, $j = 1, \ldots, N_k^i$ and $i = 1, \ldots, ||\Sigma_k||$, perform R_{k-1} replications of $g(x, \xi_x^r)$, and evaluate the sample mean and sample variance,

$$\begin{aligned} \hat{f}\left(x_{i,j}\right) &= \frac{\sum_{r=1}^{R_{k-1}} g\left(x_{i,j}, \xi_x^r\right)}{R_{k-1}} \text{ and } S_{\hat{f}}^2\left(x_{i,j}\right) \\ &= \frac{1}{(R_{k-1}-1)} \sum_{r=1}^{R_{k-1}} \left(g\left(x_{i,j}, \xi_x^r\right) - \hat{f}\left(x_{i,j}\right)\right)^2. \end{aligned} \tag{6.6}$$

Step 2. Order samples in each subregion and over all subregions: For each subregion i, $i = 1, \ldots, ||\Sigma_k||$, order the sampled points, $x_{i,(1)}, \ldots, x_{i,(N_k^i)}$, by their estimated function value so that

$$\hat{f}(x_{i,(1)}) \leq \hat{f}(x_{i,(2)}) \leq \cdots \leq \hat{f}(x_{i,(N_k^i)}).$$

Similarly, order all sampled points, $z_{(1)}, \ldots, z_{(N_k)}$, in all current subregions in Σ_k by their function values, so that

$$\hat{f}(z_{(1)}) \leq \hat{f}(z_{(2)}) \leq \cdots \leq \hat{f}(z_{(N_k)}).$$

If $g(x, \xi_x)$ is noisy, check the ordering with further replications calculated as follows.

(2.A) Calculate the differences between ordered samples, let $d_{i,j} = \hat{f}(x_{i,(j+1)}) - \hat{f}(x_{i,(j)})$, where $i = 1, \ldots, ||\Sigma_k||$ and $j = 1, \ldots, N_k^i - 1$. Determine

$$d^* = \min_{i=1,\ldots,||\Sigma_k||, j=1,\ldots,N_k^i-1} d_{i,j} \quad \text{and}$$

$$S^{*2} = \max_{i=1,\ldots,||\Sigma_k||, j=1,\ldots,N_k^i} S^2_{\hat{f}}(x_{i,(j)}).$$

(2.B) Calculate the updated replication number $R_k = \max\left\{R_{k-1}, \left(\frac{z_{\alpha_k/2}S^*}{d^*/2}\right)^2\right\}$, where $z_{\alpha_k/2}$ is the $1-\alpha_k/2$ quantile of the standard normal distribution. Perform $R_k - R_{k-1}$ more replications for each sample point. Re-estimate the performance of each sample point with R_k replications by $\hat{f}(x_{i,j}) = \frac{\sum_{r=1}^{R_k} g\left(x_{i,j}, \xi^r_{x_{i,j}}\right)}{R_k}$.

Within each subregion $\sigma_i \in \Sigma_k$, rank all the sample points $x_{i,j}$ as $x_{i,(j)}$ representing the jth best point in subregion, according to the estimated function value, and also update the entire order of all current samples with updated replications, so that

$$\hat{f}\left(x_{i,(1)}\right) \le \hat{f}\left(x_{i,(2)}\right) \le \cdots \le \hat{f}\left(x_{i,\left(N_k^i\right)}\right), \text{ and } \hat{f}\left(z_{(1)}\right) \le \hat{f}\left(z_{(2)}\right) \le \cdots \le \hat{f}\left(z_{(N_k)}\right).$$

Step 3. Build confidence interval for $y(\delta, S)$: To build the confidence interval of quantile $y(\delta, S)$, first, calculate the lower and upper bounds of δ_k as

$$\delta_{kl} = \delta_k - \frac{v\left(\widetilde{\Sigma}_k^P\right)\epsilon}{v(S)v\left(\widetilde{\Sigma}_k^C\right)} \text{ and } \delta_{ku} = \delta_k + \frac{v\left(\widetilde{\Sigma}_k^M\right)\epsilon}{v(S)v\left(\widetilde{\Sigma}_k^C\right)}. \tag{6.7}$$

Then calculate the confidence interval lower bound $CI_l = \hat{f}(z_{(r)})$ and the upper bound $CI_u = \hat{f}(z_{(s)})$, where r and s are selected by

$$\max r : \sum_{i=0}^{r-1} \binom{N_k}{i} (\delta_{kl})^i (1-\delta_{kl})^{N_k-i} \le \frac{\alpha_k}{2} \text{ and} \tag{6.8}$$

$$\min s : \sum_{i=0}^{s-1} \binom{N_k}{i} (\delta_{ku})^i (1-\delta_{ku})^{N_k-i} \ge 1 - \frac{\alpha_k}{2}. \tag{6.9}$$

The choice of r and s in constructing the confidence interval is discussed in Theorem 1.

Step 4. Find elite and worst subregions, and further sample with more replications: Step 4 identifies the indices of elite and worst subregions as e and w, representing the subregions that are likely to be maintained or pruned. The sets e and w are defined with the quantile confidence interval as

$$e = \left\{ i \middle| \hat{f}(x_{i,(N_k^i)}) < CI_l, \text{ for } i \in 1, \ldots, \left\| \Sigma_k \right\| \right\} \tag{6.10}$$

$$w = \left\{ i \middle| \hat{f}(x_{i,(1)}) > CI_u, \text{ for } i \in 1, \ldots, \left\| \Sigma_k \right\| \right\}. \tag{6.11}$$

Statistically confirm maintaining and pruning for each elite or worst subregion by sampling points up to N_k^i, where

$$N_k^i = \left\lceil \frac{\ln \alpha_k}{\ln\left(1 - \frac{\epsilon}{v(S)}\right)} \right\rceil, \text{ for all } i \in \{e \cup w\}. \tag{6.12}$$

For each new sample, perform R_k replications, and evaluate the sample mean and variances as in (6.6). Reorder the sampled points in each subregion σ_i, and update d^* and S^{*2} as in (2.A). As in (2.B), perform $\max\left\{R_k, \left(\frac{z_{\alpha_k/2} S^*}{d^*/2}\right)^2\right\} - R_k$ more replications for each new sample point, where $z_{\alpha_k/2}$ is the $1 - \alpha_k/2$ quantile of the standard normal distribution. Update $\hat{f}(x_{i,(1)}) = \min_{x_{i,j} \in \sigma_i} \hat{f}(x_{i,j})$, $\hat{f}(x_{i,(N_k^i)}) = \max_{x_{i,j} \in \sigma_i} \hat{f}(x_{i,j})$ for $i \in \{e \cup w\}$, and update $R_k \leftarrow \max\left\{R_k, \left(\frac{z_{\alpha_k/2} S^*}{d^*/2}\right)^2\right\} - R_k$.

Step 5. Maintain, Prune, and Branch: Update the maintaining indicator functions M_i, for $i \in e$, and the pruning indicator functions P_i, for $i \in w$, as

$$M_i = \begin{cases} 1, \text{ if } \hat{f}(x_{i,(N_k^i)}) < CI_l \\ 0, \text{ otherwise} \end{cases} \text{ and } P_i = \begin{cases} 1, \text{ if } \hat{f}(x_{i,(1)}) > CI_u \\ 0, \text{ otherwise.} \end{cases} \tag{6.13}$$

Update the maintained set $\widetilde{\Sigma}_{k+1}^M$ and the pruned set $\widetilde{\Sigma}_{k+1}^P$ as

$$\widetilde{\Sigma}_{k+1}^M = \widetilde{\Sigma}_k^M \bigcup \hat{\sigma}_m^k \quad \text{and} \quad \widetilde{\Sigma}_{k+1}^P = \widetilde{\Sigma}_k^P \bigcup \hat{\sigma}_p^k, \tag{6.14}$$

where $\hat{\sigma}_m^k = \bigcup_{i \in e: M_i = 1} \sigma_i$ and $\hat{\sigma}_p^k = \bigcup_{i \in w: P_i = 1} \sigma_i$, and branch the remaining current subregions in the following manner.

(1) If all subregions $\sigma_i \in \Sigma_k$ are not branchable, terminate the algorithm.
Else, if σ_i is branchable, and if σ_i, $i = 1, \ldots, \|\Sigma_k\|$, has not been maintained or pruned, then partition σ_i to $\bar{\sigma}_i^1, \ldots, \bar{\sigma}_i^B$ and update the current set of subregions

$$\Sigma_{k+1}^C = \{\bar{\sigma}_i^j : \forall i \text{ to be branched}, j = 1, \ldots, B\} \text{ and } \widetilde{\Sigma}_{k+1}^C$$

$$= \bigcup_{i \text{ to be branched}} \left(\bigcup_{j=1}^{B} \bar{\sigma}_i^j \right).$$

Determine δ_{k+1} by

$$\delta_{k+1} = \frac{\delta_k v\left(\widetilde{\Sigma}_k^C\right) - \sum_{i:M_i=1} v(\sigma_i)}{v\left(\widetilde{\Sigma}_k^C\right) - \sum_{i:P_i=1} v(\sigma_i) - \sum_{i:M_i=1} v(\sigma_i)}. \tag{6.15}$$

Set

$$k_c = \begin{cases} k_c + 1 \ , & \text{if } \sum_{i\in e} M_i + \sum_{i\in w} P_i = 0 \\ 0 \quad\quad , & \text{otherwise.} \end{cases} \tag{6.16}$$

Set $\alpha_{k+1} = \frac{\alpha_k}{B}$, $\epsilon_{k+1} = \frac{\epsilon_k}{B}$, $c_{k+1} = c_k + c$, and increment $k \leftarrow k + 1$.

(2) If $k_c \geq k_b$, set $k_c = 1$ and go to Step 1.

(3) If $k_c < k_b$, go to Step 4.

6.4 Implementation of Probabilistic Branch and Bound

Probabilistic branch and bound includes several user decisions to effectively apply the algorithm for different problems. This section provides guidelines for using PBnB for level set approximation, and discusses a variation for global optimization. The PBnB source code is available on GitHub at: https://github.com/ddlinz/PBnB.

Section 6.4.1 discusses the implementation of PBnB for level set approximation and Sect. 6.4.2 focuses on the variant for global optimization.

6.4.1 Implementation of PBnB for Level Set Approximation

The user-defined parameters of PBnB affect the computational efficiency and the quality of solutions ensured. Parameters δ, α, and ϵ impact the quality of the approximated level set provided by PBnB, and indirectly the computational effort.

The target quantile level δ, $0 < \delta < 1$, represents the desired level set, where a user could choose with several different viewpoints. It can be chosen with a particular quantile in mind, such as the best 10% solutions, or by considering how much the performance can be relaxed and still be considered good enough. In numerical results section, δ is used to represent the level set as a proportion of the domain, e.g., the top 10% solutions. However, the user can provide a specific value of performance to define a level set. The only change of the algorithm is to replace the quantile estimation with estimating the specific performance value.

The two parameters α and ϵ provide the statistical quality of the approximated level set. There is a trade-off between quality and efficiency; the higher the quality, the larger the sample size. The choice of α defines a confidence level $(1-\alpha)$, where the quality of the approximated level set is probabilistically defined with $(1-\alpha)$.

As mentioned in Sect. 6.3, the parameter ϵ is the volume of solutions that can be tolerated to be categorized incorrectly. If ϵ is interpreted as a fraction, e.g., 0.2, of the volume of the initial region, then one interpretation of the result is; if a point is randomly selected from the approximated level set, the probability it is not in the true level set is less than that fraction, e.g., 0.2. The user can choose ϵ based on the risk tolerance while implementing one solution from the approximated level set in the real world. The details of the statistical quality are discussed in Sect. 6.5.

Another major factor that can heavily influence the effectiveness of PBnB is the partitioning scheme. If a partitioning scheme can quickly identify the shape of the target level set, it significantly increases the speed of PBnB. The reason for the potential increase of efficiency is due to the possibility that the subregions are simply not small enough to capture the shape of the level set, even though PBnB may have enough samples. However, most of the time we have no information of the shape of the target level set since it is the goal intended to be approximated. There are potential research opportunities in developing a smart partitioning scheme and embedding it into PBnB algorithm. Linz et al. (2015) proposed a look-ahead approach to provide a potential lead to smart partitioning. In the numerical results section in this chapter, a straightforward empirical partitioning scheme is implemented. To be specific, it partitions the dimension of undecided subregions that has the longest length. Consequently, each variable of an undecided subregion will be partitioned into B subregions. If the user has no information about the problem, this type of empirical scheme is suggested. However, for higher dimensional problems, it may be beneficial to partition more than one variable at each iteration, since there are some many variables to be partitioned.

Other than the parameters defining the approximated level set and the partitioning scheme, the stopping condition is another user-defined setting for PBnB. In Sect. 6.3, the algorithm stops when the undecided subregions are unbranchable, where the definition of unbranchable is specified by the user. In general, an unbranchable subregion represents the smallest size of a subregion which is meaningful for its input variables. Hence, if different variables have different scales, the unbranchable subregion can be defined with a set of length for each decision variable. Other than using unbranchable as a stopping condition, there are other settings for users with different requirements. For instance, if the computational resource is limited, the stopping condition could be the total number of sample points. Also, changing the stopping condition does not affect the quality of solutions provided in the next section. The only trade-off is that the approximated level set may be smaller than desired. For high dimensional problems, it may be too computationally expensive to approximate the entire target level set and the user may simply seek a good enough solution, where the performance is bounded by the estimated quantile. In this case, a recommended stopping condition is when the first subregion is maintained.

6.4.2 Implementation of PBnB for Global Optimization

The main difference between PBnB for level set approximation and for global optimization lies in the maintained subregions. Once a subregion is considered to be maintained, i.e., within the target level set, PBnB for level set approximation stops sampling there. To apply PBnB for global optimization, it is possible to turn off the maintaining feature, and just keep sampling in the good subregions until a stopping condition is met. The details of the probabilistic analysis in this case can be found in Zabinsky et al. (2011) and Wang (2011).

The global optimization version of PBnB is implemented in a similar way to the level set approximation version with parameters δ, α, partitioning scheme and stopping condition. The major difference is the output of the global optimization version, which uses the best sampled solution as the approximated global optimal solution. The global optimization version does not provide maintained subregions of the level set, but statistically ensures the approximated global optimal solution is inside the δ level set, meaning it is good enough.

6.5 Performance Analysis

We analyze the performance of PBnB for level set approximation by deriving confidence intervals on $y(\delta, S)$ and probability bounds on the relationship of the pruned set, $\widetilde{\Sigma}_k^P$, and the maintained set, $\widetilde{\Sigma}_k^M$, with the desired level set, $L(\delta, S)$, at every iteration $k \geq 1$. This analysis is more complete than an earlier analysis that appeared in Huang and Zabinsky (2013).

First, in Sect. 6.5.1, we analyze the performance assuming the objective function $f(x)$ can be evaluated exactly, that is, there is no noise in observing $f(x)$. In Theorem 1, we derive the interval estimation of the target quantile $y(\delta, S)$ by mapping the δ_k quantile over the current region $\widetilde{\Sigma}_k^C$ at iteration k, to the δ quantile over S. This enables us to estimate the δ quantile over S even though the subregions change in size iteratively. Specifically, Theorem 1 develops bounds on $y(\delta, S)$ using $y\left(\delta_{kl}, \widetilde{\Sigma}_k^C\right)$ and $y\left(\delta_{ku}, \widetilde{\Sigma}_k^C\right)$ assuming upper bounds on the incorrect maintaining and pruning volumes and then provides a confidence interval on the target quantile, $y(\delta, S)$, using the volume of subregions maintained and pruned. We provide the quality of the level set approximation by deriving probability bounds on the volume of incorrect pruning and maintaining in Theorems 2 and 3, for a single iteration. Theorem 4 considers the *sequence* of iterations from 1 to k, and provides quantile lower and upper bounds for one-sided interval estimation, while accounting for the impact of the volume of incorrect pruning and maintaining. Theorems 5 and 6 provide probability bounds on incorrect pruning and maintaining regions of maximum volume ϵ, for the *sequence* of iterations 1 to k.

Section 6.5.2 discusses the impact of *noise* on the performance analysis. We estimate $f(x)$ with $\hat{f}(x)$, and in Theorem 7, we derive the impact of noise on the

probability of correctly ordering the samples. By incorporating the probability of correctly ordering the samples, Corollary 1, corresponding to Theorem 4, derives the probability that the sequence of interval quantile estimation up to iteration k captures the true target quantile when the function is noisy. Similarly, Corollaries 2 and 3 propose noisy versions of Theorems 5 and 6.

6.5.1 Analysis on a Function with Exact Evaluation (No Noise)

In Theorem 1, we let $\epsilon_k^M = v\left(\widetilde{\Sigma}_k^M\right) - v\left(L(\delta, S) \cap \widetilde{\Sigma}_k^M\right)$ denote the volume of $\widetilde{\Sigma}_k^M$ that is incorrectly maintained, and let $\epsilon_k^P = v\left(L(\delta, S) \cap \widetilde{\Sigma}_k^P\right)$ denote the volume of $\widetilde{\Sigma}_k^P$ that is incorrectly pruned at iteration k. Since volume is always non-negative, zero is a natural lower bound on ϵ_k^M and ϵ_k^P. Upper bounds on ϵ_k^M and ϵ_k^P are used in Theorem 1 to obtain upper and lower bounds on the target quantile and provide an interval estimation.

Theorem 1 *For any iteration $k \geq 1$, suppose there is no noise and $0 \leq \epsilon_k^P \leq \frac{\epsilon v(\widetilde{\Sigma}_k^P)}{v(S)}$ and $0 \leq \epsilon_k^M \leq \frac{\epsilon v(\widetilde{\Sigma}_k^M)}{v(S)}$. Then the bounds of the target quantile are*

$$y\left(\delta_{kl}, \widetilde{\Sigma}_k^C\right) \leq y(\delta, S) \leq y\left(\delta_{ku}, \widetilde{\Sigma}_k^C\right) \tag{6.17}$$

where δ_{kl} and δ_{ku} are from (6.7). Therefore, an interval estimate of the quantile is

$$P\left(f(z_{(r)}) \leq y(\delta, S) \leq f(z_{(s)})\right) \geq 1 - \alpha_k, \tag{6.18}$$

where $z_{(1)}, \ldots, z_{(N_k)}$ are the N_k uniform samples from the current region $\widetilde{\Sigma}_k^C$ at iteration k, ordered by function values (as in Step 2), $0 < \alpha_k < 1$, and r and s satisfy (6.8) and (6.9).

Proof The full proof is available in the appendix, and here we provide a sketch of the proof. We consider the iterative effect of δ_k on the estimation of the original δ as subregions are pruned or maintained. In the algorithm, (6.15) is used to update δ_k assuming that the maintained regions are in the level set and pruned regions are out of the level set. To incorporate the potential maximum volume error ϵ of incorrect maintaining and pruning at iteration k, we provide bounds in (6.7) such that $y\left(\delta_{kl}, \widetilde{\Sigma}_k^C\right) \leq y(\delta, S) \leq y\left(\delta_{ku}, \widetilde{\Sigma}_k^C\right)$. Since samples are independent and uniformly distributed in the current set $\widetilde{\Sigma}_k^C$, each sample acts like a Bernoulli trial and falls in a δ_{kl} or δ_{ku} level set with δ_{kl} or δ_{ku} probability, respectively. Using properties of a binomial distribution (see Conover 1999), we can build a $1-\alpha_k$ quantile confidence interval as $f(z_{(r)}) \leq y(\delta, S) \leq f(z_{(s)})$ with $y\left(\delta_{kl}, \widetilde{\Sigma}_k^C\right)$ and $y\left(\delta_{ku}, \widetilde{\Sigma}_k^C\right)$, where $f(z_{(r)})$ and $f(z_{(s)})$ are the rth and sth order samples, yielding

$$P\left(f(z^k_{(r)}) \le y\left(\delta_{kl}, \widetilde{\Sigma}^C_k\right) \le y(\delta, S) \le y\left(\delta_{ku}, \widetilde{\Sigma}^C_k\right) \le f(z^k_{(s)})\right) \ge 1 - \alpha_k. \tag{6.19}$$

□

Theorem 1 analyzes the impact of incorrect pruning and maintaining on shifting the δ_k quantile to correspond to the current set, $\widetilde{\Sigma}^C_k$. Theorem 1 assumes that ϵ^M_k is bounded by $\frac{\epsilon v(\widetilde{\Sigma}^M_k)}{v(S)}$, which we denote as event

$$A^M_k = \left\{ v\left(L(\delta, S) \cap \widetilde{\Sigma}^M_k\right) \ge v\left(\widetilde{\Sigma}^M_k\right) - \frac{\epsilon v\left(\widetilde{\Sigma}^M_k\right)}{v(S)} \right\}$$

and that ϵ^P_k is bounded by $\frac{\epsilon \widetilde{\Sigma}^P_k}{v(S)}$, which is represented by event

$$A^P_k = \left\{ v\left(L(\delta, S) \cap \widetilde{\Sigma}^P_k\right) \le \frac{\epsilon v\left(\widetilde{\Sigma}^P_k\right)}{v(S)} \right\}.$$

The event that $y\left(\delta_{kl}, \widetilde{\Sigma}^C_k\right)$ and $y\left(\delta_{ku}\widetilde{\Sigma}^C_k\right)$ are bounded correctly is denoted by $A^{CI}_k = \left\{ f(z^k_{(r)}) \le y(\delta_{kl}, \widetilde{\Sigma}^C_k) \le y\left(\delta_{ku}, \widetilde{\Sigma}^C_k\right) \le f(z^k_{(s)}) \right\}$. All three events, denoted

$$A_k = \left\{ A^M_k \cap A^P_k \cap A^{CI}_k \right\}, \tag{6.20}$$

ensure that $y(\delta, S)$ is bounded correctly on iteration k, i.e., $f(z^k_{(r)}) \le y\left(\delta_{kl}, \widetilde{\Sigma}^C_k\right) \le y(\delta, S) \le y\left(\delta_{ku}, \widetilde{\Sigma}^C_k\right) \le f(z^k_{(s)})$.

We next analyze the quality of pruned subregions and maintained subregions for a single iteration in Theorems 2 and 3 assuming event A_k. Probability bounds on interval estimations of the target quantile from iteration 1 to k are derived in Theorem 4. The final performances of the algorithm are derived in Theorems 5 and 6.

Theorem 2 *Consider any iteration k of PBnB on Problem ($\mathscr{P}$) where there is no noise, and suppose $\hat{\sigma}^k_p = \bigcup_{i \in w: \mathrm{P}_i = 1} \sigma_i$ has been pruned on the k^{th} iteration. Also, we condition on the event A_k in (6.20) being true. Then, the event that the volume of the incorrectly pruned region, i.e., $v(L(\delta, S) \cap \hat{\sigma}^k_p)$, is less than or equal to $D^P_k \epsilon_k$, where D^P_k is the number of subregions pruned at iteration k, with probability at least $1 - \alpha_k$, that is*

$$P\left(v\left(L(\delta, S) \cap \hat{\sigma}^k_p\right) \le D^P_k \epsilon_k | A_k\right) \ge 1 - \alpha_k. \tag{6.21}$$

Proof The full proof is available in the appendix, and here we provide a sketch of the proof. The proof uses the quantile definition to bound the probability that the pruned set does not incorrectly contain the target set up to a maximum error,

conditioned on the event A_k. The probability statement from Theorem 1 coupled with order statistics for the best sample in the subregion is used to further bound the desired probability. The sample size used in Step 4 is determined to achieve the desired $1-\alpha_k$ probability bound. □

Theorem 3 *Consider any iteration k of PBnB on Problem* ($\mathscr{P}$) *where there is no noise, and suppose* $\hat{\sigma}_m^k = \bigcup_{i\in w:\mathrm{M}_i=1} \sigma_i$ *has been maintained on the kth iteration. Also, suppose* A_k *in (6.20) is true. Then, the volume of the correctly maintained region is greater than or equal to* $v(\hat{\sigma}_m^k) - D_k^M \epsilon_k$, *where* D_k^M *is the number of subregions pruned at iteration k, with probability at least* $1-\alpha_k$, *or, in other words, the volume of the incorrectly maintained region is less than or equal to* $D_k^M \epsilon_k$ *with probability* $1-\alpha_k$,

$$P\left(v\left(\hat{\sigma}_m^k \setminus L(\delta, S)\right) \leq D_k^M \epsilon_k | A_k\right) \geq 1-\alpha_k. \tag{6.22}$$

Proof The proof is similar to Theorem 2.

Theorem 1 assumes $0 \leq \epsilon_k^M \leq \frac{\epsilon v(\widetilde{\Sigma}_k^M)}{v(S)}$ and $0 \leq \epsilon_k^P \leq \frac{\epsilon v(\widetilde{\Sigma}_k^P)}{v(S)}$, and Theorems 2 and 3 assume A_k. Now, in Theorem 4, we remove these conditions.

Theorem 4 *For any iteration* $k \geq 1$, *suppose there is no noise and use* δ_{kl} *and* δ_{ku} *from (6.7) to estimate target quantile* $y(\delta, S)$. *The probability that all interval estimates from iteration* 1 *to k capture the original quantile* $y(\delta, S)$ *is bounded as*

$$P\left(\cap_{i=1}^k \left\{f\left(z_{(r)}^i\right) \leq y(\delta, S) \leq f\left(z_{(s)}^i\right)\right\}\right) \geq (1-\alpha)^3 \tag{6.23}$$

where $z_{(r)}^i$ and $z_{(s)}^i$ are from Step 3, using (6.8) and (6.9) at iteration i.

Proof Without loss of generality, suppose that the algorithm builds a confidence interval every iteration, that is $k_b = 1$.

Because we estimate the target quantile using $y(\delta_{ku}, \widetilde{\Sigma}_k^C)$ and $y(\delta_{kl}, \widetilde{\Sigma}_k^C)$ as upper and lower bounds in Step 3, we consider a lower bound for a single estimation at iteration k as

$$\begin{aligned} &P\left(f\left(z_{(r)}^k\right) \leq y(\delta, S) \leq f\left(z_{(s)}^k\right)\right) \\ &\quad \geq P\left(\left\{f\left(z_{(r)}^k\right) \leq y\left(\delta_{kl}, \widetilde{\Sigma}_k^C\right) \leq y(\delta, S) \leq y\left(\delta_{ku}, \widetilde{\Sigma}_k^C\right) \leq f\left(z_{(s)}^k\right)\right\}\right) \end{aligned}$$

and when the conditions of Theorem 1 are satisfied, that is, the incorrect maintained volume is less than or equal to $\frac{\epsilon v(\widetilde{\Sigma}_k^M)}{v(S)}$, and the incorrect pruned volume is less than or equal to $\frac{\epsilon v(\widetilde{\Sigma}_k^P)}{v(S)}$, then $y\left(\delta_{kl}, \widetilde{\Sigma}_k^C\right) \leq y(\delta, S) \leq y\left(\delta_{ku}, \widetilde{\Sigma}_k^C\right)$, so we get

$$\geq P\left(\left\{v\left(L(\delta, S) \cap \widetilde{\Sigma}_k^M\right) \geq v\left(\widetilde{\Sigma}_k^M\right) - \frac{\epsilon v\left(\widetilde{\Sigma}_k^M\right)}{v(S)}\right\}\right.$$
$$\bigcap\left\{v\left(L(\delta, S) \cap \widetilde{\Sigma}_k^P\right) \leq \frac{\epsilon v\left(\widetilde{\Sigma}_k^P\right)}{v(S)}\right\} \bigcap\left\{f\left(z_{(r)}^k\right) \leq y\left(\delta_{kl}, \widetilde{\Sigma}_k^C\right)\right.$$
$$\left.\left.\leq y\left(\delta_{ku}, \widetilde{\Sigma}_k^C\right) \leq f\left(z_{(s)}^k\right)\right\}\right) = P\left(\left\{A_k^M \cap A_k^P \cap A_k^{CI}\right\}\right). \tag{6.24}$$

Now, working towards (6.23), we derive a lower bound on the probability that the quantile interval estimate captures the target quantile at every iteration from 1 to k, using (6.24), that is

$$P\left(\cap_{i=1}^k\left\{f\left(z_{(r)}^i\right) \leq y(\delta, S) \leq f\left(z_{(s)}^i\right)\right\}\right) \geq P\left(\cap_{i=1}^k\left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}\right). \tag{6.25}$$

We start proving (6.25) by using mathematical induction to show that

$$P(\cap_{i=1}^k\left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}) \geq \prod_{i=1}^{k-1}(1-\alpha_i)\prod_{i=1}^{k-1}(1-\alpha_i)\prod_{i=1}^{k}(1-\alpha_i). \tag{6.26}$$

For $k = 1$, $P(A_1^M \cap A_1^P \cap A_1^{CI}) = P(\{A_1^M \cap A_1^P | A_1^{CI}\})P(A_1^{CI}) \geq 1 \cdot (1-\alpha_1) \geq (1-\alpha_1)$, where $P\left(A_1^{CI}\right) > (1-\alpha_1)$ because the bounds are built with (6.8) and (6.9), and $P\left(\{A_1^M \cap A_1^P | A_1^{CI}\}\right) = 1$ since $\widetilde{\Sigma}_1^M$ and $\widetilde{\Sigma}_1^P$ are empty set.

Suppose $k = j$ holds, we have

$$P\left(\cap_{i=1}^j\left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}\right) \geq \prod_{i=1}^{j-1}(1-\alpha_i)\prod_{i=1}^{j-1}(1-\alpha_i)\prod_{i=1}^{j}(1-\alpha_i). \tag{6.27}$$

For $k = j+1$,

$$P\left(\cap_{i=1}^{j+1}\left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}\right) = P\left(\left\{A_{j+1}^M \cap A_{j+1}^P \cap A_{j+1}^{CI}\right\}\right.$$
$$\left.\cdot\left|\cap_{i=1}^j\left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}\right.\right) P\left(\cap_{i=1}^j\left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}\right)$$
$$= P\left(\left\{A_{j+1}^M \cap A_{j+1}^P\right\}\left|\cap_{i=1}^j\left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\} \cap A_{j+1}^{CI}\right.\right)$$
$$\cdot P\left(A_{j+1}^{CI}|\cap_{i=1}^j\left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}\right) P\left(\cap_{i=1}^j\left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}\right).$$

Note, A_{j+1}^{CI} is $\left\{f\left(z_{(r)}^{j+1}\right) \leq y\left(\delta_{j+1l}, \widetilde{\Sigma}_{j+1}^C\right) \leq y\left(\delta_{j+1u}, \widetilde{\Sigma}_{j+1}^C\right) \leq f\left(z_{(s)}^{j+1}\right)\right\}$, which is selecting bounds for $y\left(\delta_{j+1l}, \widetilde{\Sigma}_{j+1}^C\right)$ and $y\left(\delta_{j+1u}, \widetilde{\Sigma}_{j+1}^C\right)$ with a

predetermined δ_{j+1l} and δ_{j+1u}. Therefore, all A_i^M and A_i^P, for $i \leq j$, only affect the development of δ_{j+1l} and δ_{j+1u} but do not have any impact on A_{j+1}^{CI}. Furthermore, since the $j+1$ iteration reuses samples from previous iterations, we have

$$P\left(A_{j+1}^{CI} | \cap_{i=1}^{j} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}\right) > P\left(A_{j+1}^{CI}\right).$$

Therefore,

$$\begin{aligned} & P\left(\cap_{i=1}^{j+1} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}\right) \\ & \geq P\left(\left\{A_{j+1}^M \cap A_{j+1}^P\right\} | \cap_{i=1}^{j} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\} \cap A_{j+1}^{CI}\right) \\ & \quad \cdot P\left(A_{j+1}^{CI}\right) P\left(\cap_{i=1}^{j} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}\right), \end{aligned}$$

and since the elite and worse subregions are mutually exclusive and the samples used for pruning and maintaining are sampled from separate subregions, A_{j+1}^M and A_{j+1}^P are independent, thus we have

$$\begin{aligned} & \geq P\left(A_{j+1}^M | \cap_{i=1}^{j} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\} \cap A_{j+1}^{CI}\right) \\ & \quad \cdot P\left(A_{j+1}^P | \cap_{i=1}^{j} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\} \cap A_{j+1}^{CI}\right) P\left(A_{j+1}^{CI}\right) \\ & \quad \cdot P\left(\cap_{i=1}^{j} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}\right) \end{aligned}$$

and by separating the subregions that have been maintained and pruned on iteration $j+1$ from previous iterations 1 to j $\left(\widetilde{\Sigma}_{j+1}^M = \hat{\sigma}_m^j \cap \widetilde{\Sigma}_j^M \text{ and} \widetilde{\Sigma}_{j+1}^P = \hat{\sigma}_p^j \cap \widetilde{\Sigma}_j^P\right)$ and since the probabilities are conditioned on A_j^M and A_j^P, we get a lower bound by considering the volumes of incorrect pruning and maintaining only on the last iteration,

$$\begin{aligned} & \geq P\left(v(\hat{\sigma}_m^j \setminus L(\delta, S)) \leq D_j^M \epsilon_j | \cap_{i=1}^{j} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\} \cap A_{j+1}^{CI}\right) \\ & \quad \cdot P\left(v(L(\delta, S) \cap \hat{\sigma}_p^j) \leq D_j^P \epsilon_j | \cap_{i=1}^{j} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\} \cap A_{j+1}^{CI}\right) \\ & \quad \cdot P\left(A_{j+1}^{CI}\right) P\left(\cap_{i=1}^{j} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}\right), \end{aligned}$$

and since removing the conditioning on the event A_{j+1}^{CI} and on the events for prior iterations, $i < j$, $\left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}$, can allow for more incorrect pruning and maintaining on the current iteration, we have

$$\geq P\left(v(\hat{\sigma}_m^j \setminus L(\delta, S)) \leq D_j^M \epsilon_j \middle| \left\{A_j^M \cap A_j^P \cap A_j^{CI}\right\}\right)$$
$$\cdot P\left(v(L(\delta, S) \cap \hat{\sigma}_p^j) \leq D_j^P \epsilon_j \middle| \left\{A_j^M \cap A_j^P \cap A_j^{CI}\right\}\right)$$
$$\cdot P\left(A_{j+1}^{CI}\right) P\left(\cap_{i=1}^j \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}\right)$$

and using Theorems 2 and 3, (6.19) and (6.27), we have

$$\geq (1-\alpha_j)(1-\alpha_j)(1-\alpha_{j+1}) \prod_{i=1}^{j-1}(1-\alpha_i) \prod_{i=1}^{j-1}(1-\alpha_i) \prod_{i=1}^{j}(1-\alpha_i)$$
$$= \prod_{i=1}^{j}(1-\alpha_i) \prod_{i=1}^{j}(1-\alpha_i) \prod_{i=1}^{j+1}(1-\alpha_i).$$

Hence, (6.27) holds for $k = j+1$, and by mathematical induction (6.26) holds.

Using (6.26) to lower bound (6.25) yields

$$P\left(\cap_{i=1}^k f\left(z_{(r)}^i\right) \leq y(\delta, S) \leq f\left(z_{(s)}^i\right)\right) \geq \prod_{i=1}^{k-1}(1-\alpha_i) \prod_{i=1}^{k-1}(1-\alpha_i) \prod_{i=1}^{k}(1-\alpha_i)$$
$$= (1-\alpha_k) \prod_{i=1}^{k-1}(1-\alpha_i)^3$$

and in order to establish a pattern, and since $(1-\alpha_k) < 1$, we write

$$\geq (1-\alpha_k)^B (1-\alpha_k)^B (1-\alpha_k)^B \prod_{i=1}^{k-1}(1-\alpha_i)^3$$

and by applying Bernoulli's inequality, and since $\alpha_k = \frac{\alpha_{k-1}}{B}$, we get

$$\geq (1-\alpha_{k-1})(1-\alpha_{k-1})(1-\alpha_{k-1}) \prod_{i=1}^{k-1}(1-\alpha_i)^3$$
$$= (1-\alpha_{k-1})^6 \prod_{i=1}^{k-2}(1-\alpha_i)^3$$

and now, since $B \geq 2$, the pattern starts to repeat as

$$\geq (1-\alpha_{k-1})^B (1-\alpha_{k-1})^B (1-\alpha_{k-1})^B \prod_{i=1}^{k-2} (1-\alpha_i)^3$$

and by repeatedly applying Bernoulli's inequality,

$$\geq (1-\alpha)^3. \tag{6.28}$$

□

Theorem 5 *For any iteration $k \geq 0$ of PBnB on $(\mathscr{P})$ when there is no noise, the volume of the incorrectly pruned region is at most ϵ with probability at least $(1-\alpha)^4$, that is*

$$P\left(v\left(L(\delta, S) \cap \widetilde{\Sigma}_{k+1}^P\right) \leq \epsilon\right) \geq (1-\alpha)^4. \tag{6.29}$$

Proof First consider the conditional probability of the incorrectly pruned volume, given all confidence intervals capture the true quantile by A_i for all $i = 1, \ldots, k$, and by the definition of $\widetilde{\Sigma}_{k+1}^P$, $v\left(L(\delta, S) \cap \widetilde{\Sigma}_{k+1}^P\right) = v\left(\cup_{i=1}^k L(\delta, S) \cap \hat{\sigma}_p^i\right)$, we have

$$\begin{aligned} &P\left(v\left(L(\delta, S) \cap \widetilde{\Sigma}_{k+1}^P\right) \leq \epsilon \,\middle|\, \cap_{i=1}^k A_i\right), \\ &= P\left(v\left(\cup_{i=1}^k L(\delta, S) \cap \hat{\sigma}_p^i\right) \leq \epsilon \,\middle|\, \cap_{i=1}^k A_i\right) \end{aligned} \tag{6.30}$$

and at each iteration, at D_i^P subregions are pruned, yielding

$$\geq P\left(\cap_{i=1}^k \{v\left(L(\delta, S) \cap \hat{\sigma}_p^i\right) \leq D_i^P \epsilon_i\} \,\middle|\, \cap_{l=1}^k A_i\right)$$

considering a lower bound that every subregion is pruned with probability bounds as given in Theorem 2,

$$\geq \prod_{i=1}^k (1-\alpha_i) = \prod_{i=1}^k \left(1 - \frac{\alpha}{B^i}\right) \geq \left(1 - \frac{\alpha}{B^k}\right)^B \prod_{i=1}^{k-1} \left(1 - \frac{\alpha}{B^i}\right),$$

and by applying Bernoulli's inequality, $\left(1 - \frac{\alpha}{B^k}\right)^B \geq \left(1 - B\frac{\alpha}{B^k}\right)$, repeatedly

$$\geq \left(1 - \frac{\alpha}{B}\right)^B \geq 1 - \alpha. \tag{6.31}$$

Combining (6.23) from Theorem 4 with (6.31), we get

$$P\left(v\left(L(\delta, S) \cap \widetilde{\Sigma}_{k+1}^{P}\right) \leq \epsilon\right) \geq (1-\alpha)^4. \tag{6.32}$$

□

Theorem 6 *For any iteration $k \geq 0$ of PBnB on ($\mathscr{P}$) when there is no noise, the volume of the incorrectly maintained region is at most ϵ with probability at least $(1-\alpha)^4$, that is*

$$P\left(v\left(\widetilde{\Sigma}_{k+1}^{M} \setminus L(\delta, S)\right) \leq \epsilon\right) \geq (1-\alpha)^4. \tag{6.33}$$

Proof The proof is similar to Theorem 5. □

6.5.2 Analysis on a Noisy Function

In the previous analysis, we assume the objective function $f(x)$ can be evaluated exactly, without noise, and in this section we account for noise in the estimated function $\hat{f}(x)$. Theorem 7 provides probability bounds for correctly ordering the estimated function values. Theorem 8 combines all iterations from 1 to k and gives a probability bound of $1-\alpha$ on the correct ordering. Corollaries 1, 2, and 3 provide noisy versions of Theorems 4, 5, and 6.

We use the analysis of a two-stage replication approach, by Bechhofer et al. (1954), in Theorem 7, which has following assumption:

(A1) The function $g(x, \xi_x)$ itself is a random variable, due to the random variable ξ_x, and we assume $g(x, \xi_x)$ is normally distributed with an unknown common variance σ^2, and at each solution $z_i \in S$, the variance can be expressed as $a_i\sigma^2$ where a_i is a known constant for each i.

In PBnB, the constants a_i are not known, hence, we cannot exactly implement their two-stage replication approach. Our modified two-stage replication approach is described after Theorem 7.

Theorem 7 (cf. Bechhofer et al. 1954) *With Assumption (A1), the probability of correctly ordering all samples in the current region at iteration k is*

$$P\Big(\hat{f}\left(z_{(1)}^k\right) \leq \hat{f}\left(z_{(2)}^k\right) \leq \cdots \leq \hat{f}\left(z_{(N_l)}^k\right) | f\left(z_{(1)}^k\right) \leq f\left(z_{(2)}^k\right) \leq \cdots \leq f\left(z_{(N_k)}^k\right)\Big) \geq 1-\alpha_k, \tag{6.34}$$

given that we have $a_i R_0$ as the first stage replication number for each sample point to estimate the common variance by $S_0^2 = \frac{1}{N_k}\sum_{i=1}^{N_k} \frac{1}{a_i} S_{\hat{f}}^2(z_{(i)})$ and set up the minimum difference desired to be separated as d^*, and sampling $R_k =$

$\max\{a_i R_0, 2(\frac{hS_0^*}{d^*/2})^2\}$ *in the second stage of the procedure, where* h *is a value in the* H *c.d.f. of a multivariate student t's distribution with* $N_k - 1$ *dimensions such that* $H(h) = 1 - \alpha_k$.

Proof See Bechhofer et al. (1954). □

For any iteration k, we take a conservative approach to achieve the $1 - \alpha_k$ probability of correct ordering by separating each estimated performance mean $\hat{f}(z_{(i)})$ with its neighbor by the smallest difference $d^* = \min_{i=1,\ldots,N_k-1} d_i$ of any two neighbors. We also use the largest variance $S^{*2} = \max_{i=1,\ldots,N_k} S^2_{\hat{f}}(z_{(i)})$ so that all ordering is conservative. Specifically, the implemented two-stage procedure for any iteration k is in (2.A) and (2.B) in Steps 2 and 4. For the implemented two-stage procedure, we drop the assumption of a common variance with $a_i\sigma^2$ and only assume the noise is normally distributed. In Step 2 of the implemented two-stage procedure we use $z_{\alpha_k/2}$, the $1 - \alpha_k/2$ quantile of the standard normal distribution, together with d^* and S^* so that the correct ordering of two function values separated by d^* with S^{*2} variance is achieved with probability at least $1 - \alpha_k$.

The following Theorem 8 considers Assumption (A1) and Theorem 7 to derive a bound on the probability of correctly ordering the estimated function values for iteration 1 to k.

Theorem 8 *With Assumption (A1), the probability of correct ordering from iteration* 1 *to iteration* k *is*

$$P\left(\bigcap_{l=1}^{k}\left(\hat{f}\left(z^l_{(1)}\right) \le \hat{f}\left(z^l_{(2)}\right) \le \cdots \le \hat{f}\left(z^l_{(N_l)}\right)\right)\right) \ge 1 - \alpha, \tag{6.35}$$

where $z^l_{(j)}$ *is the* j^{th} *ordered sampled point at iteration* l.

Proof The probability of correct ordering from iteration 1 to iteration k can be expressed using conditional probabilities as

$$P\Bigg(\bigcap_{l=1}^{k}\left(\hat{f}\left(z^l_{(1)}\right) \le \hat{f}\left(z^l_{(2)}\right) \le \cdots \le \hat{f}\left(z^l_{(N_l)}\right) | f\left(z^l_{(1)}\right) \le f\left(z^l_{(2)}\right) \le \cdots \le f\left(z^l_{(N_l)}\right)\right)\Bigg)$$

where R_k is non-decreasing and R_k is chosen so that Theorem 7 is satisfied, hence the probability the ordering is correct on iteration l given that the ordering was correct on the previous iterations is greater than or equal to the unconditioned probability that the ordering is correct on iteration l, therefore,

$$\ge \prod_{l=1}^{k} P\left(\hat{f}\left(z^l_{(1)}\right) \le \hat{f}\left(z^l_{(2)}\right) \le \cdots \le \hat{f}\left(z^l_{(N_l)}\right) | f\left(z^l_{(1)}\right)\right.$$

$$\leq f\left(z^l_{(2)}\right) \leq \cdots \leq f\left(z^l_{(N_l)}\right)\Big)$$

and by Theorem 7 and applying Bernoulli's inequality repeatedly

$$\geq \prod_{l=1}^{k}\left(1-\frac{\alpha}{B^l}\right) \geq 1-\alpha. \tag{6.36}$$

□

With the probability bound of correct ordering, we next derive the probability that the sequence of interval estimation is correct in Corollary 1. In Corollaries 2 and 3, we derive versions of Theorem 5 and Theorem 6 with noise.

Corollary 1 *Consider any iteration $k \geq 1$ of PBnB on $(\mathscr{P})$ where (A1) is assumed. The probability that all one-sided interval estimates from iteration 1 to k capture the original quantile $y(\delta, S)$ is bounded by*

$$P\left(\cap_{i=1}^{k} f\left(z^i_{(r)}\right) \leq y(\delta, S) \leq f\left(z^i_{(s)}\right)\right) \geq (1-\alpha)^4 \tag{6.37}$$

where $z^i_{(s)}$ and $z^i_{(r)}$ are selected as in Step 3 for iteration $i = 1, \ldots, k$.

Proof Theorem 4 considers the probability of a sequence of interval estimations until iteration k with no noise in the objective function. Theorem 4 holds under the condition that each iteration's ordering of samples is correct, however when the objective function is noisy, the order of samples may be incorrect. Therefore, the probability bounds for a noisy function should include the probability of correct ordering, as in Theorem 8, which requires including a $(1-\alpha)$ probability term in the original bound. □

Corollary 2 *Consider any iteration $k \geq 1$ of PBnB on $(\mathscr{P})$ where (A1) is assumed. The probability of incorrectly pruning a volume of at most ϵ is bounded by*

$$P\left(v\left(L(\delta, S) \cap \widetilde{\Sigma}^P_{k+1}\right) \leq \epsilon\right) \geq (1-\alpha)^5. \tag{6.38}$$

Corollary 3 *Consider any iteration $k \geq 1$ of PBnB on $(\mathscr{P})$ where (A1) is assumed. The probability of incorrectly maintaining a volume of at most ϵ is bounded by*

$$P\left(v\left(\widetilde{\Sigma}^M_{k+1} \setminus L(\delta, S)\right) \leq \epsilon\right) \geq (1-\alpha)^5. \tag{6.39}$$

6.6 Numerical Results

In this section, we test PBnB for level set approximation with several test functions: sphere function, Rosenbrock's function, the centered sinusoidal function, and the shifted sinusoidal function, defined as follows.

- Sphere function $(-10 \le x_i \le 10, i = 1, \ldots, n)$

$$g_0(x) = \sum_{i=1}^{n} x_i^2. \tag{6.40}$$

  The global optimum is located at $x_* = (0, \ldots, 0)$ with $g_1(x_*) = 0$.
- Rosenbrock's function $(-2 \le x_i \le 2, i = 1, \ldots, n)$

$$g_1(x) = \sum_{i=1}^{n-1} \left[(1 - x_i)^2 + 100 \left(x_{i+1} - x_i^2 \right)^2 \right]. \tag{6.41}$$

  The global optimum is located at $x_* = (1, \ldots, 1)$ with $g_1(x_*) = 0$.
- Centered Sinusoidal function $(0 \le x_i \le 180, i = 1, \ldots, n)$

$$g_3(x) = -2.5 \prod_{i=1}^{n} \sin\left(\frac{\pi x_i}{180}\right) - \prod_{i=1}^{n} \sin\left(\frac{\pi x_i}{36}\right). \tag{6.42}$$

  The global optimum is located at $x_* = (90, \ldots, 90)$ with $g_2(x_*) = -3.5$.
- Shifted sinusoidal function $(0 \le x_i \le 180, i = 1, \ldots, n)$

$$g_3(x) = -2.5 \prod_{i=1}^{n} \sin\left(\frac{\pi (x_i + 60)}{180}\right) - \prod_{i=1}^{n} \sin\left(\frac{\pi (x_i + 60)}{36}\right). \tag{6.43}$$

  The global optimum is located at $x_* = (30, \ldots, 30)$ with $g_4(x_*) = -3.5$.

First consider the sphere function. We numerically evaluate the quality of the level set approximation and compare with theoretical results in Sect. 6.6.1. In Sect. 6.6.2, test functions (6.41)–(6.43) are tested in the "no noise" setting and the performance is evaluated for 2-, 3-, 5-, 7-, and 10-dimensional problems. Section 6.6.3 focuses on the influence of a normal noise added to the test functions. In Sect. 6.6.4, we demonstrate how the algorithm performs with integer variables by discretizing the sinusoidal function (as in Ali et al. 2005). The parameters of the algorithm are set as follows, $\delta = 0.1$, $B = 2$, $\alpha = 0.05$, $\epsilon = 0.025 v(S)$, $k_b = 2$, $c_k = 1000$, and $R_o = 20$ when the function is noisy. A subregion is unbranchable when the length of its longest side is less than four percent of the length of the domain's side. Also, we apply an upper bound of the sample size for each subregion by the density of sample points as $\frac{N_k^i}{v(\sigma_i)} \le \frac{100^n}{v(S)}$.

Table 6.1 Solution quality for noisy sphere function with 100 runs

n	Number of runs incorrect maintain > 0	Number of runs incorrect maintain $> \epsilon$	Number of runs incorrect prune > 0	Number of runs incorrect prune $> \epsilon$
2	0	0	5	0
3	0	0	0	0

6.6.1 Sphere Function with Normal Noise

The true level set of a sphere function can be easily calculated for two- and three-dimensional problems and used to compare the numerical performance with the theoretical analysis of Theorems 5, 6, Corollaries 2 and 3. The algorithm was run 100 times under no noise condition, and no run incorrectly pruned or maintained a region of volume larger than ϵ. In fact, the volume of incorrectly pruned or maintained for all 100 runs was zero—the result was perfect. Therefore, we focus on the function with $N(0, 1)$ noise. Table 6.1 lists the number of runs (out of 100) that had incorrectly maintained and pruned volumes greater than 0 and ϵ. For all 100 runs of PBnB for level set approximation on the sphere function with $N(0, 1)$ noise, no run's incorrect maintained volume exceeded the user-defined parameter, $\epsilon = 0.025v(S)$. For $n = 2$, only 5 runs have an incorrectly pruned volume greater than 0. The theoretical probability bounds of $(1 - \alpha)^5$ would allow an average of 23 runs out of 100 that could be incorrect by an ϵ amount. Since we observed zero runs that were incorrect by an ϵ amount, this suggests that the sample size used in the algorithm is conservative and the bounds in Corollaries 2 and 3 are not tight.

6.6.2 Continuous Test Functions with No Noise

In this section, we illustrate the pruning and maintaining subregions in Fig. 6.4 for the 10th iteration of PBnB on a single run on the Rosenbrock's function, the centered sinusoidal function, and the shifted sinusoidal function in two dimensions without adding noise. In Fig. 6.4, the dark gray (blue) boxes are the current undecided subregions, the white boxes represent the pruned subregions, the light gray (green) boxes are the maintained subregions, and the bold line represents the target level set (as in Fig. 6.2). The maintained subregions are clearly contained in the target level set, and the current subregions form the boundary of the level set. Practically speaking, there is a chance that some pruned subregions contain part of the target level set. However, the algorithm ensures that the volume is bounded by the user-defined parameter ϵ with probability bounds in Corollaries 2 and 3. From the two-dimensional problems, we can observe that the interaction of the test function's level set and the partition scheme affects the volume maintained and confirmed as a part of the level set at iteration 10 in Fig. 6.4.

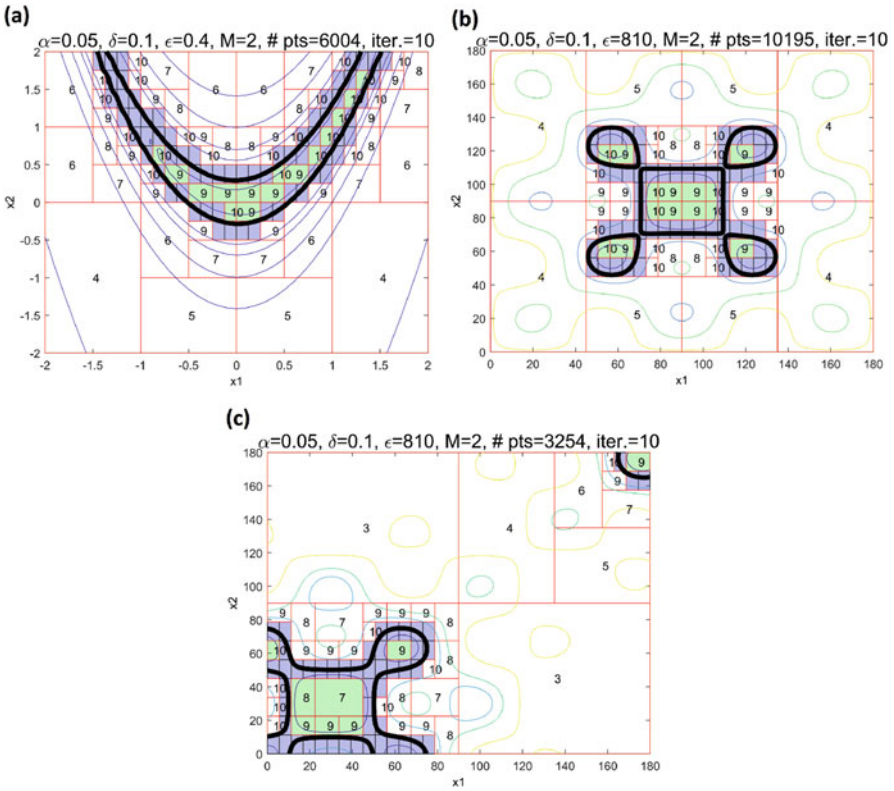


Fig. 6.4 Approximating the level set bounded by the 0.1 quantile on the tenth iteration of PBnB for two-dimensional (**a**) Rosenbrock's function, (**b**) centered sinusoidal function, and (**c**) shifted sinusoidal function by the maintained green (light gray) subregions

Figure 6.5 illustrates the updates of interval quantile estimations for the two-dimensional Rosenbrock's function, centered sinusoidal function, and shifted sinusoidal function. This interval estimation, $[\hat{f}(z_{((r))}), \hat{f}(z_{((s))})]$, represents the performance bound of the target level set as a reference for decision makers. The estimation updates for Rosenbrock's functions narrow the interval width to help pruning and maintaining. For the shifted sinusoidal function, there are subregions pruned or maintained for every iteration from $k = 3$. Hence, the quantile estimation does not update with new estimations. The centered sinusoidal function's interval estimation shrinks at the first update but loosens at the second update because more pruned and maintained subregions widen the δ_{ku} and δ_{kl}. Currently, c_k increases linearly with iterations. However, for difficult functions, a non-linear increasing of c_k may provide tighter confidence intervals considering the use of the δ_{ku} and δ_{kl}.

Figure 6.6 demonstrates the relative volume of solutions pruned, maintained, and undecided for iteration 1 to 10. In two dimensions, almost the entire target level set is captured as illustrated in Fig. 6.4. Three- and five-dimensional test functions,

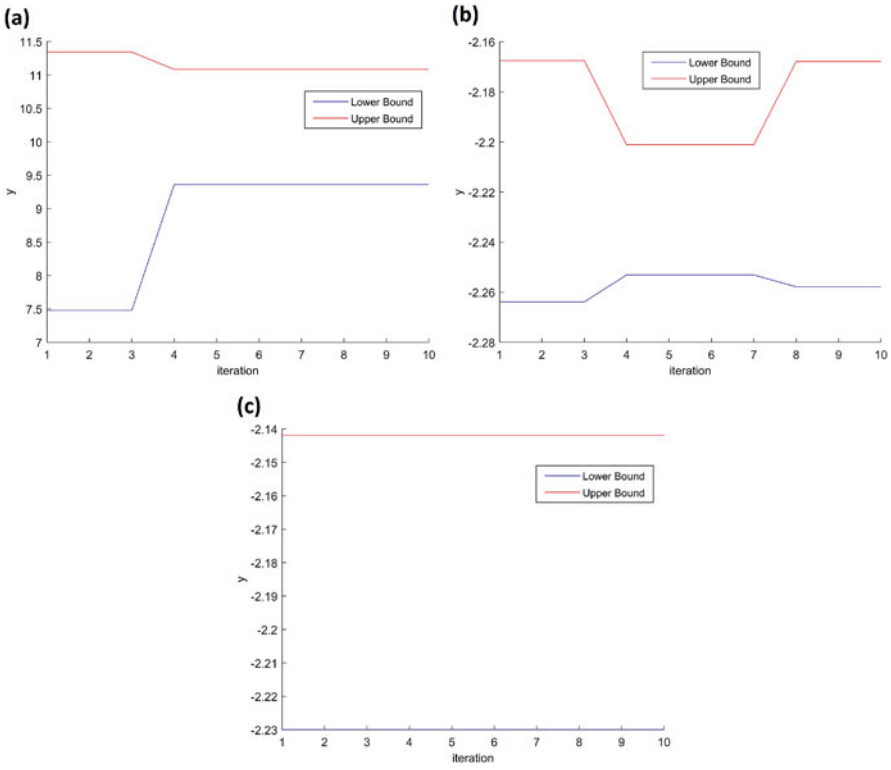


Fig. 6.5 Volume maintained and pruned by PBnB approximating the level set bounded by the 0.1 quantile for two-dimensional (**a**) Rosenbrock's function, (**b**) centered sinusoidal function, and (**c**) shifted sinusoidal function

$n = 3, 5$, are also used to test the algorithm, and the summary results are shown in Table 6.2. For all test functions, the algorithm still approximates the target 10% quantile level set and achieves between 4 and 6% maintained volume before terminating. Also, we can observe that the shifted sinusoidal function requires fewer samples for 2- and 3- dimensional problems. It is possible that the structure of the function allows the algorithm to prune a large portion of the solution space earlier than other two functions, as shown in Fig. 6.6.

Another approach to approximate a level set is to perform a grid search, where all grid points in the domain are evaluated. To achieve an approximation to a level of significance comparable to that of the numerical experiments with PBnB, we would divide each dimension into 100 points. Then the number of grid points to be evaluated is 100^n, where n is the number of decision variables. In the two-dimensional examples in Fig. 6.4, the numbers of function evaluations by PBnB are roughly between 10^3 and 10^4, which is comparable to a grid search with $100^2 = 10^4$ function evaluations. However, as dimension increases, we see that PBnB uses much fewer function evaluations than a grid search. In Table 6.3, the number of function

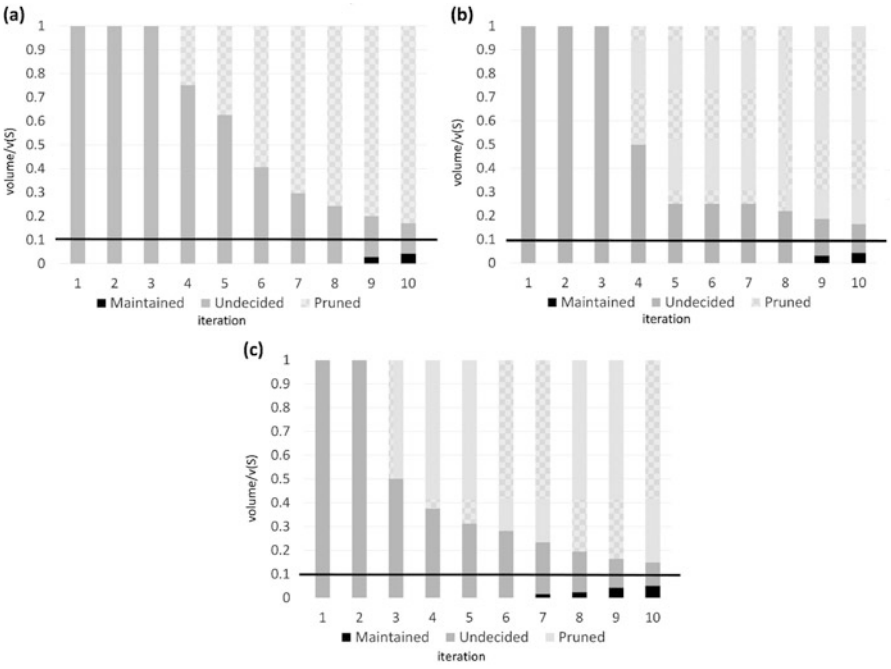


Fig. 6.6 Volume maintained and pruned by PBnB approximating the level set bounded by the 0.1 quantile for two-dimensional (**a**) Rosenbrock's function, (**b**) centered sinusoidal function, and (**c**) shifted sinusoidal function

Table 6.2 Comparison of Rosenbrock, centered and shifted sinusoidal functions with number of samples and ratio of volume maintained (VM)

Test function	Rosenbrock		Centered Sinusoidal		Shifted Sinusoidal	
Dimension	# sampled	Ratio of VM	# sampled	Ratio of VM	# sampled	Ratio of VM
$n = 2$	6004	4.30%	10,195	4.29%	3254	4.49%
$n = 3$	159,655	5.83%	212,563	4.87%	104,825	5.38%
$n = 5$	71,529,641	5.77%	126,702,225	4.14%	130,595,965	4.67%

evaluations for $n = 10$ is between 10^6 and 10^8, whereas a grid search would require $100^{10} = 10^{20}$ function evaluations. The advantage of PBnB is that it focuses on where to sample, thus requiring fewer function evaluations than enumeration as in grid search.

In higher dimensional problems ($n = 7$ and 10), approximating the entire target level set may be computationally expensive. As the dimension increases, the number of subregions increases significantly. PBnB for level set approximation allows early termination to capture a part of the target level set. As in Fig. 6.6, the volume of maintained subregions increases with more iterations and sample points. In Table 6.3, the number of iterations and sample points are shown for the first subregion maintained for each test dimensions, $n = 2, 3, 5, 7$, and 10. The result

Table 6.3 Comparing the iteration for first maintaining a subregion for Rosenbrock, centered and shifted sinusoidal functions, with number of samples and iteration of first maintained subregion

Test function	Rosenbrock		Centered Sinusoidal		Shifted Sinusoidal	
Dimension	# sampled	Iteration	# sampled	Iteration	# sampled	Iteration
$n = 2$	5707	9	11,009	9	1972	7
$n = 3$	25,531	10	33,315	10	3057	7
$n = 5$	101,880	11	473,366	13	10,970	8
$n = 7$	1,228,908	14	5,589,142	16	20,962	9
$n = 10$	92,448,129	20	229,773,645	21	108,073	11

indicates that PBnB can maintain the first subregion fairly quickly. Specifically, the centered sinusoidal function takes more iterations and sample points to first maintain a subregion. The shifted sinusoidal function consistently maintains a subregion before the other two test functions. It shows the speed of maintaining a subregion is related to the shape of the target level set of the function. Since maintaining and pruning subregions is only possible when some subregions are branched small enough to be fully in/out of the level set, the shape of the target level set (highly related to the non-convexity of the test functions) and the branching scheme are major factors impacting the algorithm's effectiveness.

6.6.3 Continuous Test Functions with Normal Noise

In order to illustrate the impact of noise, we apply $N(0, 1)$ noise to each test function. Figure 6.7 illustrates the pruned and maintained subregions in a two-dimensional solution space on the tenth iteration. The undetermined subregions for the noisy functions in Fig. 6.7 are larger than the non-noisy counterparts in Fig. 6.4. Although the shifted function performed well in the non-noisy condition, PBnB for level set approximation only captures a small part of the shifted sinusoidal function's target level set with noise, because $N(0, 1)$ noise is relatively large for the function value of the sinusoidal function. However, it is possible to maintain more subregions with further iterations that have smaller partitioned subregions and larger number of sample points.

6.6.4 Integer Test Functions

The PBnB algorithm for level set approximation can handle both integer and real-valued decision variables. The partitioning scheme on a discrete set must be adapted so that any discrete point belongs to only one subregion. The experiment in this section discretized the two-dimensional centered sinusoidal function at two levels of discretization. The centered sinusoidal function is motivated by an

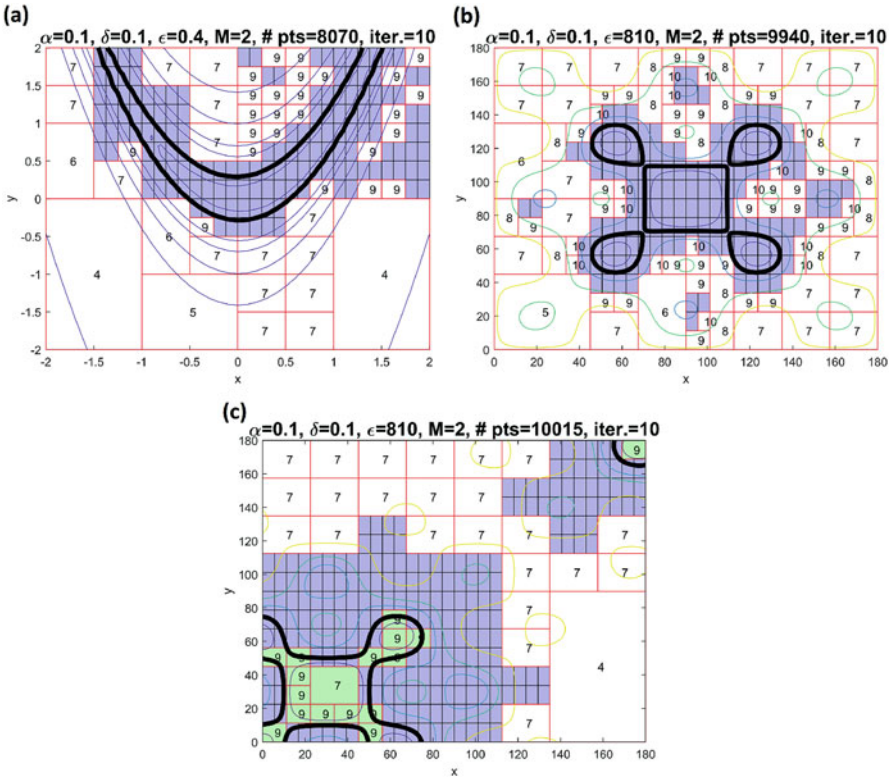


Fig. 6.7 Approximating the 0.1 level set on the tenth iteration of PBnB for two-dimensional (**a**) Rosenbrock function, (**b**) centered sinusoidal function, and (**c**) shifted sinusoidal function with N(0,1) noise by the maintained green (light gray) subregions

optimal composite structure design problem (Zabinsky 1998; Ali et al. 2005) that is interested in different levels of discretization. Specifically, we consider two discretization scenarios; the difference between points is 5 and 10 (as shown in Fig. 6.8). For the discretized sinusoidal function, we perform the same PBnB for level set approximation with a modified definition of how to branch the subregions in order to ensure that any discrete point belongs to exactly one subregion. The subregion in this discrete setting is $(l, u]$, where l and u are the lower and upper points of the box. Based on this definition, the results for the two-dimensional centered sinusoidal function are shown in Fig. 6.8. Although the maintained subregions in Fig. 6.8 seem to contain some regions outside of the level set, the discrete points in the maintained subregions actually are contained in the level set. Therefore, the PBnB algorithm is still viable for discretized problems. However, the partitioning scheme needs to be designed carefully to ensure one discrete point belongs to one subregion.

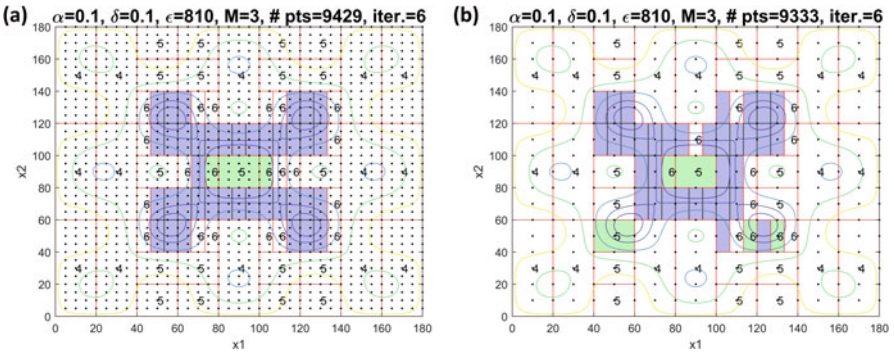


Fig. 6.8 Results of discretized sinusoidal function on the sixth iteration of PBnB with (**a**) 5 differences between points and (**b**) 10 differences between points.

6.7 Conclusions

We developed the PBnB algorithm to approximate a target level set for stochastic global optimization problems. A set of solutions allows decision makers to gain some understanding of their problem. The algorithm is capable of handling problems with mixed continuous and discrete variables. PBnB for level set approximation iteratively partitions the solution space into subregions, and prunes or maintains subregions based on the interval estimation of the target quantile. With the dynamically allocated computational resources, the algorithm provides an interval estimation of the desired quantile $[\hat{f}(z_{((r)}), \hat{f}(z_{((s)})]$, a pruned set $\widetilde{\Sigma}^P$ and a maintained set $\widetilde{\Sigma}^M$ with probability bounds and tolerable loss. The analyses of the probability bounds are provided in Sect. 6.4, where Theorems 5 and 6 assume the objective function can be evaluated exactly (no noise), and Corollaries 2 and 3 consider estimation of the objective function when it includes normal noise. The numerical results demonstrate the capability of PBnB to approximate a target level set, and the required sample size is less than grid search especially when the dimension increases since the samples are more concentrated around the target level set instead of spread out among the domain.

Currently, PBnB uses a very conservative two-stage approach to incorporate the simulation replications into the theoretical analysis (as in Theorem 8). The efficiency of PBnB has the potential to be improved by implementing existing computational budget allocation methods, such as optimal computing budget allocation (OCBA) (Chen et al. 2010; Chen and Lee 2011). A challenge is how to adapt OCBA when budget allocation is required to achieve more than one goal. For instance, in each subregion, it is important to identify the minimum sample, the maximum sample, and the rth and sth merged samples correctly at the same time. Future research in combining OCBA and PBnB may provide an improvement to the efficiency of PBnB, while preserving the theoretical analysis.

The algorithm can be further improved by adapting parameters to increase the efficiency. For example, the sample size c_k is kept constant, however, we are investigating a dynamic update for c_k during the course of the algorithm. Also, an intelligent partitioning scheme might be introduced to help PBnB branch the dimension that has more potential to maintain or prune partitioned subregions. Applying a more intelligent approach to determine the replication number could also benefit the algorithm for noisy functions.

Acknowledgements This work has been funded in part by the Department of Laboratory Medicine at Seattle Children's Hospital, and by National Science Foundation (NSF) grants CMMI-1235484 and CMMI-1632793.

Appendix

Proof of Theorem 1

Proof We consider the iterative effect on δ_k as subregions are pruned or maintained. We use the superscript k to denote the iteration that subregions are pruned $\{\sigma_i^k : P_i = 1\}$ or maintained $\{\sigma_i^k : M_i = 1\}$. By (6.15) in the algorithm, we have

$$\delta_k = \frac{\delta_{k-1} v\left(\widetilde{\Sigma}_{k-1}^C\right) - \sum_{i:M_i=1} v\left(\sigma_i^{k-1}\right)}{v\left(\widetilde{\Sigma}_{k-1}^C\right) - \sum_{i:P_i=1} v\left(\sigma_i^{k-1}\right) - \sum_{i:M_i^{k-1}=1} v\left(\sigma_i^{k-1}\right)}$$

and removing the pruned and maintained subregions from $\widetilde{\Sigma}_{k-1}^C$ yields the next current set of subregions $\widetilde{\Sigma}_k^C$, used in the denominator, then

$$= \frac{\delta_{k-1} v\left(\widetilde{\Sigma}_{k-1}^C\right) - \sum_{i:M_i=1} v\left(\sigma_i^{k-1}\right)}{v\left(\widetilde{\Sigma}_k^C\right)}$$

and invoking (6.15) in the algorithm again to replace δ_{k-1} with its equivalence in terms of δ_{k-2} (assuming that the maintained regions are in the level set and pruned regions are out of the level set), we have

$$= \frac{\frac{\delta_{k-2} v\left(\widetilde{\Sigma}_{k-2}^C\right) - \sum_{i:M_i=1} v\left(\sigma_i^{k-2}\right)}{v\left(\widetilde{\Sigma}_{k-1}^C\right)} v\left(\widetilde{\Sigma}_{k-1}^C\right) - \sum_{i:M_i=1} v\left(\sigma_i^{k-1}\right)}{v\left(\widetilde{\Sigma}_k^C\right)}$$

$$= \frac{\delta_{k-2} v\left(\widetilde{\Sigma}_{k-2}^{C}\right) - \sum_{l=k-2}^{k-1} \sum_{i:M_i=1} v\left(\sigma_i^l\right)}{v\left(\widetilde{\Sigma}_k^{C}\right)}$$

$$\vdots$$

$$= \frac{\delta_1 v\left(\widetilde{\Sigma}_1^{C}\right) - \sum_{l=1}^{k-1} \sum_{i:M_i=1} v\left(\sigma_i^l\right)}{v\left(\widetilde{\Sigma}_k^{C}\right)}$$

and by the initial setting of δ_1 and $\widetilde{\Sigma}_1^C$,

$$= \frac{\delta v(S) - \sum_{l=1}^{k-1} \sum_{i:M_i=1} v\left(\sigma_i^l\right)}{v\left(\widetilde{\Sigma}_k^{C}\right)}$$

and $\sum_{l=1}^{k-1} \sum_{i:M_i=1} v\left(\sigma_i^l\right) = v\left(\widetilde{\Sigma}_k^{M}\right)$ since it denotes the volume of all maintained subregions at the end of the $k-1$ iteration, therefore,

$$= \frac{\delta v(S) - v\left(\widetilde{\Sigma}_k^{M}\right)}{v\left(\widetilde{\Sigma}_k^{C}\right)}. \tag{6.44}$$

Based on the definition of quantile, and when X is uniformly sampled on S, we have

$$y(\delta, S) = \underset{y \in \{f(x): x \in S\}}{\arg\min} \{P(f(X) \le y | X \in S) \ge \delta\}$$

$$= \underset{y \in \{f(x): x \in S\}}{\arg\min} \left\{ \frac{v(\{x \in S : f(x) \le y\})}{v(S)} \ge \delta \right\}$$

and subtracting $\frac{v\left(\widetilde{\Sigma}_k^{M}\right) - \epsilon_k^M + \epsilon_k^P}{v(S)}$ from both sides and multiplying $\frac{v(S)}{v\left(\widetilde{\Sigma}_k^{C}\right)}$ on both sides,

$$= \underset{y \in \{f(x): x \in S\}}{\arg\min} \left\{ \frac{v\left(\{x \in S : f(x) \le y\}\right) - v\left(\widetilde{\Sigma}_k^{M}\right) + \epsilon_k^M - \epsilon_k^P}{v\left(\widetilde{\Sigma}_k^{C}\right)} \right.$$

$$\left. \ge \frac{\delta v(S) - v\left(\widetilde{\Sigma}_k^{M}\right) + \epsilon_k^M - \epsilon_k^P}{v\left(\widetilde{\Sigma}_k^{C}\right)} \right\}$$

and by (6.44), also $v\left(\{x \in \widetilde{\Sigma}_k^P : f(x) < y\}\right) = \epsilon_k^P$ and $v\left(\{x \in \widetilde{\Sigma}_k^M : f(x) < y\}\right) = v\left(\widetilde{\Sigma}_k^{M}\right) - \epsilon_k^M$,

$$= \underset{y\in\{f(x):x\in S\}}{\arg\min} \left\{ \frac{v\left(\left\{x \in S \setminus \left\{\widetilde{\Sigma}_k^P \cup \widetilde{\Sigma}_k^M\right\} : f(x) \le y\right\}\right)}{v\left(\widetilde{\Sigma}_k^C\right)} \right.$$
$$\left. \ge \delta_k + \frac{\epsilon_k^M}{v\left(\widetilde{\Sigma}_k^C\right)} - \frac{\epsilon_k^P}{v\left(\widetilde{\Sigma}_k^C\right)} \right\}$$

and since $\widetilde{\Sigma}_k^C = S\setminus\left\{\widetilde{\Sigma}_k^P \cup \widetilde{\Sigma}_k^M\right\}$, and X is uniformly distributed in $\widetilde{\Sigma}_k^C$ and $\widetilde{\Sigma}_k^C \subset S$,

$$= \underset{y\in\{f(x):x\in\widetilde{\Sigma}_k^C\}}{\arg\min} \left\{ P\left(f(X) \le y \middle| X \in \widetilde{\Sigma}_k^C \right) \ge \delta_k + \frac{\epsilon_k^M}{v\left(\widetilde{\Sigma}_k^C\right)} - \frac{\epsilon_k^P}{v\left(\widetilde{\Sigma}_k^C\right)} \right\}$$
$$= y\left(\delta_k + \frac{\epsilon_k^M}{v\left(\widetilde{\Sigma}_k^C\right)} - \frac{\epsilon_k^P}{v\left(\widetilde{\Sigma}_k^C\right)}, \widetilde{\Sigma}_k^C \right).$$

Since $0 \le \epsilon_k^M \le \frac{\epsilon \widetilde{\Sigma}_k^M}{v(S)}$ and $0 \le \epsilon_k^P \le \frac{\epsilon \widetilde{\Sigma}_k^P}{v(S)}$, an upper bound of $y\left(\delta_k + \frac{\epsilon_k^M}{v\left(\widetilde{\Sigma}_k^C\right)} - \frac{\epsilon_k^P}{v\left(\widetilde{\Sigma}_k^C\right)}, \widetilde{\Sigma}_k^C \right)$ can be achieved when $\epsilon_k^P = 0$ and $\epsilon_k^M = \frac{\epsilon v(\widetilde{\Sigma}_k^M)}{v(S)}$, yielding

$$y(\delta, S) \le y\left(\delta_k + \frac{\epsilon v\left(\widetilde{\Sigma}_k^M\right)}{v(S) v\left(\widetilde{\Sigma}_k^C\right)}, \widetilde{\Sigma}_k^C \right) = y\left(\delta_{ku}, \widetilde{\Sigma}_k^C\right). \tag{6.45}$$

Similarly, we have a lower bound when $\epsilon_k^M = 0$ and $\epsilon_k^P = \frac{\epsilon v(\widetilde{\Sigma}_k^P)}{v(S)}$, yielding

$$y(\delta, S) \ge y\left(\delta_k - \frac{\epsilon v\left(\widetilde{\Sigma}_k^P\right)}{v(S) v\left(\widetilde{\Sigma}_k^C\right)}, \widetilde{\Sigma}_k^C \right) = y\left(\delta_{kl}, \widetilde{\Sigma}_k^C\right). \tag{6.46}$$

Note, if $\epsilon_k^M = 0$ and $\epsilon_k^P = 0$, that is, there is no error in pruning and maintaining, then $y(\delta, S) = y\left(\delta_k, \widetilde{\Sigma}_k^C\right)$.

At the beginning of any iteration k, the current set $\widetilde{\Sigma}_k^C$ is uniformly sampled for $N_k = c_k$ samples. Since the samples are independent and uniformly distributed in the current set $\widetilde{\Sigma}_k^C$, each sample acts like a Bernoulli trial and falls in a δ_{kl} or δ_{ku} level set with δ_{kl} or δ_{ku} probability, respectively. Therefore, using properties of a binomial distribution, we can build a $1-\alpha_k$ quantile confidence interval as $f(z_{(r)}) \le y(\delta, S) \le f(z_{(s)})$ (Conover 1999) with $y\left(\delta_{kl}, \widetilde{\Sigma}_k^C\right)$ and $y\left(\delta_{ku}, \widetilde{\Sigma}_k^C\right)$ based on (6.45) and (6.46), where $f(z_{(r)})$ and $f(z_{(s)})$ are the rth and sth order samples that have the following binomial properties

$$P\left(f(z_{(r)}) > y\left(\delta_{kl}, \widetilde{\Sigma}_k^C\right)\right) \le \sum_{i=0}^{r-1} \binom{N_k}{i} (\delta_{kl})^i \left(1-\delta_{kl}\right)^{N_k - i} \tag{6.47}$$

$$P\left(f(z_{(s)}) \geq y\left(\delta_{ku}, \widetilde{\Sigma}_k^C\right)\right) \geq \sum_{i=0}^{s-1} \binom{N_k}{i} (\delta_{ku})^i (1-\delta_{ku})^{N_k-i}. \tag{6.48}$$

The $1-\alpha_k$ confidence interval can be approximated by two one-sided intervals. We split α_k into two halves, and allocate one half to each probability bound. Therefore, find the maximum r for which (6.47) is less than or equal to $\frac{\alpha_k}{2}$ and the minimum s for which (6.48) is greater than or equal to $1-\frac{\alpha_k}{2}$, that is

$$\max r : \sum_{i=0}^{r-1} \binom{N_k}{i} (\delta_{kl})^i (1-\delta_{kl})^{N_k-i} \leq \frac{\alpha_k}{2} \text{ and} \tag{6.49}$$

$$\min s : \sum_{i=0}^{s-1} \binom{N_k}{i} (\delta_{ku})^i (1-\delta_{ku})^{N_k-i} \geq 1-\frac{\alpha_k}{2}. \tag{6.50}$$

Combining (6.47)–(6.50), as in Conover (1999), we have

$$P\left(f(z_{(r)}^k) \leq y\left(\delta_{kl}, \widetilde{\Sigma}_k^C\right) \leq y\left(\delta_{ku}, \widetilde{\Sigma}_k^C\right) \leq f\left(z_{(s)}^k\right)\right) \geq 1-\alpha_k. \tag{6.51}$$

When there is no noise, $0 \leq \epsilon_k^P \leq \frac{\epsilon v(\widetilde{\Sigma}_k^P)}{v(S)}$ and $0 \leq \epsilon_k^M \leq \frac{\epsilon v(\widetilde{\Sigma}_k^M)}{v(S)}$, the $1-\alpha_k$ confidence interval of $y(\delta, S)$ is given by $[f(z_{(r)}), f(z_{(s)})]$ based on (6.45), (6.46), and (6.51), that is

$$P\left(f(z_{(r)}) \leq y(\delta, S) \leq f(z_{(s)})\right) \geq 1-\alpha_k.$$

□

Proof of Theorem 2

Proof We note that the event $v\left(L(\delta, S) \cap \hat{\sigma}_p^k\right) \leq D_k^P \epsilon_k$ is equivalent to the event $v\left(\left\{x : f(x) \leq y(\delta, S), x \in \hat{\sigma}_p^k\right\}\right) \leq D_k^P \epsilon_k$ by the definition of $L(\delta, S)$, and therefore, the probability of that event, that is, that the volume of the incorrectly pruned region is less than or equal to $D_k^P \epsilon_k$, can be expressed as

$$P\left(v\left(L(\delta, S) \cap \hat{\sigma}_p^k\right) \leq D_k^P \epsilon_k \middle| A_k\right) = P\left(v\left(\{x : f(x) \leq y(\delta, S), x \in \hat{\sigma}_p^k\}\right) \leq D_k^P \epsilon_k \middle| A_k\right). \tag{6.52}$$

Now, consider the probability expression of quantile in (6.4) from the main article, and let $\delta_p = \frac{D_k^P \epsilon_k}{v(\hat{\sigma}_p^k)}$. We first prove the theorem under the special case that

$y(\delta, S)$ is continuous in δ and $y(\delta_p, \hat{\sigma}_p^k)$ is continuous in δ_p, which implies that $v\left(\left\{x : f(x) = y, x \in \hat{\sigma}_p^k\right\}\right) = 0, \forall y$ and that (6.4) holds at equality. When X is a uniform sample in $\hat{\sigma}_p^k$, we have

$$P\left(f(X) < y\left(\delta_p, \hat{\sigma}_p^k\right)\right) = \frac{v\left(\left\{x : f(x) < y\left(\delta_p, \hat{\sigma}_p^k\right), x \in \hat{\sigma}_p^k\right\}\right)}{v\left(\hat{\sigma}_p^k\right)} = \delta_p = \frac{D_k^P \epsilon_k}{v\left(\hat{\sigma}_p^k\right)},$$

then multiplying $v(\hat{\sigma}_p^k)$ on both sides, we have

$$v\left(\left\{x : f(x) < y\left(\delta_p, \hat{\sigma}_p^k\right), x \in \hat{\sigma}_p^k\right\}\right) = D_k^P \epsilon_k.$$

Hence, we have

$$\begin{aligned} D_k^P \epsilon_k &= v\left(\left\{x : f(x) < y\left(\delta_p, \hat{\sigma}_p^k\right), x \in \hat{\sigma}_p^k\right\}\right) \\ &= v\left(\left\{x : f(x) \le y\left(\delta_p, \hat{\sigma}_p^k\right), x \in \hat{\sigma}_p^k\right\}\right) \\ &\quad - v\left(\left\{x : f(x) = y\left(\delta_p, \hat{\sigma}_p^k\right), x \in \hat{\sigma}_p^k\right\}\right) \end{aligned}$$

and in the special case that $v\left(\{x : f(x) = y, x \in \hat{\sigma}_p^k\}\right) = 0, \forall y$, we have

$$= v\left(\{x : f(x) \le y\left(\delta_p, \hat{\sigma}_p^k\right), x \in \hat{\sigma}_p^k\}\right). \tag{6.53}$$

We substitute the expression for $D_k^P \epsilon_k$ from (6.53) into the probability expression in (6.52), yielding

$$\begin{aligned} &P\left(v\left(\left\{x : f(x) \le y(\delta, S), x \in \hat{\sigma}_p^k\right\}\right) \le D_k^P \epsilon_k \middle| A_k\right) \\ &\quad = P\left(v\left(\left\{x : f(x) \le y(\delta, S), x \in \hat{\sigma}_p^k\right\}\right) \le v\left(\left\{x : f(x)\right.\right.\right. \\ &\qquad \left.\left.\left. \le y\left(\delta_p, \hat{\sigma}_p^k\right), x \in \hat{\sigma}_p^k\right\}\right) \middle| A_k\right) \end{aligned}$$

and from the properties of level sets, if $y(\delta, S) \le y\left(\delta_p, \hat{\sigma}_p^k\right)$, then $\{x : f(x) \le y(\delta, S), x \in \hat{\sigma}_p^k\} \subseteq \{x : f(x) \le y\left(\delta_p, \hat{\sigma}_p^k\right), x \in \hat{\sigma}_p^k\}$, therefore,

$$= P\left(y(\delta, S) \leq y\left(\delta_p, \hat{\sigma}_p^k\right) \middle| A_k \right)$$

and in the special case that $v\left(\left\{x : f(x) = y, x \in \hat{\sigma}_p^k\right\}\right) = 0, \forall y$, we have that

$$= P\left(y(\delta, S) < y\left(\delta_p, \hat{\sigma}_p^k\right) \middle| A_k \right),$$

and by the condition A_k and the pruned assumption, we have $y(\delta, S) \leq y\left(\delta_{ku}, \widetilde{\Sigma}_k^C\right) \leq f(z_{(s)}) < f(x_{(p),(1)})$, where $x_{(p),(1)}$ is the best sample out of $D_k^P N_k^p$ independent samples in $\hat{\sigma}_p^k$, therefore,

$$\geq P\left(f(x_{(p),(1)}) \leq y\left(\delta_p, \hat{\sigma}_p^k\right) \middle| A_k \right)$$
$$= 1 - P\left(f(x_{(p),(1)}) > y\left(\delta_p, \hat{\sigma}_p^k\right) \middle| A_k \right),$$

and since each of the $D_k^P N_k^p$ independent uniform samples X in $\hat{\sigma}_p^k$ satisfies

$$\geq 1 - \left(1 - \delta_p\right)^{D_k^P N_k^p}. \tag{6.54}$$

Since $N_k^p = \left\lceil \frac{\ln \alpha_k}{\ln\left(1 - \frac{\epsilon_k}{v(\sigma_i)}\right)} \right\rceil$ in Step 4, and $\delta_p = \frac{D_k^P \epsilon_k}{v(\hat{\sigma}_p)} = \frac{D_k^P \epsilon_k}{D_k^P v(\sigma_i)} = \frac{\epsilon_k}{v(\sigma_i)}$, where σ_i is a subregion pruned at the kth iteration, and $D_k^P \geq 1$, we know $N_k^p \geq \frac{\ln \alpha_k}{\ln\left(1 - \frac{\epsilon_k}{v(\sigma_i)}\right)} = \frac{\ln \alpha_k}{\ln\left(1-\delta_p\right)} \Rightarrow \ln\left(1 - \delta_p\right)^{D_k^P N_k^p} \leq \ln \alpha_k \Rightarrow \left(1 - \delta_p\right)^{D_k^P N_k^p} \leq \alpha_k$. Multiplying -1 on both sides and adding one to both sides, the inequality becomes $1 - \left(1 - \delta_p\right)^{D_k^P N_k^p} \geq 1 - \alpha_k$, hence

$$P\left(v\left(\{x : f(x) \leq y(\delta, S), x \in \hat{\sigma}_p^k\}\right) \leq D_k^P \epsilon_k | A_k \right)$$
$$\geq 1 - \alpha_k. \tag{6.55}$$

This and (6.52) yield the theorem statement in (6.21) in the special case.

Now, in the more general case where $y(\delta, S)$ and $y(\delta_p, \hat{\sigma}_p^k)$ may have discontinuities, the $v\left(\left\{x : f(x) = y, x \in \hat{\sigma}_p^k\right\}\right)$ may be positive for some y. The flow of the proof is the same, however, the possibility of discontinuities changes equalities to inequalities while accounting for $v\left(\left\{x : f(x) = y, x \in \hat{\sigma}_p^k\right\}\right)$, as follows.

When X is a uniform sample in $\hat{\sigma}_p^k$, the probability expression of quantile in (6.4) with $\delta_p = \frac{D_k^P \epsilon_k}{v\left(\hat{\sigma}_p^k\right)}$ now can be expressed as

$$P\left(f(X) < y\left(\delta_p, \hat{\sigma}_p^k\right)\right) = \frac{v\left(\left\{x : f(x) < y(\delta_p, \hat{\sigma}_p^k), x \in \hat{\sigma}_p^k\right\}\right)}{v(\hat{\sigma}_p^k)} \leq \delta_p = \frac{D_k^P \epsilon_k}{v(\hat{\sigma}_p^k)},$$

then multiplying $v(\hat{\sigma}_p^k)$ on both sides, we have

$$v\left(\left\{x : f(x) < y\left(\delta_p, \hat{\sigma}_p^k\right), x \in \hat{\sigma}_p^k\right\}\right) \leq D_k^P \epsilon_k.$$

Hence, we have

$$\begin{aligned} D_k^P \epsilon_k &\geq v\left(\left\{x : f(x) < y\left(\delta_p, \hat{\sigma}_p^k\right), x \in \hat{\sigma}_p^k\right\}\right) \\ &= v\left(\left\{x : f(x) \leq y\left(\delta_p, \hat{\sigma}_p^k\right), x \in \hat{\sigma}_p^k\right\}\right) \\ &\quad - v\left(\left\{x : f(x) = y\left(\delta_p, \hat{\sigma}_p^k\right), x \in \hat{\sigma}_p^k\right\}\right). \end{aligned} \tag{6.56}$$

We substitute the expression for $D_k^P \epsilon_k$ from (6.56) into the probability expression in (6.52), and due to the possibility of discontinuities, this is a stricter event, yielding an inequality in the probability as follows:

$$\begin{aligned} &P\left(v\left(\{x : f(x) \leq y(\delta, S), x \in \hat{\sigma}_p^k\}\right) \leq D_k^P \epsilon_k | A_k\right) \\ &\geq P\left(v\left(\left\{x : f(x) \leq y(\delta, S), x \in \hat{\sigma}_p^k\right\}\right)\right. \\ &\leq v\left(\left\{x : f(x) \leq y\left(\delta_p, \hat{\sigma}_p^k\right), x \in \hat{\sigma}_p^k\right\}\right) \\ &\quad \left. - v\left(\left\{x : f(x) = y\left(\delta_p, \hat{\sigma}_p^k\right), x \in \hat{\sigma}_p^k\right\}\right) \middle| A_k\right) \end{aligned}$$

and since $v\left(\{x : f(x) \leq y\left(\delta_p, \hat{\sigma}_p^k\right), x \in \hat{\sigma}_p^k\}\right) - v\left(\left\{x : f(x) = y\left(\delta_p, \hat{\sigma}_p^k\right), x \in \hat{\sigma}_p^k\right\}\right)$ $= v\left(\{x : f(x) < y\left(\delta_p, \hat{\sigma}_p^k\right), x \in \hat{\sigma}_p^k\}\right)$

$$\begin{aligned} &= P\left(v\left(\left\{x : f(x) \leq y(\delta, S), x \in \hat{\sigma}_p^k\right\}\right) \leq v\left(\left\{x : f(x)\right.\right.\right. \\ &\quad \left.\left.\left. < y\left(\delta_p, \hat{\sigma}_p^k\right), x \in \hat{\sigma}_p^k\right\}\right) \middle| A_k\right) \end{aligned}$$

and now comparing the level sets associated with $y(\delta, S)$ and $y\left(\delta_p, \hat{\sigma}_p^k\right)$, we see that, if $y(\delta, S) < y\left(\delta_p, \hat{\sigma}_p^k\right)$, then even in the presence of discontinuities, $\{x : f(x) \leq y(\delta, S), x \in \hat{\sigma}_p^k\} \subseteq \{x : f(x) < y\left(\delta_p, \hat{\sigma}_p^k\right), x \in \hat{\sigma}_p^k\}$, so we have the following

$$\geq P\left(y(\delta, S) < y\left(\delta_p, \hat{\sigma}_p^k\right) \middle| A_k\right),$$

and by the condition A_k and the pruned assumption, we have $y(\delta, S) \leq y(\delta_{ku}, \widetilde{\Sigma}_k^C) \leq f(z_{(s)}) < f(x_{(p),(1)})$, where $x_{(p),(1)}$ is the best sample out of $D_k^P N_k^p$ independent samples in $\hat{\sigma}_p^k$, therefore,

$$\begin{aligned} &\geq P\left(f(x_{(p),(1)}) \leq y\left(\delta_p, \hat{\sigma}_p^k\right) \middle| A_k\right) \\ &= 1 - P\left(f(x_{(p),(1)}) > y\left(\delta_p, \hat{\sigma}_p^k\right) \middle| A_k\right), \end{aligned}$$

and since each of the $D_k^P N_k^p$ independent uniform samples X in $\hat{\sigma}_p^k$ satisfies

$$P\left(f(X) > y\left(\delta_p, \hat{\sigma}_p^k\right)\right) = 1 - P\left(f(X) \leq y\left(\delta_p, \hat{\sigma}_p^k\right)\right) \leq 1 - \delta_p,$$

we have

$$\geq 1 - \left(1 - \delta_p\right)^{D_k^P N_k^p} \tag{6.57}$$

which is the same inequality as in (6.54).

As in the special case, since $N_k^p = \left\lceil \frac{\ln \alpha_k}{\ln\left(1 - \frac{\epsilon_k}{v(\sigma_i)}\right)} \right\rceil$ in Step 4, and $\delta_p = \frac{D_k^P \epsilon_k}{v(\hat{\sigma}_p)} = \frac{D_k^P \epsilon_k}{D_k^P v(\sigma_i)} = \frac{\epsilon_k}{v(\sigma_i)}$, where σ_i is a subregion pruned at the kth iteration, and $D_k^P \geq 1$, we have that

$$\begin{aligned} &P(v(\{x : f(x) \leq y(\delta, S), x \in \hat{\sigma}_p^k\}) \leq D_k^P \epsilon_k | A_k) \\ &\quad \geq 1 - \alpha_k \end{aligned} \tag{6.58}$$

which yields the theorem statement in (6.21) in the general case, too. □

References

Ali MM, Khompatraporn C, Zabinsky ZB (2005) A numerical evaluation of several stochastic algorithms on selected continuous global optimization test problems. J Glob Optim 31:635–672

Bechhofer RE, Dunnett CW, Sobel M (1954) A two-sample multiple decision procedure for ranking means of normal populations with a common unknown variance. Biometrika 41:170–176

Chen CH, He D (2005) Intelligent simulation for alternatives comparison and application to air traffic management. J Syst Sci Syst Eng 14(1):37–51

Chen CH, He D, Fu M, Lee LH (2008) Efficient simulation budget allocation for selecting an optimal subset. INFORMS J Comput 20(4):579–595

Chen CH, Yucesan E, Dai L, Chen HC (2010) Efficient computation of optimal budget allocation for discrete event simulation experiment. IIE Trans 42(1):60–70

Chen CH, Lee LH (2011) Stochastic simulation optimization: an optimal computing budget allocation. World Scientific, Singapore

Conover WJ (1999) Practical nonparametric statistics, 3rd edn. Wiley, New York

Csendes T, Pintér J (1993) A new interval method for locating the boundary of level sets. Int J Comput Math 49(1–2):53–59

Fu MC, Chen CH, Shi L (2008) Some topics for simulation optimization. In Proceedings of the 40th conference on winter simulation. IEEE, Piscataway, p 27–38

Fu MC (2015) Handbook of simulation optimization. Springer, New York

Ho YC, Cassandras CG, Chen CH, Dai L (2000) Ordinal optimisation and simulation. J Oper Res Soc 51:490–500

Ho YC, Zhao QC, Jia QS (2007) Ordinal optimization: soft optimization for hard problems. Springer, Berlin

Hu J, Fu MC, Marcus SI (2007) A model reference adaptive search method for global optimization. Oper Res 55(3):549–568

Huang H (2016) Discrete-event simulation and optimization to improve the performance of a healthcare system. Ph.D. Thesis, University of Washington

Huang H, Zabinsky ZB (2013) Adaptive probabilistic branch and bound with confidence intervals for level set approximation. In: Pasupathy R, Kim SH, Tolk A, Hill R, Kuhl ME (eds) Proceedings of the 2013 conference on winter simulation. IEEE, Piscataway, p 980–991

Huang H, Zabinsky ZB (2014) Multiple objective probabilistic branch and bound for Pareto optimal approximation. In: Proceedings of the 2014 conference on winter simulation. IEEE, Piscataway, p 3916–3927

Huang H, Zabinsky ZB, Heim JA, Fishman P (2015) Simulation optimization for medical imaging resource allocation. In: Extended abstract of the 2015 conference on INFORMS healthcare. Nashville

Huang H, Zabinsky ZB, Li Y, Liu S (2016) Analyzing hepatitis C screening and treatment strategies using probabilistic branch and bound. In: Roeder TMK, Frazier PI, Szechtman R, Zhou E, Huschka T, Chick SE (eds) Proceedings of the 2016 conference on winter simulation. IEEE Press, Piscataway, p 2076–2086

Kim SH, Nelson BL (2001) A fully sequential procedure for indifference-zone selection in simulation. ACM Trans Model Comput Simul 11(3):251–273

Linz D, Huang H, Zabinsky ZB (2015) Partition based optimization for updating sample allocation strategy using lookahead. In: Yilmaz L, Chan WKV, Roeder, TMK Moon I, Macal C, Rossetti MD (eds) Proceedings of the 2015 conference on winter simulation. IEEE Press, Huntington Beach

Nelson BL, Swann J, Goldsman D, Song W (2001) Simple procedures for selecting the best simulated system when the number of alternatives is large. Oper Res 49(6):950–963

Ólafsson S (2004) Two-stage nested partitions method for stochastic optimization. Methodol Comput Appl Probab 6:5–27

Pintér J (1990) Globally optimized calibration of environmental models. Ann Oper Res 25(1):211–221

Prasetio Y (2005) Simulation-based optimization for complex stochastic systems. Ph.D. Thesis, University of Washington

Rinott Y (1978) On two-stage selection procedures and related probability-inequalities. Commun Stat Theory Methods 7(8):799–811

Shi L, Ólafsson S (2000) Nested partitions method for stochastic optimization. Methodol Comput Appl Probab 2(3):271–291
Shi L, Ólafsson S (2009) Nested partitions method, theory and applications. Springer, New York
Tsai YA, Pedrielli G, Mathesen L, Huang H, Zabinsky ZB, Candelieri A, Perego R (2018) Stochastic optimization for feasibility determination: an application to water pump operation in water distribution networks. In: Rabe M, Juan AA, Mustafee N, Skoogh A, Jain S, Johansson B (eds) Under review in the proceedings of the 2018 conference on winter simulation. IEEE, Piscataway
Wang W (2011) Adaptive random search for noisy and global optimization. Ph.D. Thesis, University of Washington
Xu WL, Nelson BL (2013) Empirical stochastic branch-and-bound for optimization via simulation. IIE Trans 45(7):685–698
Xu J, Zhang S, Huang E, Chen CH, Lee LH, Celik N (2016) MO^{2}TOS: multi-fidelity optimization with ordinal transformation and optimal sampling. Asia Pac J Oper Res 33(3):1650017
Zabinsky ZB (1998) Stochastic methods for practical global optimization. J Glob Optim 13:433–444
Zabinsky ZB, Wang W, Prasetio Y, Ghate A, Yen JW (2011) Adaptive probabilistic branch and bound for level set approximation. In: Jain S, Creasey RR, Himmelspach J, White KP, Fu M (eds) Proceedings of the 2011 conference on winter simulation. IEEE, Piscataway, p 46–57

Zelda B. Zabinsky is currently a Professor in the Industrial and Systems Engineering Department at the University of Washington (UW), with adjunct appointments in the departments of Electrical Engineering, Mechanical Engineering, and Civil and Environmental Engineering. She received her Ph.D. from the University of Michigan in Industrial and Operations Engineering, and her bachelor's degree (magna cum laude) from the University of Puget Sound in Mathematics with a minor in Biology.

Dr. Zabinsky's interest in math modeling began as an undergraduate student when she first learned about transportation and routing optimization models and was intrigued with the "predator-prey" model. Upon graduation, she worked briefly at National Marine Fisheries Service in Seattle where she was exposed to complex models of fish migration, accounting for ocean currents and temperature. She then joined Boeing Computer Services in Seattle, where she worked with senior engineers on design problems. She moved to Ann Arbor, MI, to join her husband, and started working at Vector Research (now Altarum), where she solidified her interest in Operations Research as a blend of math modeling and engineering applications. Her career as an academician at the UW has allowed her to enjoy stimulating research and working with students. She loves being a professor; she enjoys teaching and pursuing interesting and important problems that combine mathematics with decision making.

Professor Zabinsky has published numerous papers in the areas of global optimization and algorithm complexity, and her book, Stochastic Adaptive Search in Global Optimization, describes research on theory and practice of algorithms useful for solving problems with multimodal objective functions in high dimension. She has also worked on many applications

involving optimization under uncertainty, including: optimal design of composite structures, air traffic flow management, supplier selection, supply chain and inventory management, power systems with renewable resources, water distribution networks, wildlife corridor selection, nano-satellite communications, and health care systems. She enjoys working in teams including experts in the application domain, economists, software engineers, and other operations researchers.

Her research has been funded by the National Science Foundation (NSF), Department of Homeland Security, NASA-Langley, Federal Aviation Administration (FAA), Boeing Commercial Airplane Company, Port of Tacoma, and Microsoft. She is an IIE Fellow. Professor Zabinsky teaches courses in Operations Research and has received the annual teaching award in Industrial Engineering at the University of Washington several times.

Hao Huang is an Assistant Professor in the Department of Industrial Engineering and Management at the Yuan Ze University at Taoyuan, Taiwan. He received his Ph.D. from University of Washington in the Industrial and Systems Engineering, and his M.S. and B.S. degree from National Tsing Hua University, Hsinchu, Taiwan, in Industrial Engineering and Engineering Management. His research interests include simulation optimization and healthcare applications, and his research is funded by the Ministry of Science and Technology (MOST) at Taiwan. At Yuan Ze University, he teaches multivariate analysis, time series analysis, and calculus. His e-mail address is haohaung@saturn.yzu.edu.tw.

Part III
Education

Chapter 7
Modeling Engineering Student Success Needs

Tracee Walker Gilbert, Janis Terpenny, and Tonya Smith-Jackson

Contents

7.1 Introduction 160
7.2 Purpose 160
7.3 Student Success Theoretical Perspectives 161
7.4 Engineering Student Needs Questionnaire Development and Validation Process 163
7.4.1 Step 1. Specify the Domain of Constructs 164
7.4.2 Step 2. Define Student Needs 164
7.4.3 Step 3. Develop Questionnaire Items and Layout 165
7.4.4 Step 4. Perform Expert Review 165
7.4.5 Step 5. Administer Questionnaire 166
7.4.6 Step 6. Assess Validity and Reliability 166
7.4.7 Step 7. Develop Model 167
7.4.8 Step 8. Test Model Relationships 172
7.5 Example Application 173
7.6 Summary and Key Principles 176
7.6.1 Key Principles 178
7.7 Future Work 178
7.7.1 Expanding the Operationalization of Student Success 178
7.7.2 Strengthening the Validity and Reliability of the ESNQ 178
References 179

T. W. Gilbert
System Innovation LLC, Arlington, VA, USA
e-mail: tracee.gilbert@systeminnovation1.com

J. Terpenny (✉)
The Pennsylvania State University, State College, PA, USA
e-mail: jpt5311@engr.psu.edu; jpt5311@psu.edu

T. Smith-Jackson
North Carolina A&T State University, Greensboro, NC, USA
e-mail: tlsmithj@ncat.edu

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_7

7.1 Introduction

Producing successful engineers has become increasingly important to the United States. The nation needs to respond more effectively to the existing and future needs of a technology-driven global society. This goal, of course, requires our nation's colleges of engineering to develop *and* graduate students in a variety of disciplines to meet those needs. As such, student success continues to be a pressing concern for higher education administrators, researchers, and policymakers.

In this context, the term "student success" is most often defined by graduation rates, which, overall, have not changed significantly in the United States in the last 5 years. (U.S. Department of Education DOE, National Center for Educational Statistics 2017). We note that the use of graduation rates, especially four-year graduation rates, as the definition of student success has implications for equity when applied cross-culturally (Bowman 2010; Pérez et al. 2017). For this context, however, we will use this definition with the understanding of the associated caveats.

For engineering students, it is estimated that less than half of the first year students who initially enroll in engineering go on to earn their bachelor's degree within 5 years. Moreover, underrepresented minorities (URM) (i.e., African Americans, Hispanics, and Native Americans) drop out at even higher rates than their majority peers (Chen 2013; U.S. Department of Education DOE, National Center for Educational Statistics 2017). The increasing difficulty of retaining engineering students has contributed to the decline in graduation rates. As a result, effective institutional practices to improve this trend are essential for enhancing student success in our nation's colleges of engineering (Tinto 2010).

7.2 Purpose

Since the establishment of the formal education system, researchers have been attempting to unravel the complexities associated with enhancing student success in higher education (Berger and Lyon 2005). This research has resulted in theoretical perspectives (Astin 1984; Bean 1980, 1983; Bean and Eaton 2000; Kuh 2001, 2009; Tinto 1993), which provide a better understanding of why some students decide to leave, and to some extent why others persist on to graduation. Despite a sizable body of knowledge that has identified various factors associated with student success in higher education, little work has been devoted to translating various theoretical findings into institutional action that improves student success outcomes (Tinto and Pusser 2006; Tinto 2006, 2010).

While these theories have provided most of the empirical and conceptual knowledge that has shaped institutional practices (Berger and Lyon 2005; Kuh et al. 2006; Tinto and Pusser 2006); Tinto (2006) asserted that these perspectives are not suited for guiding institutional action; largely because knowing why students leave does not directly explain why students persist. Furthermore, these perspectives

do not provide an understanding of direct actions institutions can take to help students remain in college and succeed. A lack of direct and impactful actions is further exacerbated by the abstraction levels associated with the operationalization of variables. These abstractions fall short of addressing practical outcomes that institutions can directly impact.

To overcome these limitations, this study puts forward the premise that institutional leaders must first have a well-founded understanding of the needs of the students they serve. Identifying the need is germane to most, if not all, engineering system design efforts as the basis for developing optimal solutions to satisfy an identified need. In the context of higher education, having an understanding of student success needs can form the basis from which institutional practices can be designed to meet those needs.

Specifically, the purpose of this study is to develop a statistically verified model of engineering student success needs. The basic premise of the model is that there are academic, psychological, environmental, financial, and social factors that impact student success. Based on these factors, the following seven dimensions of student success needs were developed in this study: Classroom Learning, Faculty Interaction, Sense of Community, Student Interaction, Financial, Skill, and Workload Management needs. This model identifies the dimensions and associated actionable need statements that impact student satisfaction, which is a measure of student success that is useful in determining the quality of the educational experience (Kuh et al. 2006).

The Engineering Student Needs Questionnaire (ESNQ) development and validation process is presented in order to develop the model. Based on the results of a questionnaire completed by 213 students at the University of Maryland, College Park, the relationships between model variables were tested to determine the dimensions that institutional decision-makers can target to meet the needs of their engineering students.

7.3 Student Success Theoretical Perspectives

This study uses a collection of student success theoretical perspectives as a framework to develop the model of engineering student success needs. A high level summary of the most comprehensive and influential theoretical perspectives is presented in Table 7.1. The summary provides an understanding of the research that explains why students decide to leave college and, to some extent, why students persist on to graduation.

In reviewing the array of literature, consistent patterns have emerged from across the various theoretical perspectives. First, these theoretical perspectives emphasize that student characteristics/behaviors and institutional conditions impact student success. Student behaviors include: involvement in extracurricular activities, interaction with faculty and peers, motivation, and commitment. Institutional conditions include the resources and educational practices that facilitate positive

student behavior (Kuh et al. 2006). Since institutions vary considerably in their size, culture, and student demographics, an understanding of the unique needs of the students within a particular campus environment is needed in order to allow institutions to tailor their practices to address their needs (Berger and Lyon 2005).

Secondly, a multitude of variables have been identified as a result of the theoretical perspectives identified in Table 7.1. These variables consist of both pre-entry and post-entry variables, which have been operationalized as abstract concepts that the institution does not directly impact. In terms of pre-entry variables, for example, the high school experience and socio-economic status are factors that have been identified from each of the theoretical perspectives. While this information is indirectly useful in providing insight into the student population, it does not provide actionable information that guides institutions in determining how to specifically tap into issues of socio-economic status. In terms of post-entry variables, Tinto's (1993) widely studied social and academic integration construct has been useful in informing decision-makers about the importance of integration; however, this insight does not directly provide decision-makers with guidance on what actions should be carried out to achieve academic and social integration (Tinto 2006).

Lastly, each of these theoretical perspectives (with the exception of the involvement/engagement perspective) focused on retention and persistence. Even though

Table 7.1 Relevant student success theoretical perspectives

Perspective	Theory	Source	Purpose
The Sociological Perspective	Student Integration Model	Tinto (1993)	Describes the influence of the social forces and the academic structure on student departure decisions
The Organizational Perspective	Student Attrition Model	Bean (1980, 1983)	Concentrates on the impact that the institution (i.e., organization) has on the socialization and satisfaction of students
The Psychological Perspective	Student Attrition Model	Bean and Eaton (2000)	Focuses on the role of psychological characteristics that distinguish between those students that persist and those that depart
The Financial Perspective	Financial Nexus Model	St. John et al. (2005)	Highlights the role that finances play in persistence decisions
The Minority Perspective	Student/ Institution Engagement Model	Nora (2003)	Emphasizes the unique challenges that diminish the quality of minority students' college experience
The Involvement/Engagement Perspective	Theory of Involvement; Student Engagement	Astin (1984); Kuh (2009)	Focuses on the behaviors that students engage in while in college

these perspectives have provided an immense understanding of the factors that impact college retention and persistence, few studies used the student success theoretical perspectives to directly understand student needs. Therefore, this research shifts the focus from trying to understand why students leave/stay in college to examining how to satisfy student needs in the context of the student success theoretical perspectives.

7.4 Engineering Student Needs Questionnaire Development and Validation Process

This section describes the development and validation of ESNQ, which was designed to comprehensively assess the needs of engineering students in the context of the student success theoretical perspectives. While system design methods and performance improvement tools were embedded into the questionnaire development process, the questionnaire validation process followed recommended standard procedures that are consistent with the extant literature (Churchill 1979; Clark and Watson 1995; DeVillis 1991; Netemeyer et al. 2003; Nunnally 1978). Figure 7.1 summarizes the ESNQ development and validation process.

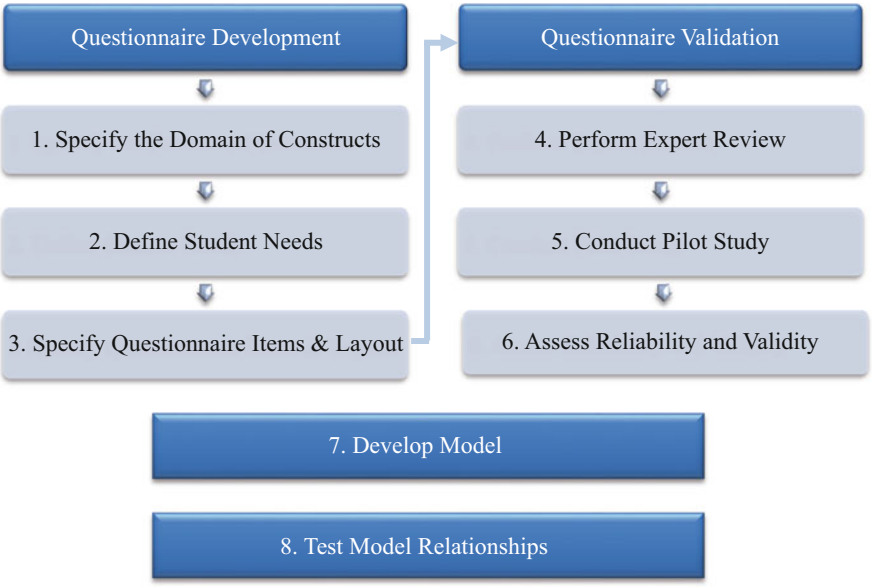


Fig. 7.1 ESNQ development and validation process

7.4.1 Step 1. Specify the Domain of Constructs

The first step in the questionnaire development process was to delineate what should be included in the domain of the construct (Churchill 1979). Determining what should be included, as well as what should be excluded, were critical steps because sources of invalidity can originate in the construct definition process (Netemeyer et al. 2003). The incorporation of this initial stage is based on a review of the literature and the importance of the student satisfaction construct. This refers to the student's satisfaction with an institution's ability to meet their needs related to aspects of the college experience that impact student success.

The multitude of variables from each of the student success theoretical perspectives reviewed led to the development of the typology of student success factors. This typology specifies the domain of constructs, which broadly categorizes variables as academic, social, psychological, environmental, and financial factors. Pre-entry variables are outside of the scope of this study since institutions have limited ability to directly impact pre-entry variables. Only post-enrollment variables were considered. By specifying the domain of the construct, it is clear that additional needs, such as transportation, housing, food, and facilities, are outside of the scope of this research.

7.4.2 Step 2. Define Student Needs

To provide actionable information that can be used for decision-making, a student success-oriented participatory design method was used to capture the voice of the student. This is the goal of any engineering design problem: to translate the voice of the customer into a description of what needs to be improved (Blanchard and Fabryky 2017). Therefore, the second step in the questionnaire development process is to define the true needs of the students.

A need is neither a solution (e.g., a summer bridge program) nor a physical measurement (e.g., number of tutoring services offered by the university). Following this assertion, a need can be defined as a detailed description of what is required of institutional practices that contribute to the success of engineering students. To capture student needs, a participatory design method was developed to identify the needs as voiced by the students. A total of 21 undergraduate engineering students from the University of Maryland, College Park participated in four participatory design meetings (i.e., 4–6 participants per meeting). Participants were asked to share their experiences in a guided group discussion over the course of a 2-h period.

The objective of these meetings was to utilize a functional decomposition process in which student needs were systematically elicited and decomposed into actionable need statements. The design skills of abstraction, persistent questioning, identifying critical performance areas, and factorization were infused into the meetings to facilitate the analysis-synthesis-evaluation cycle. As a result, the meeting process was

structured to capture ideas, which were organized such that actionable information for decision-making.

The typology of student success factors from the literature was used to identify enablers and hinders of student success. These enablers and hinderers were translated into a comprehensive pool of student need categories and student need statements for each student success factor. As a result, the dimensions of student success needs and associated need statements were developed to provide the content for the ESNQ.

7.4.3 Step 3. Develop Questionnaire Items and Layout

Based on the output from Step 2, a comprehensive pool of questionnaire items were developed for inclusion in the ESNQ. These items were cross-checked with the following sources to ensure that the domain was comprehensively covered: Pittsburgh Freshman Engineering Attitudes Survey (Besterfield-Sacre et al. 2001), Persistence in Engineering (PIE) survey instrument (Eris et al. 2005), Engineering Fields Questionnaire (Lent et al. 2005), National Survey of Student Engagement (Kuh 2001), Student Satisfaction Inventory (Schreiner and Juillerat 2010), and Institutional Integration Scales (Pascarella and Terenzini 1980).

Although existing questionnaires from both undergraduate education and engineering education provided a knowledge base of existing scales that address student success, questionnaire items could not be taken directly from the literature because the ESNQ is designed uniquely to provide actionable need statements. Similar to requirements, these need statements are short phrases that describe the functionality or actions that institutional practices should deliver in order to help students succeed.

Furthermore, the ESNQ involved an examination of students' perception of the importance of their needs, as well as their satisfaction level with the performance of the institution in meeting their needs. While traditional methods tend to examine satisfaction and importance independently, these concepts were merged and used together. Similar to the Noel Levitz Student Satisfaction Inventory (Schreiner and Juillerat 2010), a 7-point Likert-type scale format is used to capture both the importance and the satisfaction.

7.4.4 Step 4. Perform Expert Review

Following the questionnaire development process, the instrument underwent two rounds of expert reviews with students and student success researchers and practitioners. First, a focus group meeting to carry out the student expert panel review was conducted with four undergraduate *engineering students, who were considered experts based on their knowledge about the* undergraduate engineering experience.

Students were asked to complete an initial draft of the questionnaire. While completing the questionnaire, they were also instructed to edit the questionnaire items to ensure appropriateness of language, clarity, and readability. Additionally, a Q-sort procedure was used to improve the construct validity of the first round of the expert review (Moore and Benbasat 1991) with students. A group discussion followed, which allowed the participants to collectively determine whether (1) additional categories should be added or renamed; (2) additional items should be added to a particular category; and (3) if any improvements should be made to the questionnaire.

Next, a second expert panel review was conducted with eight student success researchers and practitioners. A worksheet was emailed to the expert reviewers, and returned to the researcher electronically. The goals of the practitioner expert panel review were the same as the student panel review: to review the questionnaire and provide feedback to ensure that the questionnaire comprehensively and concisely addressed the needs that are critical to engineering student success. However, the format was different to accommodate a review from participants that were in distributed locations.

7.4.5 Step 5. Administer Questionnaire

Once the two rounds of expert reviews were completed by the students and the practitioners, the questionnaire was administered electronically using the Qualtrics™ online survey tool. First, the questionnaire was administered as a pilot test to collect actual data to purify the instrument prior to its final administration. A total of 241 undergraduate engineering students from a doctoral granting institution completed the pilot questionnaire. Based on the results of pilot study, the final version of the ESNQ was administered at another doctoral granting institution. A total of 213 undergraduate engineering students completed the final version of the questionnaire.

7.4.6 Step 6. Assess Validity and Reliability

Pearson correlations were computed for each item. Then the Kaiser-Meyer-Olkin (KMO) measure of sampling was used to determine the suitability of factor analysis. The KMO value of 0.87 exceeded the recommended value of 0.60 (Kaiser 1970), and Bartlett's test of sphericity was significant, $2\ (630) = 3779$, $p < 0.001$ (Bartlett 1954). Based on the results of the correlation analysis, KMO measure, and Bartlett's test of sphericity, the ESNQ was considered suitable for exploratory factor analysis.

As a result, a factor analysis was conducted using varimax rotation. The eigenvalue greater than 1 rule was used to determine the number of factors. Nine factors had an eigenvalue greater than 1. However, the seventh, eighth, and ninth factors accounted for less than 5% of the variance, respectively (Hair et al. 1998). As a

result, a six factor model solution was retained for rotation in Table 7.1, which shows the factor loadings, eigenvalues, % variance explained, and the cumulative variance explained by the factor analysis for the rotated factor solution. As illustrated, five factors initially best represented the data in terms of variance explained (51%). The dimensions of Resources and Professional Development needs were deleted from the analysis. Furthermore, Faculty Interaction/Sense of Community needs were divided into two separate dimensions, as well as Classroom Learning/Workload Management needs. Conceptually, these dimensions represented unique aspects of student success. Moreover, items were eliminated, which were grayed out and the dimensions and associated items for the ESNQ are shown in Table 7.2.

Next, Cronbach's Alpha coefficient was used to assess the reliability of the dimensions of student success needs from Table 7.2. Following convention, a Cronbach Alpha coefficient of 0.7 or greater was the threshold for an internally consistent, reliable dimension (Nunnally 1978). If the dimension did not meet the threshold, the effect of removing each item in the dimension was investigated using the "if item deleted method." Table 7.3 depicts the results of the reliability assessment. Six of the alpha coefficients were above the 0.7 threshold, ranging from 0.75 to 0.89. However, the Classroom Learning needs dimension was slightly below the threshold ($\alpha = 0.62$). The "if item deleted method" was examined for this dimension; however, the largest increase resulted from deleting item #01–4 ($\alpha = 0.69$), which still did not meet the threshold. Therefore, none of the four items were deleted from this dimension. While this is less than the desired threshold, Devillis (1991) suggests that 0.6 is acceptable for newly developed dimensions. As a result, the four items were retained in this analysis for the Classroom Learning needs dimension. Furthermore, the results reported in Table 7.3 indicate that the dimensions of student success needs for the ESNQ are reliable and demonstrate an acceptable degree of internal consistency.

7.4.7 Step 7. Develop Model

Based on the results, evidence of statistical validity was demonstrated for the reliability and validity assessment. This indicated that the final ESNQ measured what it was intended to measure. Furthermore, a statistically verified research model of engineering student success needs (Fig. 7.2) was developed.

The basic premise of the model is that there are academic, psychological, environmental, financial, and social factors that impact student success. Based on these factors, the seven dimensions of student success needs, the dependent variables, have been refined as a result of the main study. Overall satisfaction, which is the independent variable, is used to measure student success. The definition of the research model variables are described below:

Table 7.2 Exploratory factor analysis results

Main questionnaire	Need statements	1	2	3	4	5	6
Classroom Learning (4 items)	To have classes that stimulate interest in my field of study	0.09	**0.51**	0.07	−0.06	0.07	−0.04
	To have relevant assignments (e.g., HW, labs, exams) that reinforce what I am learning in class	0.10	**0.49**	−0.09	0.03	−0.05	−0.12
	To connect what I am learning in class to the engineering profession	0.51	0.38	0.20	0.13	−0.02	0.04
	To comprehend class material	0.41	**0.53**	0.16	0.22	−0.14	−0.24
	To have class concepts communicated in a manner that I understand	0.43	**0.61**	0.24	0.17	−0.05	−0.19
Faculty Interaction (4 items)	To have approachable faculty members that I can discuss issues of interest and importance to me	**0.72**	0.15	0.12	0.07	0.07	0.19
	To have faculty members who demonstrate flexibility and responsiveness to my needs	**0.70**	0.17	0.29	0.09	0.16	0.03
	To have faculty members who are interested in engaging me in their fields	**0.73**	0.06	0.33	0.06	0.12	−0.02
	To receive timely feedback front faculty members (e.g., grades, homework, exams)	**0.52**	0.20	0.15	0.12	0.16	−0.38
Sense of Community (4 items)	To have a welcoming environment where I feel a sense of belonging	**0.66**	0.13	0.17	0.17	0.34	−0.04
	To have a supportive group of people who provide help and encouragement	**0.40**	0.20	−0.05	0.19	0.35	0.37
	To have an environment where I receive fair and unbiased treatment	**0.60**	0.03	0.13	0.25	0.22	0.01

(continued)

Table 7.2 (continued)

Main questionnaire	Need statements	1	2	3	4	5	6
	To have opportunities outside of class to cultivate relationships with the engineering community on campus	**0.59**	0.06	0.17	0.28	0.17	0.34
Student Interactions (3 items)	To have relationships with students who share my interests	0.38	−0.04	0.12	0.32	**0.55**	0.20
	To have opportunities to exchange ideas and gain knowledge from other students	0.45	−0.05	0.24	0.24	**0.59**	0.01
	To have opportunities to socialize with students from diverse backgrounds	0.24	−0.05	0.16	0.27	**0.57**	0.17
Financial (4 items)	To ease my financial burden	0.25	0.08	0.11	**0.81**	0.13	−0.10
	To have opportunities to earn money in order to offset my expenses (e.g., jobs, work study)	0.14	0.02	0.18	**0.78**	0.17	0.08
	To be informed about financial assistance opportunities	0.14	0.08	0.14	**0.84**	0.14	0.12
	To have financial assistance available to me (e.g., scholarships, grants)	0.18	0.15	0.13	**0.81**	0.12	0.01
Skills (4 items)	To develop research skills and experiences	0.05	−0.03	0.13	0.07	0.11	0.60
	To develop basic academic skills (e.g., study skills, time management)	0.16	0.33	0.31	0.15	0.52	0.01
	To develop teamwork skills	0.09	0.23	0.46	0.02	0.42	0.05
	To develop communication skills (e.g., verbal and written)	0.08	0.13	**0.72**	0.02	0.21	0.00
	To develop problem-solving skills	0.18	0.07	**0.73**	0.09	0.29	−0.15
	To develop technical skills (e.g., programming languages, software applications)	0.24	0.01	**0.55**	0.17	0.13	0.16

(continued)

Table 7.2 (continued)

Main questionnaire	Need statements	1	2	3	4	5	6
	To develop job or work-related skills and experiences	0.30	0.13	**0.59**	0.15	0.21	−0.07
Resources (4 items)	To get help with post-graduation planning (e.g., graduate school, career opportunities)	0.36	−0.08	0.50	0.22	−0.11	0.14
	To get help with academic planning (e.g., degree requirements, course scheduling)	0.24	0.03	0.66	0.17	−0.11	0.14
	To get help with personal development (e.g., personal concerns, social issues)	0.09	0.26	−0.05	−0.01	0.12	0.67
	To get help with mastering course concepts	0.04	0.68	0.13	0.11	−0.11	0.42
	To have resources available to help me (e.g., reference materials, equipment, software)	−0.03	0.59	0.11	−0.05	−0.07	0.43
Workload (4 items)	To keep up with the pace of my courses	−0.08	**0.73**	0.00	0.08	0.13	0.38
	To have a manageable workload	0.12	**0.70**	−0.01	0.05	0.26	0.13
	To cope with stress	0.04	**0.53**	0.21	0.07	0.51	0.20
	To have a balanced social personal and academic experience	0.19	**0.56**	0.09	0.15	0.36	0.01
Total		4.76	4.18	3.42	3.41	2.65	2.10
% Variance		13.21	11.61	9.50	9.49	7.36	5.83
Cumulative %		13.21	24.82	34.31	43.80	51.16	56.98

- *Classroom Learning (CL)* needs—this variable describes the extent to which students are satisfied with the institution's efforts in providing a classroom experience that enhances their ability to acquire knowledge in their field of study.
- *Faculty Interaction (FI)* needs—this variable describes the extent to which students are satisfied with the institution in providing opportunities to have quality interactions with faculty members.

Table 7.3 Cronbach alpha coefficient values for each dimension

Dimensions of student success needs	# of items	α
Classroom learning	4	0.62
Faculty interaction	4	0.81
Sense of community	4	0.77
Student interaction	3	0.82
Financial	4	0.89
Skill	4	0.76
Workload management	4	0.81

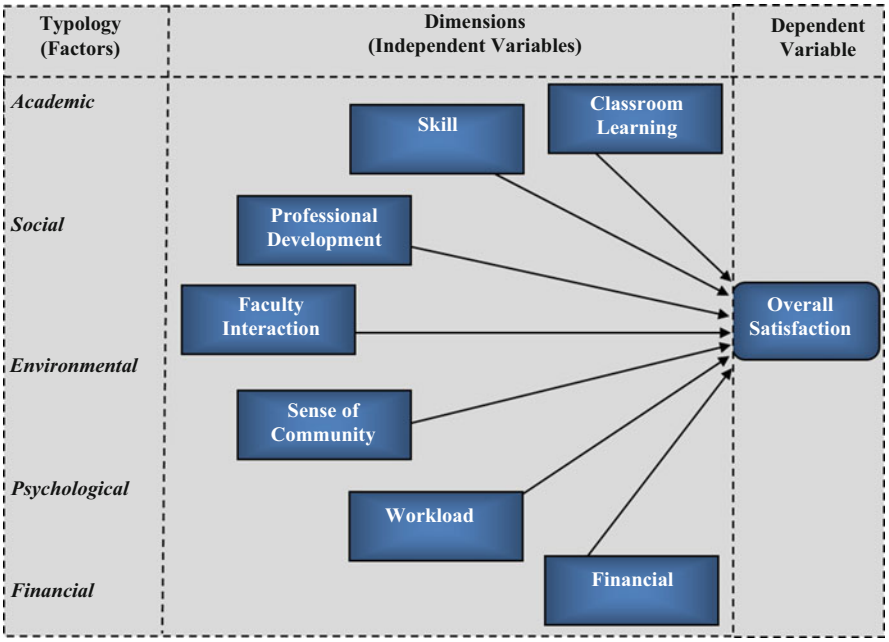


Fig. 7.2 Model of engineering student success needs

- *Student Interaction (SI)* needs—this variable describes the extent to which students are satisfied with the institution in providing opportunities to have quality interactions with other students.
- *Sense of Community (SC)* needs—this variable describes the extent to which students are satisfied with the institution's efforts to create a welcoming environment, such that students experience a sense of belonging with the engineering community on campus.
- *Financial (F)* needs—this variable describes the extent to which students are satisfied with the institution in providing available and accessible financial assistance.

- *Workload Management (WM)* needs—this variable describes the extent to which students are satisfied with the institution in helping to cope with the demands of their engineering major.
- *Overall Satisfaction (OS)*—this variable describes the extent to which students are satisfied with their overall college experience.

7.4.8 Step 8. Test Model Relationships

The third and final test of predictive validity tested the hypothesized relationships between the model variables. A hierarchical multiple regression analysis was used to understand how the dimensions of student success needs impact the students' overall satisfaction with their college experience. Before the analysis was conducted, a number of assumptions were tested to ensure that there were not any errors in the model specification. Collinearity, outliers, normality, linearity, and homoscedasticity assumptions were explored to ensure that there were no errors in the model specification. The correlation matrix, residual scatterplots, outliers, and statistics confirmed that these assumptions were not violated (Tabachnick and Fidell 2007).

A hierarchical multiple regression analysis was then used to test the hypothesized relationships in the research model, while controlling for background variables. First, the set of background variables (Gender, Ethnicity, Race, Grades, Class Level, International Student Status) were entered into block 1 in order to statistically control for these variables. Then the dimensions of student success needs, the independent variables, were entered into block 2 to determine the relationship between the dimensions of student success needs and overall satisfaction after the potential confounding relationships were removed from block 1 (Cohen et al. 2014). The results of the hierarchical regression model are summarized in Table 7.4.

The background variables were entered into block 1, which explained only 1.4% of the variance in overall satisfaction. The dimensions of student success needs were then entered into block 2, which explained 20.2% of the variance. As a result, the model as a whole accounted for 21.6% of the variance in Overall Satisfaction, $F(12, 199) = 4.192$, $p = 0.000$. However, there were only two significant predictors of Overall Satisfaction, in which the beta weights were used to determine the relative weights of the dimensions in the regression model. In order of importance, the dimensions of Classroom Learning (beta $= -0.27$, $p < 0.001$) and Workload (beta $= -0.19$, $p < 0.05$) needs were significant predictors of overall satisfaction. However, no other dimensions had a statistically significant contribution to the model at $p < 0.05$.

Table 7.4 Results of the multiple regression analysis

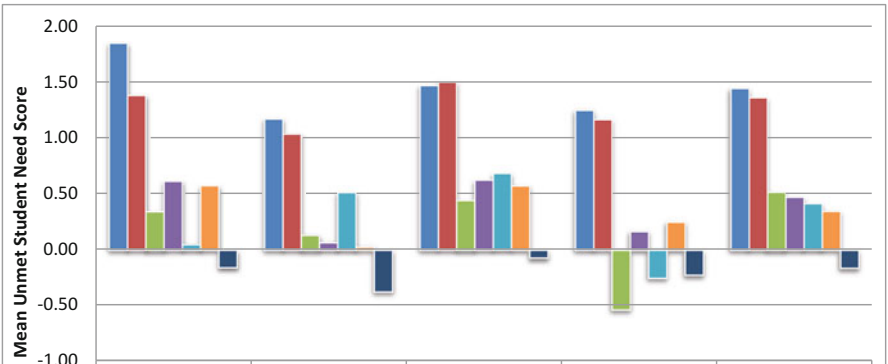
Model		Unstandardized coefficients		Standardized coefficients	T	*p*
		B	SE	Beta		
1	(Constant)	5.873	1.249		4.702	0.000
	Gender	0.073	0.273	0.019	0.266	0.791
	Ethnicity/race	0.127	0.143	0.062	0.891	0.374
	Grades	−0.065	0.051	−0.089	−1.274	0.204
	Class level	−0.033	0.124	−0.018	−0.265	0.791
	International	0.035	0.505	0.005	0.068	0.945
2	(Constant)	7.323	1.183		6.192	0.000
	Gender	−0.100	0.256	−0.026	−0.391	0.697
	Ethnicity/race	0.189	0.134	0.092	1.405	0.162
	Grades	−0.013	0.049	−0.018	−0.266	0.791
	Class level	0.039	0.117	0.022	0.337	0.737
	International	−0.279	0.485	−0.039	−0.575	0.566
	Classroom Learning	**−0.670**	**0.206**	**−0.271**	**−3.247**	**0.001****
	Workload Management	**−0.259**	**0.109**	**−0.186**	**−2.384**	**0.018***
	Faculty Interaction	0.030	0.133	0.021	0.224	0.823
	Sense of Community	−0.104	0.138	−0.073	−0.754	0.452
	Student Interaction	0.104	0.119	0.075	0.879	0.381
	Financial	0.035	0.085	0.033	0.413	0.680
	Skill	−0.199	0.140	−0.113	−1.414	0.159

*$p < 0.05$, **$p < 0.001$

7.5 Example Application

The final version of the ESNQ was used as the basis to understand student needs for a doctoral-granting institution's women in engineering (WIE) and Minorities in (Science) and Engineering Programs (MEP). Although the WIE and MEP programs are at the program level of the institution, the ESNQ can also be used as the basis for developing improvement practices across a range of student populations at various institutional levels.

Once data were collected from the ESNQ, an Importance score, a Satisfaction score, an Unmet Need score, and an Unmet Student Need Index (USNi) were calculated for each questionnaire item, as well as the means scores for each dimension. The Importance score and the Satisfaction score indicated the student's level of satisfaction and importance with the dimensions of student success needs.



	URM Females	Asian Females	White Females	Other Females	All Females
CL	1.85	1.17	1.47	1.25	1.45
WM	1.38	1.04	1.50	1.17	1.36
S	0.34	0.13	0.44	-0.53	0.52
FI	0.62	0.07	0.63	0.17	0.47
F	0.05	0.52	0.69	-0.25	0.42
SC	0.58	0.03	0.58	0.25	0.35
SI	-0.16	-0.37	-0.07	-0.22	-0.16

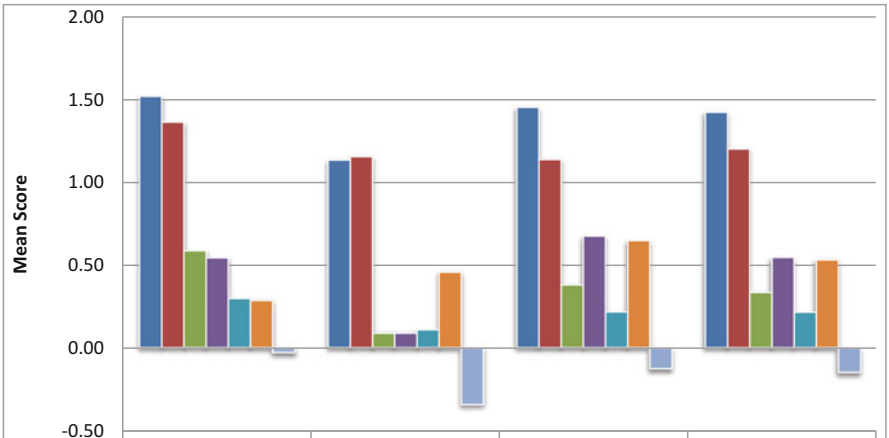
Fig. 7.3 Mean unmet need scores of female engineering students

The *Unmet Need* score was calculated by subtracting the satisfaction score from the importance score (*Unmet Need* score = *Importance* score − *Satisfaction* score).

The *Unmet Need* score provided critical information to decision-makers. This score specified how the needs of students were being fulfilled. A large unmet need score indicated that the institution was not meeting the needs of its students. Conversely, a smaller unmet need score suggested that the institution was meeting the needs of its students. Furthermore, a negative unmet need score implied that the institution was exceeding the needs of the students.

Figure 7.3 shows the results of the mean *Unmet Need* scores for the 105 female engineering students. The highest unmet need scores for the dimensions of *Classroom Learning* ($M = 1.45$, $SD = 0.98$) and *Workload Management* ($M = 1.36$, $SD = 1.24$).

These scores indicate that these needs required the greatest attention because they were not being met by the institution. Conversely, the lowest unmet need score for female engineering students—indicating that the institution was meeting the needs of female engineering students—was the *Student Interaction* Needs ($M = -0.16$, $SD = 1.43$). Moreover, the aggregated results held true across all of the subgroups. The highest unmet need scores were reported by White female engineering students for *Workload Management* ($M = 1.50$, $SD = 1.13$). All subgroups reported high unmet need scores for *Workload Management*; however, the highest reported score for URM female engineering students was *Classroom Learning* ($M = 1.85$, $SD = 1.18$).



	Minority	Asian	White	All Students
CL	1.51	1.13	1.45	1.42
WM	1.36	1.15	1.13	1.19
F	0.58	0.09	0.38	0.33
FI	0.54	0.09	0.67	0.54
SC	0.30	0.11	0.22	0.21
S	0.28	0.45	0.64	0.53
SI	-0.03	-0.34	-0.13	-0.15

Fig. 7.4 Mean unmet need scores (ethnicity/race)

Figure 7.4 shows the *unmet needs* scores across ethnicities/races for the MEP program. Similar to the female engineering students, each group reported that their needs were not being fulfilled in the needs dimensions of *Classroom Learning* and *Workload Management*. Among the ethnicities/races, URM students reported the highest levels for the needs dimensions of *Classroom Learning* ($M = 1.51$, $SD = 1.07$) and *Workload Management* ($M = 1.36$, $SD = 1.46$).

Table 7.5 provides an example of how the ESNQ was used to design institutional practices to address the needs of the students. A focus group with six MEP students and seven WIE students was conducted. The function-means tree was extended to include needs (functions), design characteristics, current practices, and improvement practices (means).

Instead of designing specific improvement practices, abstraction was used to emphasize the essential characteristics of student success practices. Therefore, students were asked to identify design characteristics of student success practices that fulfilled student needs. According to Pahl and Beitz (2013), this process provides participants an opportunity to search for and develop other solutions that contain the identified characteristics. This approach also supports systematic thinking and creativity, thereby opening up the solution space to allow participants to consider range of ideas without being fixated on traditional ideas. Additionally,

Table 7.5 Improvement practices matrix

ESNQ unmet need	To have a manageable workload (WIE)		To stimulate interest in your field (MEP)	
Characteristics	Doable work	Help coping with stress	Provide more exposure	Real world immersion
Practices	Office Hours Informal Study Groups	Advisors/Mentors Advice	Career Fairs Companies coming to campus	Study (Engineering) Abroad
Strengths	Save time by asking questions. Better information with informal interaction	Gain Knowledge from others' experience	Lots and Diverse	Well-Advertized Good program Well liked
Shortcomings	Professors are not available. Sometimes it takes longer to get something done with a larger group	Doesn't really address coping with stress directly	May not get hired just networking	Many students are not able to participate
Improvements	Offer formal and flexible tutoring to everyone	Tips sent out over email on how to cope/ stress management workshop	Help us (i.e., students) create personal relationships with industry	Field trips to research labs on campus

analyzing existing practices was also useful in stimulating new ideas. Therefore, incorporating these elements into the Improvement Practice Matrix in Table 7.5 was intended to guide the focus group through the process of synthesizing information to develop improvement practices that could address the needs of students.

7.6 Summary and Key Principles

This research is intended to bridge the gap between theory and practice. Critical to this approach is the development of tools that focuses on the conceptual system design stage of the systems engineering life cycle. Conceptual system design represents a process that begins with the identification of a need and progresses through a series of phases to generate optimum solutions to satisfy the identified need (Blanchard and Fabryky 2017). This process results in an action plan that links working steps to design phases that can be adapted in a flexible manner to solve complex problems (Pahl and Beitz 2013).

Although recent research has been conducted to improve the design process, many approaches are either too abstract or too narrowly focused. Thus, it makes it difficult to apply them to a nontechnical area (i.e., student success), which has been largely unexplored in the system design literature. Compounding the problem is the fact that there is a plethora of methods, processes, and tools that facilitate only portions of the conceptual system design. Given that these approaches were developed in technical domains, little structure, and guidance is available to institutional decision-makers who wish to employ the conceptual design process as a whole in higher education settings. These significant gaps highlight the need to integrate improved methods and tools into a unified framework that can guide institutional leaders in designing effective practices that facilitate student success, while at the same time meeting the unique needs of their students.

This study described the development, validation, and application of the ESNQ. The development process consisted of three steps. First, the domain of constructs was determined using the typology of student success factors that was developed based on a literature review. Second, a participatory design method was used to conduct four meetings with 21 students to develop the initial pool of questionnaire items. Based on the results of the meetings, the questionnaire layout and questionnaire items were developed to specify actionable need statements. Furthermore, the questionnaire layout was designed to assess the students' importance and satisfaction with the institution in meeting their needs.

Then, the questionnaire validation process was presented to assess both the validity and reliability of the ESNQ. Two expert panel reviews were conducted with four students and eight student success practitioners, respectively, to purify the instrument. As a result of these reviews, a final questionnaire emerged for pilot testing. A pilot test was conducted with 241 participants to collect data to assess the reliability and validity of the ESNQ. As a result of these steps, a conceptual research model emerged that defined the initial variables, which were subsequently refined into a research model that was developed and validated in order to identify student success needs that relate to student satisfaction. This model led to the development of a new ESNQ, which was used to assess the student success needs of engineering students. This questionnaire shifted the current paradigm of student success theoretical research from trying to understand *why* students decide to leave/stay in college to *understanding the needs of engineering students*.

In this study, the model was tested based on the responses of undergraduate engineering students at a doctoral granting institution. The dimensions of Classroom Learning and Workload Management needs were significantly related to Overall Satisfaction. However, the dimensions of Faculty Interaction, Sense of Community, Student Interaction, Financial, and Skill needs did not demonstrate a statistically significant relationship with Overall Satisfaction. These results suggest that the institutional leaders at the research site can use the findings to target improvements related to classroom learning and workload management.

7.6.1 Key Principles

At the core of the proposed approach is a design philosophy that is based on five fundamental principles that have been adapted from the systems engineering and student success theoretical perspective:

1. Student Orientation: Satisfying the needs of students is the driving force behind the design of institutional practices.
2. Analysis-Synthesis-Evaluation: Structures the improvement process based on core phases of conceptual design to ensure that student needs are identified and solutions are developed to improve institutional practices that meet their needs
3. Participatory: Requires a team approach that empowers the institutional leaders and students to be actively involved in the design of improvement efforts
4. Holistic Framework: Provides a unifying structured framework to guide institutional leaders throughout the translation of student needs into a plan of action
5. Vital to Student Success: Focuses the design process on those aspects that are critical to student success.

7.7 Future Work

Based on the results, future research will utilize the ESNQ to help decision-makers identify and prioritize the needs of engineering students as the basis for developing a framework of action to facilitate student success. Future work can be devoted to two areas

7.7.1 Expanding the Operationalization of Student Success

This research focused on student satisfaction, which is an often-overlooked outcome that is useful in determining the quality of the educational experience (Kuh et al. 2006), as a measure of student success. As Braxton et al. (2006) noted, although there are several measures of student success, the most frequently cited theories define student success in college in terms of persistence, educational attainment, and obtaining a degree (Kuh et al. 2006). Future research should investigate the relationship between satisfying student needs and additional measures of student success (e.g., attrition rates, persistence rates, GPA, graduation rates).

7.7.2 Strengthening the Validity and Reliability of the ESNQ

Although the ESNQ met the recommended thresholds for reliability and validity, additional items can be tested to strengthen the dimensions of Classroom Learning

and Skill needs. Additionally, test-retest reliability can also be used (in addition to Cronbach Alpha's coefficient used in this research) to administer the instrument to the same respondents on different occasion. By using this statistical technique to determine if the correlation between the two administrations is high, the reliability of the instrument reliability can be strengthened. Furthermore, the cultural validity of the instrument should be assessed with a larger sample size in order to assess the unique needs of women and underrepresented student populations. By doing so, future research can examine the socio-cultural influences that shape how URM and female engineering students make sense of the ESNQ items and respond to them (Solano-Flores and Nelson-Barber 2001).

References

Astin AW (1984) Student involvement: a developmental theory for higher education. J Coll Stud Pers 25:297–308

Bartlett MS (1954) A note on the multiplying factors for various x2 approximation. J R Stat Soc 16:296–298

Bean JP (1980) Dropouts and turnover: The synthesis and test of a causal model of student attrition. Res High Educ 12(2):155–187

Bean JP (1983) The application of a model of turnover in work organizations to the student attrition process. Rev High Educ 2:129–148

Bean JP, Eaton SB (2000) A psychological model of college student retention. In: Braxton JM (ed) Reworking the student departure puzzle. Vanderbilt University Press, Nashville

Berger JB, Lyon SC (2005) Past and present: a historical view of retention. In: Seidman A (ed) College student retention: formula for student success. Praeger, Westport

Besterfield-Sacre M, Moreno M, Shuman L, Atman CJ (2001) Gender and ethnicity differences in freshman engineering student attitudes: A cross-institutional study. J Eng Educ 90(4):477–490

Blanchard BS, Fabrycky WJ (2017) Systems engineering and analysis, 6th edn. Prentice Hall, Englewood Cliffs, NJ

Bowman NA (2010) The development of psychological well-being among first-year college students. J Coll Stud Dev 51(2):180–200

Braxton J, McKinney J, Reynolds P (2006) Cataloging institutional efforts to understand and reduce college student departure. In: John ES (ed) Improving academic success: using persistence research to address critical challenges, New directions for institutional research. Jossey-Bass, San Francisco

Chen X (2013) STEM attrition: college students' paths into and out of STEM fields. Statistical analysis report. NCES 2014–001. National Center for Education Statistics, Washington, DC

Churchill GA (1979) A paradigm for developing better measures of marketing constructs. J Mark Res 16:64–73

Clark LA, Watson D (1995) Constructing validity: basic issues in objective scale development. Psychol Assess 7(3):309–319

Cohen J, Cohen P (1983) Applied multiple regression/correlation analysis for the behavioral sciences, 3rd edn. Erlbaum, Hillsdale

Cohen P, West SG, Aiken LS (2014) Applied multiple regression/correlation analysis for the behavioral sciences. Psychology Press, New York

DeVillis RF (1991) Scale development: theory and applications. Sage, Newbury Park

Eris O, Chen H, Bailey T, Engerman K, Loshbaugh HG, Griffin A, Lichtenstein G, Cole A (2005) Development of the Persistence in Engineering (PIE) survey instrument. In: Proc. Amer. Soc. Eng. Educ, p 1

Hair JF, Anderson RE, Tatham RL, Black WC (1998) Multivariate data analysis. Prentice Hall, Upper Saddle River

Kaiser HF (1970) A second generation little jiffy. Psychometrika 35:401–415

Kuh GD (2001) Assessing what really matters to student learning: inside the national survey of student engagement. Change 33(3):10–17

Kuh GD (2009) The national survey of student engagement: conceptual and empirical foundations. New Directions for Institutional Research 141:1–20

Kuh GD, Kinzie J et al (2006) What matters to student success: a review of the literature. National Postsecondary Education Cooperative, Bloomington

Lent RW, Sheu H, Schmidt J, Brenner BR, Wilkins G, Brown SD, Gloster CS, Schmidt LC, Lyons H, Treisteman D (2005) Social cognitive predictors of academic interests and goals in engineering: utility for women and students at historically Black universities. J Couns Psychol 52(1):84–92

Moore GC, Benbasat I (1991) Development of an instrument to measure the perceptions of adopting an information technology innovation. Inf Syst Res 2(3):192–222

Netemeyer RG, Bearden WO, Sharma S (2003) Scaling procedures: issues and applications. Sage Publications, Thousand Oaks

Nora A (2003) Access to higher education for Hispanic students: real or illusory? In: Castellanos J, Jones L (eds) The majority in the minority: expanding representation of Latino/a faculty, administration and students in higher education. Stylus Publishing, Sterling, pp 47–67

Nunnally JC (1978) Psychometric theory. McGraw-Hill, New York

Pahl G, Beitz W (2013) Engineering design: a systematic approach. Springer Science & Business Media, Berlin

Pascarella ET, Terenzini PT (1980) Predicting freshman persistence and voluntary dropout decisions from a theoretical model. J High Educ 51:60–75

Pérez D II, Ashlee KC, Do VH, Karikari SN, Sim C (2017) Re-conceptualizing student success in higher education: reflections from graduate student affairs educators using anti-deficit achievement framework. J Excell Coll Teach 28(3):5–28

Schreiner LA, Juillerat SL (2010) Sample student satisfaction inventory: 4-year college or university version. Noel Levitt, Iowa City

Solano-Flores G, Nelson-Barber S (2001) On the cultural validity of science assessments. J Res Sci Teach 38(5):553–573

St. John EP, Paulsen MB, Carter DF (2005) Diversity, college costs, and postsecondary opportunity: an examination of the financial nexus between college choice and persistence for African Americans and Whites. J High Educ 76(5):545–569

Tabachnick BG, Fidell LS (2007) Using multivariate statistics, 5th edn. Allyn and Bacon, New York

Tinto V (1993) Leaving college: rethinking the causes and cures of student attrition. The University of Chicago Press, Chicago

Tinto V (2006–2007) Research and practice of student retention: what next? J Coll Stud Retent 8(1):1–19

Tinto V (2010) From theory to action: exploring the institutional conditions for student retention. In: Smart J (ed) Higher education: handbook of theory and research, vol 25. Springer, Dordrecht

Tinto V, Pusser B (2006) Moving from theory to action: building a model of institutional action for student success. In: Commissioned paper presented at the 2006 Symposium of the National Postsecondary Education Cooperative (NPEC)

U.S. Department of Education, National Center for Education Statistics (2017) The condition of education 2017 (NCES 2017-144), Undergraduate retention and graduation rates

Dr. Tracee Walker Gilbert has always had a passion for solving complex problems. What initially began as an interest in physics, eventually evolved into a way of life; characterized by seeing the world from a systems engineering perspective. By bridging the principles of engineering and management, systems engineering is an interdisciplinary field that provides a holistic and integrated approach to solving complex problems more efficiently and effectively. Dr. Gilbert is a systems thinker at heart, and chose this career field because she enjoys leading teams to optimize processes, design new systems, and improve practices. She was introduced to the field during her first job out of college working at Lockheed Martin in an engineering leadership development program.

Dr. Tracee Walker Gilbert is currently a passionate entrepreneur and systems engineering executive. Dr. Gilbert owns and operates System Innovation, LLC, which provides systems engineering and program management services to the Office of the Secretary of Defense (OSD). She has over 19 years of experience leading large-scale initiatives and driving strategy for systems engineering research and engineering programs across various domains including: defense, homeland security, medical and public health, commerce/census, and the education sector. She has also held various leadership positions at Lockheed Martin, MITRE, Engility, and served as a fellow for the American Association for the Advancement of Science (AAAS) Science and Technology Policy Fellowship (STPF) at Health and Human Services (HHS) and the Department of Defense (DoD).

She serves on the AAAS STPF Advisory Committee and also has a personal commitment to excellence, integrity, and motivating women and minorities to succeed in Science, Technology, Engineering, and Math (STEM) fields. She received her B.A. (Physics, Minor Japanese) from Lincoln University and her M.S. (Systems Engineering), and Ph.D. (Industrial and Systems Engineering) from Virginia Tech. She also studied abroad with the Institute for the International Students in Tokyo, Japan, while working as a staff editor for Tokyo Classifieds.

Janis Terpenny is the Peter and Angela Dal Pezzo Chair and Department Head of the Harold and Inge Marcus Department of Industrial and Manufacturing Engineering at Penn State University. She is also the director of the Center for e-Design, an NSF industry/university cooperative research center (I/UCRC) that brings 7 universities and 20+ industry and government agency members together in solving pressing problems associated with the design, manufacture, delivery and sustainment of innovative, high quality, and lower cost products.

Dr. Terpenny's career journeyand draw to STEM and engineering comes from her strong desire to understand and solve real problems and to make a difference in the world. She enjoys collaborations, particularly those that bring diverse perspectives and backgrounds together. Dr. Terpenny has always enjoyed the breadth and focus of industrial engineering always seeking to make things better with an emphasis on people, processes, and systems and the applicability of tools and methods of industrial

engineering to virtually every application domain (healthcare, manufacturing, entertainment, transportation, and more). As a professor, she enjoys bringing industry, government, and universities together, bridging research to practice, and the opportunity to educate, mentor, and inspire others.

Dr. Terpenny's research interests are focused on engineering design and smart manufacturing, including topics such as process and methods of early design; knowledge and information in design; product families and platforms; obsolescence in products and systems; complexity of products and systems; and cloud computing for design and manufacturing integration. She has also investigated topics in design education such as multidisciplinary teams; the impacts of project choice and context on engagement and learning; and the retention and success of underrepresented students. Industry and community partnerships are key to the impact of her research, teaching, and service. Prior to joining Penn State, Dr. Terpenny served as the department chair and Joseph Walkup Professor of Industrial and Manufacturing Systems Engineering at Iowa State University. She also served as the technology lead for the Advanced Manufacturing Enterprise (AME) area for one of the nation's first manufacturing institutes: Network for Manufacturing Innovation (NNMI) and the Digital Manufacturing and Design Innovation Institute (DMDII). She has also served as a program director at the National Science Foundation (NSF) and has been a professor at Virginia Tech and at the University of Massachusetts Amherst. She has 9 years of industry work experience with the General Electric Company (GE), including the completion of a 2-year corporate management program. Throughout her career, she has served as PI or co-PI on over $16 million in sponsored research and is the author of over 180 peer-reviewed journal and conference publications. She is a Fellow of IISE, a Fellow of ASME, and a member of ASEE, INFORMS, Alpha Pi Mu, and Tau Beta Pi. She is currently the chair of the Council of Industrial Engineering Academic Heads (CIEADH), Chair of the Oversight Committee for the 2018 Institute of Industrial and Systems Engineering (IISE) Annual Conference, and serves as an associate editor for the *Engineering Economist*.

Tonya Smith-Jackson was every advisor's nightmare. She loved every subject, changed her major three times, and even in her junior year, she was not certain what she wanted to do. Participation in an undergraduate summer research internship that combined social science, computer programming, and organizational design and management led her to an undergraduate senior thesis on macroergonomics, technology and work, and graduate school in psychology and industrial engineering. Dr. Smith-Jackson has been employed in industry, government, and academia in NC, NY, VA, and Germany. She was a human factors interface specialist at Ericsson Mobile Communications (now Sony Ericsson) and at IBM's human factors and network systems group. She has worked in manufacturing and in human

services as well. She chose ISE and Human Factors because these knowledge domains are interdisciplinary and demonstrate the wisdom of integrating academic cultures, applying systems thinking, and celebrating a diversity of minds. She chose academia because, above all, her vision is to facilitate the development and success of future engineers and scientists in an accessible and equitable way.

Tonya Smith-Jackson, Ph.D., CPE is a Professor, Chair, and Graduate Program Director of the Department of Industrial and Systems Engineering at NC A&T State University in Greensboro, NC. Dr. Smith-Jackson earned her graduate degrees from NC State; her Ph.D. in Psychology/Ergonomics and M.S. in Psychology/Ergonomics/Interdisciplinary Industrial Engineering from NC State University. She earned a B.A. in Psychology from UNC-Chapel Hill. She graduated from the first inaugural class of the NC School of Science and Mathematics. She is the director of the Human Factors Analytics Lab, Director of the Center for Advanced Studies in Identity Sciences, and co-director of the laboratory for Cyber-Human Analytics Research for the Internet of Things (CHARIoT). Her current research areas focus on cyber-human systems to empower caregivers, allow older people to live independently, provide support for security and privacy, and protect the nation's cyber infrastructure.

She is currently a program director at the National Science Foundation (NSF) in the Computer and Information Science and Engineering (CISE) Directorate and the Information and Intelligent Systems Division. This position is part of NSF's rotator program and allows her to serve NSF while on temporary leave from her university, NC A&T State University, where she is a professor and Chair of the Department of Industrial and Systems Engineering.

Chapter 8
A Study of Critical Thinking and Cross-Disciplinary Teamwork in Engineering Education

Hulya Julie Yazici, Lisa A. Zidek, and Halcyon St. Hill

Contents

8.1 Introduction 185
8.2 Background 187
8.3 Methodology 188
8.4 Results and Discussion 189
8.4.1 Engineering Students' Thinking Styles 189
8.4.2 Critical Thinking Performance 190
8.5 Conclusions 192
References 193

8.1 Introduction

Critical thinking is the intellectually disciplined process of actively and skillfully conceptualizing, applying, analyzing, synthesizing, and evaluating information (Facione 2000). Emerging business intelligence and analytics requires the ability to understand and interpret the burgeoning volume of data (Chiang et al. 2012; Chen et al. 2012). A report by the McKinsey Global Institute (Manyika 2011; Siemens and Long 2011) predicted that by 2018, the United States alone will face a shortage of 140,000–190,000 people with deep analytical skills (Yazici 2016). Along with mastery of subject, critical thinking ability is among the learning outcomes adopted

H. J. Yazici (✉)
Lutgert College of Business, Florida Gulf Coast University, Fort Myers, FL, USA
e-mail: hyazici@fgcu.edu

L. A. Zidek
U.A. Whitaker College of Engineering, Florida Gulf Coast University, Fort Myers, FL, USA
e-mail: lzidek@fgcu.edu

H. St. Hill
Marieb College of Health and Human Services, Florida Gulf Coast University, Fort Myers, FL, USA
e-mail: hsthill@fgcu.edu

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_8

and reinforced by ABET (2015) and AACSB (2013). Furthermore, in addition to strong analytical skills fundamental to engineering practice, a diverse set of interdisciplinary skills are required (Lattuca et al. 2017). As acknowledged in the Engineer of 2020 (National Academy of Engineering 2004), and emphasized in ABET (EC2000 criteria), the increasingly interdisciplinary nature of engineering practice calls for preparing engineers to work in cross-disciplinary teams and settings.

As engineering curriculum needs to be reassessed to focus more on skills, capabilities, and techniques as well as cultivating ethical values and attitudes, more research is needed to understand what contributes to critical thinking skills, and overall to higher academic achievement. The study of critical thinking skill in engineering education is lacking. An underlying theme is that critical thinking is not taught, rather it is developed through experiential learning and systematic approaches to problem solving.

Furthermore, previous research in engineering education brought the significance of successful collaboration of engineers with non-engineers, or social scientists (Borrego and Newswander 2008; Borrego et al. 2014; Lattuca et al. 2017). In engineering, the emphasis on multidisciplinary teamwork in accreditation criteria has contributed to the interest in multidisciplinary and interdisciplinary learning and competencies. Research in interdisciplinary learning examined the effect of interdisciplinary experiences on students' development of generic cognitive skills, such as critical thinking, problem solving, and creativity (Lattuca et al. 2017; Borrego et al. 2014). As Lattuca et al. (2017) concluded, there is a further need to investigate the value of interdisciplinary education in engineering and identify educational experiences that support the positive effects of these experiences.

Higher order thinking skills are affected by students' experiences, beliefs, and attitudes (Greeno et al. 1996; Johri et al. 2014; Terenzini and Reason 2010; Lattuca et al. 2017). Thus, learning and thinking styles may explain how an individual performs as a critical thinker, comprehends subject knowledge and constructs knowledge when faced with real-life scenarios (Yazici 2004, 2005, 2016). Although learning and thinking styles have been addressed for decades, the question arises as to how they relate to critical thinking skills.

Thus, based on the questions raised above, the purpose of this chapter is to explore the relationship of thinking styles and critical thinking performance when engineering students are exposed to an education experience within a cross-disciplined team. Students from engineering and non-engineering disciplines participated in an innovative assignment which consists of analyzing, critiquing, and re-designing a process using a case study. Process analysis methods with value stream mapping approach were applied.

This chapter describes the assignment, the critical thinking performance of engineering students in association with their thinking styles and in relation to education utilizing cross-disciplinary teamwork.

8.2 Background

Critical thinking involves scrutinizing, differentiating, and appraising information as well as reflecting on the information that will be used to make judgments (Banning 2006; Behar-Horenstein and Niu 2011). Critical thinking processes require active argumentation, initiative, reasoning, envisioning, analyzing complex alternatives (Simpson and Courtney 2002). Thus, critical thinking skills require self-correction and reflexivity (Behar-Horenstein and Niu 2011). This was earlier indicated by Sternberg (1997) by defining critical thinking as mental self-governance allowing individuals to use experienced and value judgments. Ability to raise significant questions and problems clearly and concisely, acquiring and interpreting information and formulating well-reasoned solutions and conclusions were defined as traits of critical thinker (Paul and Elder 2010).

Critical thinking is the basis for reasoned decision-making and is therefore central to engineering education and practice. Although engineers are expected to have the ability to use math and science in their thinking, this thought process is not oriented toward theory, but design and discovery (Sheppard et al. 2008). As Bonney and Sternberg (2016) pointed out, and studied in detail by Agdas (2013), one of the significant tasks of the instructor is to teach students how to learn and become critical thinkers not solely transferring the knowledge. This very well agrees with the Blooms Taxonomy. Bloom (1956) explains that mastery of concepts occurs, when learners are able to advance through the six hierarchical levels: knowledge, comprehension, application, analysis, synthesis, and evaluation. In so doing, learners are able to identify issues, define and evaluate possible solutions, and communicate problems through critical thinking. As an example, active learning, or PBL (problem-based learning) are utilized to teach students' how to learn materials and relate the learned content to problem solve and hence demonstrate the ability to think critically.

Several inventories exist to measure critical thinking. California Critical Thinking Disposition Inventory(CCTDI), California Critical Thinking Skills Test (CCTST), Watson and Glaser's Critical Thinking Appraisal, or the American Association of Colleges and Universities (AACU)'s VALUE rubric are used by previous researchers (Agdas 2013; Behar-Horenstein and Niu 2011; Ghanizadeh 2017; Yazici 2016). The AAC&U Critical Thinking VALUE Rubric (2009) measures critical thinking performance via five criteria: explanation of issues, evidence, influence of content and assumption, student's position (perspective, thesis/hypothesis), and conclusion (implication, consequence).

Another significant element of teaching critical thinking is how engineers collaborate with engineers and non-engineers incorporating their expertise relevant to the problem at hand and the importance of collaboratively seeking the most optimum solution. So, reasoned decisions are rooted in knowledge from various sources and backgrounds that are interrelated, which in turn requires critical thinking. Although critical thinking is significant within and across disciplines, there is a need to address causal links associated with thinking styles and student learning outcomes

relevant to critical thinking. Lattuca et al. (2017) concluded the critical role of the interdisciplinary thinking and habits of mind, along with the significance of co-curricular activities that bring engineering students with non-majors to build interdisciplinary competence.

Engineering education has expanded in recent years to include not only the study of interdisciplinary educational conditions and outcomes, but also studying the design of interdisciplinary courses (Newswander and Borrego 2009; Coso et al. 2010; Boden et al. 2011). Several studies examined learning experiences of limited duration and with small group of students, or in the context of an assigned performance task. Lattuca et al.'s study (Lattuca et al. 2017) focused on interdisciplinary competence in a broad cross section of engineering students (Borrego and Newswander 2008) defined engineering education collaborations as multidisciplinary and interdisciplinary. A truly interdisciplinary collaboration occurs when researchers or disciplines join to work in a common question or problem with a continuing interaction afterwards, rather than splitting apart. They reported increased satisfaction and quality of work fostering form interdisciplinary teamwork.

In engineering education, personal and social experiences, beliefs and preferences are considered as possible factors of learner success. Similar to leaning preferences, thinking styles may show the differences among learners in adapting to leaning environments, when engaged in critical thinking activities. Watson and Glaser (Watson and Glaser 1980, 2002) Critical Thinking Appraisal and California Critical Thinking Disposition Inventory are used by several researchers (Behar-Horenstein and Niu 2011; Agdas 2013; Ghanizadeh 2017). Watson–Glaser also developed thinking styles inventory (Watson and Glaser 1980). Thinking styles are positive habits that contribute to better critical thinking, problem solving, and decision-making. While no one thinking style is better than another, a balance of the various styles results in better decision-making. After completing the assessment, individuals receive a report that shows them what thinking styles they prefer and gives a clear picture of their strengths and weaknesses in decision-making. Armed with this insight, they can learn to properly balance the use of all the thinking styles and, ultimately, become better critical thinkers and decision-makers. Watson–Glaser thinking styles rank the preferred styles under seven categories: A*nalytical, insightful, inquisitive, open minded, systematic, timely, and truth seeking.*

8.3 Methodology

To explore the role of cross-disciplinary teamwork and individual thinking styles on critical thinking performance, senior bioengineering students ($n = 49$) enrolled in a healthcare engineering course are studied over the course of two semesters. The course covered several topics including applied statistics, and an overview of operations management as applied to healthcare and healthcare policy.

The critical thinking assignment involved a healthcare case study to improve the flow of maternity patients (Heizer and Render 2014, Principles of OM, 9th edition, page 293). The instructor covered the relevant material related to process analysis, value stream mapping and process management concepts and methods. An online course module is designed specifically for the assignment by the engineering and non-engineering discipline instructors. First, engineering students are asked to individually identify the issues of the current process, recognize the assumptions, and provide evidence from the case. Then, engineering students are placed randomly into group of five with non-engineering students to share their views of the issues of the process and brainstorm improvement plans via online discussion forums. Following their interaction with non-engineering students, engineering students with their cross-disciplinary group members developed an improved process flow, described the improvements made by supporting with value stream mapping metrics, facts from the case, and information from external sources. Then, students provided their final recommendation. As students from engineering and non-engineering disciplines worked on an assignment and in short term, according to Borrego and Newswander (2008), this is considered a multidisciplinary education experience, rather than interdisciplinary.

Prior to the assignment, students took the Watson and Glaser thinking styles appraisal based on 90 questions about their approach for thinking with a scale of: 3: clearly describes me; 2: somewhat describes me; 1: describes me a little; 0: does not describe me. As a result, Watson generates learners' thinking styles in ranked order, number 1: the top ranked learning style and 7: the lowest ranked learning styles, and the numbers between.

8.4 Results and Discussion

8.4.1 Engineering Students' Thinking Styles

Based on Watson–Glaser (Watson and Glaser 1980) thinking style inventory results, engineering students ranked high to medium (ranks 1–4) as *open minded* (intellectually tolerant and fair minded), systematic (conceptual, process oriented, and intuitive), *insightful* (prudent, humble, reflective, and strategic), *inquisitive* (curious, alert, and interested in the surrounding world), and *analytical* (clear thinking, orderly, and rational). Engineering students were ranked low (ranks 5–7) as *truth seeking* (independent, tough minded, and skeptical), and *timely* (efficient, reliable, and responsive).

Figure 8.1 shows the percentage of highest frequencies for all thinking styles. 78% of learners were ranked high to medium (ranks 1–4) as Open minded and as Systematic, 63% ranked high to medium as Analytical, and 56% as Inquisitive. 75% of learners ranked low (ranks 5–7) as Timely and 69% ranked low (ranks 5–

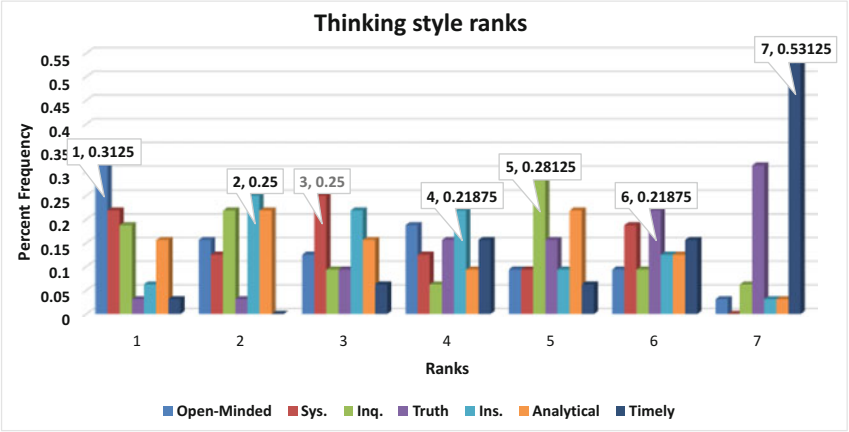


Fig. 8.1 Percent frequencies of thinking style ranks

7) as Truth Seeking. These results are comparable to previous research findings with engineering students. For instance, Agdas (2013) found that the mean score of Analyticity, Inquisitiveness, Open mindedness, and Systematicity was higher compared to Truth Seeking and Maturity of Judgment dispositions measured with CCTDI where subjects were senior Civil Engineering students.

8.4.2 Critical Thinking Performance

Learners' critical thinking performance is assessed based on AAC&U VALUE rubric. Two points were assigned for each measure of individual critical thinking performance and three points per measure for the group critical thinking. Among the critical thinking measures, learners did well (mean score of 1.91/2) in terms describing the problem, explaining the process issues, showing a correct and complete process flow, and providing evidence (1.75/2) from the case and other sources to defend their point of view. This showed that students performed well in analyzing the problem and data presented to them. However, when students worked on the improvement suggestions, citing external sources for possible solutions (1.54/2), defending their point of view on how the process can be improved (1.53/2), and providing a thorough synthesis of the problem and alternative solutions (1.46/2), a lower performance was detected. This is a deficiency consistently observed throughout the programs that needs to be improved.

Following their individual assignments, students worked with their cross-disciplined group members. Engineering students were able to discuss, brainstorm, and share information via online discussion forums. Critical thinking skills following the interdisciplinary education experience are comparable to individual

Table 8.1 Engineering students' critical thinking performance results

	Explanation of issues	Evidence	Influence of context and assumptions	Student's position	Conclusions and related outcomes
Individual mean, standard deviation	1.91, 0.19	1.75, 0.28	1.54, 0.31	1.53, 0.39	1.46, 0.38
Cross-disciplined mean, standard deviation	1.54, 0.49	1.31, 0.60	1.37, 0.49	1.37, 0.5	1.41, 0.46

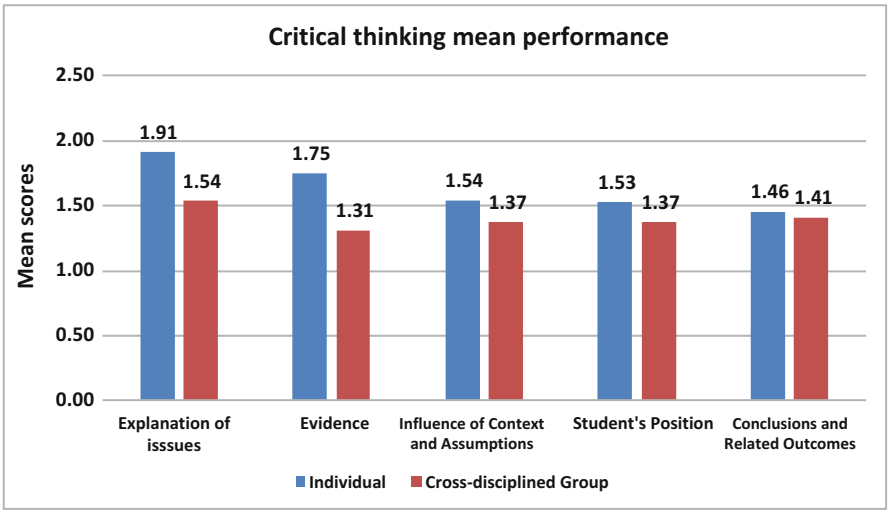


Fig. 8.2 Critical thinking individual and cross-disciplined group performances

critical thinking performance, but lower (t: 2.56, $p < 0.02$). Analyzing each measure of critical thinking, Table 8.1 and Fig. 8.2 summarize the individual and cross-disciplined group critical thinking performance mean scores.

Engineering students' individual critical thinking performance was higher compared to non-engineering students. But, when students worked with their cross-disciplined groups, this did not positively contribute to learners' performance. This may be due to the online setting of the assignment with minimum face-to-face interaction, or in some cases, group members relying on other students' work, a typical downside of student teams. It is also likely that the cross-disciplined experience was limited to one assignment only, so not a long-term interdisciplinary experience.

The challenges with cross-disciplinary teamwork were also identified in a combined engineering/business course. Junior level engineering students taking engineering entrepreneurship class teamed up with senior business students enrolled in a business planning class finding a solution to the need developer prototype and developing a business plan ensuring that there was a market for the solution. Student

team evaluations indicated a propensity to assign tasks with little to no collaborative effort. Engineering students also reported that the cross-disciplinary nature of the class detracted from the learning environment. Thus, providing further evidence that the engineering students do not positively engage in interdisciplinary teamwork.

Some of the research results were presented at previous conferences, chronologically, at ASAHP (St. Hill et al. 2015), in POMS (Yazici and St. Hill 2016b), in the IISE Annual Conference and Expo (Yazici and St. Hill 2016a), and in the IISE Annual Conference and Expo (Yazici 2017) in the Engineering Education Track. During the 2017 IISE, Interactive Education session, the audience consisting of academicians, industry practitioners/consultants, and engineering students raised important points about critical thinking:

1. At course level, educators indicated exposing students to critical thinking assignments and the use of engineering reasoning guidebook. At capstone level, this was done through thesis or capstone project report. Among the challenges, getting faculty to consistently apply the critical thinking model, and finding ways to persuade the leadership to address CT in programs school wide was reported.
2. Industry perspective emphasized strongly the importance of defining a problem, recognizing issues and current state with all the facts prior to implementing solutions, knowing how to gather evidence and how to work in teams. It is also indicated that recent grads usually need a couple of years of experience/training before they fully function as critical thinkers. Integrating critical thinking training into subject-based staff training will be useful.
3. From student perspective, important points were made: Critical thinking can be an asset and gives headway by identifying problems that may be overlooked by the management. By communicating with industry and working on research projects with industry educational institutions can adapt to industry needs.

8.5 Conclusions

The results of this study indicate a deficiency in engineering education with respect to interdisciplinary skills. Critical thinking and interdisciplinary teamwork skills are critical to professional engineers. While engineering students performed well in problem definition and process identification skills that map directly to the engineering design process, they performed poorly in collaborative solutions.

The results are similar to self-reported peer evaluations in a separate course which paired engineers and business students. In both cases, engineering students contributed to the problem definition, but took little to no credit for solution development, synthesis, and justification. While the data indicates that engineering students have well-developed critical thinking skills, interdisciplinary teamwork is lacking. As previous research showed (Ghanizadeh 2017; Lattuca et al. 2017), higher order thinking skills, reflective thinking, and self-regulation significantly influence academic achievement, thus, engineering students need to be systematically exposed to these in the curriculum. It is important that these skills are gained

over time with engaging students on long-term basis, and throughout the curriculum. This will play a role in developing interdisciplinary skills and building competence.

Acknowledgements This research was funded by the Multidisciplinary Grant Research Initiative (MDRI, Number 2014–32), Florida Gulf Coast University, Fort Myers, FL.

References

Accreditation Board for Engineering and Technology (ABET) (2015) Criteria for accrediting engineering programs [online]. http://www.abet.org/wp.../2015/05/E001-15-16-EAC-Criteria-03-10-15.pdf. Accessed 18 Dec 2015

Agdas S (2013) Effects of problem-based learning on development of critical thinking skills and dispositions in engineering. In: Abstract of Ph.D. dissertation presented to the Graduate School of the University of Florida, August 2013

Association of American Colleges and Universities (AACU) (2009) Critical thinking VALUE rubric. AACU Critical Thinking Value Matrix. https://www.aacu.org/value/rubrics/critical-thinking

Association to Advance Collegiate Schools in Business (AACSB) (2013) AACSB assurance of learning standards: an interpretation, AACSB White Paper No. 3. http://www.sdabocconi.it/sites/default/.../2_aolwhitepaper_final_11_20_07.pdf. Accessed 3 May 2013

Banning M (2006) Measures that can be used to instil critical thinking in nurse prescribers. Nurse Educ Pract 6:98–105

Behar-Horenstein L, Niu L (2011) Teaching critical thinking skills in higher education: a review of the literature. J Coll Teach Learn 8(2):25–41

Bloom BS (1956) Taxonomy of educational objectives, handbook 1: cognitive domain. Longmans Green, New York

Boden D, Borrego M, Newswander LK (2011) Student socialization in interdisciplinary doctoral education. Higher Education 62:741–755. https://doi.org/10.1007/s10734-011-9415-1

Bonney CR, Sternberg RJ (2016) Learning to think critically. In: Mayer RE, Alexander PA (eds), Handbook of research on learning and instruction, Taylor and Francis

Borrego M, Newswander LK (2008) Characteristics of successful cross-discipline engineering education collaborations. J Eng Educ 97:123–134

Borrego M, Foster MJ, Froyd JE (2014) Systematic literature reviews in engineering education and other developing interdisciplinary fields. J Eng Educ 103(1):45–76

California Critical Thinking Skills Test (CCTST)/Critical Thinking Skills Tests/Products/Home-Insight Assessment (n.d.). http://www.insightassessment.com. Accessed 4 July 2012

Chen H, Chiang R, Storey V (2012) Business intelligence and analytics: from big data to big impact. MIS Quarterly 36(4):1165–1188

Chiang R, Goes P, Stohr EA (2012) Business intelligence and analytics education, and program development: a unique opportunity for the information systems discipline. ACM Trans Manage Inf Syst 3(3):1–13

Coso AE, Bailey RR, Minzenmayer E (2010) How to approach an interdisciplinary engineering problem: characterizing undergraduate engineering students' perceptions. IEEE Frontiers in Education Conference, Washington, DC. https://doi.org/10.1109/FIE.2010.5673313

Facione PA (2000) The disposition toward critical thinking: Its character, measurement, and relation to critical thinking skill. Informal Logic 20(1):61–84

Ghanizadeh A (2017) The interplay between reflective thinking, critical thinking, self-monitoring, and academic achievement in higher education. Higher Education 74:101–114

Greeno JG, Collins AM, Resnick LB (1996) Cognition and learning. In: Berliner DC, Calfee RC (eds) Handbook of educational psychology. Macmillan, New York, pp 15–46

Heizer J, Render B (2014) Principles of operation management, 9th edn. Pearson Publishing, USA
Johri A, Olds BM, O'Connor K (2014) Situative frameworks for engineering learning research. In: Johri A, Olds BM (eds) Cambridge handbook of engineering education research. Cambridge University Press, New York, pp 47–66
Lattuca LR, Knight DB, Ro HK, Novoselich BJ (2017) Supporting the development of engineers' interdisciplinary competence. J Eng Educ 106(1):71–97
Manyika J (2011) Big data: the next frontier for innovation, competition, and productivity, executive summary. McKinsey Global Institute. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation. http://www.mckinsey.com/mgi/publications/big_data/pdfs/MGI_big_data_exec_summary.pdf
National Academy of Engineering (2004) The engineer of 2020: visions of engineering in the new century. National Academies Press, Washington, DC
Newswander LK, Borrego M (2009) Engagement in two interdisciplinary graduate programs. Higher Education 58(4):551–662. https://doi.org/10.1007/s10734-009-9215-z
Sheppard SD, Macatangay K, Colby A, Sullivan WM (2008) Educating engineers: designing for the future of the field. Jossey-Bass, San Francisco
Paul R, Elder L (2010) The miniature guide to critical thinking concepts and tools. Foundation for Critical Thinking Press, Dillon Beach
Siemens G, Long PD (2011) Penetrating the fog: analytics in learning and education. http://www.educause.edu/ero/article/penetrating-fog-analytics-learning-and-education
Simpson E, Courtney MD (2002) Critical thinking in nursing education: literature review. Int J Nurs Pract 8(2):89–98
Sternberg RJ (1997) The concept of intelligence and its role in lifelong learning and success. Am Psychol 52:1030–1037
St. Hill H, Yazici HJ, Zidek L (2015) Innovations in interprofessional education (IPE)—health professions, engineering and business: learning styles, critical thinking, and the impact on education and practice. In: IPE research paper presented at ASAHP 2015 Annual Meeting, Scottsdale, AZ, 28–30 October
Terenzini PT, Reason RD (2010) Toward a more comprehensive understanding of college effects on student learning. In: Paper presented at the Annual Conference of the Consortium of Higher Education Researchers (CHER), Oslo, Norway
Watson G, Glaser EM (1980) Watson-Glaser critical thinking appraisal: forms A and B; manual. Psychological Corporation
Watson G, Glaser E (2002) Watson-Glaser critical thinking appraisal, UK edition: practice test. Psychological Corporation, London
Yazici HJ (2004) Student perceptions of collaborative learning in operations management classes. J Educ Business 80(2):110–118
Yazici HJ (2005) A study of collaborative learning style and team learning performance. Education + Training 47(3):216–229
Yazici HJ (2016) Role of learning style and interactive response systems on student learning outcomes in undergraduate business education. Int J Oper Manag Educ 6(2):109–134
Yazici HJ (2017) Innovative assignment approach in operations course for enhancing critical thinking, engineering education track, interactive presentation. In: IISE Annual Meeting, Pittsburg, PA, 20–23 May 2017
Yazici HJ, St. Hill H (2016a) Assessment of critical thinking learning outcomes in interdisciplinary education. In: IISE Annual Meeting, IS Evaluation and Teaching Tools Track, Anaheim, CA, 21–24 May
Yazici HJ, St. Hill H (2016b) Critical thinking learning outcomes in operations with an interdisciplinary approach. In: POMS 27th Annual Conference, Teaching/Pedagogy in P/OM Track, Orlando, FL, 6–9 May 2016

Hulya Julie Yazici is currently a professor of information systems and operations management at the Florida Gulf Coast University, Lutgert College of Business. She holds a B.Sc. in mining engineering, and an M.Sc. and Ph.D. in engineering management. She worked with the manufacturing and mining industry in North America and Europe for a decade. Her research interests include project management, knowledge management systems, buyer–supplier collaboration, sustainability, and innovation in education. Some of her papers have appeared in the *Information and Management, Journal of Computer Information Systems, Project Management Journal*, *IEEE Transactions of Engineering Management, Engineering Management Journal, Journal of Manufacturing Technology Management, and Simulation and Gaming.* She is a senior member of IIE, DSI, PMI, POMS, and IEEE. Throughout her career, Dr. Yazici taught undergraduate and graduate students in STEM programs as well as in business with a dedicated focus on critical thinking, multidisciplinary learning, and student scholarship.

"I have chosen engineering and engineering management because the design of products and systems were always interesting to me, along with the human element. My industry experience in manufacturing and mining were true assets for my career and I always looked for a balance between academia and business world. I believe STEM has always great to offer to learners by providing strong analytical skills set; and we need to understand how STEM contributes to humanity and life system overall. That is where we see the influence of industrial & systems engineering and engineering management."

Lisa A. Zidek is the Associate Dean for the U.A. Whitaker College of Engineering and an Associate Professor in Bioengineering at Florida Gulf Coast University. She received her Ph.D. in Industrial Engineering Health Care Management from the University of Wisconsin. She has always been drawn to solving problems, specifically systems level challenges, making a career in Industrial Engineering a natural fit. She truly enjoys the diverse applicability of her Industrial Engineering background and has worked in a multitude of industries including health care, manufacturing, banking, non-profit services, and higher education. Dr. Zidek has served as the Vice President of Student Development for the Institute of Industrial Engineers, she currently serves as an ABET Program Evaluator and is active in ASEE. Her research interests are in engineering education, with particular emphasis on engineering entrepreneurial mindset and K-12 initiatives related to developing future engineers. She was selected to participate in the 2009–2010 Florida Campus Compact Engaged Scholarship Fellows program. Dr. Zidek has run many STEM day programs and summer camp programs with a particular emphasis on introducing K-12 students to careers in STEM, specifically Engineering. In her role as Associate Dean, she coordinates K-12 outreach events to encourage students to explore engineering and other STEM-based careers.

Halcyon St. Hill is Professor Emeritus in recognition of her exemplary contributions to the Marieb College of Health & Human Services, Florida Gulf Coast University (FGCU), and for her advancement of the FGCU academic mission through meritorious teaching, scholarship, and service. Dr. Halcyon St. Hill received her doctorate from Rutgers University, a Master of Science in Microbiology from the Waksman Institute, Rutgers University and a B.S. in Health Sciences with a Clinical Laboratory Science (CLS) major from Hunter College, New York. She holds national certification from the American Society for Clinical Pathology and a Florida license in CLS. Dr. St. Hill holds national distinction as a Fellow of the Association of Schools of Allied Health Professions (FASAHP) for outstanding leadership and contributions to health professions, and the Bio-Rad Award for Professional Achievement as a Clinical Laboratory Scientist–Generalist. Other awards include the 2009 McTarnaghan Teaching Award from the FGCU Student Government and the FGCU 2015 Presidential Award. Her academic roles at FGCU include Inaugural Assistant Dean for the Marieb College of Health & Human Services, Professor and founding chair of the Department of Health Sciences, founding director of Continual Learning, and Faculty Senate President and Trustee, FGCU Board of Trustees. Dr. St. Hill's scholarship is noted by numerous intellectual contributions, including juried articles, book chapters, model curriculum—body of knowledge documents for the American Society for Clinical Laboratory Sciences. Dr. St. Hill is recognized on the international level as an advisor and author contributor for the Clinical Laboratory Standards Institute and known for her work in specialized accreditation for the National Accrediting Agency for Clinical Laboratory Sciences. She is known for her expertise in interdisciplinary (interprofessional) education, authorship, and PI on external and internal grant funds from competitive sources including the US Department of Health and Human Services, and a multitude of refereed presentations—some sought both by invitation and peer review in prestigious arenas including Oxford University in the UK.

"I enjoy problem solving, particularly in scientific venues that impact human life, health and wellness. I also firmly believe that this can only be done through disciplinary and interdisciplinary approaches, because no discipline can achieve optimum solutions in a vacuum. This is why I entered the field of Laboratory medicine as a clinical laboratory scientist and pursued graduate education. My ultimate goal to mentor others and make a difference lead me to higher education as a professor, mentor and administrator who worked across disciplines in health sciences and engineering."

Part IV
Health

Chapter 9
Healthcare Teams Can Give Quality Patient Care, but at Lower Environmental Impact: Patient-Centered Sustainability

Janet Twomey and Michael Overcash

Contents

9.1 Patient Centered Sustainability Defined 199
9.2 Model of Analysis and Improvement 201
9.3 Roadmap for Research Community 207
9.4 Conclusions 208
References 208

9.1 Patient Centered Sustainability Defined

Every decision made by a healthcare team has first a patient-centered impact on health and second a cascading downstream (outside the hospital due to burning of fuels to make electricity) impact on that same patient's public or cumulative health. Thus, each patient-care decision has a resulting environmental shadow of energy and material use that create the secondary or unintended (and potentially suboptimal) impact on patients and the community through the environment (air, water, and land). At this time, these cascading impacts are hard to see for the healthcare providers and so they cannot participate in finding improvements. Patient-centered decisions include selection of diagnostic studies, choice of procedures, consumption of pharmaceuticals or medical consumables, and use of equipment and devices, etc. The patient's health, impacted by direct medical decisions or indirectly downstream, is directed by the Hippocratic Oath. Both impacts are governed by the Hippocratic Oath, which for hospital sustainability can be interpreted as "we seek what is good

J. Twomey (✉)
Wichita State University, Engineering Dean's Office, Wichita, KS, USA
e-mail: janet.twomey@wichita.edu

M. Overcash
Environmental Clarity, Inc., Raleigh, NC, USA
e-mail: mrovercash@earthlink.net

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_9

for the patient and that achieves energy improvement which is then good for the patient's public health."

Now it may be time for the healthcare community to explicitly understand and begin to incorporate the environment into their patient-centered decision analysis (Harper 2014; Overcash and Twomey 2013). The medical focus on the patient sitting in front of the healthcare provider must include (1) how will the patient benefit from the health decisions made by healthcare professional, but (2) also how will the patient and the community be impacted by the indirect or cascading impact from the environment born from those same healthcare decisions? This article defines a new avenue for healthcare energy and cost improvement, initiated as full conceptual approach at Wichita State University in 2012 (Overcash and Twomey 2012). This new concept is referred to as **energy improvement in patient-centered and downstream care** and is defined as

> the health interventions built on clinical decisions for patient-centered care that also improve downstream patient public health. Said differently, when the choice between 2 or more treatments or procedures have equivalent patient outcomes, is it now time for healthcare providers to consider the environment?

Other terms defined for this article are

- Life cycle analysis: The quantitative tool for calculating and comparing energy use for a device, product, drug, service, or system. This is the definitive method to establish downstream unanticipated benefits (or harm) of patient-centered care choices. Life cycle can integrate both hospital energy (electricity, natural gas, etc.) and energy used to generate hospital energy and the manufacture of consumables (disposables and reusables) into a new energy profile.
- Energy and material use for patient-centered care: Each intervention to improve patient health results that uses energy (a machine, monitor, lighting, etc.) and materials (whether disposable like an IV bag) or reusable such as a laryngeal mask airway.
- Patient-care alternatives: Choices (single or multiple) made by healthcare providers that achieve equivalent patient-care outcomes within economic limits.

Examples of the downstream impacts or medical shadows:

- The selection of a surgical anesthetic engages a long supply chain that uses fossil resources and inorganics such as fluorine in a series of chemical plants that produce emissions from energy and chemical inefficiency. The anesthesia use results in gas exhalation and venting from the hospital, another emission (Sherman et al. 2012).
- A CT study of an abdomen utilizes electrical power over fixed time of the scan as well as power through out all the 365 late night shifts in a year, while on standby. The patient CT study also consumes materials that are produced in long supply chains and the necessary packaging, all with cascading energy and chemical emissions (Twomey et al. 2012).
- The range of general surgical procedures involves direct electrical use by the instruments as well as about 20 kg of used materials per case (MacNeill et al.

2017). This entire mass of consumables are manufactured in dozens of chemical plants, converting fossil resources (oil, natural gas, etc.) into these surgical OR materials.

- In an OR, in an ICU, or for isolation requirements in a patient room, the patient-care teams utilize gowns and drapes to control hospital-acquired infections. These are manufactured globally, shipped tens of thousands of km, sterilized with steam, ethylene oxide, or gamma radiation and delivered to the hospital. The decreased energy use and chemical losses of all these steps are directly coupled to the hospital decisions for using reusable perioperative textiles thus creating an unanticipated benefit versus selecting the suboptimal disposable option (Overcash 2012).
- Long-term dialysis directly improves patient quality of life using membrane technology and substantial amounts of water to remove blood byproducts. Thus, electrical energy is used to pump fluids, consumables are required for patient safety, and water is treated before discharge to rivers. This nephrology profile identifies new supply chains with energy and chemical emissions required to deliver successful hemodialysis. This may change when delivered at home versus at a medical center (Connor et al. 2010; Soltani et al. 2015).

These five examples illustrate the direct relation between benefits of patient-centered care and the secondary effects of health impact on these same patients and society from the environment, which is a new chapter in healthcare sustainability. Most work in this field is still in research and so answers are yet to be developed. Success in achieving energy improvement and unexpected public health benefits will be a collaborative effort of engineering and medical professions over a substantial period of time and with significant resource needs.

Healthcare is growing as a percent of each national economy globally and is clearly a large economic sector. This size creates a dual responsibility, (1) achieves positive patient outcomes, but (2) society expects progress to lower healthcare impact on the environment, the downstream footprint of patient-care success (Overcash and Twomey 2013; Harper 2014). Many healthcare organizations have begun sustainability awareness programs that stimulate concepts such as energy improvement (Healthy Hospitals Initiative (HHI) is an excellent example of these awareness efforts). Society expects more of the medical community, and so why not make medical decisions that improve the patient and hospital as well as public health?

9.2 Model of Analysis and Improvement

An effective way to discuss the total consequence of medical decisions aimed at successful patient outcomes is in the context of hospitals. The complex facility of a large urban hospital consumes energy (typically electricity and natural gas), utilizes materials (reusables and disposables), and importantly delivers medical services based on patient conditions. The complexity of a hospital is the diversity of services,

Table 9.1 Healthcare service areas of hospital (U.S. EPA 2005)

1. Diagnostic services (a) Endoscopy (b) Radiology (c) Cardiac cauterization (d) Nuclear medicine (e) Sleep studies	**2. Surgical services** (a) Ambulatory out-patient (b) Surgery (c) Post-anesthesia care (d) Preoperative services (e) Anesthesia	**3. In-patient care** (a) Orthopedic (b) Neurology (c) Medical surgical (d) Urology (e) Cardiac (f) Psychiatric/behavioral (g) Geriatric (h) Palliative (i) Maternal/child care (j) Pediatric (k) Cancer (l) Rehabilitative
4. Critical care services (a) Surgical intensive care (b) Medical intensive care (c) Pediatric intensive care (d) Cardiac intensive care (e) Burn care (f) Neonatal intensive care	**5. Emergency care services**	**6. Respiratory care services**
7. Dialysis	**8. Physical therapy**	**9. Out-patient services** (a) Women's health (b) Family practice (c) General medicine (d) Rehabilitative
10. Oncology/cancer care services (a) Radiation oncology (b) Chemotherapy	**11. Laboratory services** (a) Hematology (b) Microbiology (c) Chemistry (d) Surgical pathology (e) Histology	**12. Perioperative and patient textile management**
13. Nutrition services		

and hence there are a wide range of important medical decisions. One catalogue of hospital services is given in Table 9.1. These categories are important building blocks to learn in each service area the relation of patient-centered medical decisions and the downstream energy and material impacts. With this information for each service, one can then assemble these blocks to reflect an entire hospital, where each block can achieve energy improvement in patient-centered and downstream care.

When examining hospital sustainability and energy improvements, three concepts or areas are defined, two of these have been making progress for two decades and one area, energy improvement in patient-centered and downstream care is very new, Fig. 9.1.

Area A is characterized as largely architectural and design-oriented. This community is principally focused on the hospital building envelope. As such, once the healthcare facility is completed there are few further improvements. This area is

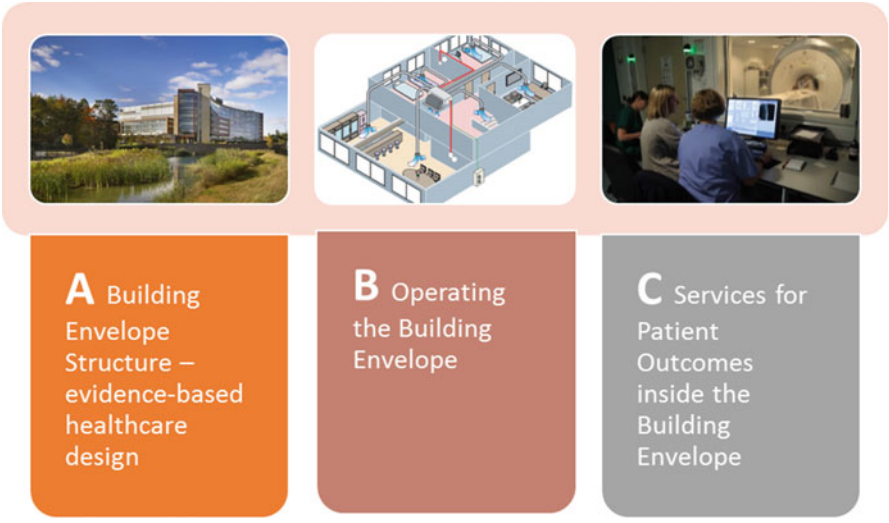


Fig. 9.1 General areas of healthcare improvement from engineering

thus referred to as construction and reconstruction and uses evidence-based design. Topics often included are as follows:

1. Improvements in materials, sitting, building design
2. Windows and views
3. Noise
4. Hand washing stations

The Center for Healthcare Design (Harris et al. 2008) and Prevention through Design (Schulte and Heidel 2009) are primary examples of this community.

Area B is the operating aspects of the building envelope and is often characterized as the electricity and fuel use of the entire building, also known as heating, ventilation, and air conditioning (HVAC). These are often referred to as non-healthcare or overhead energies. Topics often included are as follows:

1. Building design for lower energy use
2. Advances in lighting efficiency
3. Insulation and heat loss
4. Substitution of renewable sources of electricity

Important organizations working in Area B are Practice GreenHealth (2011), American Society of Healthcare Engineering (http://www.ashe.org), and Health Systems Engineering Alliance (https://www.purdue.edu/discoverypark/rche/partners/hsea.php).

Area C is new and distinctive in that the focus is on energy improvement related to health interventions by informed clinical decisions for patient-centered care. Area

C can be referred to as the actual healthcare services inside the building envelope (Area A), but not the operation of the building envelope (Area B). The work at Wichita State University is part of the national effort to improve healthcare services in Area C as a new domain for sustainability research (Overcash and Twomey 2012).

Area C is the first substantial environmental chapter for hospitals in which the healthcare providers (physicians, nurses, and administrators) have a direct opportunity to participate. Their participation can make major improvements in innovation, implementation, and as future informed decision-makers. Another important aspect of Area C research is that both direct energy (usually electricity) and all materials used (consumables) can be included together to achieve sustainability improvement. This is a distinct aspect of Area C because most hospital material use (a part of which becomes solid waste) is connected to the patient-centered care, but is not visualized in energy terms.

In order to fully improve the energy use of the entire hospital, it is necessary to subdivide this complex service facility into parts with the separate patient-centered decisions and resulting downstream impacts, Table 9.1. The services shown do not include administrative support, facilities management, dentistry, animal research, and clinical research. The overall objective and strategy is to build an understanding of the entire energy improvement in patient-centered and downstream effects, on a service block-by-service block basis, in order to then assemble these into an integrated view of healthcare facilities. The first phase at Wichita State University has covered radiology, medical textiles, and dialysis services. The Energy Information Agency data, last reported in 2003, indicate that the Area C energy for the 8000 in-patient US hospitals surveyed (comprising 1.9 billion square feet of space) is about 140 trillion BTU/year. The energy use and carbon footprint of the patient-centered care in just these 8000 hospitals is about 17 billion kg of carbon dioxide equivalent emissions per year in the USA. Thus, there may be potential gains in research for patient-centered and downstream energy improvement in Area C (about two million kg CO2eq per hospital per year). The National Health Service of Great Britain and Ireland reported that up to 30% of hospital energy is for surgical services, Fig. 9.2 (ASGBI 2011).

With the national impetus for considering healthcare energy improvement and the concept of subdividing a hospital into service blocks, what might be the healthcare sustainability roadmap in this new field? First, it is clear that the patient-care team is today substantially limited by a lack of data, in any integrated sense, regarding the relation between their direct patient-centered health decisions and the cascading environmental impacts that lead to the unanticipated effects on patient public health. Thus, the first priority is the necessary research to provide basic preliminary profiles of medical-based decisions and energy and materials use. This is not at a general level (such as the overall hospital), but at the patient-care level, where actual decisions are made. This level of granularity is a key to understanding the different impacts of equivalent patient outcomes. In a medical analogy, the needed level of granularity or information is to know the cellular functions of the liver rather than just the overall symptoms of liver disease on a patient.

1. Diagnostic
2. Surgical
3. Inpatient care
4. Critical care
5. Emergency care
6. Respiratory care
7. Oncology/cancer care
8. Food service
9. Physical therapy
10. Outpatient
11. Laboratory
12. Dialysis
13. Nutrition
14. Perioperative & patient textile management

Fig. 9.2 General service categories in hospital (ASGBI 2011)

To be effective in this research (at multiple organizations world-wide) toward the overall understanding of energy improvement in patient-centered and downstream effects, the following three principles will be important:

1. Since many of the downstream effects on patients and society (as patients) will occur outside the hospital as a result of decisions within the hospital, a life cycle approach will be essential for this research.
2. Great benefit will be gained by simplifying the first phase of research information collection regarding the hospital and the medical procedure energy impacts. The first simplification is Table 9.1 which subdivides hospital services. The second is to select small, but frequently used (or publically recognizable) examples within each service area to study. This principle will add substantially to how to conduct such research in many specialties and add to the efficiency of adding more examples later. Since it is unlikely that all medical procedures, materials, patient conditions, etc. can ever be studied with these life cycle tools, the preliminary selection is a compromise of resources and curiosity.
3. The focus on overall hospital improvement suggests that the many studies must be structured to be easily additive, thus forming the large hospital system. This is referred to as the LEGO™ principle in which various studies can be different (color, shape, function), but are known to attach together to form a structure (in this case, the hospital energy and material efficiency model). A protocol for the uniform information content would thus be beneficial.

How can we anticipate the kind of improvements that can result for the life cycle energy improvement in patient-centered and downstream care? If we look at virtually every patient-care decision, a basic framework is as follows:

(a) There are alternatives (whether procedures, studies, pharmaceuticals, linens, operational policies, etc.) that can be used to achieve equivalent patient outcomes (type of anesthesia, location of dialysis, imaging study, etc.) or
(b) There are virtually no alternatives and so the set of patient-centered care decisions is fixed.

Both of these scenarios can achieve improvement in hospital energy (and in many cases costs) and in downstream patients' public health, while managing for a successful patient outcome.

With alternative pathways for equivalent patient outcomes (case a), there are two spheres for improvement. The first sphere is to understand which of the multiple choices has a lower environmental impact (the result of life cycle research studies), while still achieving equivalent patient outcomes. Then using an additional criterion of also improving patients' public health (downstream), the better alternative can be chosen. The second sphere is research on technological improvement within any one alternative. In this way, each alternative and each technology lead to lower energy and material use. As an example, the research of Sherman et al. (2012), on anesthesia life cycle improvement illustrates case (a) the results define that in cases where morbidity and mortality are equivalent among several gases, the use of the specific anesthesia gas Desflurane should be avoided to achieve a substantial patient public health improvement. This is an example of choice among alternatives that leads to sustainability improvement. Secondly, when any of the anesthesia gases is selected, reducing the fresh gas flow rate, even with the use of a carbon dioxide absorbent, leads to a net life cycle improvement. This second example illustrates improvement within a given gas choice. Thus, a two-fold dimension of improvement is involved with patient-centered care decisions; the first is which alternative gas, and the second is what technological improvements (less fresh gas flow) to lower energy and material use.

When there is only one patient-care choice (case b) for achieving the successful patient outcome, then sustainability improvement can still be achieved by seeking to improve the technology or practices associated with the necessary alternative. So in a sense, this range of improvements is lower when there are not multiple patient-care choices. As an example, Campion et al. (2012) have used a life cycle approach of obstetric alternatives for delivering a baby. In general, the two procedures (vaginal and cesarean) are not alternatives and so each method was studied to provide information by which improvement might still be made. The use of surgical drapes and gowns for vaginal delivery was documented and so the improvement of shifting to reusable perioperative textiles (Overcash 2012) is a direct improvement in patient public health as measured by lower life cycle energy use. However, the patient-centered care choice of delivery remained the same.

9.3 Roadmap for Research Community

This roadmap for developing LEGO™-like research for the range of hospital services field is explained below. As a first step for a new research study, use Fig. 9.2 to select a hospital service for the proposed study. Next forms a collaborative team. Currently, a team with clinicians defining the medical system involving patient care (e.g., procedures for hysterectomies, imaging by ultrasound, MRSA laboratory testing) and engineers undertaking the life cycle analysis appears to be very effective. Select one or more related medical procedures that are frequently used in most hospitals (and might be viewed as alternatives) as the initial study. A goal, in each service of Fig. 9.2, is to undertake sufficient procedures, studies, devices, etc. that represent 50% of the various types of practices as a means of understanding the variability and to increase the mechanisms for improvement.

Next the team invests in power measuring tools, material use cataloguing procedures, and locations for these studies. Plug studies with monthly energy use are of little value, as more granularity to tie energy and consumables to specific patient treatment must be made. The studies have begun (Twomey et al. 2012). The purpose of these studies is not just the energy and material values, but to establish a direct link to patient-care conditions and to the variables with potential for change that reduce energy and material use thus improving public health.

As with any new initiative and research, it is valuable to plan how such information can affect hospital services behavior or technology change. For patient-centered care energy analyses, there are several mechanisms for motivating change. In the medical arena, the information on the downstream environmental benefit to the patient of alternative choices that lower environmental impact at equivalent patient-care outcomes might be adopted or featured in annual meetings of related societies (ACOG, IARS, American Society of Hematology, etc.). These professional organizations might adopt guidelines; use medical-based energy information in online guides or teaching tools; develop working groups to refine and expand the understanding of sustainability implications of such decisions for residents; or work with hospital management or insurance groups to foster change.

Within the fixed medical choices or technologies (case b), information on lowering patient exposure to environmental impacts can be used by equipment consortia, manufacturing R& D organizations, or university and other organization researchers. Thus, patient downstream health benefits can be achieved with existing medical practices. In addition, this is an excellent area for behavioral studies groups to contribute by clarifying the mechanisms for change, simplifying implementation, and encouraging early adopters.

9.4 Conclusions

Medical decisions for quality patient care represent a new dimension for hospital energy sustainability improvement. Such patient health medical decisions have both a direct energy use (in-the-hospital) plus an outside-the-hospital public health impact from the environmental emissions (air, water, and land) related to generation of energy and manufacturing of consumables. All of the areas of hospital services (Table 9.1) can contribute in energy and consumable reduction. The goal is to engage the healthcare specialists and to use their ingenuity and creativity to examine procedures, patient-based decisions, and other avenues to seek hospital sustainability improvements. Energy improvement will come from selecting patient medical alternatives that give quality care, but at lower environmental burden plus from changes to the technology and operational procedures used in all areas of the hospital services. The resulting improvements in sustainability will address societal expectations of environmental sustainability of this important economic sector.

References

American Society for Healthcare Engineering of the American Hospital Association (n.d.) http://www.ashe.org. Accessed 30 Jan 2018

Association of Surgeons of Great Britain and Ireland (2011) Consensus statement on cost-effective surgery. London, 20p

Campion N, Thiel CL, DeBlois J, Woods NC, Landis AE, Bilec MM (2012) Life cycle assessment perspectives on delivering an infant in the US. Sci Total Environ 425:191–198

Connor A, Mortimer F, Tomson C (2010) Clinical transformation: the key to green nephrology. Nephron Clin Pract 116(3):c200–c205

Harper L (2014) We should start to quantify the environmental impact of different treatments. Br Med J 348:g1997. https://doi.org/10.1136/bmj.g1997

Harris DD, Joseph A, Becker F, Hamilton K, Shepley MM, Zimring C (2008) A practitioner's guide to evidence based design. The Center for Health Design, Concord

Healthcare Engineering Alliance (n.d.) https://www.purdue.edu/discoverypark/rche/partners/hsea.php. Accessed 30 Jan 2018

MacNeill A, Lillywhite R, Brown C (2017) The impact of surgery on global climate: a carbon footprinting study of operating theatres in three health systems. Lancet Planet Health 1:e381–e388. https://doi.org/10.1016/S2542-5196(17)30162-6

Overcash M (2012) Comparison of reusable and disposable perioperative textiles. Anesth Analg 114(5):1055–1066

Overcash M, Twomey J (2012) Medical-based energy—new concept in hospital sustainability improvement. Paper presented at the NSF Workshop on Patients, Energy, and Sustainability—The Model of Decision-Making for Lower Healthcare Footprints, Wichita State University, 15–16 May 2012

Overcash M, Twomey J (2013) Green surgery—concept for the profession. Association of Surgeons of Great Britain and Ireland, 41, pp 10–12

Practice Greenhealth (2011) Best practices in energy efficiency. https://practicegreenhealth.org/topics/energy-water-and-climate/energy/best-practices-energy-efficiency. Accessed 30 Jan 2018

Schulte P, Heidel D (2009) Prevention through design. National Institute for Occupational Safety and Health. http://www.cdc.gov/niosh/programs/PtDesign. Accessed 30 Jan 2018
Sherman J, Le C, Lamers V, Eckelman M (2012) Life cycle greenhouse gas emissions of anesthetic drugs. Anesth Analg 114(5):1086–1090
Soltani SA, Overcash MR, Twomey JM, Esmaeili MA, Yildirim B (2015) Hospital patient-care and outside-the-hospital energy profiles for hemodialysis services. J Ind Ecol 19(3):504–513
Twomey J, Overcash M, Soltani S (2012) Life cycle for engineering the healthcare service delivery of imaging. In: Dornfeld D, Linke B (eds) Leveraging technology for a sustainable world. Springer, Berlin, Heidelberg
U.S. EPA (2005) Profile of the healthcare industry. EPA/310-R-05-002, Office of Enforcement and Compliance Assurance, Washington, DC

Janet Twomey is currently the Associate Dean for Graduate Education, Research, and Faculty Success in the College of Engineering at Wichita State University in Wichita, Kansas, USA. She is also a Professor of Industrial, Systems, and Manufacturing Engineering. Her areas of expertise are in sustainable engineered systems, environmental life cycle analysis (LCA), and machine learning. In the area of sustainability and LCA, Dr. Twomey together with Dr. Michael Overcash have focused their work in creating new knowledge in the fields of manufacturing, energy, and healthcare. They have published over 16 peer reviewed journal articles and received over $3.5 million in funding from the National Science Foundation, Department of Energy and industry. Their current project seeks to establish an open database of chemical life cycle information–Environmental Genome.

Dr. Twomey was drawn to industrial engineering because of her interest in data and data analysis which began while she was a Research Associate in the University of Pittsburgh's School of Medicine. As a Ph.D. student, she was able to broaden her knowledge and skill set in data analysis by focusing her research in machine learning.

Dr. Twomey has served as Program Officer for Manufacturing Enterprise Systems at the National Science Foundation from 2001 to 2004. In 1999, she was an A.D. Welliver Boeing Faculty Summer Fellow. She served a 3-year term as the Academic Vice President for Board of Trustees, Institute of Industrial and Systems Engineering.

Michael Overcash has continued contributions to the scientific evaluation of decisions leading to patient outcomes and how these can also achieve reduced environmental impact. These studies range from dialysis protocols, radiology alternatives, reusable devices and garments, and reuse of materials and products. In his tenure at Wichita State University, he and Dr. Twomey developed a wide-ranging program on the use of life cycle analysis for healthcare organizations.

Chapter 10
Improving Patient Care Transitions at Rural and Urban Hospitals Through Risk Stratification

Shan Xie and Yuehwern Yih

Contents

10.1 Care Transitions at Rural and Urban Hospitals 211
10.2 ED Utilization in CAH 214
10.2.1 Analysis 214
10.2.2 Results: ED Transfers to Short-Term General Hospitals 215
10.2.3 Results: ED Visits for Non-emergent Care 217
10.2.4 Discussion 219
10.3 Hospital and Community Characteristics Affecting 30-Day All-Cause Readmission ... 221
10.3.1 Analysis 222
10.3.2 Results 222
10.3.3 Discussion 224
10.4 Conclusions 226
References 227

10.1 Care Transitions at Rural and Urban Hospitals

During the course of an illness, patients may require care from multiple healthcare professionals and settings, and thus need transitions across different locations or different levels of care under the same location. These locations could include hospital inpatient unit, hospital emergency department, ambulatory service, post-acute nursing facilities, patient's home, primary and specialty care offices, and assisted living and long-term care facilities (Coleman and Boult 2003). During this

S. Xie (✉)
LASER PULSE Consortium, West Lafayette, IN, USA

Regenstrief Center for Healthcare Engineering, West Lafayette, IN, USA

Purdue University, West Lafayette, IN, USA

Y. Yih
School of Industrial Engineering, Purdue University, West Lafayette, IN, USA
e-mail: yih@purdue.edu

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_10

transition period, there is risk for discontinuity of care. A number of studies (Clancy 2006; Coleman and Berenson 2004; Moore et al. 2003; Coleman et al. 2005; Forster et al. 2003; Kripalani et al. 2007) have documented that care transitions can lead to an increased probability for negative impact upon patient safety, care quality, and care cost due to ineffective communication, high medication errors, and lack of follow-up care.

Critical Access Hospitals (CAHs) are certified rural hospitals by the Centers for Medicare and Medicaid Services (CMS) with cost-based Medicare reimbursement structure (CAHs n.d.). CAHs are an integral part of the US healthcare delivery system providing access to necessary care for rural residents and frequently serving as patient advocates in the coordination of regional health resources. Care transitions might be more prevalent in rural settings, since rural patients may require care that is not locally available necessitating CAHs to transfer patients to providers outside their community. Transfer has been identified as one of the top priorities for patient safety in rural community (Coburn et al. 2004). Most of the transferred patients will experience two care transitions: the CAH to larger hospitals and post-discharge from the hospital. Compared to transfer from CAHs to a higher level setting, it is more challenging when patients are discharged back to their local community due to differences in healthcare systems, barriers to communication, and lack of insurance (Prasad et al. 2011). Therefore, rural communities have additional challenges in the care transition process. Emergency department (ED) at CAHs is a critical healthcare access point for rural residents and disproportionately account for patient volumes, expenses, quality, and patient satisfaction when compared to larger urban and suburban hospitals (2013 National Rural Emergency Department Study 2013). One study has found that 54% of all ED visits to CAHs were categorized as semi/less-urgent and non-urgent (2013 National Rural Emergency Department Study 2013). Using ED for non-emergent conditions may lead to excessive healthcare spending, unnecessary testing and treatment, and risk for discontinuity of care.

While it is important to understand ED use in CAHs to reduce unnecessary health expenses and improve quality of care for rural residents, prior research in this area has been limited. Most of the existing research on rural hospitals focused on evaluating the quality of the care provided and not so much on the care coordination aspect (Casey et al. 2010, 2012, 2013; Joynt et al. 2011a; Henriksen and Walzer 2012). Among the few studies that looked at ED utilization, one study focused on examining factors associated with any ED use (Fan et al. 2011), three studies quantified the frequency of ED transfers but did not investigate situations upon discharge back to the community (National Rural Emergency Department Study 2013; De Freitas et al. 1998; Michelle Casey 2014), and only one study (National Rural Emergency Department Study 2013) identified the non-emergent use of ED in CAHs.

On the other hand, hospital readmissions may represent a more prevalent care transition issue in larger urban hospitals, especially readmissions happened within 30 days of previous hospital discharge. Twenty percent of Medicare patients were readmitted within 30 days of discharge with an estimated cost of 17 billion (Jencks et al. 2009). Thirty-day readmission is used by CMS as a quality measure and

hospitals with excess risk-standardized readmission rates are subject to financial penalties. CAHs are not considered in this readmission reduction policy. Hospitals have developed and employed various care transition interventions to reduce readmission rates. Readmission prediction models were proposed to help efficiently target at-risk patients and design tailored interventions. While most studies have focused on examining the significance of patient-related risk factors such as age, gender, race, comorbidities, clinical condition, and healthcare utilization (Hasan et al. 2009; Jiang et al. 2005; Brand et al. 2005; Allaudeen et al. 2011; Kansagara et al. 2011; Donzé et al. 2013), a number of studies suggested that factors representing hospital characteristics and community differences such as hospital teaching status, ownership, bed size, socioeconomic status, social support, access to care, and geographic location were also associated with early readmission (Herrin et al. 2015; Calvillo-King et al. 2012; Lindenauer et al. 2013; Arbaje et al. 2008; Hu et al. 2014; Weissman et al. 1994; Joynt and Jha 2013a; Joynt et al. 2011b). Fifty-eight percent of the variation in publicly reported hospital 30-day readmission rates was found attributable to county-level factors (Herrin et al. 2015). Patients living in neighborhoods with high poverty level, low education, and low household income were found to have greater readmission risk (Hu et al. 2014). Patients living alone were found to have a higher likelihood of 60-day readmission (Arbaje et al. 2008).

Despite the significant findings, the current CMS readmission measure did not adjust for hospital and community characteristics. The rationale is that risk-standardized readmission rate is a measure for quality of care, hospitals should not be treated with different standards due to variation in demographic and socioeconomic characteristics (Kansagara et al. 2011; Medicare Hospital Quality Chartbook 2012; Horwitz et al. 2012). However, some scholars have raised concerns about the level of accountability and the appropriateness of the current reimbursement policy as they argue that readmissions are largely influenced by the community environment and post-discharge care that are not within hospital's control (Herrin et al. 2015; Joynt and Jha 2013b; Axon and Williams 2011). Knowing more about the role of community characteristics could inform collaborative projects between hospitals and local communities aimed at reducing readmissions (Herrin et al. 2015). Prior research into hospital and community factors has been limited. Research findings have been from observational studies that used publicly reported readmission rates without directly adjusting for patient-level risks (Herrin et al. 2015; Joynt and Jha 2013a), they have been confined to specific diseases (Calvillo-King et al. 2012), or they examined only socioeconomic factors (Lindenauer et al. 2013; Arbaje et al. 2008; Hu et al. 2014; Weissman et al. 1994).

In this chapter, we presented two studies that focused on two important care transition issues with rural and urban hospitals. One study (Xie 2018) examined ED transfers and the use of ED for non-emergent care at CAHs. Another study assessed the impact of various hospital and community characteristics on 30-day readmission among general medicine patients while accounting for patient-level factors. Both studies aimed to provide insights to help identify patients who would be most likely to benefit and design targeted interventions to improve care transitions by stratifying patients into different risk groups.

10.2 ED Utilization in CAH

In this study, we examined two major issues with ED utilization in CAHs: (1) ED transfers to short-term general hospitals, service utilization upon discharge back to the community and factors associated with revisit within 60 days of ED transfers; (2) non-emergent ED visits and associated cost. We performed retrospective analysis using data collected from five CAHs in Indiana for 2006–2009. The data included patient utilization of four hospital service units: ED, inpatient unit (IP), observation bed (OBS), and swing bed (SWD). Our study focused on utilization for ED. Certain visits were excluded from the analyses including visits resulted in death, visits discharged to hospice care or left against medical advice, visits with incomplete records, and visits discharged to places other than short-term general hospitals or home. The resultant data set contained two subsets, visits transferred to short-term general hospitals (5029 visits for 4384 patients) and visits discharged to home (106,222 records for 52,215 patients). These two categories consisted of 99% of the ED visits. The subset of visits discharged to home was used to study non-emergent care.

10.2.1 Analysis

For patients transferred to short-term general hospitals, we analyzed the top diagnoses for transfers. Principal diagnoses code were categorized into higher level diagnosis categories based on Agency for Healthcare Research and Quality's (AHRQ) single-level Clinical Classifications Software (CCS) (Elixhauser et al. 2013). For patients who had a revisit to CAHs within 60 days of the transfer, we examined the factors associated with this revisit and the types of services utilized for this revisit. The difference in baseline characteristics between the groups with and without a 60-day revisit was tested by Chi-square test. Logistic regression was used to estimate the association. Factors serving as predictor variables included age, gender, race, marital status, AHRQ CCS categories, comorbidities, and hospital. Elixhauser comorbidity variables were created based on the secondary diagnoses provided in the data with the published algorithm (Elixhauser et al. 1998). Comorbidity variables with frequency count more than 20 were included in the model to avoid quasi-complete. As 90% of the patients had only one transfer, each transfer was considered as an independent event. The interaction effects among the predictors were tested. Due to relatively small sample size, model performance was cross-validated.

For patients discharged to home, their visits were classified into different urgency levels based on the International Classification of Diseases, 9th Revision (ICD-9) codes given by the New York University (NYU) algorithm (Faculty and Research|NYU Wagner n.d.). The algorithm also identified visits related to alcohol, drug, injury, and mental health. ICD-9 codes that did not have enough sample size

to evaluate were categorized as unclassified. Since the algorithm assigned multiple probabilities for each ICD-9 code, we chose the highest probability to represent the final category for that ICD-9 code. After identifying the categories, we also quantified charges associated with non-emergent ED visits.

10.2.2 Results: ED Transfers to Short-Term General Hospitals

Among all the ED visits, 2.74% were transferred to short-term general hospitals. Top ten diagnosis categories are presented in Table 10.1, where 12% of the transfers were for chest pain, and 5% were for paralysis. Table 10.2 compared the baseline characteristics between groups with and without 60-day revisit. No significant difference was found for gender and race. Age, marital status, and hospital were all significant. 20.6% of the transfers had a revisit within 60 days. Service types for the revisit are presented in Fig. 10.1, 86% of the revisit was for ED services, and 9% were admitted to the inpatient unit.

Results on predictor variables that were statistically significant at a significance level of 0.10 are reported in Table 10.3.

Patients between 18 and 44 years old had a higher likelihood of 60-day revisit compared to patients aged 65 and above (OR 1.48, 95% CI 1.17–1.87). Patients who were divorced or single had higher odds of revisit compared to patients who were married. CCS categories that had a higher risk of revisit as compared to the most frequently transferred diagnosis chest pain were listed. Patients transferred with complication of device were most likely to have a 60-day revisit. In addition, certain comorbidities also contribute to higher risk of revisit. Patients with neurological disorders, renal failure, or depression were more likely to visit CAHs after transfer compared to patients without these comorbidities. Finally, each hospital represented different risk profile in terms of 60-day revisit; multiple comparisons were computed for the five hospitals (Fig. 10.2). The 95% Wald confidence limits were calculated based on Wald statistic, a standard method that uses the likelihood function

Table 10.1 Top ten diagnosis categories for ED transfers

AHRQ CCS category	Frequency	Percent
Chest pain	601	11.95
Paralysis	267	5.31
Acute cerebrovascular disease	188	3.74
Acute myocardial infarction	187	3.72
Coronary atherosclerosis	181	3.6
Pneumonia	165	3.28
Dysrhythmia	155	3.08
Secondary malignancies	138	2.74
Congestive heart failure	128	2.55
Gastrointestinal (GI) hemorrhage	113	2.25

Table 10.2 Baseline characteristics of patients who had ED transfer

Variables	No revisit $n = 3991$	Revisit $n = 1038$	P value
Age, %			<0.001
0–17	10.5	6.7	
18–44	24.8	31.6	
45–64	32.1	31.5	
≥65	32.7	30.2	
Gender, %			0.844
Female	50.1	50.5	
Male	49.9	49.5	
Race, %			0.368
White	96.9	96.3	
Nonwhite	3.1	3.7	
Marital status, %			0.003
Married	45.1	40.9	
Divorced	10.5	14.3	
Single	33.4	34.3	
Widow	11.0	10.5	
Hospital, %			0.021
A	31.2	31.1	
B	16.0	16.5	
C	25.3	20.9	
D	14.2	16.9	
E	13.3	14.6	

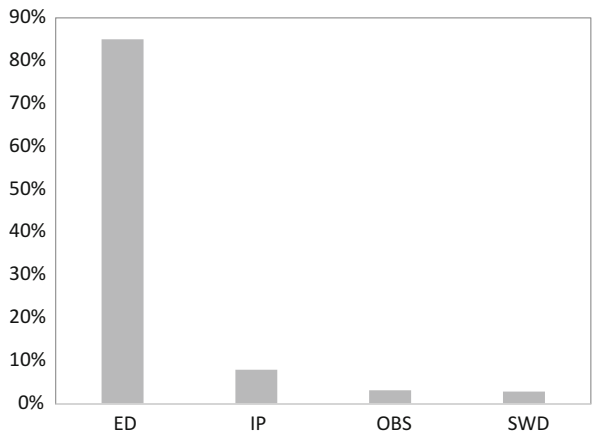


Fig. 10.1 Service utilization at CAHs for 60-day revisit

to perform statistical inference for large-sample categorical data (Agresti 2013). Hospital C had a lower risk compared to hospital A, B, D, and E; and hospital D had significantly higher risk than hospital A.

Table 10.3 Association of population characteristics with 60-day revisit

Variables	Odds ratio (95% CI)	*P* value
Age		
≥65	Reference	
0–17	0.88 (0.62–1.26)	0.496
18–44	1.48 (1.17–1.87)	0.001
45–64	1.06 (0.87–1.30)	0.549
Marital status		
Married	Reference	
Divorced	1.51 (1.20–1.90)	0.001
Single	1.18 (0.98–1.43)	0.083
Widow	1.13 (0.87–1.46)	0.374
AHRQ CCS category		
Chest pain	Reference	
Complication of device; implant or graft	3.62 (1.33–9.90)	0.012
Pancreatic disorders (not diabetes)	2.66 (1.15–6.14)	0.022
Pulmonary heart disease	2.04 (0.88–4.71)	0.096
Heart valve disorders	1.98 (0.96–4.12)	0.066
Comorbidities		
Neurological disorders	1.57 (0.94–2.62)	0.083
Renal failure	2.71 (1.61–4.56)	0.000
Depression	2.29 (1.24–4.23)	0.008
Hospital		
A	Reference	
B	1.04 (0.83–1.30)	0.753
C	0.78 (0.64–0.97)	0.022
D	1.32 (1.06–1.65)	0.015
E	1.13 (0.90–1.43)	0.300

10.2.3 Results: ED Visits for Non-emergent Care

ED visits discharged to home were stratified into different categories as shown in Fig. 10.3. For all the CAHs, non-emergent cases consisted of more than 30% of the ED visits, more than 20% of the visits could be prevented with primary care (PC treatable), and only 7% of the visits require ED care and not preventable.

We grouped both non-emergent visits and visits could be prevented with primary care as non-emergent cases and computed the average charges for those visits across 4 years as a percentage of total hospital charges and percentage of total ED charges. As reported in Fig. 10.4, more than 10% of the total hospital charges were for non-emergent ED visits, and those visits represent around 20% of the charges associated with ED utilization.

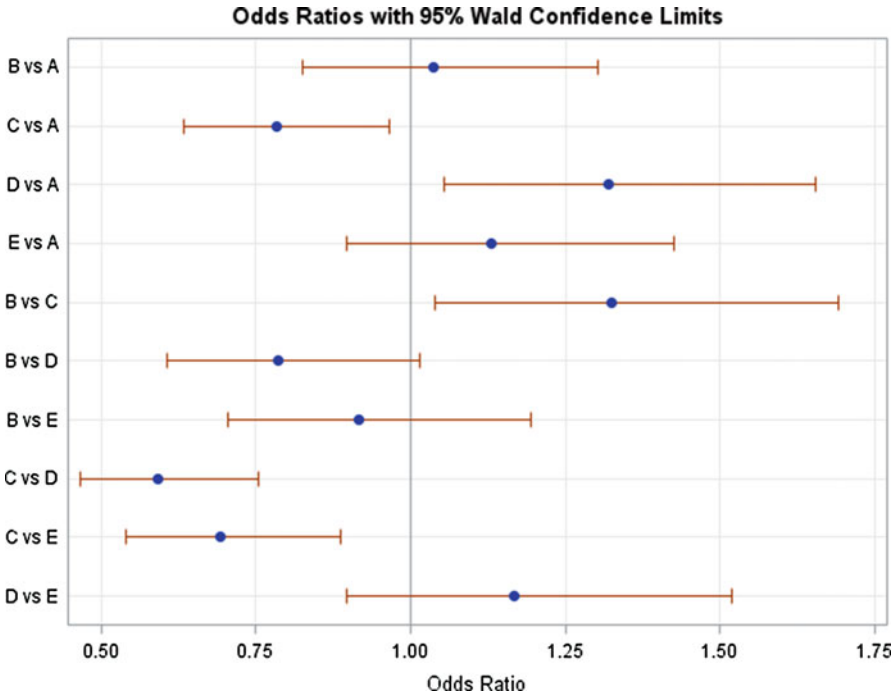


Fig. 10.2 Odds ratio for multiple comparisons of hospitals

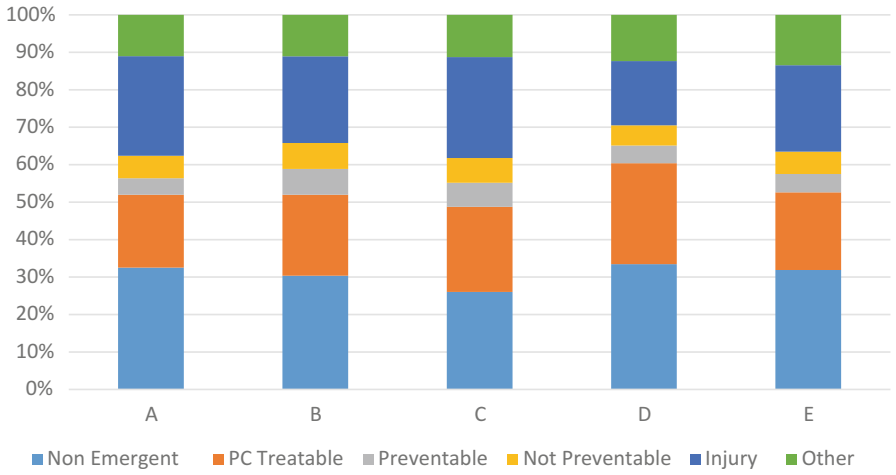


Fig. 10.3 Categories of ED utilization for five CAHs

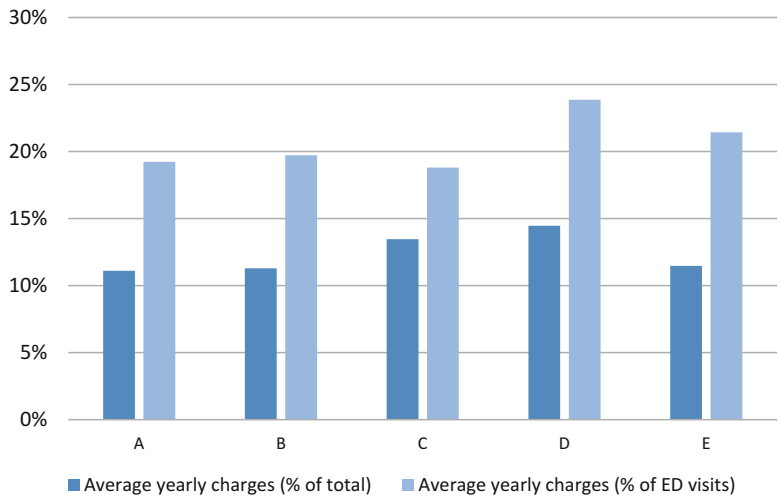


Fig. 10.4 Average yearly charges of non-emergent visits for five CAHs

10.2.4 Discussion

This stud assessed how rural residents utilized ED services. It was found that 2.74% of ED visits were transferred to short-term general hospitals for higher level care, this number is lower than the average transfer rate for rural EDs as reported by a previous study but still higher than the average transfer rate—1.8% for all US hospitals (National Rural Emergency Department Study 2013). This could be that the CAHs included in this study are all from Indiana State, while the previous study has covered a much larger and varied sample. High transfer rate was also reported to have an impact on how CAHs calculate their quality measures (Casey and Burlew 2006). The top diagnoses and other common conditions for transfers found in our study were very similar to a prior study (Michelle Casey 2014) that examined ED transfers of Medicare patients from CAHs.

While patient transfer was likely clinically appropriate, the issue is whether CAHs and larger urban hospitals have systems in place to provide coordination of care upon the patients' return to their home community. We examined the types of contact with CAHs for those patients who returned to their local Indiana CAHs within 60 days of the transfer. To the best of our knowledge, no prior research has studied this problem. It was found that 20% of the ED transfers had a revisit to CAHs within 60 days. Among those revisits, 85% were for ED care, and 9% were admitted to the inpatient unit. This observation may be an indication that patients were trying to use ED as a substitute for primary care, and there is an opportunity for improving the transition of care communication between CAHs and short-term general hospitals.

We further examined the factors associated with 60-day revisits. It was found that patients aged between 18 and 44 years old had a higher likelihood of returning to their local CAH compared to patients above 65 years old. One hypothesis could be that patients within this age group were more likely to be uninsured as compared to patients above 65 years of age who were at least covered by Medicare insurance. This partially supports the previous hypothesis that patients might be using ED as a substitute for primary care. Patients who were divorced or single had higher odds or 60-day revisit compared to patients who were married. The same findings were observed in readmission studies. A possible explanation could be that patients who are married could receive better care and support to help with their recovery.

Regarding diagnosis, patients transferred for complication of device, pancreatic disorders, pulmonary heart disease, and heart valve disorders were more likely to seek care upon return to their local community compared to patients with chest pain. This could be that these disease types subject to a higher likelihood of complications and infection thus require more intensive follow-up care or in other words are more vulnerable to ineffective transitions of care. Patients with depression and neurological disorders were at higher risk of revisit. Some studies about 30-day readmission had similar findings. This may be because these patients were more likely to have drug abuse problems or psychological issues that increased their probability of seeking care. Hospital was entered as a fixed effect to assess the impact of hospital differences. The odds ratio comparisons among hospitals suggested that patients transferred from different hospitals were subject to different risk levels. One hypothesis could be that the location of CAH, their network with regional health resources, and their adoption of health information technology all contribute to the effectiveness of care coordination at the time of transfer and upon patient discharge back to the local community.

Differentiating ED visits for emergent and non-emergent care could help reduce unnecessary health spending and improve patient care. The implicit assumption is that ED may not have the infrastructure to provide the same level of care continuity as a primary care provider. Our study has found that for each CAH, more than 30% of the ED visits were classified as non-emergent, an additional 20% were visits could be prevented with effective primary care. A previous study had the same finding where more than 50% of the ED visits in CAHs were less-urgent or non-urgent (National Rural Emergency Department Study 2013). Rural hospitals were reported to have increased ED visits due to restricted access to primary care (Hines et al. 2011; Gresenz et al. 2007). Finally, the charges associated with non-emergent care were studied. By grouping both non-emergent visits and visits could be prevented with primary care, we found that the non-emergent cases constitute more than 10% of the total hospital charges and 20% of the ED charges. While reducing non-emergent ED visits could reduce unnecessary healthcare spending and improve patients' continuity of care, it challenges the financial viability of the CAHs as those ED visits represent potential revenue for the hospitals.

Our study focused on CAHs in Indiana State. More data on ED utilization from other geographic areas would help validate the findings and make the conclusion more generalizable. Also, we did not have information on activities after ED transfers. The gaps in transitions of care could be better evaluated if more information were available on patients' stay at short-term general hospitals and how they have utilized local health resources upon return to their local community.

10.3 Hospital and Community Characteristics Affecting 30-Day All-Cause Readmission

In this study, we used the Healthcare Cost and Utilization Project (HCUP) state inpatient data for Arkansas and Washington State from 2008 to 2009 to identify risk factors that associate with 30-day all-cause hospital readmissions. We chose these states because they both participated in HCUP and provided unique patient identifiers that enabled identification of readmission; and, the two states represented very different health profiles; Washington was ranked among the top ten while Arkansas was ranked among the bottom ten (America's Health Rankings n.d.). Forty-one and 43 short-term general hospitals were identified in Arkansas and Washington, respectively. This study included adult patients aged 18 years or above. Patients discharged to hospice care were excluded. We also excluded certain admissions based on the CMS hospital-wide all-cause unplanned readmission measure (YNHHSC/CORE 2014), which include admissions that resulted in death, admissions for psychiatric diagnoses, admissions for cancer treatment, admissions for transplant, chemotherapy or radiotherapy, childbirth-related hospitalizations, admissions for rehabilitation, and admissions left against medical advice. We also excluded admissions that transferred to another hospital on the same day of admission. For same-day transfers happened within a hospital, admissions were merged as extended inpatient hospitalization rather than two separate admissions. Admissions during December were excluded due to insufficient follow-up period. The final data set contained 1,049,330 admissions for 773,727 patients. We split the data into 70% training (734,531) and 30% testing (314,799).

We selected the common patient risk factors that associated with readmission risk as identified from the relevant literature. We also obtained hospital and county related characteristics through American Hospital Association (AHA) database and Health Resources and Services Administration's Area Resource File (ARF) using Federal Information Processing Standard (FIPS) county code. Overall, factors serving as predictor variables can be grouped into three categories: (1) patient-level factors including age, gender, Charlson comorbidity index score (Quan et al. 2005, 2011) (calculated based on the secondary diagnoses provided in the data), length of stay, utilization of emergency service, and discharge disposition; (2) hospital characteristics including number of beds, teaching status, and ownership; (3)

community characteristics including state, median household income represented as state quartile, number of primary care physicians per capita (include general practice, general internal medicine, and family medicine physicians), and percent of residents with high school education. Primary care physician ratio was adjusted based on county population ($\times$100,000).

10.3.1 Analysis

The study outcome was all-cause 30-day readmission or admission to the same hospital within 30 days of discharge from the index admission. Since the unit of analysis is admission, a single individual in the data set could have multiple admissions and readmissions. The difference in baseline characteristics between the groups with and without 30-day readmission was tested by Chi-square test for categorical variables and Wilcoxon test for continuous variables (Table 10.4). We used multilevel logistic regression to estimate the risk of 30-day readmission. Admissions were nested within hospitals and hospitals were nested within counties. As 83% of the patients had only one admission, each admission was considered as an independent event. The prediction model was constructed based on the training data and evaluated using the testing data.

10.3.2 Results

As shown in Table 10.4, all predictor variables had a significant difference between the readmitted and not readmitted groups. Among the training cohort, 9.3% of admissions resulted in 30-day readmission. We then estimated the association between risk predictors and 30-day readmission by controlling for other covariates using multilevel logistic regression. The model was evaluated using a separate testing set and had an area under the curve (AUC) of 0.64. Results on predictor variables that were statistically significant at a significance level of 0.10 are reported in Table 10.5.

All patient-level factors were significant in predicting readmission risk. Female patients were less likely to be readmitted than male. Age had a positive relationship with risk of 30-day readmission, where the older the individual, the higher likelihood of readmission. Variables that indicative of disease severity including the length of stay, Charlson index score, and utilization of emergency services were positively associated with 30-day readmission. Patients discharged to skilled nursing facility, or home health care had higher readmission risk compared to patients discharged to home (OR 1.08 and 1.29, respectively). For hospital and county level factors, individuals treated at major teaching hospitals were more likely to be readmitted compared to patients treated at minor or non-teaching hospitals. Patients from areas with higher median household income had lower likelihood of

Table 10.4 Baseline characteristics of training cohort

Variables	No readmission $n = 671{,}954$	Readmission $n = 62{,}577$	p value
Female sex, %	60.1	56.0	<0.001
Age (years), mean (SD)	57.3 (20.5)	62.1 (19.0)	<0.001
Length of stay, mean (SD)	3.9 (4.8)	5.2 (5.9)	<0.001
Charlson index score, mean (SD)	1.1 (1.6)	1.8 (1.9)	<0.001
Emergency service used, %	51.8	65.1	<0.001
Discharge disposition, %			
Routine discharge	76.6	68.1	<0.001
Skilled nursing facility	9.7	13.9	
Home health care	8.2	11.8	
Other	5.5	6.2	
Bed size, %			<0.001
<200	29.2	27.9	
200–399	36.1	35.4	
≥400	34.7	36.7	
Teaching status, %			<0.001
Major	6.6	8.3	
Minor	63.1	60.9	
Non-teaching	30.3	30.8	
Ownership, %			<0.001
Government, non-federal	14.1	14.8	
Non-government, not-for-profit	74.9	74.5	
Private, for-profit	11.0	10.7	
Median household income (state quartile), %			<0.001
1	27.7	29.0	
2	27.3	28.5	
3	24.7	24.4	
4	20.3	18.1	
State, %			<0.001
Arkansas	35.8	39.8	
Washington	64.2	60.2	
Primary care physicians/100 k, mean (SD)	48.9 (14.0)	48.8 (14.1)	<0.001
Percent of residents with high school education, mean (SD)	87.9 (5.2)	87.4 (5.4)	<0.001

Median household income: 1–4, indicating the poorest to wealthiest populations

readmission. Patients admitted in Washington State were less likely to have 30-day readmission as compared to patients from Arkansas State (OR 0.84). Also, patients from communities with higher primary care physician ratio and higher percentage

Table 10.5 Factors associated with 30-day hospital readmission

Variables	Odds ratio (95% CI)
Female sex	0.97 (0.95–0.99)**
Age (per 10 years)	1.05 (1.04–1.05)**
Length of stay	1.02 (1.02–1.03)**
Charlson index score	1.16 (1.16–1.17)**
Emergency service used	1.51 (1.49–1.54)**
Discharge disposition	
Routine discharge	Reference
Skilled nursing facility	1.08 (1.05–1.11)**
Home health care	1.29 (1.25–1.32)**
Other	0.88 (0.84–0.91)**
Teaching status	
Major	Reference
Minor	0.69 (0.50–0.95)**
Non-teaching	0.64 (0.45–0.91)**
Median household income	
1	Reference
2	0.98 (0.96–1.00)**
3	0.95 (0.93–0.97)**
4	0.91 (0.88–0.94)**
State	
Arkansas	Reference
Washington	0.84 (0.71–0.99)**
Primary care physicians/100 k (per 10 physicians)	0.96 (0.92–1.00)**
Percent of residents with high school education	0.99 (0.98–1.00)*

Significance level: *$p < 0.1$; **$p < 0.05$
CI confidence interval

of residents with high school education were less likely to be readmitted within 30 days.

10.3.3 Discussion

Using data from multiple hospitals across two very different states, this study examined the impact of hospital and community characteristics on 30-day readmission while controlling for patient demographics and illness severity. Arkansas and Washington represent two extremes of the health profile such as smoking, drinking, obesity, and mortality. Most patient-level factors identified as significant predictors were consistent with prior literature. Older adults and patients with more severe conditions, as indicated by longer length of stay, higher Charlson comorbidity index score and use of emergency service before admission, tend to have higher

readmission risk (Hasan et al. 2009; Donzé et al. 2013; Arbaje et al. 2008; Van Walraven et al. 2010; Coleman et al. 2004; Halfon et al. 2006). The effect of gender was mixed, where some studies (Halfon et al. 2006; Krumholz et al. 1997) found that female patients were less likely to be readmitted than male patients, which are consistent with our finding, others (Yazdany et al. 2014; Bohannon and Maljanian 2002) found opposite results. Patients discharged to home health care were found at greater risk of readmission compared to patients discharged to home, which was similar to a previous study (Philbin and DiSalvo 1999) about patients with heart failure. This finding is somewhat surprising as home health care is supposed to provide additional support. It is possible that home health care could introduce additional risk if transitions of care were not properly coordinated. On the other hand, patients discharged to home health care may have more severe conditions than patients discharged to home and the variation was not captured by other predictors in our model.

After controlling for patient-level risks, we found that major teaching hospitals were associated with increased readmission risk. This finding was consistent with a prior study (Joynt and Jha 2013a) that examined the relationship between hospital characteristics and 30-day readmission. Although the underlying cause is not clear, one explanation could be that major teaching hospitals are more likely to encounter patients with medically complex conditions and diverse mix of socioeconomic background than minor or non-teaching hospitals.

Patients in our study living in a community with lower median household income were found at greater risk of being readmitted. Previous studies (Hu et al. 2014; Amarasingham et al. 2010) that investigated a community-level socioeconomic effect had a similar conclusion—patients from communities with low household income or in the lowest socioeconomic quintile had higher odds of 30-day readmission. We also found that patients from communities with higher proportion of residents with high school education had lower likelihood of 30-day readmission. This finding is similar to a previous study (Hu et al. 2014) that looked at the association between neighborhoods with low education and risk of readmission. The effects of these socioeconomic variables measured at the community level could be explained by the availability of healthcare resources, access to primary or post-discharge care services, and presence of social support.

Finally, primary care physician ratio was found to have a negative association with readmission risk, where patients living in community with higher ratio of primary care physicians were less likely to have 30-day readmission. One possibility is that with more available primary care physicians, patients may receive better follow-up care and hence reduce the likelihood of readmission. Patients from Washington State were found to have lower readmission risk than patients from Arkansas State. One hypothesis could be that Washington has heavy managed care penetration. As reported by Kaiser Family Foundation (State HMO Penetration Rate n.d.), the health maintenance organization (HMO) penetration rate in Washington was more than two times higher than Arkansas (31.5% vs. 14.2% in January 2016). A previous study (Zhan et al. 2004) has found that increased HMO penetration was significantly associated with reduced preventable hospitalizations.

Our study had several limitations. The data was from 2008 and 2009, but readmission concerns still exist as indicated by high readmission rates. In addition, hospital and community factors are understudied, and thus findings from this study can still provide insights towards a more accurate readmission measure. Patient illness severity was captured by conventional measures such as length of stay, Charlson comorbidity index, and emergency service usage. However, our data lacked clinical and medical information such as blood pressure, hemoglobin level, and adverse drug events, which were found to be associated with risk of readmission (Donzé et al. 2013; Forster et al. 2005). Having said that, this information may be more applicable for disease-specific studies than a study focuses on diverse medical conditions. We were not able to cluster admissions within patients due to convergence issues, but 83% of the patients in our data had only one admission, and thus we assumed each admission as an independent event. Finally, our use of county as a unit of analysis may fail to take into account considerable variation in access to care, socioeconomic status, or other key differences between subunits (cities or neighborhoods) within these counties.

10.4 Conclusions

In summary, we examined important care transition issues at rural and urban hospitals and provided insights to help target at-risk patients. Efforts and resources targeted towards ED transfers and non-emergent ED visits may improve the quality of care for rural residents. Rural areas need a better model to provide primary care access, which is the key to improve patients' continuity of care. The quality of ED transfers need careful evaluation, and necessary protocols and guidelines should be implemented. Hospital characteristics and patients' environment following hospital discharge are significantly associated with readmission risk. This finding underscores the concerns about the current public reporting and reimbursement policies and to what extent hospitals should be held accountable for higher readmission rates that could be attributed at least in part to their community characteristics. Financial incentives may lead to unintended consequences and exacerbate health disparities if hospitals try to avoid readmissions by limiting access for low-income patients or patients with complex health needs. Understanding more about the influence of community characteristics could inform collaborative programs between hospitals and local communities and deliver effective readmission reduction strategies.

Acknowledgments The CAH study was funded, in part, with support from the Indiana Clinical and Translational Sciences Institute funded, in part by National Institutes of Health, National Center for Advancing Translational Sciences, Clinical and Translational Sciences Award. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was also supported in part by Regenstrief Center for Healthcare Engineering and Regenstrief Foundation. We thank Dr. Steve Witz and Dr. Greg Arling for their valuable suggestions.

References

2013 National Rural Emergency Department Study (2013.) https://www.ivantagehealth.com/wp-content/uploads/2013/09/6th-Annual-ED-Study-vF2.pdf. Accessed 9 Aug 2016

Agresti A (2013) Categorical data analysis, 3rd edn. Wiley, Hoboken

Allaudeen N, Vidyarthi A, Maselli J, Auerbach A (2011) Redefining readmission risk factors for general medicine patients. J Hosp Med 6(2):54–60

Amarasingham R, Moore BJ, Tabak YP et al (2010) An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. Med Care 48(11):981–988

America's Health Rankings (n.d.) United Health Foundation. http://www.americashealthrankings.org/states. Accessed 7 July 2016

Arbaje AI, Wolff JL, Yu Q, Powe NR, Anderson GF, Boult C (2008) Postdischarge environmental and socioeconomic factors and the likelihood of early hospital readmission among community-dwelling Medicare beneficiaries. Gerontologist 48(4):495–504. https://doi.org/10.1093/geront/48.4.495

Axon R, Williams MV (2011) Hospital readmission as an accountability measure. J Am Med Assoc 305(5):504–505. https://doi.org/10.1001/jama.2011.72

Bohannon RW, Maljanian RD (2002) Hospital readmissions of elderly patients hospitalized with pneumonia. Conn Med 67(10):599–603

Brand C, Sundararajan V, Jones C, Hutchinson A, Campbell D (2005) Readmission patterns in patients with chronic obstructive pulmonary disease, chronic heart failure and diabetes mellitus: an administrative dataset analysis. Intern Med J 35(5):296–299. https://doi.org/10.1111/j.1445-5994.2005.00816.x

Calvillo-King L, Arnold D, Eubank KJ et al (2012) Impact of social factors on risk of readmission or mortality in pneumonia and heart failure: systematic review. J Gen Intern Med 28(2):269–282. https://doi.org/10.1007/s11606-012-2235-x

Casey M, Burlew M (2006) Analysis of CAH inpatient hospitalizations and transfers: implications for National Quality Measurement and Reporting. Flex Monitoring Team Briefing Paper. http://www.flexmonitoring.org/publications/bp13/. Accessed 14 Aug 2013

Casey M, Burlew M, Moscovice I (2010) Critical access hospital year 5 hospital compare participation and quality measure results. Flex Monitoring Team Briefing Paper. http://rhrc.umn.edu/wp-content/files_mf/policybriefno.15.pdf. Accessed 13 Aug 2013

Casey M, Hung P, Barton B, Moscovice I (2012) Hospital compare quality measures: 2010 National and Washington Results for Critical Access Hospitals. Flex Monitoring Team Briefing Paper. http://www.flexmonitoring.org/wp-content/uploads/2013/08/Washington2012.pdf. Accessed 13 Aug 2013

Casey MM, Moscovice I, Klingner J, Prasad S (2013) Rural relevant quality measures for critical access hospitals. J Rural Health 29(2):159–171

Clancy CM (2006) Care transitions: a threat and an opportunity for patient safety. Am J Med Qual 21(6):415–417

Coburn AF, Wakefield M, Casey M, Moscovice I, Payne S, Loux S (2004) Assuring rural hospital patient safety: what should be the priorities? J Rural Health 20(4):314–326

Coleman EA, Berenson RA (2004) Lost in transition: challenges and opportunities for improving the quality of transitional care. Ann Intern Med 141(7):533–536. https://doi.org/10.7326/0003-4819-141-7-200410050-00009

Coleman EA, Boult C (2003) Improving the quality of transitional care for persons with complex care needs. J Am Geriatr Soc 51(4):556–557. https://doi.org/10.1046/j.1532-5415.2003.51186.x

Coleman EA, Min S, Chomiak A, Kramer AM (2004) Posthospital care transitions: patterns, complications, and risk identification. Health Serv Res 39(5):1449–1466

Coleman EA, Smith JD, Raha D, Min S (2005) Posthospital medication discrepancies: prevalence and contributing factors. Arch Intern Med 165(16):1842–1847

Critical Access Hospitals (CAHs) (n.d.) Introduction—rural health information hub. https://www.ruralhealthinfo.org/topics/critical-access-hospitals. Accessed 11 Aug 2016

De Freitas TL, Spooner GR, Szafran O (1998) Admissions and transfers from a rural emergency department. Can Fam Physician 44:789–795

Donzé J, Aujesky D, Williams D, Schnipper JL (2013) Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. JAMA Intern Med 173(8):632–638. https://doi.org/10.1001/jamainternmed.2013.3023

Elixhauser A, Steiner C, Harris DR, Coffey RM (1998) Comorbidity measures for use with administrative data. Med Care 36(1):8–27

Elixhauser A, Steiner C, Palmer L (2013) Clinical classifications software (CCS)

Faculty & Research|NYU Wagner (n.d..) http://wagner.nyu.edu/faculty/billings/nyued-background. Accessed 11 Aug 2016

Fan L, Shah MN, Veazie PJ, Friedman B (2011) Factors associated with emergency department use among the rural elderly. J Rural Health 27(1):39–49. https://doi.org/10.1111/j.1748-0361.2010.00313.x

Forster AJ, Murff HJ, Peterson JF, Gandhi TK, Bates DW (2003) The incidence and severity of adverse events affecting patients after discharge from the hospital. Ann Intern Med 138(3):161–167

Forster AJ, Murff HJ, Peterson JF, Gandhi TK, Bates DW (2005) Adverse drug events occurring following hospital discharge. J Gen Intern Med 20(4):317–323. https://doi.org/10.1111/j.1525-1497.2005.30390.x

Gresenz CR, Rogowski J, Escarce JJ (2007) Health care markets, the safety net, and utilization of care among the uninsured. Health Serv Res 42(1p1):239–264. https://doi.org/10.1111/j.1475-6773.2006.00602.x

Halfon P, Eggli Y, Pretre-Rohrbach I, Meylan D, Marazzi A, Burnand B (2006) Validation of the potentially avoidable hospital readmission rate as a routine indicator of the quality of hospital care. Med Care 44(11):972–981. https://doi.org/10.1097/01.mlr.0000228002.43688.c2

Hasan O, Meltzer DO, Shaykevich SA et al (2009) Hospital readmission in general medicine patients: a prediction model. J Gen Intern Med 25(3):211–219. https://doi.org/10.1007/s11606-009-1196-1

Henriksen M, Walzer N (2012) Illinois critical access hospitals: enhancing quality of care in rural Illinois. Center for Governmental Studies, DeKalb. http://www.cgsniu.org/services/CWED/Health_Care_Policy/ICAHN/icahn-04.10.2012-Issue1.pdf. Accessed 14 Aug 2013

Herrin J, St Andre J, Kenward K, Joshi MS, Audet A-MJ, Hines SC (2015) Community factors and hospital readmission rates. Health Serv Res 50(1):20–39

Hines A, Fraze T, Stocks C (2011) Emergency department visits in rural and non-rural community hospitals, 2008. http://www.ncbi.nlm.nih.gov/books/NBK56307/. Accessed 10 Sept 2013

Horwitz L, Partovian C, Lin Z et al (2012) Hospital-wide all-cause unplanned readmission measure: final technical report. Centers for Medicare and Medicaid Services, Baltimore

Hu J, Gonsahn MD, Nerenz DR (2014) Socioeconomic status and readmissions: evidence from an urban teaching hospital. Health Aff 33(5):778–785. https://doi.org/10.1377/hlthaff.2013.0816

Jencks SF, Williams MV, Coleman EA (2009) Rehospitalizations among patients in the Medicare fee-for-service program. N Engl J Med 360(14):1418–1428

Jiang HJ, Andrews R, Stryer D, Friedman B (2005) Racial/ethnic disparities in potentially preventable readmissions: the case of diabetes. Am J Public Health 95(9):1561–1567. https://doi.org/10.2105/AJPH.2004.044222

Joynt KE, Jha AK (2013a) Characteristics of hospitals receiving penalties under the hospital readmissions reduction program. J Am Med Assoc 309(4):342–343. https://doi.org/10.1001/jama.2012.94856

Joynt KE, Jha AK (2013b) A path forward on Medicare readmissions. N Engl J Med 368(13):1175–1177. https://doi.org/10.1056/NEJMp1300122

Joynt KE, Harris Y, Orav EJ, Jha AK (2011a) Quality of care and patient outcomes in critical access rural hospitals. J Am Med Assoc 306(1):45–52

Joynt KE, Orav E, Jha AK (2011b) Thirty-day readmission rates for Medicare beneficiaries by race and site of care. J Am Med Assoc 305(7):675–681. https://doi.org/10.1001/jama.2011.123

Kansagara D, Englander H, Salanitro A et al (2011) Risk prediction models for hospital readmission: a systematic review. J Am Med Assoc 306(15):1688–1698. https://doi.org/10.1001/jama.2011.1515

Kripalani S, Jackson AT, Schnipper JL, Coleman EA (2007) Promoting effective transitions of care at hospital discharge: a review of key issues for hospitalists. J Hosp Med 2(5):314–323

Krumholz HM, Parent EM, Tu N et al (1997) Readmission after hospitalization for congestive heart failure among Medicare beneficiaries. Arch Intern Med 157(1):99–104

Lindenauer PK, Lagu T, Rothberg MB et al (2013) Income inequality and 30 day outcomes after acute myocardial infarction, heart failure, and pneumonia: retrospective cohort study. BMJ 346:f521. https://doi.org/10.1136/bmj.f521

Medicare Hospital Quality Chartbook 2012: Performance Report on Outcome Measures (2012) Prepared by Yale New Haven Health Services Corporation Center for Outcomes Research and Evaluation. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-AssessmentInstruments/HospitalQualityInits/Downloads/MedicareHospitalQualityChartbook2012.pdf. Accessed 7 July 2016

Michelle Casey MS (2014) Which Medicare patients are transferred from rural emergency departments?. http://rhrc.umn.edu/wp-content/files_mf/whichmedicarepatientsaretransferred.pdf. Accessed 11 Aug 2016

Moore C, Wisnivesky J, Williams S, McGinn T (2003) Medical errors related to discontinuity of care from an inpatient to an outpatient setting. J Gen Intern Med 18(8):646–651

Philbin EF, DiSalvo TG (1999) Prediction of hospital readmission for heart failure: development of a simple risk score based on administrative data. J Am Coll Cardiol 33(6):1560–1566. https://doi.org/10.1016/S0735-1097(99)00059-5

Prasad S, Klingner JM, Moscovice I (2011) Care transitions: "time to come home". http://ruralhealth.und.edu/pdf/umrhrc_finalreport0311.pdf. Accessed 23 Sept 2013

Quan H, Sundararajan V, Halfon P et al (2005) Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care 43:1130–1139

Quan H, Li B, Couris CM et al (2011) Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. Am J Epidemiol 173(6):676–682. https://doi.org/10.1093/aje/kwq433

State HMO Penetration Rate. Kaiser Family Foundation (n.d..) http://kff.org/other/state-indicator/hmo-penetration-rate/. Accessed 2 Aug 2016

Van Walraven C, Dhalla IA, Bell C et al (2010) Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. Can Med Assoc J 182(6):551–557

Weissman JS, Stern RS, Epstein AM (1994) The impact of patient socioeconomic status and other social factors on readmission: a prospective study in four Massachusetts hospitals. Inquiry 31(2):163–172

Xie S (2018) Risk stratification and prediction for tailored population health management interventions. Dissertation, Purdue University

Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation (YNHHSC/CORE) (2014) 2014 Measure Updates and Specifications Report: Hospital-Wide All-Cause Unplanned Readmission—Version 3.0

Yazdany J, Marafino BJ, Dean ML et al (2014) Thirty-day hospital readmissions in systemic lupus erythematosus: predictors and hospital and state-level variation. Arthritis Rheumatol 66(10):2828–2836. https://doi.org/10.1002/art.38768

Zhan C, Miller MR, Wong H, Meyer GS (2004) The effects of HMO penetration on preventable hospitalizations. Health Serv Res 39(2):345–361. https://doi.org/10.1111/j.1475-6773.2004.00231.x

Shan Xie recently received her Ph.D. in Industrial Engineering from Purdue University. Her research area is in healthcare delivery system and data analytics. She worked at the Regenstrief Center for Healthcare Engineering at Purdue University as a graduate research assistant for 5 years. During this period, she focused on research relating to population health management, risk stratification modeling, and health outcome evaluation. Shan has many years of experience working with different types of healthcare data. In 2012, she received her M.S. in Industrial Engineering—with a concentration in Human Factors—from Purdue University. Shan always had passion for STEM and was determined to study Engineering since high school. Her undergraduate degree was in Mechanical Engineering. She enjoys applying engineering skills to solve practical problems and especially appreciates the opportunity to learn healthcare issues and help improve healthcare system. She is motivated and excited to continue working in STEM fields and develop her career in healthcare.

Dr. Yuehwern Yih is the Academic Director of LASER PULSE Consortium, Associate Director of Regenstrief Center for Healthcare Engineering at the Discovery Park, and Professor of Industrial Engineering at Purdue University. Her expertise resides in system and process design, monitor, and control to improve its quality and efficiency for complex systems, such as health systems, manufacturing, and supply chains. Dr. Yih published over 150 scientific articles and book chapters, four edited books, and a patent on system engineering and management. Her Handbook of Healthcare Delivery Systems is the first handbook covering the wide arrays of sectors in healthcare delivery systems to provide a holistic view of healthcare delivery as an integrated system. Her contributions in this area have been recognized by a National Science Foundation Young Investigator Award (NYI), a Dell K. Allen Outstanding Young Manufacturing Engineer Award, GE Faculty Fellow, NEC Faculty Fellow, Institute for Industrial (and Systems) Engineers (IIE) Fellow, Purdue Engagement Faculty Fellow Award, and Executive Leadership in Academic Technology and Engineering (ELATE) Fellow. Dr. Yih received the highest honor at Purdue in engagement, the Faculty Engagement Fellow Award, based on her work at AMPATH that designs and implements a nutrition information system and a food distribution system for HIV patients in Western Kenya. This integrated nutrition system was deployed in 2005 and provided food for over 38,000 HIV patients and their families each year. Her work was also featured in Industrial Engineer Magazine cover article in January 2014. Currently, Dr. Yih is working on connecting patient demands to health supply chain operations to improve maternal and child health in Uganda, funded by the Melinda and Bill Gates Foundation Grand Challenge Grant.

"I started my engineering training at the National Tsing Hua University where I received my B.S. degree in Industrial Engineering. Later, I continued to pursue in advance degree in Industrial Engineering in the States and received the Ph.D.

degree from the University of Wisconsin-Madison. Since I was young, I always enjoyed solving logic puzzles and analyzing the causal effects. I am extremely lucky because Industrial Engineering is a perfect match for me. With both parents being educators, I was encouraged and given the opportunity to be my best and do what I love, thanks to my parents. I went to the best girls' high school in Taiwan and my classmates all went to colleges to do majors in engineering, math, physics, or chemistry, many with Ph.Ds. Although in college I was one of the very few women on campus because it was an engineering school, it didn't really strike me that being a woman in engineering comes with additional challenges until I started my faculty career. Regardless of all the challenges, I love what I do because every day I know I have opportunities to make a positive difference in people's lives."

Chapter 11
To Be Healthy, Wealthy, and Wise: Using Decision Modeling to Personalize Policy in Health, Hunger Relief, and Education

Julie Simmons Ivy, Muge Capan, Karen Hicklin, Nisha Nataraj, Irem Sengul Orgut, Amy Craig Reamer, and Anita Vila-Parrish

Contents

11.1 Introduction: Dr. Julie Simmons Ivy 234
11.2 To Be Healthy 236
11.2.1 Mode of Delivery: Dr. Karen Hicklin 236
11.2.2 The Care of Complex Patients: Using Simulation Frameworks to Study Decisions—Dr. Nisha Nataraj 241
11.2.3 Optimal Patient-Centered Response to Acute Physiological Deterioration of Hospitalized Patients: Dr. Muge Capan 247
11.2.4 Using Operations Research Methods to Improve the Medication Supply Chain and Its Connection to Patient Care: Dr. Ana Vila-Parrish 252
11.3 To Be Wealthy 255
11.3.1 To Be Equitable and Effective: Using Mathematical Modeling to Reduce Food Insecurity—Dr. Irem Sengul Orgut 255
11.4 To Be Wise 260

J. S. Ivy (✉) · N. Nataraj
North Carolina State University, Raleigh, NC, USA
e-mail: jsivy@ncsu.edu; nnatara@ncsu.edu

M. Capan
Drexel University, Philadelphia, PA, USA
e-mail: Muge.Capan@drexel.edu

K. Hicklin
University of North Carolina—Chapel Hill, Chapel Hill, NC, USA
e-mail: kthickli@ncsu.edu

I. S. Orgut
University of Alabama, Tuscaloosa, AL, USA
e-mail: isengul@ncsu.edu

A. C. Reamer
North Carolina State University, Raleigh, NC, USA

University of North Carolina-Wilmington, Wilmington, NC, USA
e-mail: aecraig@ncsu.edu

A. Vila-Parrish
Gartner, Stamford, CT, USA

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_11

11.4.1 Student Performance in Mathematics over Time: A Social Application of a Markovian Model to an Education System—Dr. Amy Craig Reamer 260
11.5 Conclusions 265
References 266

11.1 Introduction: Dr. Julie Simmons Ivy

This chapter presents an overview of non-traditional applications of industrial and systems engineering that have served as the basis for my research and is a collaborative effort with six of my female former doctoral students. Their research is highlighted in the form of vignettes that summarize the decision approached used to address human-centered problems in the areas of health, humanitarian logistics, and education. The National Academy of Engineering's Grand Challenges challenged the engineering community to "Engineer better medicines," "Improve health informatics," and "Advance personalized learning," our research begins to address these challenges and extends them to include engineering a better future for our poorest citizens. This research develops theory as derived from these real-world problems particularly in the areas of stochastic modeling, Markov decision processes (MDPs), semi-Markov decision processes (SMDPs), partially observable Markov decision processes (POMDPs), and Bayesian decision analysis. Simulation, deterministic optimization, robust optimization, and stochastic programming methods are also incorporated. These methods are used to personalize solutions to the needs of the individual.

The theme of this chapter is to be healthy, wealthy, and wise. Each section reflects the voice of one of the researchers in the form of a case study. Each case study introduces a complex societal issue related to health, poverty, or education, presents a summary of the researcher's contributions in that area focusing on the role of decision modeling, highlighting unique challenges associated with addressing the particular challenge, and summarizes the researcher's vision for future work in the area. These vignettes are designed to stand-alone.

In the context of health, four vignettes are presented in the areas of (1) medical/clinical decision-making related to birth and the care of complex patients, (2) care delivery related to bedside patient rescue, and (3) clinical operations related to pharmaceutical inventory management. Current policies are "one size fits all," the goal of this research is to develop personalized screening, intervention, treatment, and disease management policies and strategies. Unfortunately, we do not have a magic recipe for wealth so instead in the context of wealth, we present the fifth vignette in which we consider the inverse problem of satisfying hunger need in a community. In the context of education, we present the sixth vignette in which we characterize elementary and middle school student performance on standardized exams over time optimizing student outcomes. Each of these societal challenges involves complex decision-making with multiple attributes under conditions of uncertainty. The research presented seeks to inform this decision-making, improve

decision quality, and hopefully improve outcomes in the process. A brief overview of each section is presented in the discussion that follows.

Section 11.2.1 introduces the research of Dr. Karen Hicklin who has modeled decision-making related to birth with a focus on the decision to have a cesarean section as a function of the patient's current state and future risk. This research develops various decision models to capture decision-making during labor. It takes a novel look at the value of information using Bayesian decision theory to determine not only the value of information but also the conditions under which the information has most value. In addition, a discrete event simulation model is developed to replicate birth progression for a population of women. This simulation model served as the data seed for the Bayesian decision model and a MDP model developed to capture the dynamic nature of decision-making during labor.

Section 11.2.2 introduces the research of Dr. Nisha Nataraj, which considers the management of complex patients, patients with one or more comorbid conditions, in the context of diabetes management and sepsis diagnosis and treatment. This research integrates cluster analysis, logistic regression, and simulation: (1) to characterize the impact of diabetic medication management on diabetic women's risk for developing breast cancer; and (2) to model the stochastic evolution of the sepsis trajectory of comorbid patients is modeled, to evaluate the impact of interventions such as fluids and anti-infectives, and to identify those comorbidities that behave similarly in patients along sepsis. This research was conducted in collaboration with Mayo Clinic and Christiana Care Health System in partnership with the NSF-sponsored S.E.P.S.I.S. research collaborative (NSF IIS1522107).

Section 11.2.3 introduces the research of Dr. Muge Capan and is at the intersection of medical decision-making and care delivery. This research in collaboration with Mayo Clinic models decision-making related to in-hospital patient deterioration. This research developed personalized models to capture patient deterioration in order to optimize response. In particular, this research focused on decision-making as it relates to Rapid Response Teams (RRT), these are teams of physicians with a higher level of expertise from the intensive care unit who may be called to the general floor to assist with patient care in response to patient deterioration. This research used electronic medical records to develop patient-specific SMDP models to capture patient condition dynamics and the significant role time plays in response to deterioration to identify personalized policies for RRT activation. This research has motivated and served as a testbed for an NSF-sponsored Smart and Connected Health collaborative research project with Mayo Clinic and Christiana Care Health System on sepsis (NSF IIS1522107).

Section 11.2.4 introduces the research of Dr. Ana Vila-Parrish in the area of healthcare delivery and logistics focusing on pharmaceutical inventory management and policy development. This research identifies hospital-based pharmaceutical inventory management policies considering patient condition dynamics and drug perishability. Additional studies incorporate the impact of patient condition dynamics during an influenza outbreak on hospital pharmaceutical inventory management and the impact of centralized and decentralized pharmacy structure on pharmaceutical inventory management.

Section 11.3.1 introduces the research of Dr. Irem Sengul Orgut that focuses on humanitarian logistics particularly as it relates to hunger relief. This research is based on a close to 8 year collaboration with the Food Bank of Central and Eastern North Carolina that has led to an NSF-funded collaborative research project with Dr. Lauren Davis of North Carolina A&T University (CMMI 1000828) and a new NSF-supported smart service system project, F.E.E.E.D. This research considers the equitable and effective distribution of donated food considering capacity constraints. Through this research the critical role that capacity constraints play in equitable food distribution was identified and the concept of a "bottleneck" county that constrains distribution if there are equity requirements was introduced.

Section 11.4.1 introduces the research of Dr. Amy Craig Reamer in the area of education. This research develops stochastic models of the evolution of elementary and middle school student end-of-grade (EOG) exam performance in mathematics over time. EOG exam data from 1996 to 2009 for students in third through eighth grade in North Carolina is used to inform the understanding of student performance on annual standardized tests in mathematics over time with the goal of determining strategies for when and how to optimally intervene to improve performance. Further, the research considers the impact of student attributes on performance to inform how to personalize mathematics education.

This collection of six vignettes highlights the potential for using industrial and systems engineering methods and tools to address some of society's greatest challenges in health care, hunger relief, and education. Analytical thinking, mathematical and computer modeling, and optimization methods provide a structural underpinning necessary for addressing these complex decision problems. With the increasing availability of data, the rise of analytics, and advances in computational capability, it is now more possible for modeling to capture the complexities associated with these problems and to influence how researchers and practitioners address complex societal issues, such as health and education disparities, hunger relief, and personalized medical decision-making.

11.2 To Be Healthy

11.2.1 Mode of Delivery: Dr. Karen Hicklin

Introduction

Thirty-three-year-old Robin and her husband, Randy, are expecting their first child. They went to parenting classes, consulted with the obstetrician about the labor process, and felt prepared. At 38 weeks gestation, they knew that any day now they would be the proud parents of a baby boy. One night at approximately 10:00 pm, Robin began to experience contractions. Not too worried, she decided that if the contractions persisted, they would head to hospital. A few hours later, around 2:00 am, they head to the hospital and upon arriving, they were informed that Robin

was 1–2 cm dilated and since she was having persistent contractions they would admit her to the hospital. Around 7:00 am, there were no further dilation changes and the doctor decided to break her water and start her on pitocin (a drug used to speed up labor). She received an epidural shortly after and had no pain until afternoon. At 1:00 pm, Robin started to experience more intense and continuous contractions but was measured to be only 3 or 4 cm dilated. Robin and Randy were told that Robin was not progressing, and they should think about a C-section. Although she would be given the opportunity to labor an additional 10 h, Robin was informed that there may not be much progress. Around 1:30 pm, although afraid and unprepared, it was decided that Robin would undergo a cesarean delivery.

This story provokes questions of whether that decision was the best and was the timing of the decision appropriate.

Each day millions of women enter labor and in 2016, there were approximately 3.9 million births in the United States (Hamilton et al. 2017). When a woman enters labor, she will deliver in one of two ways: vaginal delivery (or normal delivery) or cesarean delivery (C-section). A C-section is a surgical procedure in which an incision is made in the mother's abdomen and uterus for the delivery of a fetus. The current rate of cesarean delivery is 31.9% (Hamilton et al. 2017), which is more than triple the rate suggested by the World Health Organization who believe a rate between 10 and 15% leads to the best health outcomes for the mother and child (World Health Organization Human Reproduction Programme 2015). The first cesarean rate was 5.0% and was recorded in 1965 (Taffel et al. 1987). Despite the increase in the number of C-sections over the years, there has not been a decrease in the number of morbidities as one might expect.

For some women, a C-section is the most appropriate delivery mode due to factors such as multiple gestations, hypertensive disorders, fetal distress, or labor arrest. However, there are concerns regarding the number of unnecessary C-sections. Those would be C-sections given to patients who were ideal candidates for vaginal delivery but received a C-section instead. A large proportion of these births are due to the notion of failure-to-progress or labor arrest. A C-section due to a failure-to-progress diagnosis means it was determined that the patient will not progress to full cervical dilation within a particular time frame deemed necessary for safe delivery. Dr. Emmanuel Friedman was one of the first obstetricians to divide labor into phases and stages with the goal of identifying abnormal labor. His work led to the Friedman Curve, which was established in the mid-1950s. The two major outcomes of this curve stated that (1) nulliparous women (women giving birth for the first time) should have cervical dilation progression of 1.2 cm per hour and multiparous women (women who have given birth before) (Hamilton et al. 2017) should have cervical dilation progression of 1.5 cm per hour and (2) no cervical dilation change in 2 h is an indication of labor arrest. Although the Friedman Curve had been used to diagnose abnormal labor for many years, many believe its use may be outdated and that following such a curve may have played a part in the increase of C-sections due to failure-to-progress (Zhang et al. 2010a, b). A retrospective study across 19 US hospitals conducted by the Consortium on Safe Labor showed that

women are not laboring as once believed and some women need more time in labor before deciding a C-section is needed (Zhang et al. 2010a).

Labor is often described in three stages. The first stage is the time period from the onset of labor until 10 cm, the second stage is the period from 10 cm to the delivery of the baby, and the third stage is the delivery of the placenta. At any time during the first or second stages of labor, a C-section may be performed to ensure safe delivery of the fetus. For women who indeed need a C-section, performing the procedure as early as possible is best. This timely decision is necessary to avoid adverse complications to the mother or child by prolonging labor for a woman whose best delivery mode is to have a C-section. However, for women where the decision of remaining in labor is unclear, allowing more time in labor provides additional insight into how the patient is progressing. Deciding to perform a C-section for such a patient could lead to an unnecessary C-section and also increase the risk of future complications as well as increase the need for subsequent C-sections in future pregnancies. The decision may also change depending on the stakeholder and his or her preferences, experience, and valuation of the two delivery modes (i.e., utility). Due to this complicated decision process, we have developed a simulation model and stochastic decision models to evaluate the conditions for when a C-section is needed considering a failure-to-progress diagnosis.

Modeling Approaches

Discrete Event Simulation Discrete event simulation is used to model the natural progression of labor. This model has two goals—to model the natural progression of labor in the absence of C-sections and to model various stopping rules for ending labor for a C-section and the effect those rules have on the C-section rate and rate of expected complications. In order to model the natural labor process, we used percentiles of how long a woman is expected to be at a particular dilation state as reported in publicly available literature sources (Harper et al. 2012; Zhang et al. 2010b). Using these values in a percentile matching procedure, we were able to develop probability distributions for each dilation state for three different laboring types: (1) spontaneous labors, (2) augmented labors, and (3) induced labors. By modeling the natural progression of labor, we were able to identify the average length of active labor (i.e., time from 3 cm to vaginal delivery). This work also provided insight into the length of labor for different labor patterns (i.e., spontaneous, augmented, and induced labors). By evaluating the expected risk of complications for remaining in labor and the C-section rate as a function of various stopping rules, we have been able to identify stopping rules that could lead to lower C-section rates while also lowering the rate of complications (Hicklin 2016). The discrete event simulation not only provides insight into labor progression but was also used as input into stochastic decision models, which will be discussed in the subsequent sections.

Bayesian Decision Analysis In the theory of Bayesian statistics, there is a belief regarding the "true value" of a particular parameter and that belief is updated as

more information is collected. In the context of the mode of delivery decision, there is a true mode of delivery and the decision-maker has a belief as to what that mode should be. Each woman who enters labor will either deliver vaginally or through cesarean delivery. However, the optimal mode of delivery is not always apparent during the labor process. In a Bayesian decision model, we determine the optimal mode of delivery as a function of the belief that the patient will have a successful vaginal delivery. We evaluate the trade-off between making the decision of delivery mode with information currently available (prior) or prolong labor in order to learn more about the patient before making a decision (posterior). The decision-maker learns more about the patient by taking observations of labor progression. This model provides insight into the value of information in the context of waiting in labor.

We tested our modeling framework using three different methodologies. In the first methodology, we assume labor progression (in the context of cervical dilation changes) can be modeled as an exponential probability distribution and define an observation to be whether there is change in dilation from 1 h to the next. In the second approach, we still define an observation to be change or no change but model labor progression according to a lognormal probability distribution. The state space of the first and second approaches are more simplistic than the third, in which an observation is not only whether there is change in dilation but how much change occurred in an hour. That is, for a patient currently reported to be 7 cm, that patient may remain at 7 cm, move to 8, 9, or 10 cm, or delivery during the next hour. This model uses the same probability distribution as the second approach, but accounts for a more dynamic observation space and provides additional insight to better inform the model. The probability is calculated based on the discrete event simulation discussed previously. In each approach, the results provide the belief values in which it is optimal to (1) continue labor with routine monitoring, (2) continue labor with caution, and (3) end labor for a C-section. In the case where the decision is to continue labor with caution, the decision of best delivery mode is not certain and more information is needed before for delivery mode is determined (Hicklin et al. 2017).

Markov Decision Process In order to evaluate the trade-off between continuing labor and performing a C-section considering preferences and risk, we develop an infinite-horizon MDP. In this model, the objective is to maximize the expected utility of health outcomes for the mother and child as a function of time in labor and delivery mode. The state is defined by three elements: (1) current dilation state (i.e., 3 cm, 4 cm, ..., 10 cm) or delivery mode (i.e., vaginal delivery, C-section, emergency C-section), (2) time spent in the first stage of labor, and (3) time spent in the second stage. Similar to the Bayesian model, the probability of transitioning from one state to another is derived from the discrete event simulation model, which assumes labor progression is log-normally distributed. We calculate the expected value of health outcomes considering four different scenarios. The first scenario is when both the mother and child are healthy and have no complications, the second scenario is when the mother is healthy, but the child has a complication, the third

scenario is when the mother has a complication and the child is healthy, and the last scenario is when both the mother and child have complications. Using probability estimates of complications for each patient and allocating the highest preference (value of 1) to the first scenario, lowest preference (value of 0) to the last scenario, and equal weighting for the second and third scenario (value of 0.5), we calculate the expected utility of health outcomes. From this utility value, we subtract a disutility value that represents delivery mode. Once delivery is achieved in this model, the expected rate of complications as a function of total time in both stages is subtracted from the net utility. The action (i.e., continue labor or C-section) which results in the maximum value of utility provides the optimal decision for that time period. The results provide the total number of hours a patient should remain in labor at each cervical dilation state for varying disutility values for delivery modes. If we fix the disutility of vaginal delivery to be the lowest (value of 0) and the disutility of emergency C-section to be the highest (value of 1) and vary the disutility for C-section, the decision is to allow patients to remain in labor longer for higher values of C-section disutility.

Implications and Next Steps

Each model provides a significant connection to understanding what factors prompt the decision-maker to decide a laboring woman needs a C-section due to the notion that she is not progressing in a manner in which she will achieve vaginal delivery within a time frame deemed for safe delivery. The simulation model allows us to understand how long labor can last if not interrupted and if labor is not interrupted what are the resulting complications. In turn, this leads to the question of whether those probable complications outweigh the short- and long-term complications for the mother and child of performing a C-section. The Bayesian and Markov models both evaluate trade-offs of allowing labor to continue versus ended labor for a C-section. The Bayesian model provides insight into the value information associated with waiting in labor and the MDP provides an understanding of how long a patient can be expected to remain in labor in order to achieve a particular utility value considering the resulting disutility of expected complications that may arise. Given the value each model provides, we have also developed a Bayesian MDP model that looks at the total time a woman can safely remain in labor as a function of the decision-maker's belief of the patient's success as a candidate for vaginal delivery.

Our contributions in this area have provided many platforms to help stakeholders make wiser and better decisions for women's health. It has also provided a means to facilitate shared decision-making in which both patients and providers are able to share their concerns and determine the best decision together. The ultimate goal of this work has been to develop interfaces that allow providers to consult real-time for assistance in decision-making or to be used between patients and providers to understand the potential risk and complications that may arise before labor occurs.

11.2.2 The Care of Complex Patients: Using Simulation Frameworks to Study Decisions—Dr. Nisha Nataraj

Introduction

With advances in personalized medicine, accounting for heterogeneity in patient care is becoming more attainable. Still, an aging population, changing lifestyles and social structures, and the growing burden of chronic disease (Anderson and Horvath 2004) have resulted in a growth in the number of complex patients. Complexities may arise due to a variety of factors, including biological, environmental, demographic, socioeconomic, behavioral, and cultural factors (Safford et al. 2007; Schaink et al. 2012; Loeb et al. 2015), making diagnosis and treatment more difficult.

One of the most common manifestations of complexity is multimorbidity (Schaink et al. 2012), the presence of multiple conditions in an individual. Often used synonymously with comorbidity (the simultaneous presence of one or more conditions with respect to an index disease), multimorbidity is highly prevalent—Approximately one in four US adults have two or more chronic conditions (Ward et al. 2014). Consequently, the economic burden of multimorbidity and associated resource utilization are also high. In fact, 71% of all total healthcare spending in the United States in 2010 was on patients with multiple chronic conditions (MCC) (Gerteis et al. 2014). In 2010, individuals with MCC contributed to 88% of all home health care visits, 83% of all prescriptions, 70% of all inpatient stays, and 64% of all clinician visits (Gerteis et al. 2014). This burden extends beyond the strain on healthcare spending and resources. Individuals with MCC have a higher risk for preventable hospitalizations and complications (Wolff et al. 2002), generally lower quality of life (Vogeli et al. 2007), and greater mortality risk (Chang et al. 2012; Nunes et al. 2016). The management of MCC can also strain primary care-givers and family, partly due to limited mobility and increased expenditures (Giovannetti et al. 2012). Polypharmacy is also a major concern because individuals are often on several different medications at once, making keeping track of adverse reactions and medication interactions challenging (Marengoni and Onder 2015; Mannucci et al. 2014).

Decision Uncertainty

There is a pressing need to evaluate how comorbidities impact prognosis and outcomes for patients with multiple conditions. Shifting to a multiple disease perspective will enable a more accurate representation of the impact and the interaction of the diseases on costs, resource allocation, treatment priorities, etc. However, there are several difficulties from a decision-making standpoint when trying to capture the impact of more than one disease in individuals. From a clinical perspective, heterogeneity in patients, comorbidities, difficulty translating

disease-specific guidelines for applicability in a multi-disease setting, and lack of data are some of the barriers present (Caughey et al. 2011). From a modeling perspective, adequately representing more than one disease poses unique challenges, including determining how to account for comorbidity, interactions between the diseases, and cause of death for multiple diseases, as well as limitations in the availability of comprehensive data sources. While statistical models can be a useful way of characterizing comorbidity (Zhang et al. 2013; Nataraj et al. 2018), they have limited applicability for decision-making and studying disease trajectories at an individual level. Simulation has several advantages that make it particularly amenable to the modeling of complex patients, such as being able to account for (1) multiple sources of stochasticity, (2) complex systems with numerous moving pieces, and (3) the critical role played by humans (Brailsford 2007).

In this case study, we present two perspectives outlining the use of simulation frameworks for modeling complexity. The details of these studies are available in Nataraj (2017). We examine multimorbidity from a systems standpoint by discussing how simulation models can be used to study the impact of comorbidity from two different index-disease perspectives, diabetes and sepsis. We restrict our attention to these conditions since patients with diabetes and sepsis are both complex, partly due to increased prevalence of comorbidities and the risk of complications. While there are some similarities, complexity can manifest differently in these two conditions, making for challenging disease management and care. Diabetes, a metabolic disease that affects the way in which the body can process glucose, is a chronic condition where decisions are typically made over a longer time horizon, over months and, possibly, years. Conversely, sepsis, the body's inflammatory response to infection, is an acute complication that can quickly result in organ failure and a high likelihood of death (Liu et al. 2014). With sepsis, decisions must be made swiftly—on an hourly and daily basis. Therefore, the separate study of these conditions can offer valuable insights into decision-making for complex patients.

Perspective 1: Sepsis as an Index Disease

One particularly challenging aspect of sepsis is the difficulty in detecting the condition. This is primarily due to the lack of a gold-standard diagnostic test. Additionally, the lines between the transitions along the sepsis trajectory are often not clear-cut, for the purposes of clinical decision-making as well as definition. As a result, a significant focus of the international community has been on addressing the challenge of defining sepsis clinically (Czura 2011; Vincent et al. 2013; Singer et al. 2016) using physiological (patient vitals) and cellular (lab results) markers of the condition. Patients can deteriorate rapidly through states along the sepsis spectrum making timely intervention both critical and difficult—sepsis can progressively worsen to severe sepsis, i.e., organ dysfunction and septic shock. The condition can present and progress very differently in patients depending on their comorbidities

(Wang et al. 2012; Iskander et al. 2013) as well as their age, race, and gender (Esper et al. 2006).

Objective Recent efforts to better define the sepsis spectrum clinically have resulted in multiple criteria for the identification of the onset of sepsis (Vincent et al. 2013; Singer et al. 2016). The trajectory of a sepsis patient is difficult to clearly map, particularly because fragmented care makes it challenging for providers to take into account a patient's medical history (Cox and Wysham 2015). Using patient and visit-level electronic health records (EHR) data from a large hospital system, the study uses a simulation framework to understand how relevant dynamic physiological and cellular attributes evolve over time after a therapeutic intervention in patients with sepsis, while considering patient heterogeneity through age, gender, race, and comorbidity. The framework is designed to allow the comparison and quantification of the impact of different clinical definitions on the timely identification of sepsis states, given the implementation of recommended therapeutic interventions.

Methods Inpatient EHR data from a large hospital system was used to develop a comprehensive discrete event simulation (DES)-based framework to study patient trajectories along the sepsis spectrum in a cohort of individuals with ICD-9-CM coded sepsis. The framework allows us to study the evolution of patient trajectories, specified by different sepsis definitions and driven by recommended therapeutic interventions (Fig. 11.1). Inputs to the framework consist of (1) a database of septic patients, derived from EHR data, (2) definitions for sepsis identification with clinical

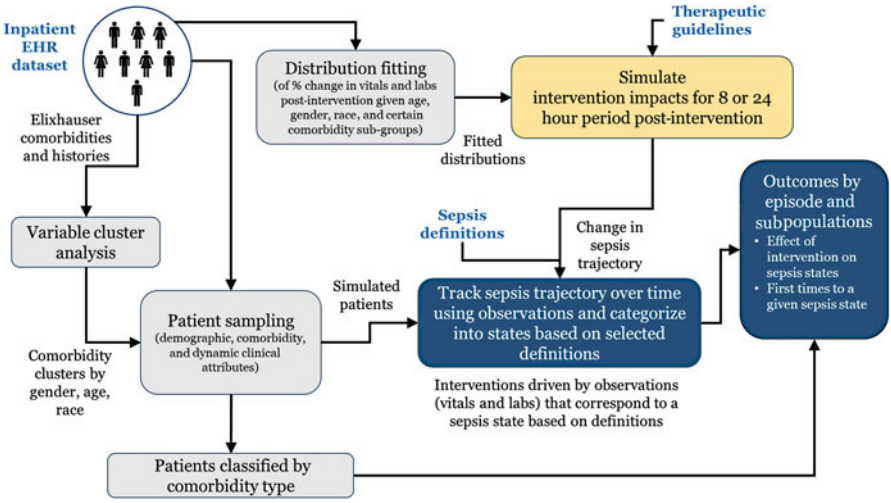


Fig. 11.1 Overview of the comprehensive DES framework modeling the impact of definition and interventional guidelines on sepsis patient trajectories. Text in blue represents inputs to the framework, gray boxes indicate inputs to the simulation, navy boxes represent model components, and the yellow box indicates the simulation component

markers for component states distinctly described, and (3) Timing-based guidelines for therapeutic interventions. First, heterogeneity in the population is addressed by using hierarchical clustering to identify important comorbidities in subpopulations stratified by age, gender, and race. Patients at the visit level are then simulated from the EHR data by sampling patients' static and dynamic clinical, demographic, and comorbidity-related attributes in their entirety. Patient episodes, defined as a portion of a patient's visit specified by a hospital unit such as the ED or ICU, are tracked. The EHR data is also employed to estimate percent changes in patients' vitals and labs in the hours prior to and following an intervention, given their demographic and comorbidity-related features. These are then fit to distributions to determine the impact of a therapeutic intervention in each subpopulation.

The simulated patients, comorbidity clusters, and fitted distributions then become inputs to the DES component of the framework. Using dynamic clinical attributes from the EHR data, the simulation model (1) tracks the trajectory of patients over time, as categorized into states by the different clinical definitions specified; (2) simulates the first impact of a therapeutic intervention for each state within an episode (timed in accordance with the specified guidelines) on vitals and labs by generating a random number from the fitted distributions; and (3) determines the new post-intervention state using the specified definitions. Outcomes studied include state-changes in patients after an intervention and times taken to first identify a given sepsis state under each definition at the visit or episode levels.

Key Findings Results indicated that sepsis progressed differently in individuals with a larger number of comorbidities, underscoring the need to tailor interventions for this population. Of note was the finding that histories and current comorbidities behaved similarly in patients with sepsis. Hypertension, anemia, and fluids and electrolyte disorders were commonly found comorbidity clusters across all gender, race, and age subpopulations. The model of patient trajectories revealed that some definitions were able to identify sepsis states quicker than others, which is critical given the rapid deterioration. When examining the impact of interventions, the differences in definitions also translated to very different outcomes in patients, despite the same guideline being used. This suggests that the definition considered has a real impact on outcomes, signaling the need for a more precise consensus definition so as to provide improve outcomes for patients.

Perspective 2: Diabetes as an Index Disease

Diabetes is associated with several comorbidities which can worsen outcomes such as length of stay and cost amongst others in patients (Maddigan et al. 2005; Kerr et al. 2007). Examples include cardiovascular disease, nephropathy, neuropathy, retinopathy, and certain common cancers such as breast, pancreatic, and colorectal (Mazze et al. 2011). Here, we focus on a specific comorbidity of diabetes, breast cancer, and consider the complexity associated with decision-making when medication for the index disease influences the comorbidity. Recent

research suggests that diabetes itself, as well as medications such as metformin and insulin can affect the incidence of breast cancer. While diabetes (Tsilidis et al. 2015) and insulin glargine (Karlstad et al. 2013) have been found to increase an individual's risk of breast cancer, metformin is believed to reduce this risk (Franciosi et al. 2013). However, the exact nature of this relationship is yet to be ascertained and substantial uncertainty remains (Badrick and Renehan 2014).

Objective In light of this uncertainty, we explore the impact of different assumptions regarding these relationships on breast cancer prognosis and treatment recommendations in women with diabetes. These uncertainties also reflect a need to re-evaluate treatment guidelines for women with diagnosed diabetes. While early insulin initiation can aid in more controlled A1c (blood glucose) levels and diabetes, we hypothesize that there may be a trade-off in terms of cancer and hypoglycemia risk. The goals of this study are to build a detailed, DES model of the two diseases and associated vascular complications to (1) examine the impact of the uncertain association between glycemic-control medication and breast cancer incidence on prognosis in women with type 2 diabetes and (2) provide recommendations for initiation of insulin therapy under different assumptions, using combined data from the National Health and Nutrition Examination Survey (NHANES) and the Breast Cancer Surveillance Consortium (BCSC).

Methods The framework consisted of four sub-models to study disease progression, medication policies, and death in an individual (Fig. 11.2). These sub-models describe (1) breast cancer progression, adapted from the Tejada et al. models (Tejada

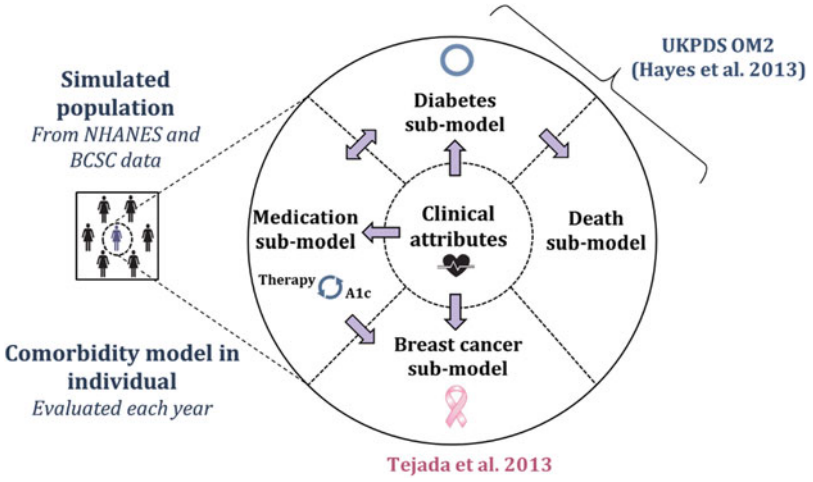


Fig. 11.2 Combined breast cancer—diabetes DES framework overview: A comorbidity model consisting of breast cancer, diabetes, and death sub-models as well as a medication regimen is evaluated on an annual basis in individuals whose clinical attributes are simulated. One-way and two-way interactions between the sub-models are represented by arrows

et al. 2013, 2014); (2) diabetes progression, based on the UKPDS outcomes model 2 (UKPDS OM2) (Hayes et al. 2013); (3) the diabetes medication regimen based on the Standards of Diabetes Care (American Diabetes Association 2017); and (4) all-cause death determined by the equations in UKPDS OM2.

Women with diagnosed diabetes are modeled as entities who flow through the disease progression and medication sub-models on an annual basis. Each woman is treated for diabetes based on her A1c levels, which are in turn impacted by glycemic-control agents in the medication sub-model. Over the 20-year simulation horizon, a woman may experience one or more of the following annually: diabetes-related micro- and macro-vascular complications and hypoglycemia, development of breast cancer, a change in medication and A1c levels, or death. Each woman's A1c level is evaluated on an annual basis, and a decision is made about the diabetes medications she should be prescribed. Within the medication sub-model, metformin and unspecified dual therapy are the first and second-line therapies in women with A1c levels less than 10%. With increasing A1c levels over time, insulin is initiated depending on the scenario under consideration—at 9% as dual therapy or between 10 and 12% as triple therapy. Using 16 scenarios, the DES framework allows us to study the impact of insulin initiation at A1c levels varied between 9% and 12% on outcomes such as cancer prevalence in the population, death, and diabetes and hypoglycemic events, under four different assumptions. The first assumption serves as the control group of scenarios in which none of the factors, diabetes, metformin, or insulin influences the risk of breast cancer incidence. The second, third, and fourth assumptions model scenarios where diabetes alone increases breast cancer risk, metformin alone decreases breast cancer risk, and metformin and insulin both independently affect breast cancer risk, respectively.

Key findings Our model showed that the risk for hypoglycemia was a significant driver of the timing of insulin initiation. Metformin improved cancer outcomes as well as lowered the risk of hypoglycemia in women. Results showed that the protective effect of metformin is strong enough to negate the potential negative impact of insulin on cancer incidence. The absence of the protective anti-cancer effect of metformin in scenarios assuming that diabetes alone impacts cancer saw an increased cancer prevalence when compared to those assuming that insulin and metformin both impact cancer. This underscores two important considerations: (1) it is important to control for treatment when studying breast cancer risk and diabetes to fully understand these relationships and (2) in the long-term, there are additional benefits to administering metformin concomitantly in women on insulin.

Implications and Next Steps

This case study provides simulation framework-based approaches to address one of the most frequent manifestations of complexity—multimorbidity, and its role in patient treatment and outcomes. Despite data limitations and modeling challenges associated with multiple diseases, our findings highlight the need to move from a single-disease to a multi-disease focus. The framework for breast cancer and

diabetes can be extended to study the relationship between diabetes and other comorbid cancers, such as prostate cancer. The framework for the evolution of sepsis can serve as a testbed for future definitions of sepsis under a variety of treatment guidelines. With the availability of additional data, both frameworks can be used to inform personalized decision-making in patients. There is a strong consensus for the need to move from a single-disease to a multi-disease focus (Piette and Kerr 2006; Geraci et al. 2005; Boyd and Fortin 2010). Systematic approaches to studying the interactions between multiple diseases can improve clinical decision-making. From a systems perspective, a thorough understanding is required of what data and methods are needed to build better models that account for patient heterogeneity and MCC.

11.2.3 Optimal Patient-Centered Response to Acute Physiological Deterioration of Hospitalized Patients: Dr. Muge Capan

Introduction to Early Warning Scores in Critical Care

Patient safety concerns due to delayed or inadequate response to *acute physiological deterioration* (APD), defined as acute instability in health condition triggered by abnormality in clinical observations (Smith et al. 2006; DeVita et al. 2006), have led to initiatives to avoid preventable harm in hospitals (Committee on Quality of Health Care in America, Institute of Medicine 2001; Wachter and Pronovost 2006). Timely detection of APD is crucial in preventing further decline in the health of hospitalized patients. Delay or omission of response to APD have been reported as one of the leading causes of preventable in-hospital death in United States (McGloin et al. 1999; Smith and Wood 1998), corresponding to over 12% of all inpatient deaths (CDC/NCHS 2010) and up to $29 billion in healthcare cost per year nationwide (Kohn et al. 1999).

In the domain of critical care, timely detection of APD has been associated with early signals of vital sign abnormalities, e.g., extremely high or low values of heart rate, that occur hours prior to harm events, e.g., in-hospital cardiac arrest (Ludikhuize et al. 2012). Sustainable *resuscitation*, i.e., stabilization and recovery, relies on continuous surveillance to identify high-risk patients and facilitate appropriate interventions. Early Warning Scores (EWSs) have been widely utilized by healthcare systems to risk-stratify patients and provide ongoing system surveillance (Bynd et al. 2004; Gray et al. 2002). An EWS is an algorithm that quantifies a patient's risk of deterioration, or harm events, based on observed abnormalities in physiological measures. The EWSs dynamically calculate a score as the weighted sum of routinely collected physiological measures (e.g., blood pressure, respiratory rate). Higher scores indicate worse health conditions. A number of EWSs were developed and implemented during the last decades such as the Modified Early Warning Score (MEWS) (Subbe et al. 2001), the VitalPAC Early Warning Score

(VIEWS) (Prytherch et al. 2010), and the National Early Warning Score (NEWS) (Royal College of Physicians 2012).

Decision Problem

In addition to risk stratification of patients, EWSs are commonly utilized in decisions regarding critical care assessment outside of an intensive care unit (ICU) provided by a Rapid Response Team (RRT). RRTs were first introduced in 1989 as a patient safety initiative in order to rapidly provide critical care to seriously ill patients (Bristow et al. 2000; Hillman 2008). RRTs are limited resources typically composed of critical care nurses, physicians, and respiratory therapists that are tasked with the assessment and treatment of patients that exhibit signs of APD based on the premise that early intervention may promote improved patient safety (Chan et al. 2010; Buist et al. 2002). While signs of APD have been studied individually and as part of EWSs, evaluating clinical interventions, e.g., RRT activation, based on EWS thresholds in an individualized manner remains an understudied area. To date, there is no single EWS system implemented across healthcare systems. Within the same healthcare system, EWSs may be implemented in a "one-size-fits-all" fashion without considering the patient or provider characteristics, e.g., utilizing RRT activation policies with the same EWS thresholds for all patient types.

The objectives of this study are: (1) to better understand the dynamics of the APD and recovery processes during hospitalization represented by a stochastic decision model, (2) to determine optimal EWS-based RRT policies that minimize the time spent in distressed health conditions by taking patient and provider heterogeneity into consideration. In the remainder of this section of the chapter, we explore the relationship between patient characteristics and APD utilizing large-scale Electronic Health Records (EHR) data. We discuss a novel methodology for developing Markovian models based on EWSs with the goal to personalize RRT activation decisions. We conclude with discussion of key findings and implications for healthcare delivery systems.

Methods

Study Population This is a retrospective study using patient-level data including demographics, vital signs, location of care, clinical events, and health outcomes extracted from the EHRs provided by the study hospital. The study cohort includes 55,385 adult patients hospitalized at the study hospital from January 2011 to December 2012. The inclusion criteria are age at admission ($\geq$18 years) and care location. Statistically significantly different patient subpopulations are identified to determine individualized RRT activation decisions, and rely on two patient characteristics: risk of deterioration (ROD) during hospitalization with three levels (i.e., low, moderate, or high ROD) defined using the Braden skin score which is

a risk assessment tool used at admission and admission type with two levels (i.e., medical or surgical patients) (Capan et al. 2015a).

Baseline Decision Model The objective of the baseline decision model (Capan et al. 2015b) is to find optimal subpopulation-specific RRT policies that minimize the total expected time associated with APD, recovery, and stabilization processes during hospitalization. We develop a semi-Markov Decision Process (SMDP) model that allows the patient health condition, as measured by the NEWS value and care location, to evolve as a stochastic process while the time spent in each health state follows a probability distribution. States are ordered such that lower NEWS is associated with better health. The absorbing state represents the patient leaving the general ward, including transfer to a higher level care, discharge, and death. The continuous-time infinite-horizon model dynamics are presented in Fig. 11.3 using a hypothetical patient trajectory.

In Fig. 11.3, each circle represents a newly observed NEWS value derived from the EHR data entry of the NEWS components. Decision epochs, represented by

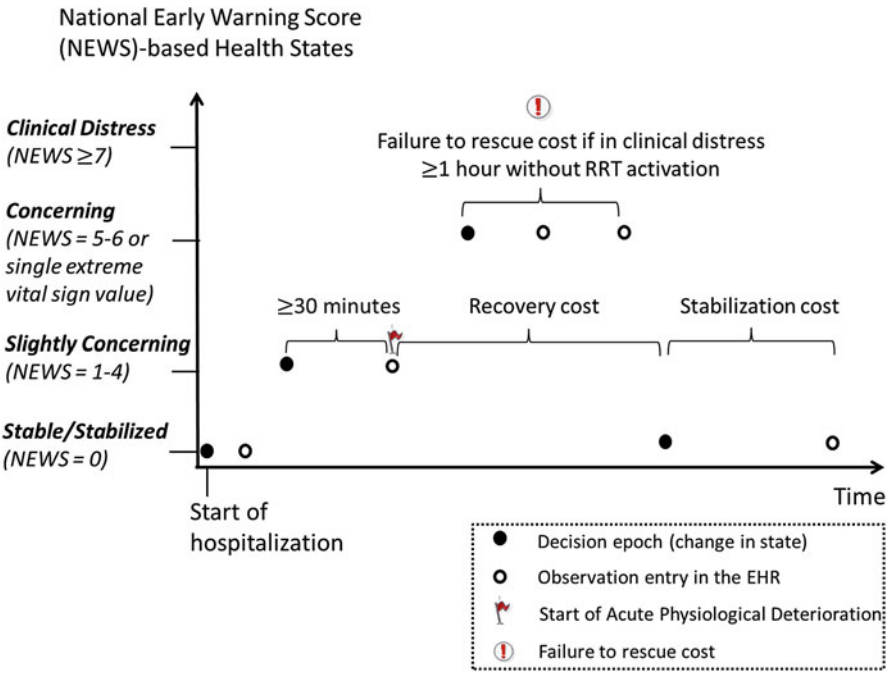


Fig. 11.3 The continuous-time infinite-horizon model dynamics using a hypothetical patient trajectory. Patient health evolves stochastically over time beginning with the start of hospitalization. EHR stands for Electronic Health Records. RRT stands for Rapid Response Team. Cost is measured in minutes and accumulated depending on time spent in recovery-, stabilization-, and clinical distress-related resuscitation processes. Actions can be taken in decision epochs and have the potential to impact patient's health trajectory

full circles, refer to time points of the actions which coincide with the health state changes. As Fig. 11.3 shows, recovery-related cost (i.e., time to return to the stable health state represented by the lowest possible NEWS value 0), stabilization-related cost (i.e., time remaining in stabilized health states for a clinically appropriate time), and failure to rescue cost (i.e., time the patient spends in clinical distress state represented by NEWS ≥ 7 without an RRT activation for at least an hour or longer) are considered as the costs in the model since time is critical in the response to APD. Actions include waiting or activating the RRT immediately. In this context, waiting means that the decision-maker provides necessary care for the patient without using RRT resources at that time. For each subpopulation, the optimal RRT activation policies are computed using the Policy Iteration Algorithm (Puterman 1994).

Modified Decision Model The modified decision model (Capan et al. 2015b) aims to capture the provider heterogeneity by grouping bedside providers with similar characteristics based on providers' perception of costs associated with health states and RRT resource needs, i.e., how much time the provider thinks it will take to care for a patient in this condition. The reason for including provider heterogeneity is that the bedside provider team may consist of providers with different perception of patient's health condition and resource needs—which may result in variations in response to APD. We use hierarchical clustering to classify the providers into unique clusters according to their cost perception measures and solve cluster-specific linear programming formulation of the modified SMDP model to obtain the optimal RRT activation policies.

Key Findings

Results of the baseline model showed that the total recovery- and stabilization-related cost significantly increased with worsening health condition for each subpopulation. Two optimal RRT activation thresholds for the considered subpopulations were identified. Specifically, surgical patients with a moderate or high ROD and medical patients with any level of ROD may benefit from an RRT activation when their NEWS is between 1 and 4, whereas the threshold is 5–6 or a single extreme vital sign value for surgical patients with a low ROD.

The modified SMDP model allowed the providers to assign weights to the baseline model cost function and to a conservative cost function representing the maximum stabilization-related cost to incorporate evaluation of the workload associated with patient needs. The total expected costs remained nonincreasing in health state, as was the case in the baseline model results. However, the optimal RRT policy structure changed compared to the policies in the baseline SMDP model, especially in the distress state if the provider selected a high value for a high RRT time perception value. Results showed that providers with conservative cost perceptions in clinically critical states preferred waiting over activating an RRT immediately.

Implications and Next Steps

Patient physiology can change dynamically over the course of a hospitalization. APD, characterized by an acute disturbance in physiological measures, can result in harm events if there are delays in recognition and response. EWSs provide a tool for quantifying the extent of APD. Existing EWSs rely on fixed thresholds to trigger an RRT activation, without considering patient- and provider-level characteristics. We developed a novel stochastic decision model to represent the uncertainty in a patient's health progression. In addition, heterogeneity in provider teams may impact the RRT activation policies which we incorporated into a modified SMDP model formulation. The results of baseline and modified models provided optimal subpopulation-specific RRT activation policies.

By integrating statistical analysis, decision analytical modeling, and mathematical programming, this study has the potential to transform the response to APD in non-ICU settings. The novelty of the methods lies in the computation of clinical stabilization-related cost as a function of the patient subpopulation and providers' assessment of patient needs. The identified optimal RRT policies utilize physiological measures that are readily available during routine hospital rounding. The findings of this study suggest that EWS-based resuscitation systems should encourage and enable frontline providers to proactively monitor and quantify physiologic decline, and communicate the concerns about patients in a systematic way. Based on this research, clear and enforceable guidelines can be developed to implement individualized patient rescue systems. The findings of this study were successfully implemented to guide a tiered individualized rapid response system for patients who present with or develop APD during a hospitalization (Capan et al. 2017).

Design and implementation of EWS-based and personalized clinical decision support systems represent a promising future research area. However, every healthcare system is unique, and EWSs need to be validated and fine-tuned for the appropriate use in the given healthcare setting. Health system-specific analysis has the potential to transform the implementation of EWS-based decision support systems while satisfying the system performance metrics by optimizing resuscitation interventions resulting in sustainable improvement in patient safety.

Acknowledgements This research was performed in collaboration with Dr. Jeanne Huddleston, Mayo Clinic, the Mayo Clinic Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery and North Carolina State University, Raleigh, NC.

11.2.4 Using Operations Research Methods to Improve the Medication Supply Chain and Its Connection to Patient Care: Dr. Ana Vila-Parrish

Introduction

Pharmaceutical expenses have continued to increase alongside the cost of health care in the United States (ISMP 2010). While health systems focus on price variation and product formulary decisions, there are understudied areas that can contribute to significant cost savings. For example, operational inefficiencies plague many facets of healthcare delivery including cost, quality, and outcomes. From a cost perspective, suboptimal inventory and logistics strategies result in a low number of inventory turns each year and waste from expired and obsolescent medications. Shortages cause resources to be focused on tracking supplies or making shipment expediting decisions at an increased cost. In addition to improvements through reduction in supply chain costs, clinical resources, and patients are impacted negatively when operations are misaligned with clinical care requirements. Clinicians spend non-value added time looking for products instead of on patient care. Insufficient supply can cause delays in patient care or product substitutions.

These realities inspired our work which aimed to develop operations research models, focused on the pharmacy supply chain, which reflected the inherent trade-offs between cost and the quality of patient care.

The objective of our work was to demonstrate the value of applying operations research methods to represent the clinical and operational impacts of various pharmacy supply chain strategies. Figure 11.4 summarizes the interaction between these elements in our research.

This section will summarize four components of our work.

1. Use simulation modeling to understand the trade-offs of various medication inventory strategies and scenarios
2. Develop analytical models that would predict medication inventory requirements based on patient data using Markov Decision Processes
3. Evaluate medication inventory decisions during surges in demand
4. Develop analytical models to understand how important factors such as criticality, demand variability, and costs of different supply chain structures could play a role in medication location decisions

In all cases, the research is focused on developing supply chain models that represent the relationship between patient mix and the pharmacy requirements to provide care at a sustainable cost.

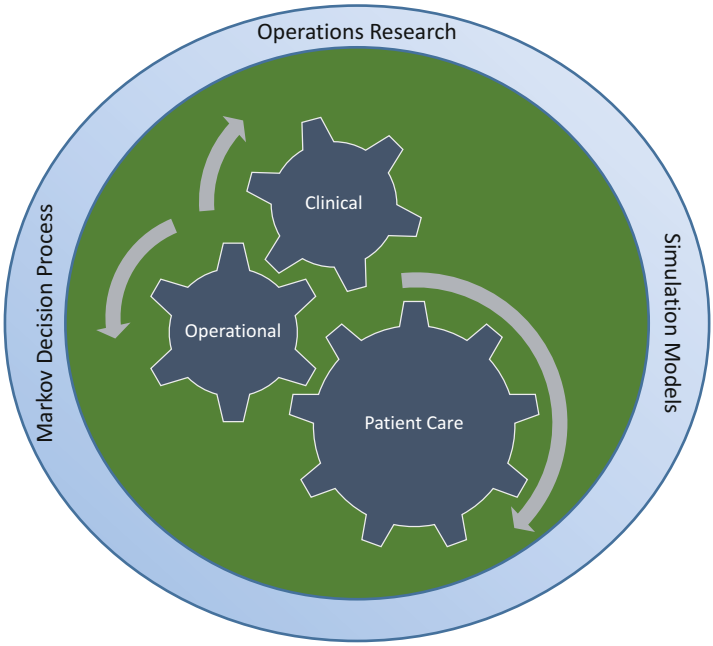


Fig. 11.4 Methods for modeling clinical, operational, and patient care attributes of health supply chain

Research Implications for the Patient-Driven Medication Supply Chain

Use simulation modeling to understand the tradeoffs of various medication inventory strategies and scenarios (Vila-Parrish et al. 2008)

In this work, we developed a simulation model to understand the performance of a medication inventory system under various costs (holding, expiration, stockout). Two inventory policies were tested. A fixed policy that holds the base stock level static and independent of actual patient mix. This policy is common in hospital pharmacies since it is easy to implement because it is based on historical usage. The second policy is an adaptive policy that uses knowledge of patient mix to manage inventory. The optimal inventory levels in the second policy were found using a MDP where the states reflected the type of patient in the system. Each patient type (1–3) has a different demand distribution for a specific medication. The results showed that the adaptive policy outperformed the fixed policy in the experimental study.

Develop analytical models that would predict medication inventory requirements based on patient data using Markov Decision Processes (Vila-Parrish et al. 2012)

This research extended the simulation-based work to create a MDP model with two forms of inventory: raw material and finished (e.g., intravenous medications that must be processed). A further complication is the perishability of the finished good

stage state. We use a stochastic "demand state" as a surrogate for patient condition and develop a MDP to determine optimal, state-dependent two-stage inventory and production policies. A Markov chain describes the changes in patient condition and therefore demand. Using real-world data, we presented a case study that applied our modeling framework to ordering and production strategies for the drug Meropenem. The results showed that with a greater understanding of demand, advanced preparation of medications can improve system performance—lowering stockouts and costs while improving pharmacy resource efficiency. *Agency for Healthcare Research and Quality (AHRQ) under the grant number 1R36HS017756-01.*

Evaluate medication inventory decisions during surges in demand (Vila-Parrish et al. 2015)

The outbreak of an infectious disease may put significant strain on hospitals, especially when these diseases cause a sharp increase in patient admissions. This research focused on the development of two simulation models: (1) a disease outbreak model and (2) a medication inventory model. The goal was to identify the inventory strategy for medications under these surge conditions. These two models are used to identify inventory policies for managing medication during disease outbreaks. As a case study, we use historical influenza data, including the disease spread, as the demand stream into the inventory simulation model. Due to the computational complexities of the model, we formulated a dynamic program for a reduced version of this model. We compare three different simulation-based policies for managing inventory. These ranged from newsvendor-type policies that are static to a policy based on patient mix and driven by a Markov process. Given the uncertainty in any disease model, we performed sensitivity analysis on several parameters such as the gross attack rate and inventory holding cost parameters. Under these various conditions, the policy that was derived through the use of OptQuest in our simulation package Arena led to the best and easiest to implement results.

Develop analytical models to understand how important factors such as criticality, demand variability, and costs of different supply chain structures could play a role in medication location decisions (Smith et al. 2016)

Medication inventory location strategies are not frequently revisited based on demand data. Medications may be stored in various locations such as in automated dispensing cabinets at the hospital unit level, in a central pharmacy within a hospital or an offsite location. Often a medication is stored in a combination of centralized and decentralized locations. For medications with higher demand variability or infrequent use this strategy results in increased inventory at each echelon. Unnecessarily high inventory levels can result in increased waste costs and high holding costs across the health system. In this research, we develop an optimization model that determines inventory levels at each point of storage based on attributes such as demand variability, product cost, criticality to the patient, and cost of the stockout. We assume that inventory can be shifted from any point of storage to another but that there is a cost to do so—at a minimum, a resource cost but in the case where the product must be shipped from an offsite location a greater penalty is assumed. The results show that there is a relationship between stocking strategy and cost

but that there is also a relationship to patient care considerations due to stockout-related delays of care. The simulation models suggest that developing product classifications that consider various attributes can aid inventory policy development that meets both cost and service level requirements.

Key Considerations for the Future

As the move towards value-based performance and reimbursement continues, the role of the pharmacist in providing clinical care will continue to evolve. The cost of the supply chain is a significant percentage of the total cost of patient care. Initiatives focused on risk stratification will require supply chain organizations to align themselves with supporting the operational requirements. The opportunity for supply chain leaders to use data and analytics to efficiently do so will be paramount. Delivering the "right medication, at the right time, and right place" in a way that improves outcomes is no longer the sole focus of clinicians. Services such as meds-to-beds and home delivery services will require that these handoffs between clinical and operational workflows to become further intertwined. To do so at scale and at a sustainable cost is the challenge for today's health system organizations and supply chain and analytics are enablers in achieving these requirements.

11.3 To Be Wealthy

11.3.1 To Be Equitable and Effective: Using Mathematical Modeling to Reduce Food Insecurity—Dr. Irem Sengul Orgut

Introduction

The amount of food produced in the world is more than enough to feed the world's population and yet, millions are hungry every day (Food and Agriculture Organization of the United Nations (FAO) 2018). According to statistics by FAO (2018), food insecurity has been on the rise and 815 million people worldwide were undernourished in 2016. In the United States, 42 million people, including 13 million children, suffered from food insecurity in 2016 (Feeding America 2018). *Food insecurity* is defined as "the limited or uncertain availability of nutritionally adequate and safe foods or limited or uncertain ability to acquire acceptable foods in socially acceptable ways" by the United States Department of Agriculture (USDA 2018). The term "hunger" has been replaced by the term "food insecurity" as hunger is considered to be an individual-level physiological condition that may result from food insecurity and is highly subjective (USDA 2018). In this research, we address the problem of the equitable and effective distribution of food donations by food banks so that the maximum number of people benefit from these donations, and minimum amount of food goes to waste.

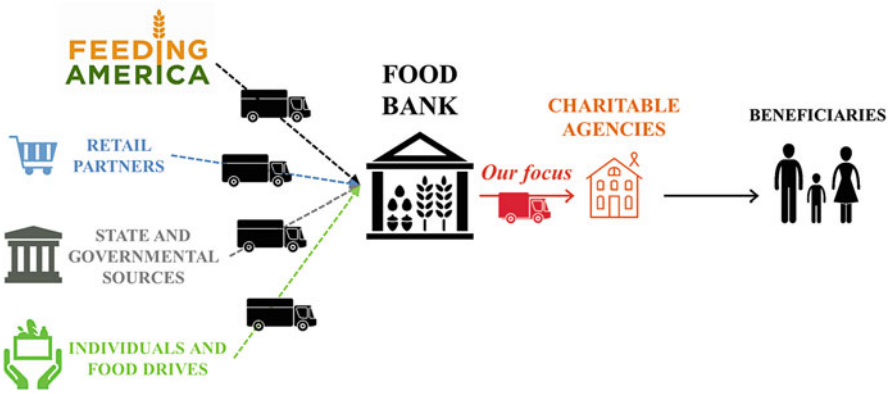


Fig. 11.5 Nonprofit food distribution chain

Feeding America is the United States' largest nonprofit hunger-relief organization, providing food to more than 46 million people in need each year through a nationwide network of 200 food banks and 60,000 food pantries and meal programs (Feeding America 2018). Feeding America collects surplus but nourishing food from farmers, manufacturers, and retailers and distributes the food to its member food banks. The food banks are responsible for distributing the food that they receive from Feeding America and other sources such as local grocers, government and individuals, to partner agencies in their allocated service region. Agencies are usually small nonprofits such as churches, food pantries, and soup kitchens that operate with very limited budgets and are, in turn, responsible for distributing the food to food-insecure people in their local area. In this study, we focus on the distribution of food donations received by the food bank to the charitable agencies at the tactical level. This supply chain is illustrated in Fig. 11.5.

For this study, we collaborate with the Food Bank of Central and Eastern North Carolina (FBCENC), a Feeding America affiliate located in Raleigh, NC. FBCENC operates from six main branch locations where food donations are collected before they are distributed to charitable agencies in a 34-county service region. During the fiscal year 2016–2017, FBCENC distributed 70 million pounds of food to 629,410 beneficiaries through 845 partner agencies (FBCENC 2018). Through our close partnership, we aim to provide FBCENC representatives optimal or near-optimal and implementable policies to achieve the equitable and effective distribution of food donations considering capacity limitations of the charitable agencies. We discuss these objectives and constraints in detail in the next section.

Decisions and Uncertainty

There are two significant differences between nonprofit supply chains, such as the one we consider in this work, and for-profit supply chains, which are commonly

studied in Industrial Engineering literature. The first difference is related to the objectives of the decision-makers. In a typical for-profit supply chain, monetary objectives such as maximizing profit, or time-based objectives such as minimizing lead time are common. However, in a nonprofit supply chain, such as FBCENCs, the primary objectives are reaching the maximum number of people, distributing food in an equitable manner and minimizing waste; making monetary and time-based objectives secondary. Under its "Fairshare" program (FBCENC 2018), FBCENC aims to distribute food in proportion to the food-insecure population so that, ideally, each beneficiary receives the same amount of donated food, by weight, over a specified period, which they report back to Feeding America. However, historical data suggests that food donations received per beneficiary in one county may be much less than those in another county; also, food waste may be observed in some counties, while others face stockouts. In this study, we aim to address these objectives through generating policies for the equitable and effective distribution of food donations such that ideally, each person in FBCENC's service region receives the same amount of food (*perfect equity*) and the amount of food waste is minimized (i.e., *effectiveness is maximized*).

The second difference between a for-profit and nonprofit supply chain originates from the respective sources of uncertainty. In a for-profit supply chain, demand is usually the main source of uncertainty. However, in a nonprofit supply chain, demand is a function of the county-level population of food-insecure individuals as measured by the United States Census Bureau (2016), which is relatively stable over time. Feeding America and FBCENC also use these demand estimates to measure the level of equity. Considering FBCENC's supply chain, the main source of uncertainty arises from the receiving capacities of the partner agencies. There are several factors that affect an agency's capacity, i.e., the amount of food an agency can receive in each time period. Some of these factors are the agency's budget (to pay shared maintenance fees to food banks), transportation capabilities (trucks and drivers to get food from the food bank), workforce (for administrative and IT purposes) and storage (for storing food, both at ambient temperature and refrigerated). Supply is also an uncertain factor as it depends on the amount of ad hoc donations that the food bank receives. However, in our work, we focus on the decision made by the food bank after food donations have been received, hence consider the capacity uncertainty under known supply. We also aggregate the agencies based on the county they are located in and assume counties to be the smallest distribution locations. Because of this, if new agencies are recruited to the network or current agencies are suspended due to the failure to satisfy reporting requirements, the total capacity of the county also changes.

The objectives of equity and effectiveness are in direct conflict with each other. This trade-off can be explained in a simple example. If the decision-maker's sole objective is to satisfy an equitable distribution, an optimal but trivial solution would be to distribute no food, which results in an equitable solution since each beneficiary receives the same amount of food, yet not an effective solution as all the food supply is sent to waste. Alternatively, if the sole objective is to minimize waste, the decision-maker can send food to counties arbitrarily at maximum capacity, which

would be an effective solution maximizing food shipment; however, this distribution would not consider the "need" at each county, which is measured by the food-insecure population, and hence would not be an equitable solution.

Methods and Key Findings

We first develop a deterministic linear programming model (Sengul Orgut et al. 2015) to minimize the amount of undistributed food while maintaining a user-specified upper bound on the deviation from perfect equity and derive closed-form optimal solutions. This model assumes that county capacities are deterministic and known and aims to understand the structural properties of this supply chain and obtain easily implementable policies. We prove that locations with low capacity-to-demand ratios, *bottlenecks*, constrain the total amount of food distributed in the service region owing to the equity requirement. Therefore, counties' capacities, which in practice are uncertain, strongly influence the optimal solution. We also extend this model to obtain optimal policies for the allocation of additional receiving capacity to counties in the service area. In practice, some examples of this additional capacity may be additional transportation capabilities or mobile pantries. Hence, these deterministic models can be used both for benchmarking the performance of FBCENC by exploring the trade-off between equity and the total distributed food, and for obtaining managerial insights into how capacity investments can be made in collaboration with local agencies to improve FBCENC's ability to meet its goals. We use historical data obtained from FBCENC to illustrate our results and perform a probabilistic sensitivity analysis. The sensitivity analysis also shows that the results are highly dependent on county capacities, which are considerably uncertain.

To address stochastic capacities, we develop a two-stage stochastic model to address the equitable and effective food distribution problem in a single period under capacity uncertainty (Sengul Orgut et al. 2017). In the first stage, the decision-maker is required to make shipments to the counties without knowing the realized capacities of the counties. The only restrictions imposed on the first stage shipments are that (1) they should be perfectly equitable; and (2) the total food shipment cannot exceed the supply. In the second stage, county capacities are observed, and recourse actions are taken in the form of additional shipments from the food bank or sending surplus food to waste at the counties, incurring additional cost on the objective function. We prove that this model has a newsvendor-type closed-form optimal solution. We present numerical results using data from FBCENC and perform a sensitivity analysis on the cost coefficients.

We extend the single period model to multiple periods where supply is received at the beginning of each period and shipments are made prior to observing the capacities for that time period (Sengul Orgut 2015). After capacities are observed at each period, recourse actions can be taken by either shipping extra food from the food bank to the counties or sending surplus food to waste at the counties. Any unshipped supply in the food bank is transferred to the following period as starting inventory. Since it is not possible to obtain a closed-form solution for

the multi-period model, we develop heuristics and test their performance by using a Multi-Stage Stochastic Programming Model as a benchmark through extensive computational experiments. The results show the promising performance of the heuristics.

Lastly, we develop a robust optimization model that maximizes the total food distribution while allowing the capacity parameters to vary within pre-defined ranges (Sengul Orgut et al. 2018). This model considers the same problem as the stochastic single-period model but assumes a different approach where instead of the probability distributions of the county capacities, it is sufficient to have a range for each county's capacity. The trade-off between these two models is that the robust model is easier to implement but may not be able to capture the exact details of probability distributions of county capacities whereas the stochastic single-period model requires knowledge of the probability distributions but is more difficult to implement by food banks. The model also includes a robustness control parameter (Bertsimas and Sim 2004) to control the trade-off between the total amount of food distributed and the robustness of the solution towards these capacity deviations. If such a parameter is not introduced, we obtain overly-conservative solutions that are protected against the lowest possible capacity values at each county. We derive structural properties of this model and develop an efficient algorithm that solves this model optimally and show that the optimal solution to this model preserves the bottleneck structure of the deterministic problem studied in Sengul Orgut et al. (2015). We use data from FBCENC to illustrate the trade-off between effectiveness and robustness for varying levels of the robustness control parameter.

Implications and Next Steps

The models we develop in this study focus on a single food type and do not consider any combinations of different food types. FBCENC categorizes food in four main groups: dry goods, produce, frozen food, and refrigerated food. As a part of our future work, we would like to explore the correlations and interactions between different food groups. Feeding America's 2025 goal is to not only provide food to the food-insecure population but "to provide access to enough nutritious food for people struggling with hunger" (Feeding America 2018). Hence, as an extension of our work, we will build models that consider the nutritional requirements of the food-insecure population as well as the correlations between the capacities of multiple food groups for satisfying equitable and effective food distribution. Another possible extension of our work is to consider stochastic supply for the single- and multi-period food shipment problems.

Our study addresses a problem faced by food banks constantly: how to distribute food donations as equitably as possible while minimizing food waste when the capacities of the charitable agencies are uncertain and change over time. We develop closed-form optimal solutions, algorithms, or heuristics that can easily be implemented to provide policies for food distribution. We show, through numerical studies based on the data from our partner food bank, that the food bank operations

can be significantly improved by using our developed policies. Food banks in the United States and other developed countries face similar challenges and can benefit from our results. For example, per our discussions with another food bank in southern California, we learned that they face similar problems. Their capacity challenges are mostly related to the lack of drivers since they are responsible for distributing food to the agencies in their service region. Further, they are required to pay a tax per pound of food wasted, making the minimizing waste objective even more significant.

Our models can also be used by larger organizations such as Feeding America, that face similar challenges on a larger scale since they also collect food donations and distribute these donations to their food bank affiliates nationwide. Additionally, any system which aims to achieve the equitable and effective distribution of scarce resources under stochastic capacity can utilize our results. Some examples of such problems are the allocation of ambulance dispatch centers within counties and the distribution of donated blood to blood banks.

11.4 To Be Wise

11.4.1 Student Performance in Mathematics over Time: A Social Application of a Markovian Model to an Education System—Dr. Amy Craig Reamer

Introduction

Education has a significant impact on an individual's health and prosperity which, in turn, affects the nation's population health and economy (Zimmerman et al. 2015). The merits of this statement are well justified in the literature and the relationship between education, health and wealth is the overarching theme of this chapter. The US K-12 education system is an environment ripe with opportunities for exploring complex problems with uncertain conditions. Of particular concern is transforming the delivery of science, technology, engineering, and mathematics (STEM) education to prepare students to succeed in the ever-changing, technologically advanced, global workforce (National Science Board 2007; National Academy of Science et al. 2007; National Center for Education Statistics 2006; United States Department of Education 2008). In its 2007 report calling for STEM education reform, the National Science Board emphasized a need for the "vertical alignment" of STEM learning across the K-16 spectrum, prognosticating that a solid foundation in the STEM disciplines in early education is essential to a student's future mastery of advanced STEM material (National Science Board 2007). Although mathematics is but one discipline of the ubiquitous "STEM" acronym, it is perhaps the most visible since its concepts and methods cut across all science, technology, and engineering disciplines (National Research Council 2001). Numerous reports have documented

the decline in mathematics proficiency of US students, jeopardizing the nation's ability to prepare qualified STEM professionals for a workforce that so desperately needs them (Nation's Report Card 2015, National Academy of Science et al. 2007). The 2015 US national report card in mathematics reveals that only 33% of eighth graders are proficient in the subject (Nation's Report Card 2015). In this section, we use a Markovian model to track the evolution of students' performance in mathematics in the elementary and middle grades. This research, summarized from Reamer 2012 and Reamer et al. 2015, marks a first step in developing a Markovian structure for modeling an education system and offers an application of Operations Research in a non-traditional, social context.

The K-12 public education system is a complex, uncertain environment. Education administrators and policy makers rely in large part on standardized assessments to measure a student's cognition at a point in time along their dynamic learning trajectory. In this research, we use data from the North Carolina Department of Public Instruction (NCDPI), particularly student performance on standardized assessments known as End-of-Grade (EOG) tests (NC DPI 2018a). Since 1995, the EOG tests have been administered annually in mathematics and reading comprehension to students in third grade through the eighth grade (NC DPI 2018b). At the time of this research, a student's raw score on the EOG test was categorized into one of four achievement levels corresponding to a level of proficiency as defined by the state of North Carolina. The EOG test is mapped to the curriculum standards adopted by NCDPI which have also changed since the original publication of this study. A description of the achievement levels circa 2012 is shown in Table 11.1.

Table 11.1 Description of North Carolina End-of-Grade (EOG) test achievement levels (NC DPI 2018b; Reamer 2012)

Achievement level	Descriptor	NC proficiency status
I	Students performing at this level do not have sufficient mastery of knowledge and skills in this subject area to be successful at the next grade level	Non-proficient in mathematics
II	Students performing at this level demonstrate inconsistent mastery of knowledge and skills in this subject area and are minimally prepared to be successful at the next grade level	Non-proficient in mathematics
III	Students performing at this level consistently demonstrate mastery of grade-level subject matter and skills and are well prepared for the next grade level	Proficient in mathematics
IV	Students performing at this level consistently perform in a superior manner clearly beyond that required to be proficient at grade-level work	Proficient in mathematics

The third to eighth grade offers an important window of data collection in characterizing the stochastic progression of a student's performance in mathematics over time and has implications for their future achievement in STEM courses. In this research, we develop and use a Markovian model to forecast a student's proficiency in mathematics, as defined by NCDPI, by the eighth grade given their initial level of proficiency in the third grade. We refer the reader to Reamer 2012 and Reamer et al. 2015 for a review of relevant literature and motivation for use of the Markovian model.

Methods

Data for this research was obtained from the North Carolina Education Research Data Center (NCERDC) at Duke University (NCERDC 2018). The NCERDC maintains data on students in the North Carolina public school system, collected annually by NCDPI. We used the NCERDC's encrypted student master identification numbers to longitudinally track cohorts of students from the third to eighth grade, documenting their EOG achievement level scores from one grade level to the next. We obtained statewide test results from all 100 North Carolina counties from 1995 to 2009, the last year of data that was available for the original study. The 2004–2009 cohort, for example, consisted of nearly 75,000 student records. We made a series of assumptions in analyzing the data and those assumptions are detailed in Reamer 2012.

Demographic covariates of interest included student gender, ethnicity, and exceptionality status, which identified students as academically/intellectually gifted (AIG) and/or learning disabled. We were also interested in their participation in a free or reduced price school lunch program as a measure of socioeconomic status. We conducted a series of descriptive statistical analyses on one aggregate cohort of students to investigate potential differences in EOG mathematics test performance amongst the aforementioned student populations.

To model the data, we used a first-order Markov chain model, which is a discrete-time stochastic process that satisfies the Markov property and can be defined by a set of states and transition probabilities amongst the states (Howard 1960). The Markov property, often referred to as the "memoryless" property, assumes that a future state of the system depends only on the present state, and not on states or events that transpired in the past (Howard 1960). We proffer that in the sequential and cumulative process of a student's learning of mathematics over time the student arrives to a current state of cognition that is independent of the pathway they took to accumulate said knowledge, and this claim is validated in the learning theory literature (Smallwood 1971; Simon and Tzur 2004; Confrey and Maloney 2010). We define the state of our model as a student's EOG mathematics test achievement level score at the end of a grade. A state-transition diagram for one cohort of students studied is presented in Fig. 11.6. In the diagram, the nodes represent the states of the system and the arrows represent the probabilistic transitions in EOG test achievement level scores from one grade level to the next. From Fig. 11.6,

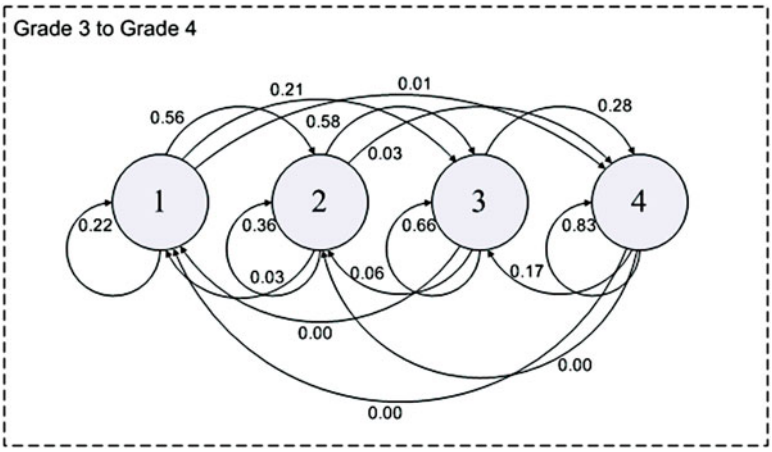


Fig. 11.6 State-transition diagram (Reamer 2012)

we see that a student that scores within achievement level I (state 1) on the EOG mathematics test in third grade has a 21% chance of scoring within achievement level III (state 3) in fourth grade. This data was organized into transition probability matrices showing the possible movements between states for all ten cohorts studied. Our ultimate objective is to calculate the probability that a student will score within achievement levels III or IV by eighth grade. Per Table 11.1, North Carolina students that score within levels III or IV are classified as "proficient" in mathematics. Refer to Reamer 2012 and Reamer et al. 2015 for a description of the structure, notation, and assumptions associated with the model. We developed aggregate Markov chain models of ten cohorts of students, that we longitudinally tracked from third to eighth grade. We used the Chapman-Kolmogorov equations to calculate the probability of proficiency by eighth grade conditional upon the initial level of achievement in third grade. Other Markov chains were developed based on the demographic factors of interest.

Key Findings

A full report of the statistical hypothesis testing and analysis of the various Markov chains can be found in Reamer et al. 2015. We present a summary of key results here.

From the statistical analysis, male and female students' achievement level scores on the EOG mathematics test were comparable across all grade levels. With regard to ethnicity, using the variable names defined by the NCERDC, Asian and White students scored higher than all other ethnic populations. Black students scored the lowest of all ethnic groups considered. The small sample sizes of other ethnic groups including those that identified as Hispanic, American Indian, and Multi-Racial

limited our ability to claim significance of further comparisons. Not surprisingly, AIG students scored higher than non-AIG students. Students that participated in their school's free or reduced price lunch program scored lower than students that did not participate in the program. As a result of the statistical findings, we developed separate Markov chain models based on the demographic covariates, presented in full in Reamer et al. 2015, in addition to the aggregate Markov chain models of the "typical" student.

The transition probably matrices for the "typical" student revealed that the probability of advancing to a higher level of achievement on the EOG test or regressing to a lower level is dependent upon the grade-level transition, i.e., third to fourth or fourth to fifth and so on. However, select results were independent of the grade transition, for example, it is very unlikely for a student that scores within achievement level I in any grade to score within achievement level IV in the next grade and vice versa. The threshold that separates proficiency (states III or IV) from non-proficiency (states I or II) was of particular interest. We found that students had a significant chance of improving from an achievement level II to III within one grade level, but were unlikely to make the transition from a II to a IV. In terms of regressing, the probability that a student would transition from a III to a II increased over time with the eighth graders being the most likely to do so. Not surprisingly, students that earn a IV in a given grade are very likely to earn a IV in the following grade, confirming their mastery of grade-level concepts.

In our multi-step Markovian model estimating future proficiency, we found that students that initially scored within the proficiency threshold will likely remain proficient by the eighth grade. We noted that the likelihood of regressing from the proficient states is the highest in the middle grades. Estimated percentage proficiency by eighth grade for male and female students is practically equally likely, for all initial levels in third grade.

Implications and Next Steps

For decades, education researchers and social scientists have longitudinally tracked student cognition and learning using regression analyses and a host of other statistical techniques. In this work, we offer a proof of concept Markovian model, introducing a quantitative method for analyzing the expansive data available in large-scale public education systems, with implications for extensions of this work. The Markov chain models describe the progression of a student's natural performance in mathematics over time. Future work will focus on the development of a decision support system where decision-makers can optimally identify the timing and placement of learning interventions in hopes of altering and advancing a student's future performance.

The fear that a student can fall behind and stay behind in mathematics in the early grades is corroborated by this research, underscoring the importance of the scaffolding required to build a solid foundation in mathematics early in formal education. Our results suggest that once a student advances to a level of proficiency,

they are likely to remain there, but the risk of regressing from proficiency is the highest in middle school. Thus, the middle grades represent a critical opportunity for educators to identify and mitigate obstacles to a student's learning process. Since the plan or track for taking high school mathematics courses is typically defined during middle school, hinging upon the decision of when to take Algebra I, the middle grades become a critical time in the differentiation of student's trajectories. While male and female students performed similarly in this study, we know the disparities amongst gender widen with age and are confounded by factors, such as confidence and interest, which could be incorporated into a future model.

Analyzing demographic factors that impact student performance in the manner presented here could lead to strategies that focus on more personalized learning, one of the National Academy of Engineering's (NAE) Grand Challenges for Engineering (NAE 2018). Software systems that pace the delivery of content based on a student's learning style and value-added models used by administrators to identify high-risk students and teacher efficiency are already commonplace in public school systems (NCDPI 2018c). The Markovian models presented in this section may supplement these efforts by introducing a structure for the application of optimization modeling techniques to enhance student learning when used in tandem with an educator's judgment. In this non-traditional application of Operations Research, we are essentially modeling how a child thinks and learns, and the humanistic interaction between educators and students is an important component, not to be overlooked. As a result, another avenue of future work presents an analysis of the explicit relationship between an educator's judgment of a student's performance on the EOG tests and their actual performance.

Mathematical knowledge and persistence have been shown to impact student success in STEM professions, particularly engineering. This work is the first step in developing a Markovian framework for optimizing the vertical alignment of mathematics education. Today's engineers can play a valuable role in solving challenging problems associated with the K-12 learning system in hopes of educating the engineers of tomorrow.

11.5 Conclusions

This collection of vignettes use the framework of decision theory to provide structure for addressing these societal challenges. This multi-step framework first identifies the decision-maker—often there may be more than one and these decision-makers may not have the same objectives. For a given decision-maker, it is next important to classify those things that the decision-maker controls (i.e., has the power to make a choice) from those things that may influence the decision-maker's outcomes but they cannot control (i.e., critical uncertainties). Third, it is important to identify the goals of the decision-maker (e.g., utility maximization, cost minimization, minimizing risk), in these types of complex problems, decision-makers often have multiple often conflicting goals and trade-offs have to made.

Modeling provides a framework for explicitly evaluating these trade-offs (e.g., equity and effectiveness, health and cost minimization). Fourth, understanding the information available to accurately represent the attributes of the decision problem is often a challenge. The modeling framework can be used to combine information from various sources and to explore the impact of these types of modeling assumptions on decisions.

Modeling for societal decision-making has a few implicit requirements: (1) Metrics for measuring a "Good" decision; (2) Methods for capturing progression—this could be disease progression, patient health progression, the progression of student performance (or learning), the evolution of inventory/supply/demand over time; and (3) Methods for defining and characterizing the "state" of the system, process, or patient.

There are several challenges associated with this: (1) Data—data is often messy, there may be too much data of one type or too little data of another, data often involve time series, longitudinal data with nonrandom missingness, sparse data, and a mixture of qualitative and quantitative measures. (2) All the information needed for decision-making is rarely available in one place and requires using data from different sources. There are consequences associated with combining data from different information sources, including the potential for inaccuracy due to differences in the attributes of the sample populations, potential for bias due to differences in data collection and study design. However, despite these limitations, bringing the information from different data sources together can provide a structure and framework necessary to represent the complex reality driving the decision problem and in doing so inform decision-making. (3) These complex societal challenges require integration of social science methods with systems modeling to better capture human behavior, individual preferences, and to inform behavior modeling.

The research presented in this chapter explores societal challenges in health, hunger relief, and education in a siloed fashion. Each societal challenge is presented independently—in isolation, however, these problems are intertwined. Hunger affects health and student performance, health affects student performance and can impact wealth, and education affects both health and wealth. The future of this work is to truly take a systems approach that explicitly captures these relationships and models the complex interactions at the individual and societal levels of health, wealth, and education.

References

American Diabetes Association (2017) Pharmacologic approaches to glycemic treatment. In: Standards of medical care in diabetes-2017. pp S64–S74

Anderson G, Horvath J (2004) The growing burden of chronic disease in America. Public Health Rep (Washington, DC: 1974) 119(3):263–270

Badrick E, Renehan AG (2014) Diabetes and cancer: 5 years into the recent controversy. Eur J Cancer 50(12):2119–2125

Bertsimas D, Sim M (2004) The price of robustness. Oper Res 52(1):35–53
Boyd CM, Fortin M (2010) Future of multimorbidity research: how should understanding of multimorbidity inform health system design? Public Health Rev 32(2):451–474
Brailsford SC (2007) Proceedings of the 2007 winter simulation conference. In: Henderson SG, Biller B, Hsieh M-H, Shortle J, Tew JD, Barton RR (eds). pp 1436–1448
Bristow PJ, Hillman KM, Chey T et al (2000) Rates of in-hospital arrests, deaths and intensive care admissions: the effect of a medical emergency team. Med J Aust 173:236–240
Buist MD, Moore GE, Bernard SA et al (2002) Effects of a medical emergency team on reduction of incidence of and mortality from unexpected cardiac arrests in hospital: preliminary study. Br Med J 324(7334):387–390
Bynd S, Sen A, Subbe C, Gemmell L (2004) Modified early warning score: one for all and A&E? Br J Anaesth 92(4):611P–612P
Capan M, Ivy JS, Rohleder T, Hickman J, Huddleston JM (2015a) Individualizing and optimizing the use of early warning scores in acute medical care for deteriorating hospitalized patients. Resuscitation 93:107–112. https://doi.org/10.1016/j.resuscitation.2014.12.032
Capan M, Ivy JS, Wilson JR, Huddleston JM (2015b) A stochastic model of acute-care decisions based on patient and provider heterogeneity. Health Care Manag Sci 20(2):187–206. https://doi.org/10.1007/s10729-015-9347-x
Capan M, Wu P, Campbell M, Mascioli S, Jackson EV (2017) Using electronic health records and nursing assessment to redesign clinical early recognition systems. Health Syst 6(2):112–121
Caughey G, Roughead E, Roughead EE (2011) Multimorbidity research challenges: where to go from here? J Comorb 1(1):8–10
CDC/NCHS (2010) National hospital discharge survey, 2000–2010. http://www.cdc.gov/nchs/data/databriefs/db118.htm
Chan PS, Jain R, Nallmothu BK et al (2010) Rapid response teams: a systematic review and meta-analysis. Arch Intern Med 170:18–26
Chang C-H, Lin J-W, Wu L-C, Lai M-S, Chuang L-M (2012) Oral insulin secretagogues, insulin, and cancer risk in type 2 diabetes mellitus. J Clin Endocrinol Metab 97(7):E1170–E1175
Committee on Quality of Health Care in America, Institute of Medicine (2001) Crossing the quality chasm: a new health system for the 21st century. National Academies Press, Washington, DC
Confrey J, Maloney A (2010) The construction, refinement, and early validation of the equipartitioning learning trajectory. In: Proceedings of the 9th international conference of the learning sciences, Chicago, vol 1, pp 968–975
Cox CE, Wysham NG (2015) Untangling health trajectories among patients with sepsis. Ann Am Thorac Soc 12(6):796–797
Czura C (2011) Merinoff symposium 2010: sepsis—speaking with one voice. Mol Med 17(1–2):2–3
DeVita MA, Bellomo R, Hillman K, Kellum J, Rotondi A, Teres D, Auerbach A, Chen W, Duncan K, Kenward G, Bell M, Buist M, Chen J, Bion J, Kirby A, Lighthall G, Ovreveit J, Braithwaite RS, Gosbee J, Milbrandt E, Peberdy M, Savitz L, Young L, Harvey M, Galhotra S (2006) Findings of the first consensus conference on medical emergency teams. Crit Care Med 34:2463–2478
Esper AM, Moss M, Lewis CA, Nisbet R, Mannino DM, Martin GS (2006) The role of infection and comorbidity: factors that influence disparities in sepsis. Crit Care Med 34(10):2576–2582
FBCENC (2018) Food Bank of Central & Eastern North Carolina. http://www.foodbankcenc.org/site/PageServer?pagename=FBCENCHome. Accessed 2018
Feeding America (2018). https://hungerandhealth.feedingamerica.org/. Accessed 2018
Food and Agriculture Organization of the United Nations (2018). http://www.fao.org/state-of-food-security-nutrition/en/
Franciosi M, Lucisano G, Lapice E, Strippoli GFM, Pellegrini F, Nicolucci A (2013) Metformin therapy and risk of cancer in patients with type 2 diabetes: systematic review. PLoS One 8(8):1–12

Geraci JM, Escalante CP, Freeman JL, Goodwin JS (2005) Comorbid disease and cancer: the need for more relevant conceptual models in health services research. J Clin Oncol 23(30):7399–7404

Gerteis J, Izrael D, Deitz D, LeRoy L, Ricciardi R, Miller T et al (2014) Multiple chronic conditions chartbook. Agency for Healthcare Research and Quality, Rockville

Giovannetti ER, Wolff JL, Xue Q-L, Weiss CO, Leff B, Boult C et al (2012) Difficulty assisting with health care tasks among caregivers of multimorbid older adults. J Gen Intern Med 27(1):37–44

Gray L, Smyth K, Palmer R, Zhu X, Callahan J (2002) Heterogeneity in older people: examining physiologic failure, age, and comorbidity. J Am Geriatr Soc 50(12):1955–1961

Hamilton BE, Martin JA, Osterman MJK, Driscoll AK, Rossen LM (2017) Births: provisional data for 2016. NVSS Vital Statistics Rapid Release 2(2):1–21. https://www.cdc.gov/nchs/data/vsrr/report002.pdf

Harper LM, Caughey AB, Odibo AO, Roehl KA, Zhao Q, Cahill AG (2012) Normal progress of induced labor. Obstet Gynecol 119(6):1113–1118. https://doi.org/10.1097/AOG.0b013e318253d7aa

Hayes AJ, Leal J, Gray AM, Holman RR, Clarke PM (2013) UKPDS outcomes model 2: a new version of a model to simulate lifetime health outcomes of patients with type 2 diabetes mellitus using data from the 30 year United Kingdom Prospective Diabetes Study: UKPDS 82. Diabetologia 56(9):1925–1933

Hicklin K (2016) Decision models for mode of delivery combining patient and clinician risk perceptions and preferences. North Carolina State University

Hicklin K, Ivy J, Cobb Payton F, Viswanathan M, Myers E (2018) Exploring the value of waiting during labor. Serv Sci 10(3):334–353. https://doi.org/10.1287/serv.2018.0205

Hillman K (2008) Rapid response systems. Indian J Crit Care 12(2):77–81

Howard R (1960) Dynamic programming and Markov processes. M.I.T. Press, Cambridge

Institute of Safe Medication Practices (2010) National survey on drug shortages reveals serious impact of patient safety [WWW document]. http://www.ismp.org/pressroom/PR20100923.pdf. Accessed 19 Feb 2012

Iskander KN, Osuchowski MF, Stearns-Kurosawa DJ, Kurosawa S, Stepien D, Valentine C et al (2013) Sepsis: multiple abnormalities, heterogeneous responses, and evolving understanding. Physiol Rev 93(3):1247–1288

Karlstad O, Starup-Linde J, Vestergaard P, Hjellvik V, Bazelier MT, Schmidt MK et al (2013) Use of insulin and insulin analogs and risk of cancer—systematic review and meta-analysis of observational studies. Curr Drug Saf 8(5):333–348

Kerr EA, Heisler M, Krein SL, Kabeto M, Langa KM, Weir D et al (2007) Beyond comorbidity counts: how do comorbidity type and severity influence diabetes patients' treatment priorities and self-management? J Gen Intern Med 22(12):1635–1640

Kohn LT, Corrigan JM, Donaldson MS (1999) To err is human: building a safer health system. National Academy Press: Institute of Medicine Report, Washington

Liu V, Escobar GJ, Greene JD, Soule J, Whippy A, Angus DC et al (2014) Hospital deaths in patients with sepsis from 2 independent cohorts. JAMA 312(1):90

Loeb DF, Binswanger IA, Candrian C, Bayliss EA (2015) Primary care physician insights into a typology of the complex patient in primary care. Ann Fam Med 13(5):451–455

Ludikhuize J, Smorenburg SM, de Rooij SE et al (2012) Identification of deteriorating patients on general wards; measurement of vital parameters and potential effectiveness of the modified early warning score. J Crit Care 27(4):424.e7–424.13

Maddigan SL, Feeny DH, Johnson JA (2005) Health-related quality of life deficits associated with diabetes and comorbidities in a Canadian National Population Health Survey. Qual Life Res 14(5):1311–1320

Mannucci PM, Nobili A, REPOSI Investigators (2014) Multimorbidity and polypharmacy in the elderly: lessons from REPOSI. Intern Emerg Med 9(7):723–734

Marengoni A, Onder G (2015) Guidelines, polypharmacy, and drug-drug interactions in patients with multimorbidity. BMJ (Clinical Research Ed) 350(4):h1059

Mazze RS, Strock ES, Bergenstal RM, Criego A, Cuddihy R, Langer O et al (2011) Staged diabetes management. Wiley-Blackwell, Oxford

McGloin H, Adam SK, Singer M (1999) Unexpected deaths and referrals to intensive care of patients on general wards. are some cases potentially avoidable? J R Coll Physicians Lond 33:255–259

Nataraj N (2017) Modeling for the care of complex patients. North Carolina State University, Raleigh

Nataraj N, Ivy JS, Payton FC, Norman J (2018) Diabetes and the hospitalized patient: a cluster analytic framework for characterizing the role of sex, race and comorbidity from 2006 to 2011. Health Care Manag Sci 21(4):534–553

Nation's Report Card (2015) The nation's report card: mathematics. https://www.nations report-card.gov/reading_math_2015/#?grade=4. Acceseed 2.1.2018

National Academy of Engineering (2018) NAE grand challenges for engineering in the 21st century: advance personalized learning. http://www.engineeringchallenges.org/challenges/learning.aspx. Accessed 2.1.2018

National Academy of Sciences, National Academy of Engineering & Institute of Medicine (2007) Rising above the gathering storm: energizing and employing America for a brighter economic future. The National Academies Press, Washington, DC

National Center for Education Statistics (2006) The nation's report card: mathematics 2005. http://nces.ed.gov/nationsreportcard/pdf/main2005/2006453.pdf. Accessed 2.1.2018

National Research Council Center for Education: Mathematics Learning Study Committee (2001) Adding it up: helping children learn mathematics. The National Academies Press, Washington, DC

National Science Board (2007) National action plan for addressing the critical needs of U.S. science, technology, engineering, and mathematics education system. http://www.nsf.gov/nsb/documents/2007/stem_action.pdf. Accessed 2.1.2018

North Carolina Education Research Data Center (2018) Homepage. Duke University Center for Child and Family Policy. https://childandfamilypolicy.duke.edu/research/nc-education-data-center/. Accessed 2.1.2018

Nunes BP, Flores TR, Mielke GI, Thumé E, Facchini LA (2016) Multimorbidity and mortality in older adults: a systematic review and meta-analysis. Arch Gerontol Geriatr 67:130–138

Piette J, Kerr E (2006) The impact of comorbid chronic conditions on diabetes care. Diabetes Care 29(3):725–731

Prytherch DR, Smith GB, Schmidt PE, Featherstone PI (2010) ViEWS—towards a national early warning score for detecting adult inpatient deterioration. Resuscitation 81(8):932–937

Public Schools of North Carolina (2018a) North Carolina end-of-grade tests at grades 3-8. http://www.ncpublicschools.org/accountability/testing/eog/. Accessed 2.1.2018

Public Schools of North Carolina (2018b) ABC program information. http://www.dpi.state.nc.us/accountability/reporting/abc/2000-01/history. Accessed 2.1.2018

Public Schools of North Carolina (2018c) Educator effectiveness model: EVAAS. http://www.ncpublicschools.org/effectiveness-model/evaas/. Accessed 2.1.2018

Puterman ML (1994) Markov decision processes: discrete stochastic dynamic programming. Wiley-Interscience, New York

Reamer A (2012) Characterizing the progression of performance in mathematics over time: the application of Markovian models to an education system. Doctoral Dissertation. Retrieved from North Carolina State University Libraries: http://www.lib.ncsu.edu/resolver/1840.16/8728. Accessed 2.1.2018

Reamer A, Ivy J, Vila-Parrish A, Young R (2015) Understanding the evolution of mathematics performance in primary education and the implications for STEM learning: a Markovian approach. Comput Hum Behav 47:4–17

Royal College of Physicians (2012) National early warning score (NEWS): standardizing the assessment of acute illness severity in the NHS. Report of a working party. RCP, London

Safford MM, Allison JJ, Kiefe CI (2007) Patient complexity: more than comorbidity. The vector model of complexity. J Gen Intern Med 22(Suppl 3):382–390

Schaink A, Kuluski K, Lyons R, Fortin M (2012) A scoping review and thematic classification of patient complexity: offering a unifying framework. J Comorb 2(1):1–9

Sengul Orgut I (2015) Modeling for the equitable and effective distribution of food donations under capacity constraints. Doctoral dissertation. https://repository.lib.ncsu.edu/bitstream/handle/1840.16/10469/etd.pdf?sequence=2

Sengul Orgut I, Ivy JS, Uzsoy R, Wilson JR (2015) Modeling for the equitable and effective distribution of donated food under capacity constraints. IIE Trans 48(3):252–266

Sengul Orgut I, Ivy JS, Uzsoy R (2017) Modeling for the equitable and effective distribution of food donations under stochastic receiving capacities. IISE Trans 49(6):567–578

Sengul Orgut I, Ivy JS, Uzsoy R, Hale C (2018) Robust optimization approaches for the equitable and effective distribution of donated food. Eur J Oper Res 269(2):516–531

Simon M, Tzur R (2004) Explicating the role of mathematical tasks in conceptual learning: an elaboration of the hypothetical learning trajectory. Math Think Learn 6(2):91–104

Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M et al (2016) The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 315(8):801–810

Smallwood R (1971) The analysis of economic teaching strategies for a simple learning model. J Math Psychol 8:285–301

Smith AF, Wood J (1998) Can some in-hospital cardio-respiratory arrests be prevented? A prospective survey. Resuscitation 37:133–137

Smith GB, Prytherch DR, Schmidt P et al (2006) Hospital-wide physiological surveillance—a new approach to the early identification and management of the sickpatient. Resuscitation 71:19–28

Smith KN, Vila-Parrish AR, Ivy JS, Abel SR (2016) A simulation approach for evaluating medication supply chain structures. Int J Syst Sci Oper Logist 4(1):13–26

Subbe C, Kruger M, Rutherford P, Gemmel L (2001) Validation of a modified early warning score in medical admissions. QJM 94(10):521–526

Taffel SM, Placek PJ, Liss T (1987) Trends in the United States cesarean section rate and reasons for the 1980-85 rise. Am J Public Health 77(8):955–959

Tejada JJ, Diehl K, Ivy JS, Wilson JR, King RE, Ballan MJ et al (2013) Combined DES/SD simulaton model of breast cancer screening for older women: an overview. In: 2013 winter simulations conference (WSC). IEEE, pp 41–53

Tejada JJ, Ivy JS, King RE, Wilson JR, Ballan MJ, Kay MG et al (2014) Combined DES/SD model of breast cancer screening for older women, II: screening-and-treatment simulation. IIE Trans 46(7):707–727

Tsilidis KK, Kasimis JC, Lopez DS, Ntzani EE, Ioannidis JPA (2015) Type 2 diabetes and cancer: umbrella review of meta-analyses of observational studies. BMJ 350:g7607

United States Census Bureau (2016) Small area income and poverty estimates. http://www.census.gov/did/www/saipe/data/interactive/saipe.html?s_appname=saipe&map_yearselector=2014&map_geoselector=aa_c&s_appName=saipe&map_yearSelector=2014&map_geoSelector=aa_c. Accessed 2016

United States Department of Agriculture (2018). https://www.ers.usda.gov/topics/food-nutrition-assistance/food-security-in-the-us/measurement.aspx

United States Department of Education (2008) Foundations for success: the final report of the national mathematics advisory panel. http://www2.ed.gov/about/bdscomm/list/mathpanel/report/final-report.pdf. Accessed 2.1.2018

Vila-Parrish AR, Ivy JS, King RE (2008) A simulation based approach for inventory modeling of perishable pharmaceuticals. In: Proceedings of the winter simulation conference, pp 1532–1538

Vila-Parrish AR, Ivy JS, King R, Abel SR (2012) Patient-based pharmaceutical inventory management: a two-stage inventory and production model for perishable products with Markovian demand. Health Syst 1(1):69–83

Vila-Parrish AR, Ivy JS, He B (2015) Impact of the influenza season on a hospital from a pharmaceutical inventory management perspective. Health Syst 4(1):12–28. Special Issue: Public Health Preparedness; Abingdon

Vincent JL, Opal SM, Marshall JC, Tracey KJ (2013) Sepsis definitions: time for change. Lancet 381(9868):774–775

Vogeli C, Shields AE, Lee TA, Gibson TB, Marder WD, Weiss KB et al (2007) Multiple chronic conditions: prevalence, health consequences, and implications for quality, care management, and costs. J Gen Intern Med 22(S3):391–395

Wachter RM, Pronovost PJ (2006) The 100,000 lives campaign: a scientific and policy review. Jt Comm J Qual Patient Saf 32(11):621–627

Wang HE, Shapiro NI, Griffin R, Safford MM, Judd S, Howard G et al (2012) Chronic medical conditions and risk of sepsis. PLoS One 7(10):e48307. Gold JA (ed)

Ward BW, Schiller JS, Goodman RA (2014) Multiple chronic conditions among US adults: a 2012 update. Prev Chronic Dis 11:E62

Wolff JL, Starfield B, Anderson G et al (2002) Prevalence, expenditures, and complications of multiple chronic conditions in the elderly. Arch Intern Med 162(20):2269

World Health Organization Human Reproduction Programme, 10 April 2015 (2015) WHO statement on caesarean section rates. Reprod Health Matters 23(45):149–150. https://doi.org/10.1016/j.rhm.2015.07.007

Zhang J, Landy HJ, Branch DW, Burkman R, Haberman S, Gregory KD, Hatjis CG et al (2010a) Contemporary patterns of spontaneous labor with normal neonatal outcomes. Obstet Gynecol 116(6):1281–1287

Zhang J, Troendle J, Reddy UM, Laughon SK, Branch DW, Burkman R, Landy HJ et al (2010b) Contemporary cesarean delivery practice in the United States. Am J Obstet Gynecol 203(4):326.e1–326.e10. https://doi.org/10.1016/j.ajog.2010.06.058. Elsevier Inc.

Zhang S, Payton FC, Ivy JS (2013) Characterizing the impact of mental disorders on HIV patient length of stay and total charges. IIE Trans Healthcare Syst Eng 3(3):139–146

Zimmerman E, Woolf S, Haley A (2015) Understanding the relationship between education and health. U.S. Department of Health and Human Services Agency for Healthcare Research and Quality. https://www.ahrq.gov/professionals/education/curriculum-tools/population-health/zimmerman.html. Accessed 2.1.2018

Julie Simmons Ivy is a Professor in the Edward P. Fitts Department of Industrial and Systems Engineering and Fitts Faculty Fellow in Health Systems Engineering. She previously spent several years on the faculty of the Stephen M. Ross School of Business at the University of Michigan. She received her B.S. and Ph.D. in Industrial and Operations Engineering at the University of Michigan. She also received her M.S. in Industrial and Systems Engineering with a focus on Operations Research at Georgia Tech. She is a President of the Health Systems Engineering Alliance (HSEA) Board of Directors. She is an active member of the Institute of Operations Research and Management Science (INFORMS), Dr. Ivy served as the 2007 Chair (President) of the INFORMS Health Applications Society and the 2012–2013 President for the INFORMS Minority Issues Forum. Dr. Ivy was encouraged to pursue engineering because she was good at math and science in high school. She chose Industrial Engineering because she felt it was the field of engineering that dealt most with people. Her research interests are mathematical modeling of stochastic dynamic systems with emphasis on statistics and decision analysis as applied to healthcare, public health, and humanitarian logistics. Dr. Ivy loves the idea that engineering tools can be used to improve people's lives so she has sought to work on problems that are human-

centered and affect people's well-being through her research in healthcare, hunger relief, and education.

Muge Capan, Ph.D., is an Associate Clinical Professor of Decision Sciences and Management Information Systems at LeBow College of Business at Drexel University. She has received her Ph.D. in Industrial and Systems Engineering from North Carolina State University. She has expertise in predictive analytics and optimization applied in healthcare, e.g., decision analytical modeling, forecasting, and mathematical modeling with applications to medical decision-making and optimization of healthcare systems. Dr. Capan's current research projects include analytical models for risk stratification in inpatient settings, early warning score-based clinical decision support systems, development and implementation of analytical models for personalization of sepsis diagnosis and treatment in inpatient settings, and patient flow optimization. She is currently the Primary Investigator on a multi-institutional National Science Foundation grant and serves as a Co-Investigator on an RO1 NIH National Library of Medicine grant. Both grants are focusing on sepsis identification and management by integrating industrial engineering and computer science in patient care and advancing the scientific knowledge to predict and prevent sepsis-induced health deterioration.

Karen Hicklin is currently a postdoctoral fellow in the Department of Statistics and Operations Research at the University of North Carolina at Chapel Hill as a part of the Carolina Postdoctoral Program for Faculty Diversity. Karen received her Ph.D. in Industrial Engineering from the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University under the direction of Julie Ivy. Her research interests are mathematical modeling of stochastic systems with an emphasis on statistical and decision analysis as applied to healthcare, service environments, and public health. In particular, her research area focuses on decision-making under uncertainty with a concentration in decision-making in healthcare and humanitarian logistics. In her thesis work, Karen developed stochastic decision models to analyze the mode of delivery decision for expectant mothers through the evaluation of tradeoffs between prolonging labor and performing a cesarean section considering the various short- and long-term health outcomes for the mother and child. In addition to her mode of delivery work, Karen is interested in developing decision models for HIV testing and treatment and uterine fibroids management and treatment. Karen is originally from Washington, DC, and completed her undergraduate studies at Spelman College where she received a B.S. in Mathematics. Upon graduating from Spelman, Karen returned to Washington, DC, and received an M.S. in Mathematics and Statistics from Georgetown University. As a child, Karen participated in various STEM activities on the weekends and during the summers

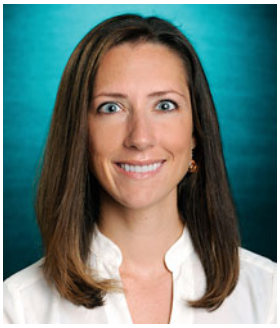
that helped foster her desire to learn more and explore different fields. Karen attributes her journey into STEM from a curiosity to understand how things work and an ambition to work on difficult problems.

Nisha Nataraj is a Prevention Effectiveness Fellow in the Division of Unintentional Injury Prevention, National Center for Injury Prevention and Control at the Centers for Disease Control and Prevention in Atlanta, GA. She is from Bangalore, India, where she received a bachelor's degree in 2010 from M.S. Ramaiah Institute of Technology in Industrial Engineering and Management. She then completed a master's in Industrial and Systems Engineering in 2012 from Rochester Institute of Technology, in Rochester, New York. This research was conducted when she was a graduate student in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University in Raleigh, NC, where she received her Ph.D. in 2017. Nisha is motivated to use her educational background to work in healthcare predominantly because efficient, effective and, just as importantly, equitable healthcare is a universal need. She believes that industrial engineering methods are uniquely poised to help address distinctive healthcare challenges, such as the significant stochasticity within the human body and associated care processes, heterogeneous populations with behavior that is difficult to predict, and the constant evolution in medicine as evidenced by novel drugs and changing policies. Her research interests include the analytical and predictive modeling of health systems research and delivery using simulation and applied statistical and stochastic methods. She is interested in the study of disease progression, management, and outcomes, with a particular interest in comorbidities and health disparities. Her research has been published in *Health Care Management Science* and the *Asia-Pacific Journal of Operational Research*. She is a member of the *Institute for Industrial and Systems Engineers* (IISE) and the *Institute for Operations Research and the Management Sciences* (INFORMS).

Irem Sengul Orgut received her Ph.D. in Industrial Engineering with a minor in Statistics in 2015 from the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. She now works at Lenovo as the Analytics Project Manager where she has been using operations research and analytics methods to improve customer experience and quality. Prior to starting her doctoral studies, she received her B.S. degrees in Industrial Engineering and Mechanical Engineering from Bogazici University, Istanbul, Turkey in 2010. She received various awards for her teaching and research. She is a member of INFORMS and IIE. Her web address is https://iremsengul.wordpress.com/.

"My initial interest in STEM fields started during high school years when I realized that engineering, mathematics and statistics can be used in combination to address the biggest

challenges faced by humanity and make a difference in the world. Since then, I have been especially passionate about utilizing STEM methods on complex problems that involve multiple objectives and conflicting decision makers and generating implementable solutions that make a significant and positive impact on people's lives."



Dr. Amy Craig Reamer is the Program Director of the NC State Engineering 2+2 Transfer Program at the University of North Carolina Wilmington. She earned a Bachelor of Science degree in Industrial Engineering, a Master's degree in Industrial Engineering, and a Ph.D. in Industrial and Systems Engineering (ISE), all from North Carolina State University. Prior to assuming her current position, she served as the Project Manager for a $3.21 million National Science Foundation funded grant, entitled Mathematics Instruction using Decision Science and Engineering Tools (MINDSET), housed in the NC State Edward P. Fitts Department of Industrial and Systems Engineering. Previously, Dr. Reamer worked as a Cost Engineer at Lenovo International, formerly the IBM Personal Computing Division.

Dr. Reamer developed a love for mathematics at an early age, in large part because of her mother's interest in the subject. A series of fortunate encounters with mentors at NC State University, to include the ISE advisor Mr. Clarence Smith, led to her decision to pursue ISE as a major. The idea of using mathematics and engineering to organize and optimize solutions still excites Dr. Reamer and she enjoys promoting engineering education in her community.

Dr. Anita Vila-Parrish is a Research Director in the Gartner Healthcare Supply Chain group. Dr. Vila-Parrish's focus is on the pharmacy supply chain, supply chain analytics, and inventory—all from the perspective of the healthcare provider. This involves the impact of these process improvements on patient outcomes due to a more efficient delivery of care. Dr. Vila-Parrish has over 15 years of experience in consulting, research and product development settings. Prior to Gartner her roles involved managing and conducting original, applied research as a Teaching Associate Professor and Director of Undergraduate Programs at North Carolina State University's Department of Industrial and Systems Engineering and was also an independent consultant focused on the use of healthcare data to improve supply chains, logistics, and delivery. She found her path to Industrial Engineering when her mother, an Electrical Engineer suggested it as a good fit—as always mothers are right! This interest that was sparked in her senior year of high school was ignited at North Carolina State University and became the pivotal foundation for her career.

Chapter 12
Improving Patient Safety in the Patient Journey: Contributions from Human Factors Engineering

Pascale Carayon and Abigail R. Wooldridge

Contents

12.1 Patient Safety: Progress and Opportunity for Improvement 275
12.2 Patient Journey 278
12.3 Contributions of Industrial and Systems Engineering to the Patient Journey 282
12.4 Modeling the Patient Journey for Improving Patient Safety: Examples 287
12.4.1 Modeling Medication Management Process 287
12.4.2 Modeling Care Coordination for Chronic Patients 289
12.5 Conclusions 293
References 293

12.1 Patient Safety: Progress and Opportunity for Improvement

The landmark publication of the US Institute of Medicine report "To Err is Human: Building a Better Health System" (Kohn et al. 1999) sparked interest in improving patient safety and quality of care, with the argument that healthcare delivery needed to be redesigned to satisfy patient needs and provide safe, effective, and efficient care (Institute of Medicine 2001). Almost 20 years later, patient safety has improved, but remains a significant problem.

P. Carayon (✉)
Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI, USA

Center for Quality and Productivity Improvement, University of Wisconsin-Madison, Madison, WI, USA
e-mail: pcarayon@wisc.edu

A. R. Wooldridge
Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA
e-mail: arwool@illinois.edu

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_12

On the 10th anniversary of the IOM report on To Err is Human, Wachter (2010) gave a "B-" to patient safety improvements. The 2013 AHRQ-funded report on patient safety practices (Shekelle et al. 2013a) identified 100 patient safety practices and strongly encouraged the implementation of 10 of them, such as hand hygiene, interventions for preventing venous thromboembolism, and checklists to prevent operative and post-operative events (Shekelle et al. 2013b). The 2014 National Healthcare Quality and Disparities Report by the Agency for Healthcare Research and Quality (AHRQ 2014) indicates that about half of the patient safety measures improved between 2010 and 2013, which led to 17% reduction in rates of hospital-acquired conditions. In 2013, 12% of hospital discharges had any number of hospital-acquired conditions. The most frequent hospital-acquired complications were: adverse drug events, pressure ulcers, catheter-associated urinary tract infections, and falls. Vincent et al. (2008) reviewed changes in various patient safety indicators in the past 20–30 years, which produced a mixed picture with both improvement and deterioration in patient safety. Multiple reasons are at play in this mixed review of patient safety progress, including challenge of reliable and valid patient safety measurement (Vincent et al. 2008), and insufficient investment in patient safety improvement and research (Wachter 2010). The call by the National Academy of Engineering and the Institute of Medicine to create new partnerships between engineering and healthcare (Reid et al. 2005) was issued in 2005, but still remains a work in progress. Additional attention needs to be dedicated to systems approaches to health, healthcare delivery, and patient safety, including the one proposed here that focuses on the patient journey.

In a discussion paper sponsored by the Institute of Medicine and the National Academy of Engineering, Kaplan et al. (2013) argue for the increasing use of systems approaches to improve healthcare delivery and other factors influencing health. They define the systems approach to health as follows: "A systems approach to health is one that applies scientific insights to understand the elements that influence health outcomes; models the relationships between those elements; and alters design, processes, or policies based on the resultant knowledge in order to produce better health at lower cost." The application of the systems approach to patient safety recognizes the web of factors that influence the multiple encounters and interactions between patients and healthcare delivery. This naturally leads to the expanded concept of patient safety recently proposed by Vincent and Amalberti (2016), which argues for looking at what happens to patients over a longer period of time. A narrow focus on specific patient/clinician encounters ignores many important factors influencing patient safety; it ignores the complexity of patients, their lives, needs, aspirations, motivations, and what occurs between specific patient/clinician encounters.

Figure 12.1 describes the approach proposed by Vincent and Amalberti (2016) for examining patient safety events along the patient journey. The patient journey is a temporal series of encounters with a healthcare facility, a hospital unit, a primary care clinic, a specialist clinic, or a home health agency. Each encounter is characterized by "active care," which is influenced by latent failures (management decision, organizational processes) and working conditions (e.g., workload, equipment). For

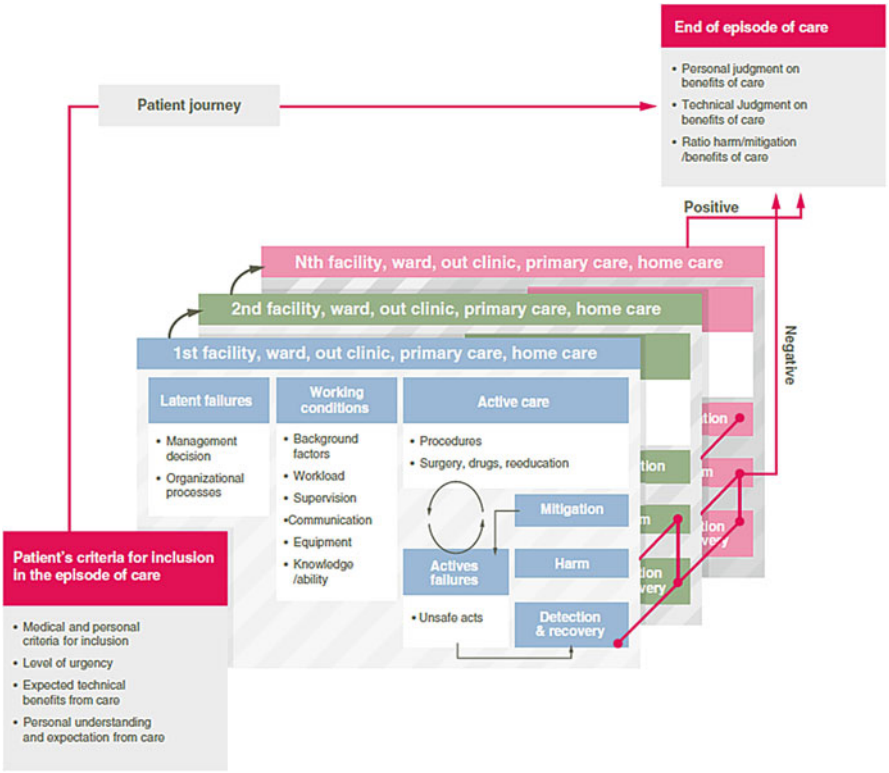


Fig. 12.1 Patient safety in the patient journey (Vincent and Amalberti 2016)

each encounter, unsafe acts could occur, which may be mitigated or may create harm. Vincent and Amalberti argue that this longitudinal analysis of encounters that occur over time with various positive and negative feedback loops is critical to better understand the final safety event and outcome.

An important element of the patient journey approach is incorporating the patient's perspective of safety; for example, consider a patient who underwent a knee replacement surgery, despite not reacting well to anesthesia and having "white coat syndrome." The first active care encounter with the surgeon was an appointment where knee replacement was selected as the best course of treatment, since their knee was immobilized at approximately 24-degree flexion for 2 years. The next active encounter was a pre-operative appointment, in which the patient's overall health was assessed for surgery. Then, the surgery took place; before hospital discharge, the knee was at 12-degree flexion and was reduced further to 7-degree within 2 days of at-home physical therapy. Within 2 weeks of the surgery, the patient went to a post-operative appointment in the surgical clinic, during which the surgeon examined the patient and their range of motion, which ranged from 0- to 88-degree flexion. The surgeon was not satisfied with this progress and bluntly said further

surgery would be required if the patient did not achieve 105-degree flexion. The surgeon seemingly forgot the baseline health of the patient when evaluating their progress after the procedure, leading to an inaccurate judgement of progress; further, the patient left feeling threatened about having a second surgery. This then could impact the patient's judgments on benefits of care—despite having technically good care, the complexity of their personal situation was not considered. The patient may be disheartened and disappointed by that interaction, possibly resulting in decreased adherence to the surgeon's recommendations (luckily, this particular patient is a bit more stubborn and resolved to do extra physical therapy). Further, they may also feel the nuances of their situation were not considered, leading to a perception of compromised, possibly unsafe, care. This case illustrates that a surgeon can provide technically excellent care—there were no active failures in any of these encounters and the transition from surgical clinic to hospital to home actually went quite well—but the patient can still be left with a feeling their care was compromised. Further analysis may identify a latent failure contributing to this situation, such as staffing decisions made by management may have affected workload and other working conditions and led to the surgeon rushing through their appointments without time to completely review each patient's chart, hence forgetting the baseline of that patient versus the average patient.

12.2 Patient Journey

In order to more effectively address patient safety, we need to understand the patient journey as challenges often occur at the interfaces when patients transition across care settings and environments. This process perspective of the patient journey is in line with the well-known Structure-Process-Outcome (SPO) model of Donabedian (1988). According to the SPO model, the "process of care" is "what is actually done in giving and receiving care" (page 1745). The SPO model revolutionized approaches to healthcare quality by introducing three domains of assessment for healthcare quality, i.e., structural (e.g., organization of care teams), process (e.g., follow-up treatment after test), and outcome (e.g., adverse drug events, healthcare-associated infections) measures. Donabedian also drew attention to the actual interactions between patients and the care delivery system as ways of improving healthcare quality. This process perspective on patient care has been operationalized in many different ways. Different terms are used to describe the care process or interactions between patients and the healthcare delivery system (see Table 12.1).

Many studies of care processes focus on the healthcare delivery side. For instance, studies have examined workflow of primary care physicians (Holman et al. 2016), intensive care physicians (Carayon et al. 2015a), and nurses (Douglas et al. 2012), and described activities and time spent on various activities. These approaches are important, but they tend to ignore the perspective of the patient, i.e., the patient's role, what the patient does, their interactions with the healthcare delivery system, and associated barriers and facilitators in getting high-quality

Table 12.1 Terminology related to patient journey

Terminology	Description and examples
Patient flow	Analysis of patient flow in hospital in Ghana to identify inefficiencies, including flow mapping with description of process and quantitative analysis of time spent on various process steps (Dixon et al. 2015)
Patient pathway	Participatory co-design process for designing ED patient pathway, i.e., what happens to patients in the ED (Reay et al. 2017) Sequence of steps between two points of the care process, e.g., series of consecutive steps or events from admission to discharge (Trebble et al. 2010)
Patient healthcare trajectory	Steps a patient with MI goes through based on their disease management and care, including socio-economic aspects (Pinaire et al. 2017)
Patient journey	Patient journey mapping to describe patient experience, including tasks within encounters, the emotional journey, the physical journey, and the various touch points (McCarthy et al. 2016) Analysis of patient safety events that patients may experience due to an accumulation of problems over a long period of time and across multiple care settings (Vincent et al. 2017)

care. In his milestone paper on the Structure-Process-Outcome model, Donabedian (1988) provided a balanced perspective as he described the care process as including both the patient and the healthcare professional: the care process "... includes the patient's activities in seeking care and carrying it out as well as the practitioner's activities in making a diagnosis and recommending or implementing treatment" (page 1745).

The patient journey concept is based on the perspective of patients, i.e., patients' experience of their care. Patient experience is critical for patient safety and clinical effectiveness (Doyle et al. 2013). As described by Carayon and Wood (2009), "Care to patients is provided through a myriad of interactions between various individuals: the patients themselves, their families and friends, healthcare providers, and various other staff" (page 29). These interactions occur over time and across multiple organizations, including hospitals, primary care clinics, specialist clinics, home, long-term care, etc. The spatio-temporal interactions and care transitions are the essence of the patient journey (Carayon and Wood 2009). As shown in Fig. 12.2, the patient often transitions from one care setting to another care setting or environment (e.g., home). Between care transitions, the patient interacts with multiple people in specific physical and organizational environment(s); these activities may rely on the use of multiple tools and technologies. We define the patient journey as the spatio-temporal distribution of patients' interactions with multiple care settings over time.

Ben-Tovim et al. (2008) described the patient journey during an hospitalization: the patient goes from one unit or service to another unit or service with different groups involved (see Fig. 12.3). During a hospital stay, patients are physically

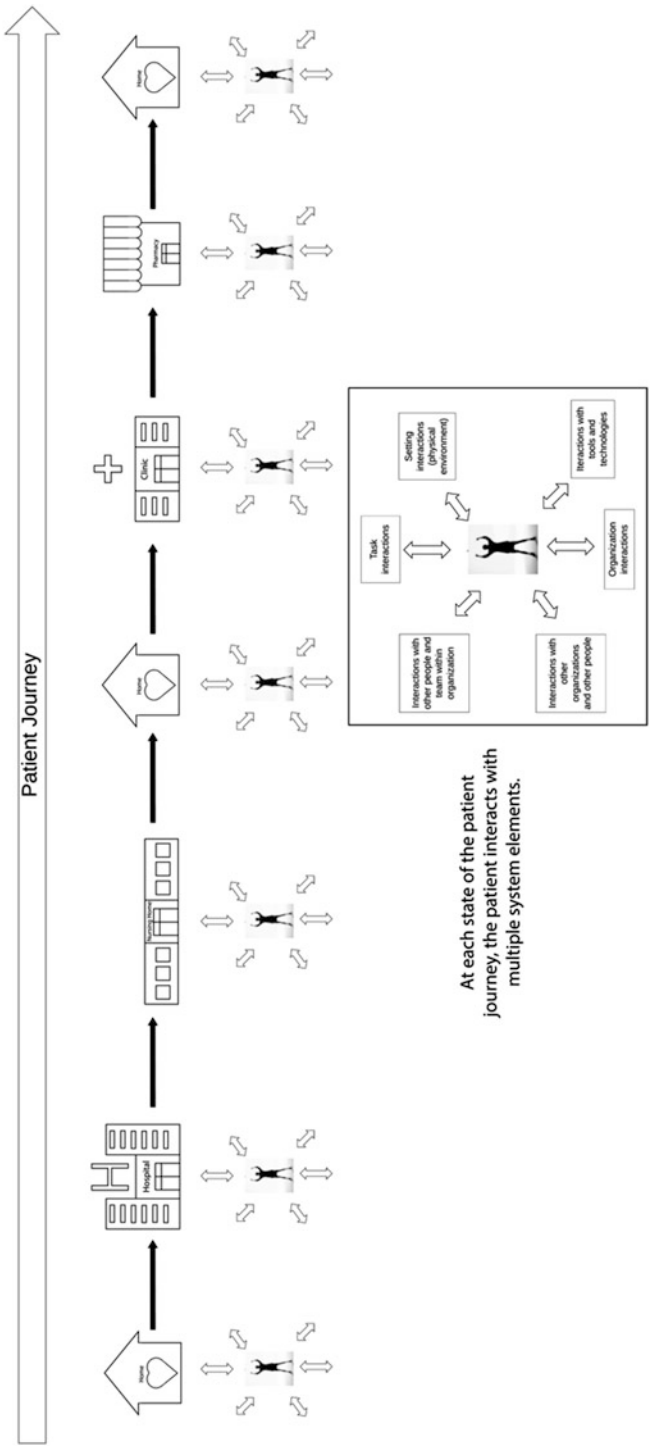


Fig. 12.2 Patient journey as a set of interactions and transitions; Adapted from Carayon and Wood (2009)

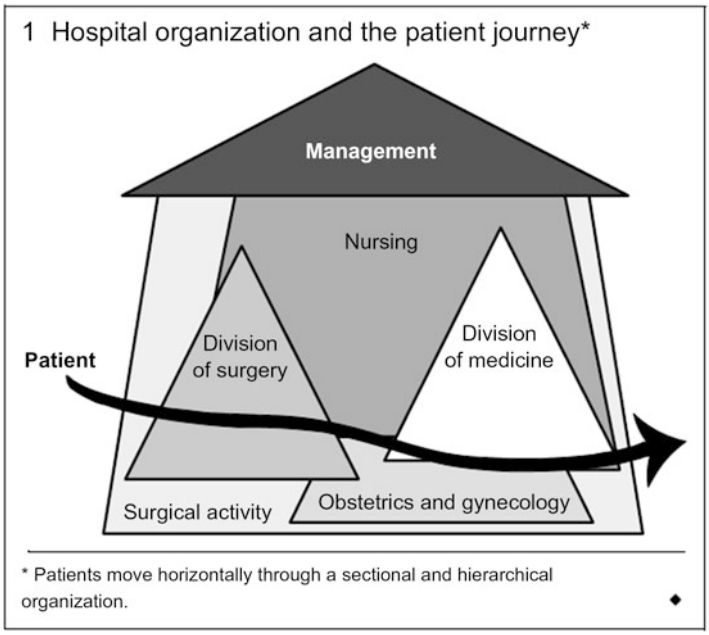


Fig. 12.3 Patient journey in the hospital (Ben-Tovim et al. 2008) (page S14)

transferred from hospital unit to hospital unit and receive care from different groups. The patient is the only "constant" in the patient journey. Healthcare professionals who care for the patient during their journey typically see only the segment of the journey that they are involved in and/or responsible for; therefore, they have a limited understanding of the interfaces between the different segments of the patient journey.

Vincent and Amalberti (2016) suggested that we need to look at the complex sequence of transitions and interfaces along the patient journey. The patient journey should be a combined representation of what happens to the patient, in particular across transitions. Therefore, the patient journey should not be a representation of separate diseases or issues experienced by the patient, but their cumulative impact. In line with Ben-Tovim et al. (2008), Vincent and Amalberti indicated that: "Each healthcare professional involved with a patient will only have a partial view of the patient journey" (page 44). There is, therefore, a need to develop approaches and methods for representing the patient journey and the full perspective of care as experienced by patients; this could then be shared with healthcare professionals across different care settings that participate in the patient journey. The result would be to enhance awareness of the patient journey and how safety develops over time (Schultz et al. 2007).

Jones (2013) contrasts the concept of "health journey" to that of patient journey. The health journey is more comprehensive than the patient journey as it includes

all of the activities and experiences of people that can enhance their healthy behaviors. In a manner similar to our conceptualization of the patient journey, Jones emphasizes the role of the person who makes decisions, seeks health, etc. This person-centered focus challenges healthcare delivery organizations and other organizations providing health services to reposition their processes to address the person's needs and characteristics. In this chapter, we focus on the concept of patient journey and recognize the need to address the concept of health journey in future research; this will help to bridge gaps between healthcare and health.

12.3 Contributions of Industrial and Systems Engineering to the Patient Journey

As previously described, progress in improving patient safety has been slower than expected (Agency for Healthcare Research and Quality 2015) and renewed efforts to improve patient safety have ensued. This resulted in the expanded conceptualization of patient safety throughout the patient journey (Vincent and Amalberti 2016; Carayon and Wood 2009). New tools, concepts, and methods are required in order to examine safety in the patient journey.

Industrial and Systems Engineering can make significant contributions by providing concepts, frameworks, and methods to model and evaluate the patient journey (Reid et al. 2005; Kaplan et al. 2013; President's Council of Advisors on Science and Technology 2014). Specifically, human factors and systems engineering (HF/SE) has been increasingly recognized as incorporating needs and desires of stakeholders (i.e., healthcare professionals and patients) as well as important sociotechnical aspects of healthcare (Reid et al. 2005; Gurses et al. 2012). Human factors, also known as ergonomics, is "the scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data and methods to design in order to optimize human well-being and overall system performance" (International Ergonomics Association (IEA) 2018). Systems engineering encompasses a broad range of approaches to understanding the performance of large, complex systems by understanding the performance of elements of that system and the interactions between those elements (Reid et al. 2005). HF/SE provides methods to model processes, such as the patient journey; these models can be used to establish shared understanding of processes as well as identify opportunities for improving safety and efficiency (Jun et al. 2009; Siemieniuch and Sinclair 2005). In healthcare, this can result in improved patient safety (Walker and Carayon 2009; Xie and Carayon 2015).

In healthcare, process models have been used to describe processes such as medication ordering and administration (Beuscart-Zéphir et al. 2007; Johnson and Fitzhenry 2006) and previsit planning, panel management, and checkout in primary care clinics (Wooldridge et al. 2017). These process descriptions are useful when

they are shared among the process stakeholders because they increase individual, team, and organizational awareness (Schultz et al. 2007). Process modeling methods have been used to study information flow, for example in chronic disease care (Unertl et al. 2009) and health information exchange use (Unertl et al. 2012). Process modeling has also been used to study communication, specifically interprofessional communication (Kummerow Broman et al. 2017) and interruptions in trauma care (Brixey et al. 2008). Process models can provide the input necessary for other methods, such as those that help to identify failure modes, e.g., failure modes and effects analysis (FMEA) of chemotherapy ordering and administration (van Tilburg et al. 2006) and IV medication administration using smart infusion pump technology (Wetterneck et al. 2006, 2009). Eason et al. (2012) used process modeling to study interdependencies in health information technology use across organizational boundaries. For a more complete review of process modeling methodologies, see Wooldridge et al. (2017). Recently, Werner et al. (2017) performed a process analysis of medication management during transition of elderly adults from hospital to skilled nursing home care transition, including developing an integrated model of medication management completed by the hospital, home health agency, and patient/caregiver(s). This provides a useful example of examining processes across organizational boundaries and lays a foundation for studying how barriers and facilitators propagate through distributed processes, such as the patient journey. Another important feature of this example is including the work that the patient and/or their caregiver perform outside of interaction(s) with healthcare professionals—they must understand, procure, follow instructions, and implement changes as part of medication management, in addition to communicating with home healthcare and hospital providers in this distributed process.

These process models must meet certain criteria in order to be useful in determining potential solutions for improving patient safety along the patient journey. While not focused on modeling the patient journey, Jun et al. (2009, 2010) evaluated several process and systems modeling approaches in light of how healthcare professionals perceive these methods. In their 2009 paper (Jun et al. 2009), eight modeling methods were evaluated:

- Stakeholder diagrams
- Information diagrams
- Process content diagrams
- Flowcharts
- Swim lane activity diagrams
- State transition diagrams
- Communication diagrams
- Data flow diagrams

The authors constructed each type of model for three healthcare processes: patient discharge from a ward, diabetic patient care in primary care, and an inpatient prostate cancer diagnostics procedure. The 2010 paper (Jun et al. 2010) added two methods: sequence diagrams and Integrated Definition for Function Modeling (IDFM). All ten methods were characterized by the focus (i.e., activity, stake-

holder, information) and linkage type (i.e., hierarchical, sequential, information). Healthcare professionals rated each method on familiarity with, usability (i.e., easily understandable) and utility (i.e., helpful in better understanding the system). They were most familiar with flowcharts. The researchers found that all methods could not capture the complexity inherent in healthcare processes (Jun et al. 2009), highlighting the need for human factors researchers to extend existing methods and develop new methods. They argue that greater use of these, and new, methods will improve our understanding of the complex and diverse healthcare system (Jun et al. 2010).

Process models are not only useful to *describe* processes in healthcare, but they can also be used to *evaluate* those processes. Simsekler et al. (2018a) proposed a systems-based framework to support the use of risk identification practices to patient safety in the following three stages: system familiarization, identification of risks, and presentation of risks. The first step of system familiarization is system description, using textual/graphic descriptions and process modeling approaches. The models are an input for brainstorming to identify risks that have not been identified through incident reporting. The system-based risk identification (RID) framework was applied in a gastroenterology unit; healthcare professionals involved felt that the framework was usable (e.g., was easy to use, would like to use the framework, has appropriate level of detail, helped to imagine new risks and hazards) and useful (e.g., were more aware of system-wide safety risks, would not identify same risks without the framework) (Simsekler et al. 2018a). In another study (Simsekler et al. 2018b), the authors evaluated six process modeling methods that can be used in the RID framework:

- Organizational diagrams
- Information diagrams
- Task diagrams
- Flow diagrams
- Communication diagrams
- System diagrams

The authors constructed each diagram for the treatment of adult patients with Attention Deficit Hyperactivity Disorder (ADHD) at a specialty clinic in Cambridge, UK. Eight healthcare professionals individually evaluated each of the six diagrams on general usefulness in risk identification and usefulness in identifying seven specific types of risk (i.e., task-related, equipment-related, organizational, environmental, staff-related, and patient-related risks). System and flow diagrams were found to be the most generally useful. Each diagram had strengths and weaknesses in identifying specific types of risks; in other words, using multiple models more completely identifies risks. Pragmatic considerations often result in the application of only one type of model, which the authors conclude should be carefully selected based on the risk sources that are most relevant to the process being evaluated.

Here we propose three criteria for evaluating HF/SE methods for modeling the patient journey:

1. The methods must be able to represent complex processes distributed over time and space.
2. The methods must model the process accurately.
3. The methods and their outputs must be usable and useful to the full range of users (e.g., HF/SE professionals, patient safety researchers, clinicians and other healthcare professionals, patients and their family/caregivers).

First, patient journey modeling methods must be able to handle the complex, interconnected processes in healthcare. These processes are often distributed across space and time and are completed by very large teams, such as those studied by Werner et al. (2017). These process models must, therefore, be able to depict the work completed in multiple locations by many people. In addition, these models must have flexible timescales, depending on the scale of time examined in the analysis. This timescale depends on the conditions of the patient, problems identified or expected to be identified, and complexity of the patient journey (Vincent and Amalberti 2016). For example, the timescale in studies of errors in medication administration may be very small, such as focusing on prescribing, dispensing, administering, and monitoring; alternatively, the timescale in studies considering safety in care transitions experienced by patients with chronic conditions will be much larger, as care coordination for these patients unfolds over weeks, months, or years. Further, they must not only describe the patient journey, but also support process evaluation, such as the investigation in propagation of errors in medication ordering and administration (Carayon et al. 2015b) and risk identification (Simsekler et al. 2018b).

The models of the patient journey must account for the actual process by including a complete and accurate description of the process. For the process model to be complete, it should include communication and coordination work—i.e., articulating work that is required to enable other work (Strauss 1985; Sawyer and Tapia 2006)—in addition to the tasks related to the actual provision of care, tasks performed by nonclinical staff and tasks performed by the patient/caregiver. Another aspect of completeness is the representation of all of the elements of sociotechnical work systems. The Systems Engineering Initiative for Patient Safety (SEIPS) model (Carayon et al. 2006, 2014a) provides a description of *five work system elements* and situates them in the context of Donabedian's SPO model (Donabedian 1988). An accurate model of the patient journey should describe who (*person*) is doing what (*tasks*) with *tools and technologies*; all of these activities happen in a *physical* and *organizational* environment.

In response to the finding of Jun and colleagues that existing methods do not effectively capture all aspects of complex healthcare systems, we have developed a method that incorporates sociotechnical systems theory while relying on the flowchart and swim lane diagram structure that is familiar to clinicians (Wooldridge et al. 2017). The SEIPS-based process modeling method incorporates all work system elements in process maps. It is important to consider all work system elements involved in a process, as well as the interactions between those elements, when optimizing the design of the system (Smith and Carayon-Sainfort 1989); these elements must be balanced in order to avoid negative outcomes for both patients

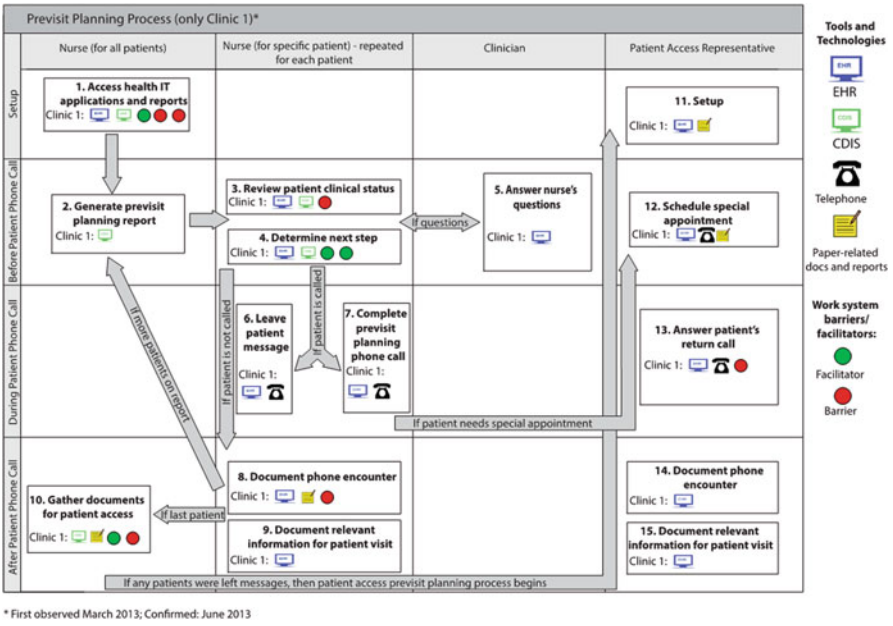


Fig. 12.4 Example of SEIPS-based process model of previsit planning in primary care (Wooldridge et al. 2017)

and workers (Carayon et al. 2013). Finally, for the process model to be accurate, not only should the entire process and all involved roles be represented, it should represent what workers, including healthcare professionals as well as patients and their family/caregivers, actually do (i.e., the *activity* (Leplat 1989; Daniellou and Rabardel 2005) or *work-as-done* (Hollnagel 2015)). This is different than what administrators and others believe that the workers do (i.e., the *task* (Leplat 1989; Daniellou and Rabardel 2005) or *work-as-imagined* (Hollnagel 2015)).

Figure 12.4 shows an example of a SEIPS-based process model. This process is a previsit planning process (Sinsky et al. 2013) at a primary care clinic. In this process, a nurse determines what tests and/or procedures should be done before each patient's scheduled appointment, then calls the patient to inform them of those tests and arrange for the patient to complete them. In this process map, roles completing each task are represented by columns; rows and arrows incorporate the temporal aspects of the process. Each box represents a higher level task, and the icons within that box represent the tools and technologies used. Organizational factors (such as communication and coordination mechanisms) are found in the arrows. In addition to describing the work system involved in this process, red and green circles indicate if there are work system barriers and facilitators associated with each process step; a detailed table provides additional data on tasks as well as work system barriers and facilitators.

HF/SE research tells us that tools must be usable and useful for them to be used; process models are no different. The process models must be clear, accurate, and easy to understand in order for HF/SE professionals and patient safety experts to be able to utilize them when designing or evaluating processes. They must also be usable and useful to clinicians and other stakeholders in the process. The objective is that process maps document the processes as currently performed and are used to increase shared understanding and organizational awareness. As noted earlier, Jun et al. (2009, 2010) evaluated the usability and utility of and familiarity with ten process modeling methods from clinician perspectives; they provide a framework to include other stakeholders in the process. In our experience, visual models of the patient journey can be easier to understand and encourage discussion between stakeholders required to increase shared understanding and awareness. HF/SE can help us ensure that these methods are usable and useful, as well as provide the basis to understand how to optimize performance throughout the patient journey.

12.4 Modeling the Patient Journey for Improving Patient Safety: Examples

12.4.1 Modeling Medication Management Process

Medication safety remains a major source of preventable medical errors. The rate of adverse drug events per 1000 hospital discharges has decreased from 49.5 in 2010 to 40.3 in 2013; but adverse drug events remain a major patient safety issue (AHRQ 2014). In particular, medication errors or complications are a major patient safety issue in the outpatient setting. A study of 2248 patients in 11 ambulatory clinics showed that 18% of them reported a drug complication, such as gastrointestinal complication, sleep issues, and fatigue. A 2007 systematic review examined 29 studies: 14 studies in the ambulatory setting and 15 studies in the hospital (Thomsen et al. 2007). The median ADE (Adverse Drug Event) incidence was 15 per 1000 person-months and the preventable ADE incidence was 5.6 per 1000 person-months. Medication errors that resulted in preventable ADEs often occur in the prescribing and monitoring stages. This systematic review clearly outlines that medication safety is an issue in the ambulatory setting as well.

Errors can occur at multiple stages of the medication management process: prescribing, dispensing, administration, or monitoring. Many studies focus on medication errors at specific stages, e.g., prescribing or ordering (Abramson et al. 2012; Dean et al. 2002; Lesar et al. 1990) or administration (Berdot et al. 2013; Calabrese et al. 2001; Westbrook et al. 2011). Other studies such as the original studies by Bates et al. (1993, 1995) assess errors at multiple stages of the medication management process. They found that medication errors resulting in preventable ADEs most often occur at the ordering and administration stages. These studies are critical to understand errors at various stages of the medication management

process; they also begun to identify stages where errors are intercepted. For instance, Bates et al. (1995) found that 48% of the ordering errors were intercepted at a later stage of the medication management process. It is important to recognize the temporal interdependencies between stages of the medication management process.

We have examined medication safety along the entire medication management process in two ICUs (Intensive Care Units) of a large medical center (Carayon et al. 2014b). We assessed a total of 630 admissions to a hybrid medical/surgical ICU and a cardiac ICU between October 2006 and March 2007. We used the methods developed by Bates et al. (1995) with four trained nurse data collectors who gathered data on medication errors and adverse drug events. We found a rate of 0.4 events (preventable or potential adverse drug events) per patient-day. Similarly to Bates et al. (1995), we found that the majority of medication errors occurred at the ordering and administration stages: 32% at the ordering stage and 39% at the administration stage. Our data collection method allowed us to capture related medication errors, in particular sequences of errors. A sequence of errors occurred when an error at a stage led to an error at another stage. A typical sequence occurred when an error at the dispensing stage was followed by an error at the medication administration stage: the nurse was unable to administer the medication (administration error) because it was not available in the medication room due to a delay or issue in dispensing the medication (dispensing error). Another frequent sequence involved a transcription error (e.g., medication order not transcribed on the medication administration record), which led to a medication not being administered. Figure 12.5 displays the frequent sequential errors that we found in this study of ICU patients.

In order to develop a deeper understanding of sequences of medication errors, we conducted an additional analysis to assess propagation of errors in the medication management process (Carayon et al. 2015c). We conducted a secondary analysis of 1732 medication events that were identified in our study of medication safety in ICUs (Carayon et al. 2014b). We developed a Markov Chain model of error propagation through the various stages of medication management: ordering, transcription, preparation, dispensing, administration, and monitoring. Errors at the preparation stage often led to dispensing errors. Errors at the dispensing stage led to administration errors (see also Fig. 12.5).

Another important aspect of understanding patient safety in the temporal medication management process is the identification of error detection and recovery (Kanse et al. 2006; Wetterneck 2012). In a study of care coordination for patients with chronic conditions, we found that care managers identified a total of 608 medication errors, which included missing or omitted information, dosage errors, medication orders that were not restarted or renewed, and medication orders that were not discontinued (Carayon et al. 2016). Care managers were able to correct many of these errors in collaboration with other care team members, such as primary care physicians. Error detection and recovery mechanisms are important to understand in the patient journey (see Fig. 12.1); they are often key in ensuring that negative outcomes and harm do not result for patients.

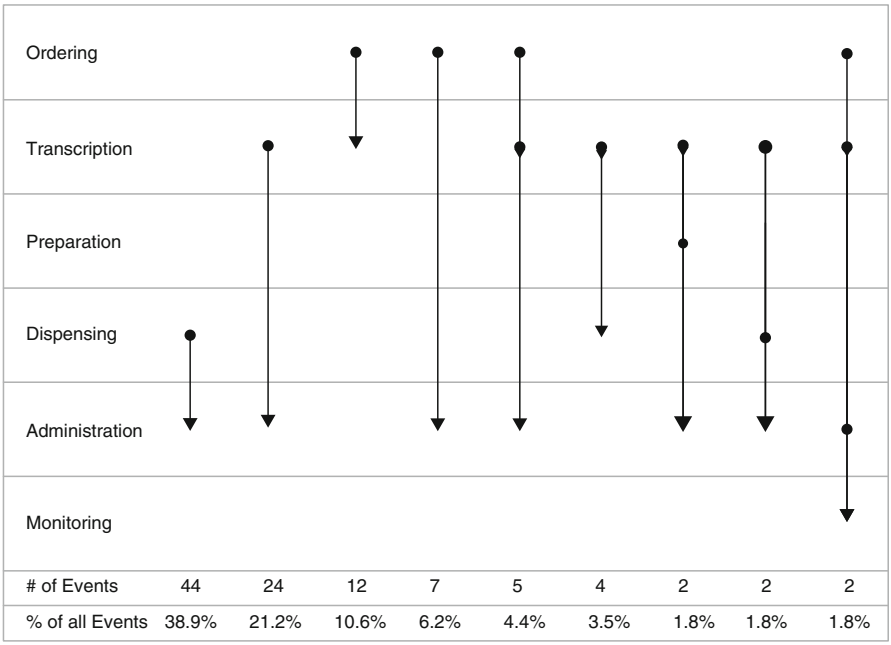


Fig. 12.5 Sequential medication errors in the intensive care unit (Carayon et al. 2014b)

12.4.2 Modeling Care Coordination for Chronic Patients

Patients with chronic conditions account for about half of the US population (Ward et al. 2014) and more than 80% of healthcare costs (Anderson 2010). Heart failure (HF) (Hollenberg and Heitner 2012) and chronic obstructive pulmonary disease (COPD) (Pauwels et al. 2001) are serious chronic conditions that contribute to hospitalization, morbidity, and mortality of adults in the USA. Patients with chronic conditions often transition through different care settings, e.g., hospital, outpatient clinics, skilled nursing facilities, and receive care from a variety of healthcare professionals (Bodenheimer 2008). When these transitions are not safe, patients may be (re-)admitted to hospitals or die (Jencks et al. 2009); care coordination interventions can improve patient safety in these care transitions (Marcotte et al. 2015; McDonald et al. 2007). One such intervention is the use of care managers to support patient care by providing patient education, performing medication reconciliation, organizing follow-up visits, assessing patient needs, and monitoring patient status (Maliski et al. 2004; Tomcavage et al. 2012).

Geisinger Health System, funded by the Office of the National Coordinator for Health Information Technology (IT), developed and implemented a program to improve the health of chronically ill patients in central Pennsylvania by

improving care coordination within and across healthcare organizations, e.g., hospitals, skilled nursing facilities, home health agencies, and primary care clinics (http://keystonebeaconcommunity.org/). The Center for Quality and Productivity Improvement (CQPI) at the University of Wisconsin-Madison evaluated various aspects of the project from a human factors engineering perspective (http://cqpi.wisc.edu/keystone-beacon-community-rural-community-wide-retirement-home-proposal/). A variety of methods (e.g., interviews, observations, surveys, and focus groups) were used to study the work system, process and both patient and clinician outcomes in order to assess and improve the design and implementation of health IT-supported care management; results of some of these analyses have been reported elsewhere (Alyousef et al. 2012, 2017; Carayon et al. 2012, 2015d; Hoonakker et al. 2014).

In the Keystone Beacon Project, new and existing health IT applications were used to identify patients who have HF or COPD. These patients were then assigned nurse care managers (CM) to work with them and complement their care team. Identified patients were managed by inpatient CMs when they were hospitalized and by outpatient CMs when they were not. The CMs were supported by medical management assistants (MMAs) to help with their workload. A key component of the evaluation conducted by CQPI included describing the work of care coordination. Given our sociotechnical systems approach—which focuses on the need to consider both the technical and social subsystems within a work environment to jointly optimize performance (Kleiner 2008)—and the program's focus on health IT-supported care management, it was important for us to understand not only what tasks were completed by CMs and MMAs, but also the tools and technologies they used.

We developed process maps to model these processes, shown in Fig. 12.6. In Fig. 12.6, columns distinguish between work CMs and MMA do when they are (right column) and are not (left column) interacting with patients or their family. The top half of the figure focuses on when the patient is not hospitalized, and the bottom focuses on the patient in the hospital. The temporal nature of the process, i.e., the patient journey, is captured in the arrows as well as rows of the process model (i.e., the tasks to the left and top of the figure are completed first; then the rest of the tasks on that line going right are completed before the next row of tasks). The rows separated by dots indicate work done for the first interaction with a newly identified patient (first), the final patient interaction before a transition to another care setting (last) and all interactions in between (middle). The tools and technologies used are indicated by different icons; different roles use different tools and technologies, as indicated by the color highlighting: green for inpatient CMs, blue for outpatient CMs, and pink for MMAs.

Outpatient CMs and MMAs first work in parallel to identify patients eligible to participate in the program using a variety of technologies: a case management software (i.e., Wisdom), electronic health records (EHR, both internal in that clinic and external to other healthcare organizations), the health information exchange

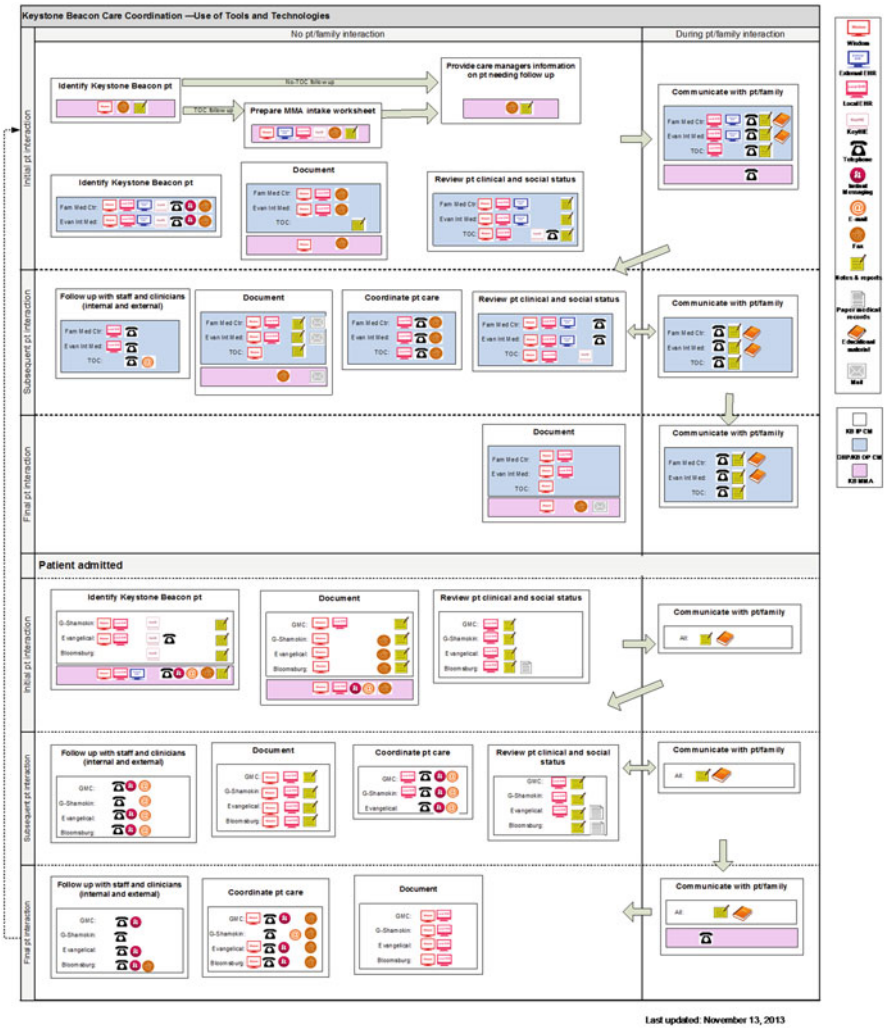


Fig. 12.6 Inpatient and outpatient care managers' activities during the patient journey

(HIE) developed for that project, telephones, instant messaging, and paper notes and reports. Depending on whether the patient was receiving transition of care (TOC) follow-up from others, the MMA might complete an intake form, using the same technologies in addition to other documentation, while the CM completes documentation. The MMA then gives information to the CM, and the CM reviews the patient's status before communicating with the patient and their family; during that communication they use EHRs, phones, notes and reports, and educational material. For subsequent patient interactions, the CM iterates between following up with clinicians and staff, documenting their work, coordinating patient care,

reviewing the patient's status, and interacting with the patient and their family. The MMA helps with documentation; in each of these steps, between two and four technologies are used. For the final patient interaction, the CM and MMA complete documentation (one to three technologies) and communicate with the patient and/or family.

When an eligible patient is admitted to a hospital, the inpatient CM, working with an MMA, identifies them using Wisdom, the internal and external EHRs, the HIE, phone, instant messaging, e-mail, fax, and written notes and reports. There may be communication between outpatient and inpatient CMs (Kianfar 2016). When a patient is identified, the CM and MMA complete documentation using various technologies and then the CM reviews the patient's status (using the hospital EHR, paper notes and reports, and paper medical records). The CM then communicates with the patient and/or their family. For subsequent patient interactions while the patient is hospitalized, the inpatient CM iterates between following up with clinicians and staff, documenting their work, coordinating patient care, reviewing the patient's status and communicating with the patient and their family. After the last communication with a patient and/or their family, the inpatient CM follows up with clinicians and staff, coordinates care as the patient returns to the community setting, and completes documentation.

The dashed line from the bottom row going back to the beginning row indicates that an outpatient CM must recognize that the patient is back in the community setting and resume managing the care of that patient. Thus, the process begins again, until the patient is admitted again or no longer requires CM services. Again, there may be communication between outpatient and inpatient CMs (Kianfar 2016).

This process model clearly demonstrates that the care coordination process between two care settings (i.e., community setting and hospital) unfolds over time. This is congruent with work by Werner et al. (2018a) that highlighted that hospital-to-home care transitions occur over a period of time, up to 3 months after discharge. Further, the healthcare professionals use a variety of tools and technologies, electronic and otherwise, to accomplish their work, which is distributed between interactions with the patient/family and work outside of those interactions. This aligns with the shift to think of patient safety across the patient journey rather than only during the time the patient interacts with the healthcare system; patient safety could be compromised during the time the patient and CM are not interacting. However, a missing component is the work that the patient and/or their family/caregiver must complete outside of those interactions, such as going from being a passive recipient of medication in the hospital to managing their medications, and communicating with their care team (Werner et al. 2018a, b). In order to fully understand safety throughout the patient journey, we must strive to incorporate the perspective of the patient, including the work patients do outside of their interactions with healthcare professionals.

12.5 Conclusions

Progress in improving patient safety and quality of care has been slow. We, like others (Kaplan et al. 2013; Vincent and Amalberti 2016), believe that adopting a systems approach to patient safety, including the patient's perspective and their journey over time and space, is key to achieving breakthroughs in patient safety. This should involve the use of various process modeling methodologies created by HF/SE to model and analyze the patient journey as a process. These models must accurately represent complex processes that are distributed over time and space and be useful to a variety of stakeholders. To this end, we have developed and are refining the SEIPS-based process modeling method (Wooldridge et al. 2017). We presented two examples of our work using HF/SE techniques to model the patient journey, as well as the work of others that emphasize the concept of patient journey. One of our examples focused on inpatient medication management through medication prescribing, dispensing, administration, and monitoring. The second example focused on care coordination for patients with chronic conditions. Future research should consider the concept of patient journey and further develop modeling approaches for capturing the patient journey, including the patient's perspective and work. This is critical to improve patient safety.

Acknowledgments This research was supported by the Clinical and Translational Science Award (CTSA) program, through the NIH National Center for Advancing Translational Sciences (NCATS), Grant UL1TR000427. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Abramson EL, Bates DW, Jenter C, Volk LA, Barron Y, Quaresimo J et al (2012) Ambulatory prescribing errors among community-based providers in two states. J Am Med Inform Assoc 19(4):644–648

Agency for Healthcare Research and Quality (2015) 2014 National Healthcare Quality & disparities report. Agency for Healthcare Research and Quality, Rockville. Contract No.: AHRQ Pub. No. 15-0007

AHRQ (2014) National healthcare quality and disparities report. Agency for Healthcare Research and Quality, Rockville, p 2015

Alyousef B, Carayon P, Hoonakker P, Hundt AS, Cartmill R, Tomcavage J, et al (2012) Care managers' challenges in using multiple health IT applications. In: The Human Factors and Ergonomics Society (ed) Proceedings of the human factors and ergonomics society 56th annual meeting. The Human Factors and Ergonomics Society, Santa Monica, pp 1748–1752

Alyousef B, Carayon P, Hoonakker P, Hundt AS, Salek D, Tomcavage J (2017) Obstacles experienced by care managers in managing information for the care of chronically ill patients. Int J Human Comput Interact 33(4):313–321

Anderson G (2010) Chronic care: making the case for ongoing care. Robert Wood Johnson Foundation, Princeton

Bates DW, Leape LL, Petrycki S (1993) Incidence and preventability of adverse drug events in hospitalized adults. J Gen Intern Med 8(6):289–294

Bates DW, Cullen DJ, Laird N, Petersen LA, Small SD, Servi D et al (1995) Incidence of adverse drug events and potential adverse drug events: implications for prevention. J Am Med Assoc 274(1):29–34

Ben-Tovim DK, Dougherty ML, O'Connell TJ, McGrath KM (2008) Patient journeys: the process of clinical redesign. Med J Aust 188(6):S14–SS7

Berdot S, Gillaizeau F, Caruba T, Prognon P, Durieux P, Sabatier B (2013) Drug administration errors in hospital inpatients: a systematic review. PLoS One 8(6):e68856

Beuscart-Zéphir M-C, Pelayo S, Anceaux F, Maxwell D, Guerlinger S (2007) Cognitive analysis of physicians and nurses cooperation in the medication ordering and administration process. Int J Med Inf 76:S65–S77

Bodenheimer T (2008) Coordinating care—a perilous journey through the health care system. N Engl J Med 358(10):1064–1071

Brixey JJ, Tang Z, Robinson DJ, Johnson CW, Johnson TR, Turley JP et al (2008) Interruptions in a level one trauma center: a case study. Int J Med Inf 77(4):235–241

Calabrese AD, Erstad BL, Brandl K, Barletta JF, Kane SL, Sherman DS (2001) Medication administration errors in adult patients in the ICU. Intensive Care Med 27(10):1592–1598

Carayon P, Wood KE (2009) Patient safety: the role of human factors and systems engineering. In: Rouse WB, Cortese DA (eds) Engineering the system of healthcare delivery. IOS Press, Amsterdam, pp 23–46

Carayon P, Hundt AS, Karsh B-T, Gurses AP, Alvarado CJ, Smith M et al (2006) Work system design for patient safety: the SEIPS model. Qual Saf Health Care 15(Suppl I):i50–ii8

Carayon P, Alyousef B, Hoonakker PLT, Hundt AS, Cartmill R, Tomcavage J et al (2012) Challenges to care coordination posed by the use of multiple health IT applications. Work (Reading, Mass) 41(2):4468–4473

Carayon P, Karsh B-T, Gurses AP, Holden RJ, Hoonakker P, Hundt AS et al (2013) Macroergonomics in health care quality and patient safety. Rev Hum Factors Ergon 8:4–54

Carayon P, Wetterneck TB, Rivera-Rodriguez AJ, Hundt AS, Hoonakker P, Holden R et al (2014a) Human factors systems approach to healthcare quality and patient safety. Appl Ergon 45(1): 14–25

Carayon P, Wetterneck TB, Cartmill R, Blosky MA, Brown R, Kim R et al (2014b) Characterising the complexity of medication safety using a human factors approach: an observational study in two intensive care units. BMJ Qual Saf 23(1):56–75

Carayon P, Wetterneck TB, Alyousef B, Brown RL, Cartmill RS, McGuire K et al (2015a) Impact of electronic health record technology on the work and workflow of physicians in the intensive care unit. Int J Med Inf 84(8):578–594

Carayon P, Ju F, Ri C, Hoonakker P, Wetterneck TB, Li J (2015b) Medication error propagation in intensive care units. Proc Human Factors Ergon Soc Annu Meeting 59(1):518–521

Carayon P, Ju F, Cartmill R, Hoonakker P, Wetterneck TB, Li J (2015c) Medication error propagation in intensive care units. In: The Human Factors and Ergonomics Society (ed) Proceedings of the human factors and ergonomics society annual meeting, 59. The Human Factors and Ergonomics Society, Santa Monica, pp 518–521

Carayon P, Hundt A, Hoonakker P, Kianfar S, Alyousef B, Salek D et al (2015d) Perceived impact of care managers' work on patient and clinician outcomes. Eur J Person Centered Healthcare 3(2):158–167

Carayon P, Cartmill R, Hoonakker P, Hundt AS, Salek D, Walker J, Tomcavage J (2016) Collaborative processes of care managers in the detection and recovery of medication errors. In: Mollo V, Falzon P (eds) Proceedings of the healthcare systems ergonomics and patient safety conference, Toulouse, France, pp 87–92

Daniellou F, Rabardel P (2005) Activity-oriented approaches to ergonomics: some traditions and communities. Theor Issues Ergon Sci 6(5):353–357

Dean B, Schachter M, Vincent C, Barber N (2002) Causes of prescribing errors in hospital inpatients: a prospective study. Lancet 359(9315):1373–1378

Dixon CA, Punguyire D, Mahabee-Gittens M, Ho M, Lindsell CJ (2015) Patient flow analysis in resource-limited settings: a practical tutorial and case study. Global Health Sci Pract 3(1): 126–134

Donabedian A (1988) The quality of care. How can it be assessed? J Am Med Assoc 260(12):1743–1748

Douglas S, Cartmill R, Brown R, Hoonakker P, Slagle J, Van Roy KS et al (2012) The work of adult and pediatric intensive care unit nurses. Nurs Res 62(1):50–58

Doyle C, Lennox L, Bell D (2013) A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. BMJ Open 3(1):1–18

Eason K, Dent M, Waterson P, Tutt D, Hurd P, Thornett A (2012) Getting the benefit from electronic patient information that crosses organisational boundaries. Final report. NIHR Service Delivery and Organisation Programme, Department of Health, London

Gurses AP, Ozok AA, Pronovost PJ (2012) Time to accelerate integration of human factors and ergonomics in patient safety. BMJ Qual Saf 21(4):347–351

Hollenberg S, Heitner S (2012) Congestive heart failure. Cardiology in family practice: a practical guide. Humana Press, New York, pp 91–111

Hollnagel E (2015) Chapter 18: Why is work-as-imagined different from work-as-done? In: Wears RL, Hollnagel E, Braithwaite J (eds) Resilient health care, Volume 2: The resilience of everyday clinical work. Ashgate Publishing, Farnham

Holman GT, Beasley JW, Karsh BT, Stone JA, Smith PD, Wetterneck TB (2016) The myth of standardized workflow in primary care. J Am Med Inform Assoc 23(1):29–37

Hoonakker PLT, Carayon P, Alyousef B, Cartmill RS, Hundt AS, Kianfar S, et al (2014) Build it and they will come? Assessment of use, usability, and usefulness of the Keystone Health Information Exchange. In: Organizational design and management (ODAM) conference; 2014 Aug 17-20. IEA Press, Copenhagen

Institute of Medicine (2001) Crossing the quality chasm: a new health system for the 21st century. National Academy Press, Washington, DC

International Ergonomics Association (IEA) (2018) Defintiion and domains of ergonomics. [cited 2018 February 19, 2018]. http://www.iea.cc/whats/

Jencks SF, Williams MV, Coleman EA (2009) Rehospitalizations among patients in the medicare fee-for-service program. N Engl J Med 360(14):1418–1428

Johnson KB, Fitzhenry F (2006) Case report: activity diagrams for integrating electronic prescribing tools into clinical workflow. J Am Med Inf Assoc 13(4):391–395

Jones PH (2013) Design for care—innovating healthcare experience. Rosenfield Media, Brooklyn

Jun GT, Ward J, Morris Z, Clarkson J (2009) Health care process modelling: which method when? Int J Qual Health Care 21(3):214–224

Jun GT, Ward J, Clarkson P (2010) Systems modelling approaches to the design of safe healthcare delivery: ease of use and usefulness perceived by healthcare workers. Ergonomics 53(7): 829–847

Kanse L, Van der Schaaf TW, Vrijland ND, Van Mierlo H (2006) Error recovery in hospital pharmacy. Ergonomics 49(5–6):503–516

Kaplan GS, Bo-Linn G, Carayon P, Pronovost P, Rouse W, Reid P et al (2013) Bringing a systems approach to health. Institute of Medicine and National Academy of Engineering, Washington, DC

Kianfar S (2016) Understanding care coordination activities performed for chronically ill patients. ProQues: University of Wisconsin-Madison, Madison, Wisconsin

Kleiner BM (2008) Macroegonomics: work system analysis and design. Hum Fact 50(3):461–467

Kohn LT, Corrigan JM, Donaldson MS (eds) (1999) To err is human: building a safer health system. National Academy Press, Washington, DC

Kummerow Broman K, Kensinger C, Hart H, Mathisen J, Kripalani S (2017) Closing the loop: a process evaluation of inpatient care team communication. BMJ Qual Saf 26(1):30–32

Leplat J (1989) Error analysis, instrument and object of task analysis. Ergonomics 32(7):813–822
Lesar TS, Briceland L, Delcoure K et al (1990) Medication prescribing orders in a teaching hospital. J Am Med Assoc 263:2329–2334
Maliski SL, Clerkin B, Litwin MS (2004) Describing a nurse case manager intervention to empower low-income men with prostate cancer. Oncol Nurs Forum 31(1):57–64
Marcotte L, Kirtane J, Lynn J, McKethan A (2015) Integrating health information technology to achieve seamless care transitions. J Patient Saf 11(4):185–190
McCarthy S, O'Raghallaigh P, Woodworth S, Lim YL, Kenny LC, Adam F (2016) An integrated patient journey mapping tool for embedding quality in healthcare service reform. J Decis Syst 25(Suppl 1):354–368
McDonald KM, Sundaram V, Bravata DM, Lewis R, Lin N, Kraft SA et al (2007) Closing the quality gap: a critical analysis of quality improvement strategies (vol. 7: Care coordination). Agency for Healthcare Research and Quality (US), Rockville
Pauwels RA, Buist AS, Calverley PM, Jenkins CR, Hurd SS (2001) Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. NHLBI/WHO Global Initiative for Chronic Obstructive Lung Disease (GOLD) Workshop summary. Am J Respir Crit Care Med 163(5):1256–1276
Pinaire J, Azé J, Bringay S, Landais P (2017) Patient healthcare trajectory. An essential monitoring tool: a systematic review. Health Inf Sci Syst 5(1):1
President's Council of Advisors on Science and Technology (2014) Report to the President, better health care and lower costs : accelerating improvement through systems engineering. Executive Office of the President, President's Council of Advisors on Science and Technology, Washington, DC
Reay SD, Collier G, Douglas R, Hayes N, Nakarada-Kordic I, Nair A et al (2017) Prototyping collaborative relationships between design and healthcare experts: mapping the patient journey. Design Health 1(1):65–79
Reid PR, Compton WD, Grossman JH, Fanjiang G (2005) Building a better delivery system. A New engineering/health care partnership. The National Academies Press, Washington, DC
Sawyer S, Tapia A (2006) Always articulating: theorizing on mobile and wireless technologies. Inf Soc 22(5):311–323
Schultz K, Carayon P, Hundt AS, Springman SR (2007) Care transitions in the outpatient surgery preoperative process: facilitators and obstacles to information flow and their consequences. Cogn Technol Work 9(4):219–231
Shekelle PG, Wachter R, Pronovost P, Schoelles K, McDonald K, Dy S et al (2013a) Making health care safer II: an updated critical analysis of the evidence for patient safety practices. Agency for Healthcare Research and Quality, Rockville. Contract no.: AHRQ publication no.13-E001-EF
Shekelle PG, Pronovost PJ, Wachter RM, McDonald KM, Schoelles K, Dy SM et al (2013b) The top patient safety strategies that can be encouraged for adoption now. Ann Intern Med 158(5 Pt 2):365–368
Siemieniuch CE, Sinclair MA (2005) The analysis of organisational processes. Evaluation of human work, 3rd edn. CRC Press, New York, pp 977–1008
Simsekler MCE, Ward JR, Clarkson PJ (2018a) Design for patient safety: a systems-based risk identification framework. Ergonomics 61(8):1046–1064
Simsekler MCE, Ward JR, Clarkson PJ (2018b) Evaluation of system mapping approaches in identifying patient safety risks. Int J Qual Health Care 30(3):227–233
Sinsky CA, Willard-Grace R, Schutzbank AM, Sinsky TA, Margolius D, Bodenheimer T (2013) In search of joy in practice: a report of 23 high-functioning primary care practices. Ann Fam Med 11(3):272–278
Smith MJ, Carayon-Sainfort P (1989) A balance theory of job design for stress reduction. Int J Ind Ergon 4(1):67–79

Strauss AL (1985) Work and the division of labor. Sociol Quart 26(1):1–19
Thomsen LA, Winterstein AG, S∅ndergaard B, Haugb∅lle LS, Melander A (2007) Systematic review of the incidence and characteristics of preventable adverse drug events in ambulatory care. Ann Pharmacother 41(9):1411–1426
van Tilburg CM, Leistikow IP, Rademaker CMA, Bierings MB, van Dijk ATH (2006) Health care failure mode and effect analysis: a useful proactive risk analysis in a pediatric oncology ward. Qual Saf Health Care 15(1):58–63
Tomcavage J, Littlewood D, Salek D, Sciandra J (2012) Advancing the role of nursing in the medical home model. Nurs Admin Q 36(3):194–202
Trebble TM, Hansi N, Hydes T, Smith MA, Baker M (2010) Process mapping the patient journey: an introduction. Br Med J 341:394–401
Unertl KM, Weinger MB, Johnson KB, Lorenzi NM (2009) Describing and modeling workflow and information flow in chronic disease care. J Am Med Inf Assoc 16(6):826–836
Unertl KM, Johnson KB, Lorenzi NM (2012) Health information exchange technology on the front lines of healthcare: workflow factors and patterns of use. J Am Med Inf Assoc 19(3):392–400
Vincent C, Amalberti R (2016) Safer healthcare—strategies for the real world. Springer Open, New York
Vincent C, Aylin P, Franklin BD, Holmes A, Iskander S, Jacklin A et al (2008) Is health care getting safer? Br Med J 337(7680):1205–1207
Vincent C, Carthey J, Macrae C, Amalberti R (2017) Safety analysis over time: seven major changes to adverse event investigation. Implement Sci 12(1):151
Wachter RM (2010) Patient safety at ten: unmistakable progress, troubling gaps. Health Aff (Millwood) 29(1):165–173
Walker JM, Carayon P (2009) From tasks to processes: the case for changing health information technology to improve health care. Health Aff 28(2):467–477
Ward BW, Schiller JS, Goodman RA (2014) Multiple chronic conditions among US adults: a 2012 update. Prev Chronic Dis 11:E62
Werner NE, Malkana S, Gurses AP, Leff B, Arbaje AI (2017) Toward a process-level view of distributed healthcare tasks: medication management as a case study. Appl Ergon 65:255–268
Werner NE, Tong M, Borkenhagen A, Holden RJ (2018a) Performance-shaping factors affecting older adults' hospital-to-home transition success: a systems approach. Gerontologist
Werner NE, Jolliff AF, Casper G, Martell T, Ponto K (2018b) Home is where the head is: a distributed cognition account of personal health information management in the home among those with chronic illness. Ergonomics 61(8):1065–1078
Westbrook JI, Rob MI, Woods A, Parry D (2011) Errors in the administration of intravenous medications in hospital and the role of correct procedures and nurse experience. BMJ Qual Saf 20(12):1027–1034
Wetterneck TB (2012) Error recovery in health care. In: Carayon P (ed) Handbook of human factors and ergonomics in health care and patient safety, 2nd edn. Taylor & Francis, Boca Raton, pp 763–774
Wetterneck TB, Skibinski KA, Roberts TL, Kleppin SM, Schroeder M, Enloe M et al (2006) Using failure mode and effects analysis to plan implementation of smart intravenous pump technology. Am J Health Syst Pharm 63:1528–1538
Wetterneck TB, Hundt AS, Carayon P (2009) FMEA team performance in health care: a qualitative analysis of team member perceptions. J Patient Saf 5(2):102–108
Wooldridge AR, Carayon P, Hundt AS, Hoonakker PLT (2017) SEIPS-based process modeling in primary care. Appl Ergon 60:240–254
Xie A, Carayon P (2015) A systematic review of human factors and ergonomics (HFE)-based healthcare system redesign for quality of care and patient safety. Ergonomics 58(1):33–49

Pascale Carayon, Ph.D., is Procter & Gamble Bascom Professor in Total Quality in the Department of Industrial and Systems Engineering, the Director of the Center for Quality and Productivity Improvement and the Founding Director of the Wisconsin Institute for Healthcare Systems Engineering at the University of Wisconsin-Madison. Being good at math and physics, Pascale was advised by her high-school teachers to get into the "preparatory classes," i.e., the 2-year educational program that allows French students to compete for entry to top engineering schools. She was admitted to the Ecole Centrale de Paris, the top 2 engineering school in France. The training at the Ecole Centrale de Paris was a great fit for Pascale as it allowed her to explore various disciplines, including statistics, management, and financing, which led her to career in Industrial Engineering. Pascale received her Engineer diploma from the Ecole Centrale de Paris, France, in 1984 and her Ph.D. in Industrial Engineering from the University of Wisconsin-Madison in 1988. After completing her Ph.D., she was hired as an assistant professor of Industrial Engineering at the University of Wisconsin-Madison. Between 1995 and 1999, Pascale returned to France and spent 4 years at an engineering school, the Ecole des Mines de Nancy, where she developed multiple educational and research activities on quality and productivity improvement. In 1999, she returned to Madison where she was asked to lead the Center for Quality and Productivity Improvement. In the past 15–20 years, she developed the SEIPS or Systems Engineering Initiative for Patient Safety program, which is internationally recognized for its human factors and systems engineering contribution to patient safety. She recently created the Wisconsin Institute for Healthcare Systems Engineering, which aims at transforming healthcare and achieving the quadruple aim through interdisciplinary research involving multiple engineering disciplines and the health sciences. The ability to make an impact is what drives Pascale, whether it is impact on undergraduate and graduate students, on junior faculty in engineering and health sciences, or on patients and clinicians. For almost 20 years, Pascale has dedicated her research effort to improving patient safety and healthcare quality; what a great cause and a great opportunity for industrial engineers to make a difference!

Abigail R. Wooldridge is an Assistant Professor in Industrial and Enterprise Systems Engineering at the University of Illinois at Urbana-Champaign and leader of the Human Factors in Sociotechnical Systems Laboratory. From a young age, she desired a career that would provide the opportunity to impact society positively to make the world a better place and was drawn to science/math as a possible way of doing this due to female role models in the sciences (e.g., her mother in dietetics, math and science teachers) and her own inclination. She decided to enter engineering because it was the application of science and math to solve practical problems; this seemed to fit with her goal of making the world better! By that time, she had refined her goal a bit to focus on improving healthcare for both consumers (i.e., patients and their families) and providers

(i.e., clinicians and other healthcare professionals). She settled on industrial engineering specifically because it provided the tools, methods, and theories to account for the people in the healthcare system in addition to the technical subsystem. She completed her Bachelor of Science and Master of Engineering in Industrial Engineering at the University of Louisville in 2011 and 2012, respectively, and then completed a Master of Science in Industrial and Systems Engineering at the University of Wisconsin-Madison in 2013. Between 2013 and 2015, she managed decision support and surgical scheduling at the Anne Bates Leach Eye Hospital, part of the Bascom Palmer Eye Institute at the University of Miami Miller School of Medicine, where she was able to see the impact industrial engineers can have on patients and healthcare professionals firsthand. She returned to University of Wisconsin-Madison to complete her doctoral training in Industrial and Systems Engineering at the University of Wisconsin-Madison from 2015 to 2018, where she also worked as a research assistant in the Center for Quality in Productivity Improvement. She is looking forward to continuing her research and beginning to train the future generation of industrial engineers to improve healthcare quality and patient safety.

Chapter 13
Advanced Medical Imaging Analytics in Breast Cancer Diagnosis

Yinlin Fu, Bhavika K. Patel, Teresa Wu, Jing Li, and Fei Gao

Contents

13.1 Introduction 301
13.2 Literature Review 304
13.2.1 Gray Level Co-occurrence Matrix 304
13.2.2 Local Binary Pattern 305
13.2.3 Gabor Filters Bank 306
13.2.4 Deep Learning 307
13.3 Comparison Experiments 308
13.3.1 Experiment 308
13.4 Conclusions 314
References 314

13.1 Introduction

Cancer is the leading cause of death for Americans aged 40–79 years. In 2017, more than 1.68 million new cancer cases were diagnosed in the USA. The lifetime probability of developing cancer for males is 1 out of 2 and females is 1 out of 3 (https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2017.html). Among women, breast (30%), lung (12%), and colorectal (8%) cancers are the most common. Breast cancer has the most number of new cases per population, and it has been the second leading cause of the cancer death in women (Siegel et al. 2016). But it is also one of the most treatable malignancies if detected early. According to American Cancer Society, breast cancer

Y. Fu · T. Wu (✉) · J. Li · F. Gao
School of Computing, Informatics, Decision Systems Engineering,
Ira Fulton Schools of Engineering, Arizona State University, Tempe, AZ, USA
e-mail: yinlinfu@asu.edu; Teresa.Wu@asu.edu; jing.li.8@asu.edu; fgao16@asu.edu

B. K. Patel
Division of Breast Imaging, Department of Radiology, Mayo Clinic, Phoenix, AZ, USA
e-mail: Patel.Bhavika@mayo.edu

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_13

incidence rates in females have been increasing slightly over the 10-year period from 2004 to 2013, with the trend entirely driven by rising rates among non-white women. Yet, the 5-year relative survival rates of breast cancer significantly increased from 75% (1975–1977) to 84% (1987–1989) to 91% (2006–2012). One contributing factor is detecting cancers early from the implementation of the population-based breast cancer screening mammography program in the late 1970s (Törnberg et al. 2006).

Modern imaging technologies allow for visualization of multi-dimensional and multi-parametric data. Imaging is increasingly used to measure physical parameters such as tissue properties and to glean temporal insight on biological function. Three primary imaging modalities are used to diagnose breast cancer: Digital Mammography (DM), Ultrasound (US), and Magnetic Resonance Imaging (MRI). Digital Mammography (DM) is a specialized modality for breast imaging. It uses low-dose X-rays to detect breast cancer and has been adopted as the mainstay technique in the breast cancer screening program. Ultrasound is an imaging technology that uses high-frequency sound waves to characterize tissues. It does not use harmful ionizing radiation and is substantially lower in cost. However, ultrasound suffers from a high number of false positives, leading to excessive biopsies (Tosteson et al. 2014). MRI utilizes a strong magnetic field, typically 1.5 or 3 T, for human scanners. MRI has proven to be highly useful in diagnosing many conditions by showing the differences between healthy and diseased soft tissues of the body including breast, heart, and kidney. However, MRI is costly and has limited accessibility (Breast Cancer MRI 2018).

Interpreting breast images is a challenging task for radiologists. Current breast cancer screening suffers from a relatively high false positive recall rate (i.e., $\geq$10%). On average, the chance of having a false positive result after 10-year mammogram screening is about 50–60% (https://ww5.komen.org/BreastCancer/AccuracyofMammograms.html; Hubbard et al. 2011). To address this issue, researchers attempt to develop computer-aided detection (CADe) and diagnosis (CADx) (Tan et al. 2014; do Nascimento et al. 2013; Muramatsu et al. 2016; Gao et al. 2016) schemes as a "second reader" in hopes of helping improve radiologists' performance in breast cancer diagnoses. A notable effort in CADe and CADx scheme is to screen the imaging features to identify patterns of tumor phenotype prediction (Zacharaki et al. 2009; Davnall et al. 2012; Kassner and Thornhill 2010; Skogen et al. 2013).

In medical images, "texture" has been studied as the local characteristic pattern of image intensity. Therefore, texture analysis—a general methodology of extracting multiple features on the pixel/voxel basis to evaluate the local spectral or frequency contents—has attracted great attention. Multi-parametric models using these derived texture features have been extensively studied to assist disease diagnosis and staging. Broadly speaking, texture analysis can be categorized into four methods: (1) structural methods: texture is characterized by feature primitives and their spatial arrangements. Take a round circle as an example, the primitives are the center point and the radius. Apparently, the biggest advantage of a structural method is

its simplicity. However, this is also being recognized as its limitation. As in reality, medical images often have irregular shaped objects and cannot be easily represented by structural methods. (2) Model-based methods: texture is represented by the use of generative image models (e.g., fractals). Fractal models have the property of "self-similarities," that is, any object can be decomposed into smaller copies of itself. Fractal dimensional metrics are derived to define the object. Specifically, the higher the dimensional value is, the more complicated the object is. Model-based approaches are known to require extensive computing efforts in estimating the large number of model parameters. (3) Statistical methods: spatial distributions of the intensity values from each pixel/voxel in the images are first generated. A set of second or higher order statistics from the distributions are derived as texture features. Some well-studied texture algorithms are Gray Level Co-occurrence Matrix (GLCM) and Local Binary Patterns (LBP). (4) Transform-based method: using spatial frequency properties of the pixel/voxel intensity variation, each image is first converted into a new form. For instance, wavelet transformation analyzes the frequency content of an image within different scales and frequency directions. Wavelet coefficients corresponding to the scales and directions are derived as texture features. For example, algorithms in this category include Gabor wavelet transform, Fourier transform, and S-transform.

As research on texture analysis for medical imaging continues, another emerging field of medical imaging is deep learning using the convolutional neural network (CNN). CNN is a feedforward neural network model which has been implemented in the computer vision community for decades (Lecun et al. 1998, 2015). As computing technology advances, training CNN with a large number of layers (a.k.a. deep CNN) is now possible. The applicability of deep CNN (DCNN) in imaging was first explored in the ImageNet competition (Russakovsky et al. 2014), and the big success motivated researchers to study its potentials in other applications ranging from natural language processing, image segmentation to medical imaging analysis (Tajbakhsh et al. 2016; Cha et al. 2016). The uniqueness of a DCNN is to use a large number of trainable parameters from different layers to achieve a predictive power. The trained DCNN often can then be used to extract discriminative features at varying levels of abstraction (Bar et al. 2015). Similar to texture analysis research in medical imaging, these features are used to develop predictive models. Some successes of DCNN in medical applications are reported, including chest pathology identification (Samala et al. 2016) and breast mass detection and classification (Krizhevsky et al. 2012).

In summary, CADe/CADx schemes have initially focused on texture analysis and more recently deep learning (e.g., DCNN) approaches to extract features for predictive model development. In this chapter, our interest is to conduct comparison studies to evaluate the performance of texture features vs. DCNN features using DM images for breast cancer diagnosis. We downloaded a dataset of 88 subjects (45 malignant vs. 43 benign) from the public full-field digital mammographic database INbreast (Moreira et al. 2012). Three texture algorithms and a DCNN are applied to the images to extract features. Gradient Boosted Tree Classifier is implemented

for both feature selection and classification. Based on a tenfold cross-validation, the model developed using texture features from GLCM has an accuracy of 0.82 (sensitivity = 0.87, specificity = 0.77), using texture features from LBP has an accuracy of 0.72 (sensitivity = 0.74, specificity = 0.80), using texture features from Gabor has an accuracy of 0.72 (sensitivity = 0.71, specificity = 0.72), and using the texture features from all three algorithm has an accuracy of 0.81 (sensitivity = 0.89, specificity = 0.72). Using the DCNN features, the model provides an accuracy of 0.89 (sensitivity = 0.91, specificity = 0.87). We conclude that DCNN, as an emerging field, has the potential for improving the diagnosis and can be adopted in the CADx/CADe systems.

In what follows, we provide a comprehensive literature review on the applications of texture algorithms as well as DCNN in mammographic imaging in Sect. 2. Section 3 presents the details of the comparison of the experiments followed by the conclusion in Sect. 4.

13.2 Literature Review

In this section, we first focus on three popular texture algorithms: GLCM, LBP, and Gabor filters and review their applications on DM for breast cancer diagnosis followed by a review of deep CNN.

13.2.1 Gray Level Co-occurrence Matrix

GLCM was first proposed by Haralick and Shanmugam in 1973 (Haralick et al. 1973). The co-occurrence matrix is created to extract statistical information about the pixel-pair distribution from the image with a specific spatial relationship defined by the direction (e.g., 45°, 90°) and the distance (e.g., 1-pixel separation). Given the particular relationship, the element (i, j) from the matrix indicates the number of co-occurrences of pixel values of i and j. Thirteen texture features can be derived from the matrix: angular second moment, contrast, correlation, sum of squares (variance), inverse difference moment, sum of average, sum of variance, sum of entropy, entropy, difference variance, difference entropy, and two types of information measures of correlation. For details of the features, interested readers can refer to (Haralick et al. 1973). GLCM has been extensively studied in medical imaging, and we summarize its applications on breast imaging in Table 13.1.

Table 13.1 GLCM application to breast cancer diagnosis using digital mammography

Authors	Description	Dataset
Li et al. (2017)	Texture features of histogram, GLCM, and run length matrix (RLM) are investigated to differentiate malignant vs. benign patients.	Institution dataset including 302 patients (226 malignant vs. 76 benign).
Purwadi et al. (2016)	Three feature extraction methods, GLCM, PCA, and SIFT are investigated and compared using screening mammographic images.	2796 images from Image Retrieval in Medical Applications (IRMA) dataset.
Sharma and Khanna (2014)	Features from Zernike moments of different orders are compared with GLCM for the breast cancer diagnosis.	800 images from IRMA and 200 images from DDSM.
Jona and Nagaveni (2014)	GLCM features are investigated to assess microcalcifications from breast images.	320 images(206 normal, 63 benign, and 51 malignant) are taken from MIAS.
Mascaro et al. (2009)	GLCM, Sum histogram, and LBP are studied for breast tumor segmentation.	322 mammographic images from mini-MIAS database.
Rangayyan et al. (2010)	The impacts of pixel resolution on GLCM texture features are evaluated for breast cancer diagnosis.	Institution dataset including 111 breast masses (65 benign vs. 46 malignant).

13.2.2 Local Binary Pattern

LBP is a texture descriptor introduced by Ojala et al. (2001). For each pixel in an image, the LBP algorithm compares the gray levels of its neighborhood pixels. If the neighborhood pixel has a higher value, one is assigned; otherwise, zero is assigned. After assigning 1 or 0 to all the neighborhood pixels, a vector consisting of binary values is generated. Taking a 3×3 patch from an image as an example, the central pixel has eight neighborhood pixels, and thus the LBP will generate a vector with eight digits. Each pixel from the image will create such a vector. A histogram of 256 bins is then derived from the vectors to describe the image properties. As a guideline, most clinical research uses 10–20 bins depending on the dataset as well as the computational power of the systems used to run the algorithms. The patch size can be extended, e.g., 5×5 or more. If the neighborhood size is 2, the surrounding pixels will be 24 pixels. Thus, the number of bins and the bin sizes should be adjusted accordingly. In Table 13.2, we summarize some research on applying LBP for breast cancer study.

Table 13.2 LBP application to breast cancer diagnosis using digital mammography

Authors	Description	Dataset
Kashyap et al. (2018)	LBP features as one type of texture features are extracted to classify the suspicious regions as abnormal vs. normal.	332 samples from MIAS and 500 sample from DDSM datasets
Kashyap et al. (2017)	Features are extracted by binarized statistical image features (BSIF) and LBP for predictive model development.	332 samples from mini-MIAS and 300 randomly chosen samples from DDSM database.
Reyad et al. (2014)	Features extracted using LBP, statistical measures, multi-resolution frameworks, and contourlet transform are used to develop SVM classifier.	512 images (256 normal + 256 malignant) from DDSM database.
Gargouri et al. (2012)	Based on LBP, a new local pattern model named gray level and local differences (GLLD) based on LBP feature are developed for cancer diagnosis.	1000 ROIs (note one image may have multiple ROIs) from DDSM database.
Choi and Ro (2012)	Multi-resolution LBP features based on LBP are derived for classification model development.	89 single view mammograms from MIAS database and 303 from DDSM database.
Mascaro et al. (2009)	Features from the Sum Histogram, GLCM, and LBP are studied. A modification of LBP is proposed for a better distinction of the tissues.	322 mammographic images from mini-MIAS database.
Lladó et al. (2009)	The basic LBP histogram is extended to a spatially enhanced histogram which encodes both the local region appearance and the spatial structure of the masses.	1792 ROIs extracted from the DDSM database.
Oliver et al. (2007)	Features from LBP are used to represent salient micro-patterns and preserve at the same time the spatial structure of the masses. The features are then used for SVM model development.	1792 ROIs extracted from DDSM database.

13.2.3 Gabor Filters Bank

Both GLCM and LBP reviewed above are statistical methods. Gabor filters provide a transformation-based method. It considers an image in a space whose coordinate system is related to texture characteristics, such as frequency content or spatial resolution. Some example transform-based methods are Gabor filters, Fourier,

Table 13.3 Gabor filter bank application to breast cancer diagnosis using digital mammography

Authors	Descriptions	Dataset
Abbey et al. (2012)	The response of a Gabor filter at locations in the breast interior is used to quantify and characterize non-Gaussian statistics in breast images.	Institute dataset of 26 women with positive mammographic findings and biopsy verification of disease status.
Malar et al. (2012)	Gabor features and GLCM features are compared with proposed wavelet-based texture features for microcalcification detection.	120 ROIs are extracted from 55 mammogram images from MIAS database.
Ferrari et al. (2004)	Convolving a group of Gabor filters to enhance the identification of the pectoral muscle from MLO view of DM images.	84 MLO mammograms from the mini-MIAS database.
Ferrari et al. (2001)	Using a multi-resolution representation based upon Gabor wavelets to detect left-right asymmetry as a way for cancer diagnosis.	80 images from 20 normal cases, 14 asymmetric cases, and 6 architectural distortion cases from mini-MIAS database.

and S-transform, among which Gabor filters are known to provide better spatial localization (Larroza et al. 2016) and have proven advantages in edge detection and texture segmentation problems (Jain et al. 1997). Often, Gabor Filters Bank—a set of predefined Gabor filters with the combination of frequencies and orientations of the sinusoidal wave and the variances of the Gaussian kernel is constructed to address the localization issue (Materka 2004). For a sinusoidal wave in a 2D space, the direction parameter describes the angle of the wave function and four commonly used directions are: 0°, 45°, 90°, and 135°. The frequency parameter decides the location of a series of peaks on the sinusoidal wave, and these peaks are the band passes position on the wave and should be selected based on the size of the image as well as the potential texture embedded in the images. Taking Gaussian kernel function in Gabor filters as an example, the variance is another feature extracted as it will affect the width of the band pass. Gabor filters have also been studied for medical images for disease diagnosis. We have its application to breast cancer summarized in Table 13.3.

13.2.4 Deep Learning

Deep learning is a type of machine learning algorithm inspired by the structure and function of brain called artificial neural network to train a computer to perform human-like tasks. Deep learning sets up parameters for the network model and studies patterns using multi-layered processing. The emergence of deep learning

brings a new paradigm for machine learning and thus has attracted significant attention. As seen in our literature review (Table 13.4), an extensive research investigates the application of deep learning on medical imaging in the last 2 years (2016–2018). There are many interesting problems deep learning can tackle and feature extraction is of interest in this chapter. Deep learning has revolutionized feature extraction as the features from low to deep layers can be extracted describing different imaging properties. Some details of the literature on deep learning on breast imaging are shown in Table 13.4.

As reviewed, enormous efforts are invested to study the imaging features for disease diagnosis in the CADe/CADx schema development. The recent trend on deep learning appears to be strong. In this chapter, we are interested in comparing the performance of traditional texture feature extraction algorithms with that of deep learning in this arena.

13.3 Comparison Experiments

Dataset This dataset is obtained from the online accessible full-field digital mammographic database named INbreast. The database was acquired by researchers at the Breast Centre in CHJKS, Porto, under permission of both the Hospital's Ethics Committee and the National Committee of Data Protection. The images in the database were acquired with the system MammoNovation Siemens FFDM. Images are with the pixel size of 70 mm (microns), and 14-bit contrast resolution. For each subject, both craniocaudal (CC) and mediolateral oblique (MLO) views are available. For each image, the annotations of the region of interests (ROIs) were made by a specialist in the field, and validated by a second specialist. The masks of ROIs were also available. In this study, a dataset of 88 subjects is extracted from the database by including all subjects that have Breast Imaging Reporting and Data System (BIRADS) (American College of Radiology 1998) scores of 1, 2, 5, and 6. Subjects with BIRADS 1 and 2 are regarded as benign, and subjects with BIRADS 5 and 6 are regarded as the suspicious or malignant tumor. As an initial attempt, only CC view images are used in the comparison experiment. Some example images for benign and malignant (cancer) cases are shown in Fig. 13.1.

13.3.1 Experiment

Before feature generation, an imaging pre-processing procedure is launched. First, for each image, we identify a minimum-area bounding box that contains the tumor region. The bounding box size varies case by case due to different sizes of tumors (ranging from 65×79 to 514×457 in this study). This rectangle is extracted and saved as one image. The second step is to normalize the image intensity to be between 0 and 1 using the min-max normalization. The normalized images are then

Table 13.4 Deep learning application to breast cancer diagnosis using digital mammography

Authors	Descriptions	Dataset
Li et al. (2018)	Develop a supervised learning approach for automated estimation of percentage density (PD) on digital mammograms (DMs). The conclusion is that DCNN approach is significantly better and more robust than feature-based learning approach for automated PD estimation on DMs.	661 DMs with only CC views being used. 478 DMs were used for training and 92 DMs were used for blind testing.
Mohamed et al. (2018a)	Investigate a deep learning-based breast density classifier to consistently distinguish two categories: "scattered density" and "heterogeneously dense."	Institution dataset with 22,000 DMs from 1427 women (2005–2016).
Samala et al. (2017)	Investigate the training of DCNN for computer-aided classification of malignant and benign masses on mammograms with transfer learning from nonmedical images (digitized-screen film mammography).	Institution dataset: 1655 SFM views and 310 DM views 277 SFM views from DDSM.
Kooi and Karssemeijer (2017)	Investigate the addition of symmetry and temporal context information to a DCNN to detect malignant soft tissue lesions in mammography.	Public dataset of 18366 patients with one or more exams.
Mohamed et al. (2018b)	Develop DCNN for BIRADS density categorization. The results show that the classification of density categories using MLO view image is significantly higher than that using the CC view.	Public dataset 963 women (2005–2016).
Carneiro et al. (2017)	Use DCNN to classify normal, benign, and malignant.	INbreast, DDSM, and Imagenet.
Antropova et al. (2017)	Develops a methodology that extracts and pools low-to mid-level features using a pre-trained CNN and fuses them with hand-crafted radiomic features computed using conventional CADx methods.	Institution dataset: 690 dynamic contrast-enhanced MRI, 245 DM, 1125 US.

(continued)

Table 13.4 (continued)

Authors	Descriptions	Dataset
Teare et al. (2017)	Utilize dual deep CNN at different scales for classification of full mammogram images.	6000 DMs images from DDSM and Zebra Mammography Dataset (ZMDS).
Becker et al. (2017)	Evaluate the diagnostic accuracy of deep learning for the detection of breast cancer in an independent, dual-center mammography dataset. The results show that the performance of using DCNN was not significantly different from the human performance.	Institute dataset of 143 patients.
Jadoon et al. (2017)	The paper proposes a novel classification technique for large dataset of mammograms using deep learning method. The method targets a three-class classification problem (normal, malignant, and benign classes).	Institute dataset.
Kooi et al. (2017a)	The paper studies a CNN to classify cyst and mass patches. It shows good performance can be obtained on a small dataset by pre-training the network on a large dataset of a related task.	A large collection of screening mammograms obtained from the Netherlands.
Samala et al. (2016)	The paper develops a CAD system for masses in digital breast tomosynthesis(DBT) volume using a deep CNN with transfer learning from mammograms.	2282 digitized film and DMs are 324 DBT volumes were collected.
Huynh et al. (2016)	To alleviate the need of large datasets, the study uses the transfer learning technique to extract tumor information from medical images via CNN originally pre-trained for nonmedical tasks. The conclusion is transfer learning can improve current CADx methods while also providing standalone classifiers without large datasets.	Institution dataset: 219 breast lesions (607 ROIs with 261 benign and 346 malignant).
Kooi et al. (2017b)	The paper provides a head-to-head comparison between a CNN and a CAD system relying on an exhaustive set of manually designed features and shows the CNN outperforms a state-of-the-art mammography CAD system.	About 45,000 images collected from a large-scale screening program in the Netherlands.

(continued)

Table 13.4 (continued)

Authors	Descriptions	Dataset
Wang et al. (2016)	The study evaluates the performance of deep learning-based models on large datasets on microcalcification.	Institution dataset. The training dataset includes 1000 images (677 benign vs. 323 malignant). The test group has 204 images (97 benign vs. 107 malignant).
Kallenberg et al. (2016)	The main idea of the model is to learn a deep hierarchy of increasingly more abstract features from unlabeled data. Once the features have been learned, a classifier is trained to map the features to the labels of interest.	3 datasets: (1) 493 images (healthy) from Dutch breast cancer screening program; (2) 668 images from Mayo Clinic; (3) 394 cancers and 1182 healthy controls from Dutch breast cancer screening program.
Arevalo et al. (2016)	The paper integrates deep learning techniques to automatically learn discriminative features avoiding the design of specific hand-crafted image-based feature detectors. The results show that deep learning method improved the performance significantly compared to state-of-the-art image descriptors.	Institution dataset: 344 patients with 736 images containing 426 benign and 310 malignant lesions.

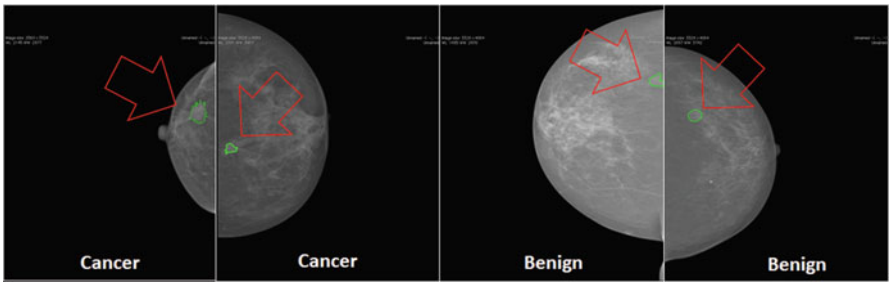


Fig. 13.1 Sample images from INbreast database (CC View image only)

fed into the texture algorithms (GLCM, LBP, and Gabor) for feature exaction. For DCNN feature extraction, one more step is needed. Specifically, the normalized images are resized to 224 × 224 to take advantage of pre-trained DCNN (in this study, ResNet). Note that as DCNN evolves, different network models have been developed, including AlexNet, GoogLeNet, and VGGNet, just to name a few. The known issue of these models is the gradient vanishing when the number of layers increases significantly. ResNet is chosen in this research as it has a "short-cut" architecture to address this issue and has outperformed the earlier DCNN

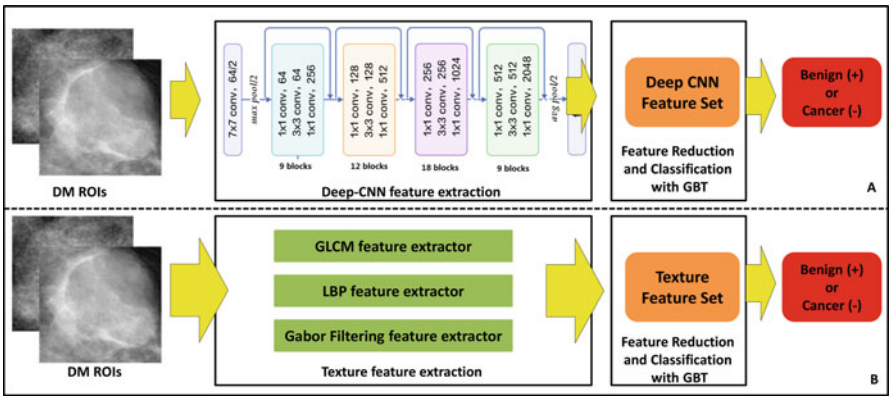


Fig. 13.2 Architecture of experiment

models (AlexNet, GoogLeNet, and VGGNet) in the recent ImageNet competition. Figure 13.2 illustrates the workflow of the comparison experiments.

For DCNN feature generation (Fig. 13.2a), the extracted ROIs are fed into the pre-trained ResNet and feature maps of the last layer are retained. For each feature map, a numerical value is obtained by taking the average of means. This pipeline extracts 1024 DCNN features for each subject. For texture feature generation (Fig. 13.2b), the ROIs are fed into an in-house developed texture pipeline (Kooi et al. 2017b), and a total of 89 texture features are extracted. Specifically, for GLCM algorithm, we extract 65 features (13 features for each direction: 0°, 45°, 90°, 135°, and average). For LBP, we extract 10 features using 10 as bin size. For Gabor filtering, we extract 12 features. The parameter setting for each texture algorithm is shown in Table 13.5.

Once the features are derived, we implement Gradient Boosted Tree (GBT) for feature selection first (threshold = 0.01). A second GBT is implemented on the reduced feature set to classify the case as cancer vs. benign. The procedures are implemented with a python library named "sklearn." Model settings such as loss function, learning rate, and max depth are the default. Different settings to prevent the model from overfitting are used, for instance: we set maximum depth of individual trees to be 3 and use early stopping strategy. The experimental results based on tenfold cross-validation are shown in Table 13.6.

From Table 13.6, we observe that among the three texture algorithms, GBT model on GLCM features performs the best (0.82) while the models using LBP and Gabor features both have an accuracy of 0.72. Note that 27 out of 69 GLCM, 9 out of 10 LBP, and 13 out of 14 Gabor features are used for GBT classifier. When all 89 features are used jointly, 21 GLCM, 3 LBP, and 3 Gabor features are selected for the classification with the model performance being 0.81. We conclude GLCM features, in general, contribute more to the classification with better performances than the other texture feature algorithms. Our second observation is models using all three texture features, jointly or individually, have low performance in specificity

Table 13.5 Parameter setting for texture analysis algorithms

Algorithm	Parameter	Description	Guideline for selection
GLCM	Direction	In which direction the co-occurrence should be considered	[0, 45, 90, 135]
	Distance	How far between two points the co-occurrence should be considered	1
LBP	Distance	Radius of the neighborhood	1, 3
	Number of surrounding points	Number of points that should be selected on the circle of neighborhood.	8, 24
	Number of bins of histogram	The granularity of the texture spectrum	10
Gabor	Frequency	The frequency of the sinusoidal wave	Depends on the properties of images (0.1, 0.3, and 0.5 are used in our case)
	Direction	The direction of the sinusoidal wave	[0, 45, 90, 135]
	Variance	The variance of the Gaussian kernel	Depends on the properties of images (1 and 2 are used)

Table 13.6 Comparison results

	Classifier with GLCM features	Classifier with LBP features	Classifier with Gabor features	Classifier with all (GLCM + LBP + Gabor) texture features	Classifier with deep CNN features
Accuracy	0.82	0.72	0.72	0.81	0.89
Sensitivity	0.87	0.74	0.71	0.89	0.91
Specificity	0.77	0.70	0.72	0.73	0.87
Feature number after reduction	27	9	13	27 (21 GLCM + 3 LBP + 3 Gabor)	15

which has been recognized as the clinical challenge in breast imaging. In looking at DCNN-based classification modeling, among 1024 DCNN features, 15 features are selected and the classifier has an overall accuracy of 0.89 with the sensitivity being 0.91 and specificity being 0.87. The performance on specificity is promising which deserves more in-depth investigation. The general conclusion we could draw from this comparison experiment is DCNN has the great potential over traditional texture algorithms in extracting imaging feature for cancer diagnosis.

13.4 Conclusions

The importance of imaging in disease screening, early diagnostics, and treatment assessments has been long recognized. As imaging technology advances, in parallel, researchers from computing and informatics are keeping the pace in developing advanced CADe/CADx systems to process, analyze big imaging data to discover imaging biomarkers to assist disease diagnosis. In this chapter, we comprehensively review the applications of three well-adopted texture feature algorithms: GLCM, LBP, and Gabor on breast cancer diagnosis. Noting the strong wave in deep learning, we also review the application of deep learning in breast cancer studies. The comparison of the texture algorithm vs. deep learning is conducted using a publicly available Digital Mammography dataset. While our initial experiments show deep learning has advantages in term of accuracy, sensitivity, and specificity over traditional texture algorithm, we intend to assess the robustness of the models as one of the future tasks.

References

Abbey CK, Nosrateih A, Sohl-Dickstein J, Yang K, Boone JM (2012) Non-Gaussian statistical properties of breast images. Med Phys 39(11):7121–7130

American College of Radiology (1998) BI-RADS Committee, "breast imaging reporting and data system". Radiol Clin North Am 40:409–430

Antropova N, Huynh BQ, Giger ML (2017) A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. Med Phys 44(10):5162–5171

Arevalo J, González FA, Ramos-Pollán R, Oliveira JL, Guevara Lopez MA (2016) Representation learning for mammography mass lesion classification with convolutional neural networks. Comput Methods Prog Biomed 127:248–257

Bar Y, Diamant I, Wolf L, Greenspan H (2015) Deep learning with non-medical training used for chest pathology identification. In: Medical imaging 2015: computer-aided diagnosis, vol 9414, p 94140V. International Society for Optics and Photonics

Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A (2017) Deep learning in mammography diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. Investig Radiol 52(7):434–440

Breast Cancer MRI—magnetic resonance imaging | MRI Scan | Imaginis—the women's health & wellness resource network. [Online]. http://www.imaginis.com/mri-scan/magnetic-resonance-breast-imaging-mri-mr-2. Accessed 23 May 2018

Cancer Facts & Figures 2017 (2017) [Online]. https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2017.html. Accessed 31 Jan 2018

Carneiro G, Nascimento J, Bradley AP (2017) Automated analysis of unregistered multi-view mammograms with deep learning. IEEE Trans Med Imaging 36(11):2355–2365

Cha KH, Hadjiiski L, Samala RK, Chan H-P, Caoili EM, Cohan RH (2016) Urinary bladder segmentation in CT urography using deep-learning convolutional neural network and level sets. Med Phys 43(4):1882–1896

Choi JY, Ro YM (2012) Multiresolution local binary pattern texture analysis combined with variable selection for application to false-positive reduction in computer-aided detection of breast masses on mammograms. Phys Med Biol 57(21):7029–7052

Davnall F et al (2012) Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice? Insights Imaging 3(6):573–589

do Nascimento MZ, Martins AS, Neves LA, Ramos RP, Flores EL, Carrijo GA (2013) Classification of masses in mammographic image using wavelet domain features and polynomial classifier. Expert Syst Appl 40(15):6213–6221

Ferrari RJ, Rangayyan RM, Desautels JEL, Frère AF (2001) Analysis of asymmetry in mammograms via directional filtering with Gabor wavelets. IEEE Trans Med Imaging 20(9):953–964

Ferrari RJ, Rangayyan RM, Desautels JEL, Borges RA, Frère AF (2004) Automatic identification of the pectoral muscle in mammograms. IEEE Trans Med Imaging 23(2):232–245

Gao F, Zhang M, Wu T, Bennett KM (2016) 3D small structure detection in medical image using texture analysis. In: 2016 38th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp 6433–6436

Gargouri N, Dammak Masmoudi A, Sellami Masmoudi D, Abid R (2012) A new GLLD operator for mass detection in digital mammograms. Int J Biomed Imaging 2012:765649

Haralick RM, Shanmugam K, Dinstein I (1973) Textural features for image classification. IEEE Trans Syst Man Cybernet SMC-3(6):610–621

Hubbard RA, Kerlikowske K, Flowers CI, Yankaskas BC, Zhu W, Miglioretti DL (2011) Cumulative probability of false-positive recall or biopsy recommendation after 10 years of screening mammography. Ann Intern Med 155(8):481

Huynh BQ, Li H, Giger ML (2016) Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. J Med Imaging 3(3):34501

Jadoon MM, Zhang Q, Haq IU, Butt S, Jadoon A (2017) Three-class mammogram classification based on descriptive CNN features. Biomed Res Int 2017:3640901

Jain AK, Ratha NK, Lakshmanan S (1997) Object detection using gabor filters. Pattern Recogn 30(2):295–309

Jona JB, Nagaveni N (2014) Ant-cuckoo colony optimization for feature selection in digital mammogram. Pak J Biol Sci 17(2):266–271

Kallenberg M et al (2016) Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. IEEE Trans Med Imaging 35(5):1322–1331

Kashyap KL, Bajpai MK, Khanna P (2017) Globally supported radial basis function based collocation method for evolution of level set in mass segmentation using mammograms. Comput Biol Med 87:22–37

Kashyap KL, Bajpai MK, Khanna P, Giakos G (2018) Mesh-free based variational level set evolution for breast region segmentation and abnormality detection using mammograms. Int J Numer Method Biomed Eng 34(1):e2907

Kassner A, Thornhill RE (2010) Texture analysis: a review of neurologic MR imaging applications. Am J Neuroradiol 31(5):809–816

Kooi T, Karssemeijer N (2017) Classifying symmetrical differences and temporal change in mammography using deep neural networks. J Med Imaging (Bellingham) 4(4):044501

Kooi T, van Ginneken B, Karssemeijer N, den Heeten A (2017a) Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network. Med Phys 44(3):1017–1027

Kooi T et al (2017b) Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal 35:303–312

Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th international conference on neural information processing systems, vol 1. Curran Associates Inc., pp 1097–1105

Larroza A, Bodí V, Moratal D (2016) Texture analysis in magnetic resonance imaging: review and considerations for future applications. In: Assessment of cellular and organ function and dysfunction using direct and derived mri methodologies. InTech

Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324

Lecun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444

Li Z et al (2017) Diagnostic performance of mammographic texture analysis in the differential diagnosis of benign and malignant breast tumors. Clin Breast Cancer 18(4):e621–e627
Li S et al (2018) Computer-aided assessment of breast density: comparison of supervised deep learning and feature-based statistical learning. Phys Med Biol 63(2):25005
Lladó X, Oliver A, Freixenet J, Martí R, Martí J (2009) A textural approach for mass false positive reduction in mammography. Comput Med Imaging Graph 33(6):415–422
Malar E, Kandaswamy A, Chakravarthy D, Giri Dharan A (2012) A novel approach for detection and classification of mammographic microcalcifications using wavelet analysis and extreme learning machine. Comput Biol Med 42(9):898–905
Mascaro AA, Mello CAB, Santos WP, Cavalcanti GDC (2009) Mammographic images segmentation using texture descriptors. In: Proceedings of the 31st annual international conference of the IEEE engineering in medicine and biology society: engineering the future of biomedicine, EMBC 2009, pp 3653–3656
Materka A (2004) Texture analysis methodologies for magnetic resonance imaging. Dialog Clin Neurosci 6(2):243–250
Mohamed AA, Berg WA, Peng H, Luo Y, Jankowitz RC, Wu S (2018a) A deep learning method for classifying mammographic breast density categories. Med Phys 45(1):314–321
Mohamed AA, Luo Y, Peng H, Jankowitz RC, Wu S (2018b) Understanding clinical mammographic breast density assessment: a deep learning perspective. J Digit Imaging 31(4):387–392
Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS (2012) INbreast: toward a full-field digital mammographic database. Acad Radiol 19(2):236–248
Muramatsu C, Hara T, Endo T, Fujita H (2016) Breast mass classification on mammograms using radial local ternary patterns. Comput Biol Med 72:43–53
Ojala T, Pietikäinen M, Mäenpää T (2001) A generalized local binary pattern operator for multiresolution gray scale and rotation invariant texture classification. Springer, Berlin, pp 399–408
Oliver A, Lladó X, Freixenet J, Martí J (2007) False positive reduction in mammographic mass detection using local binary patterns. In: Medical image computing and computer-assisted intervention : MICCAI ... international conference on medical image computing and computer-assisted intervention, vol 10, no. Pt 1, pp 286–93
Purwadi NS, Atay HT, Kurt KK, Turkeli S (2016) Assessment of content-based image retrieval approaches for mammography based on breast density patterns. Stud Health Technol Inf 228:727–731
Rangayyan RM, Nguyen TM, Ayres FJ, Nandi AK (2010) Effect of pixel resolution on texture features of breast masses in mammograms. J Digit Imaging 23(5):547–553
Reyad YA, Berbar MA, Hussain M (2014) Comparison of statistical, LBP, and multi-resolution analysis features for breast mass classification. J Med Syst 38(9):100
Russakovsky O et al (2014) International journal of computer vision. Kluwer Academic, Boston
Samala RK, Chan H-P, Hadjiiski L, Helvie MA, Wei J, Cha K (2016) Mass detection in digital breast tomosynthesis: deep convolutional neural network with transfer learning from mammography. Med Phys 43(12):6654–6666
Samala RK, Chan H-P, Hadjiiski LM, Helvie MA, Cha K, Richter C (2017) Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms. Phys Med Biol 62(23):8894–8908
Sharma S, Khanna P (2014) Computer-aided diagnosis of malignant mammograms using Zernike moments and SVM. J Digit Imaging 28(1):77–90
Siegel RL, Miller KD, Jemal A (2016) Cancer statistics, 2016. CA Cancer J Clin 66(1):7–30
Skogen K, Ganeshan B, Good C, Critchley G, Miles K (2013) Measurements of heterogeneity in gliomas on computed tomography relationship to tumour grade. J Neurooncol 111(2):213–219
Tajbakhsh N et al (2016) Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans Med Imaging 35(5):1299–1312
Tan M, Pu J, Zheng B (2014) Reduction of false-positive recalls using a computerized mammographic image feature analysis scheme. Phys Med Biol 59(15):4357–4373

Teare P, Fishman M, Benzaquen O, Toledano E, Elnekave E (2017) Malignancy detection on mammography using dual deep convolutional neural networks and genetically discovered false color input enhancement. J Digit Imaging 30(4):499–505

The Susan G. Komen Breast Cancer Foundation, "Mammogram accuracy—accuracy of mammograms | Susan G. Komen®," *Susan G. Komen*, 2017. [Online]. https://ww5.komen.org/BreastCancer/AccuracyofMammograms.html. Accessed 5 Feb 2018

Törnberg S et al (2006) Breast cancer incidence and mortality in the Nordic capitals, 1970-1998. Trends related to mammography screening programmes. Acta Oncol 45(5):528–535

Tosteson ANA et al (2014) Consequences of false-positive screening mammograms. JAMA Intern Med 174(6):954–961

Wang J, Yang X, Cai H, Tan W, Jin C, Li L (2016) Discrimination of breast cancer with microcalcifications on mammography by deep learning. Sci Rep 6:1–9

Zacharaki EI et al (2009) Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. Magn Reson Med 62(6):1609–1618

Yinlin Fu is a Ph.D. student supervised by Professor Teresa Wu in School of Computing, Informatics, Decision Systems Engineering at Arizona State University. She received her B.S. in Statistics from USTC, China, in 2013. Her research interest is in statistical modeling, feature selection in unsupervised learning, and computer-aided diagnosis in the healthcare application.

Yinlin Fu began to show an interest in STEM fields since middle school. She liked mathematics and believed that engineering could help to solve significant problems to develop new technologies, and even set the tone for the future. During the 4 years in college, she received a rigorous education in advanced mathematics and statistics. Then in the Ph.D. program, she gained much engineering training under the supervision of Teresa Wu. After being involved in many healthcare-related projects, she learned how to apply her mathematical and engineering skill sets to solve the real-world problem. Notably, she has a great passion for solving healthcare problems such as disease diagnosis and hospital's staffing justification from the aspect of data analysis and statistical modeling.

Bhavika K. Patel, M.D., completed Radiology residency and a Breast Imaging fellowship from Emory University in Atlanta, GA. She is currently working as a Senior Associate Consultant and Assistant Professor at Mayo Clinic, Arizona. Research interests include: (1) advancing breast imaging technologies to improve sensitivity and specificity of breast cancer detection, (2) exploring personalized medicine to better detect early and pre-cancerous states, and (3) utilizing computer aided detection tools to better enable prediction and assessment of breast cancer treatment response.

Dr. Patel was interested in being a doctor since the ripe age of 4. She loved the problem-solving, patient interaction and desired to make a difference in medicine field. In addition, she was naturally interested in learning about science and math. After starting medical school, she appreciated how radiology was the main center point of nearly every patient interaction and specifically, enjoyed the problem-solving in diagnostic radiol-

ogy. Completing her breast imaging fellowship, she becomes passionate about improving the accuracy of the current state of breast imaging studies using advanced imaging analytics. Dr. Patel believes that combining advanced imaging tools with the current state of radiology imaging will lead to a better and earlier detection of clinically relevant cancers as well as decreased anxiety and costs caused by false positive findings.

Teresa Wu is a Professor of Industrial Engineering Program, School of Computing, Informatics, Decision Systems Engineering at Arizona State University. She received her Ph.D. in Industrial Engineering from the University of Iowa in 2001. Her current research interests include: swarm intelligence, distributed decision support, distributed information system, health informatics. Professor Wu has published more than 80 journal articles such as *IEEE Transactions on Evolutionary Computation, IEEE Transactions on Pattern Analysis and Machine Intelligence, Information Science*. She is currently serving as the editor-in-chief for IIE Transactions on healthcare systems engineering.

Dr. Wu was born in an engineering family. When she was in elementary school, she was highly motivated by Dr. Yu Wei's achievements, the first female to do Ph.D. in engineering in China. Since then, Dr. Wu got attracted to engineering and later decided to take engineering as her career path. Dr. Wu did majors in mechanical engineering and started her first job as a system administrator for the J8 Flight Simulator. This working experience got Dr. Wu fascinated with informatics-focused research and she started her Ph.D. study on information system from industrial engineering afterward. Since 2008, Dr. Wu began the collaboration with Mayo Clinic with a special focus on the informatics applications in radiology which includes operational workflow improvement, and medical decision-making for disease diagnosis and treatment assessment. This chapter is one example of imaging informatics applications to breast cancer diagnosis.

Jing Li is an Associate Professor in Industrial Engineering at Arizona State University. She received her B.S. from Tsinghua University in China and an M.A. in Statistics and a Ph.D. in Industrial and Operations Engineering from the University of Michigan in 2005 and 2007, respectively. Her research interests are applied statistics, data mining, and quality and systems engineering. She is a founding co-director for an ASU-Mayo Clinic Imaging Informatics Lab. She is a recipient of NSF CAREER award. She is a member of IIE, INFORMS, and ASQ.

Dr. Li developed the keen interest of integrating mathematical/statistical models and engineering principles for problem-solving in college. During the time of getting her Ph.D. in Michigan, she got rigorous training in statistical modelling skills and engineering logic thinking. Upon graduation, she got particularly passionate by using her skill sets to support decision-making in healthcare, in particular in the area of

cancer diagnosis by leveraging and integrating multi-source heterogeneous data including diagnostic imaging. She has been working on collaborative research projects intersecting between data-driven statistical modelling, health systems informatics, and imaging-based clinical decision-making. She is hoping to bring computerized intelligence driven by the availability of big data in healthcare to further the diagnosis, treatment, and care of patients with cancer and other challenging diseases.

Fei Gao was born in Hebei, China, in 1990. He received his B.S. and M.S degree in Computer Science from BUAA, China, in 2013 and Arizona State University in 2015, respectively. He is currently pursuing the Ph.D. degree in Industrial Engineering under the instruction of Prof. Teresa Wu. His research interests include image processing, machine learning, and deep learning.

Chapter 14
Decision-Making in Sequential Adaptive Clinical Trials, with Implications for Drug Misclassification and Resource Allocation

Alba C. Rojas-Cordova, Ebru K. Bish, and Niyousha Hosseinichimeh

Contents

14.1 Introduction 321
14.1.1 The Relevance of Pharmaceutical R&D 321
14.1.2 Clinical Trial Design: Traditional Versus Sequential Adaptive Trials 323
14.2 Background and Overview of Current Research 325
14.2.1 Background on Sequential Adaptive Trials 325
14.2.2 Stopping Rules for Sequential Trials with Binary Response Under the Triangular Test 327
14.2.3 Overview of Research on Resource Allocation in the Pharmaceutical Industry 330
14.3 Drug Misclassification Risk and Optimal Resource Allocation in Sequential Adaptive Clinical Trials 332
14.3.1 Quantifying the Likelihood of Drug Misclassification in Current Practices 332
14.3.2 Optimal Resource Allocation and Trial Termination Policies 334
14.4 Conclusions 339
References 341

14.1 Introduction

14.1.1 The Relevance of Pharmaceutical R&D

The output of pharmaceutical research and development (R&D) has the potential to impact the quality of human life. The detrimental effects of severe diseases such as cancer, hepatitis C, and AIDS have been significantly mitigated, thanks to new medications and therapies. As reported by the Pharmaceutical Research

A. C. Rojas-Cordova (✉)
Department of EMIS, Southern Methodist University, Dallas, TX, USA
e-mail: alba@smu.edu

E. K. Bish · N. Hosseinichimeh
Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA, USA

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_14

and Manufacturers of America (Pharmaceutical Research and Manufacturers of America 2016a), cancer fatality rates have declined 23% since their peak in the 1990s (National Cancer Institute 2014), with approximately 83% of survival gains attributable to new treatments and new medications (Sun et al. 2008). Unlike a few years ago, a wide range of treatment options are now available to cure hepatitis C, with cure rates of more than 90% over a course of treatment as short as 8 weeks (Pharmaceutical Research and Manufacturers of America 2016b). The battle against HIV/AIDS has also been successful, with a fatality rate that declined 87% (National Center for Health Statistics 2015) and 862,000 premature deaths avoided in the United States (US) alone, since the introduction of highly active antiretroviral treatment (HAART) (Lacey et al. 2014).

Pharmaceutical R&D also represents a significant component of the US and the worldwide economy. Ding et al. (2013) state that pharmaceutical R&D holds a share of 19% of all business spending on R&D worldwide, and that the US alone finances about 36% of the global expenses in pharmaceutical R&D. Despite these high expenditures, the rate of success in drug discovery remains steadily low. The development of new drugs is a complex, lengthy, and costly process, and is surrounded by high levels of uncertainty. In addition to the inherent uncertainty in the outcome of clinical testing, the highly regulated environment in which the new drug development process takes place amplifies the firms' risk of incurring unrecoverable expenses, as the ultimate decision to commercialize the drug in the US depends entirely on the Food and Drug Administration (FDA). These and other elements set drug development apart from the innovation processes in other technology-intensive industries.

DiMasi et al. (2016) have recently estimated the time between the start of clinical testing and submission of a new drug application to the FDA to be around 80.8 months, or slightly short of 7 years, with clinical trials accounting for the majority of this time. This same report indicates that the estimated out-of-pocket cost of a new drug receiving FDA approval was $1.39 billion in 2013, representing a 166% increase over the cost of drugs approved in the 1990s, which were reported in DiMasi et al. (2003). In addition, the capitalized cost estimate of an approved drug, which includes the research and development costs of those drugs that failed to get approval, rose to $2.56 billion—145% higher than the capitalized cost of drugs approved in the 1990s (DiMasi et al. 2016). The authors also state that the overall change in the risk profile for new drug development accounted directly for a 47% increase in costs. In the early 2000s, the overall likelihood that a drug that is clinically tested will eventually be approved was estimated to be 21.50%. This probability has dropped down to 11.83% in 2016, which translates into a higher risk of developing drugs that will not receive regulatory approval, and whose development costs may not be recovered by the trial sponsor (DiMasi et al. 2016).

In 2014, the R&D spending of US-based pharmaceutical companies reached $53.3 billion (Pharmaceutical Research and Manufacturers of America 2016a), with clinical trials representing at least 50% of the total R&D cost (Ding et al. 2013; DiMasi et al. 2003; Halliday et al. 1997). Clinical trials typically consist of three

phases, and the increasing costs of Phase 2 and 3 trials, along with late-stage failures (i.e., drugs that undergo all three clinical trial phases and ultimately fail due to safety issues or lack of efficacy), are the main components of R&D expenditures (Paul et al. 2010; David et al. 2009; Orloff et al. 2009), and have motivated academicians and practitioners to devise new methodologies and change business practices in order to reduce costs and investment risks during these phases.

Throughout, we use the terms "drug" and "therapy" interchangeably. In this chapter, our focus is on Phase 3 trials with binary response, that is, the patient's response to a given therapy can take one of two possible values, which are generically referred to as *success* and *failure* (Whitehead 1997). For instance, acute bleeding may be controlled within a specific time frame or not, the patient may survive after receiving a certain therapy or not, the blood pressure may fall to a specified level or not. In Phase 3, the candidate drug is tested within a large group of patients to establish its safety and efficacy; and this constitutes the longest and most resource-intensive phase of a clinical trial (Ding et al. 2013). We consider trial designs in which the candidate drug is tested comparatively against a control (i.e., current standard treatment or placebo) to understand the risk–benefit tradeoff associated with its administration.

14.1.2 Clinical Trial Design: Traditional Versus Sequential Adaptive Trials

We first discuss traditional trial designs, also known as "fixed sample size" trials, in which the number of patients to be enrolled in the trial is determined a priori, based on the desired statistical significance level and power, as well as the drug's characteristics and the treatment advantage to be observed. Then, the trial is conducted and patient responses are collected to determine, via statistical analyses, whether or not the candidate drug is superior to the control. From that perspective, the design in the traditional, fixed sample size trials is rigid, because the Phase 3 trial needs to be brought to completion, with the pre-determined number of patients enrolled, independently of the patient responses observed while the trial is ongoing, except for the case where safety concerns emerge and the trial is immediately terminated. The rigidity in the design of fixed sample size trials has contributed to high costs and long trial durations, as the trial needs to be completed even when the interim patient responses may indicate that the new drug may not be sufficiently effective.

Adaptive clinical trials emerged as a promising path to faster, more economical, and more efficient drug R&D. Research shows that adaptive trials have the potential to offer significant benefits to both the trial's sponsor (the "firm") and the patients (Berry 2012, 2011; Berry et al. 2010; Barker et al. 2009). In 2012, the United States President's Council of Advisors on Science and Technology (PCAST) released a report including recommendations to advance innovation in drug discovery,

development, and evaluation. The PCAST Report states that incorporating new efficiencies into clinical trials is essential for achieving a higher innovation productivity (President's Council of Advisors on Science and Technology 2012). Adaptive trials have also triggered changes in the regulatory process performed by the FDA; the 21st Century Cures Act, signed into law in December of 2016, mandates the FDA to issue guidance on the implementation and review of novel clinical trial designs (U.S. Congress 2016).

Adaptive clinical trials offer various flexibilities to the decision-maker (e.g., the flexibility to stop early for benefit or futility, or drop arms or doses, among others) and enable researchers to efficiently learn about the drug's efficacy (David et al. 2009) through tools such as Bayesian methodologies (Berry 2012; Orloff et al. 2009), and to manage the remainder of the trial accordingly. In particular, sequential adaptive designs, also known as group sequential designs, allow for early trial termination due to anticipated futility or established efficacy at various interim analysis points (decision epochs). At every interim analysis point, the Data and Safety Monitoring Board (DSMB) analyzes the patient responses obtained from the trial thus far, and decides whether to terminate or continue the trial based on various factors, such as the drug's characteristics, the treatment expected to be observed, and external information (Chow and Chang 2008).

This flexibility presents decision makers with new opportunities: an early termination based on interim results reduces the commitment of financial resources, and if the drug is already deemed successful, then it also shortens the time-to-market—options that fixed sample size trials do not offer. However, this additional flexibility also complicates the resource allocation and trial termination decision in an adaptive trial, i.e., at each interim analysis point, the decision maker learns more about the drug, and she can either continue the trial, in which case she needs to determine an optimal patient enrollment plan for the subsequent periods of the trial, or she can terminate the trial.

The decision to terminate a trial is very important, because it not only affects the current clinical study, but also might affect future trials in the same therapeutic area. Further, continuing a trial for too long delays the dissemination of important information and puts participants at unnecessary risk (Todd et al. 2001). Moreover, in any clinical trial, drug misclassification (i.e., terminating the development of an effective drug for futility, or terminating the development of an ineffective drug for benefit) is possible, and the impact of adaptive trial designs on drug misclassification risk is not well-understood.

In summary, adaptive trials are different from their traditional counter-parts that have been used for several decades, and their implications at different levels of the R&D and policy making processes are not fully understood. In today's drug development and clinical trial planning, practitioners will have to define investment and resource allocation strategies that address these flexibilities, and to evaluate drugs that are tested not only traditionally, but also adaptively (David et al. 2015). In this chapter, we analyze how interim analyses affect resource allocation policies and the performance of the drug R&D process, and provide practitioners and policy makers with insights that inform and support the corresponding decision-making.

The remainder of this chapter is organized as follows. In Sect. 14.2, we provide an overview of the characteristics of sequential adaptive clinical trials and the stopping rules that we consider, as well as a summary of current research on resource allocation in the pharmaceutical industry. In Sect. 14.3, we study the current testing and trial termination practices in sequential adaptive clinical trials with binary response, and provide an overview of an optimal resource allocation and trial termination policy. We conclude, in Sect. 14.4, with a discussion of our findings and concluding remarks, and with suggestions for important future research directions.

14.2 Background and Overview of Current Research

In this section, we first present important background information on sequential adaptive clinical trials, and then provide a brief overview of the research on different resource allocation decisions in the pharmaceutical industry, ranging from project pipeline management to capacity investments.

14.2.1 Background on Sequential Adaptive Trials

Clinical trials conducted to establish a new drug's safety and efficacy typically consist of three phases: in Phases 1 and 2, the drug is tested on small groups of patients, while in Phase 3, statistically significant results need to be generated by testing the drug on a larger group of patients. Once the safety and efficacy of a candidate drug are established, a new drug application (NDA) can be submitted to regulatory agencies for approval, and upon approval, the drug can be commercialized. As discussed in Sect. 14.1.2, traditionally, Phase 3 trials involve fixed sample sizes, that is, a fixed number of patients to test in the trial. This fixed sample size is determined before the start of the trial, based on the drug's characteristics, the treatment advantage to be detected, and the desired statistical significance level and power. Once data on the full patient sample are collected and analyzed, the candidate drug is classified as successful (i.e., effective) or futile (i.e., ineffective) based on comparison of the drug's performance with a control therapy. Subsequent approval is then granted if the regulatory agency deems that the evidence provided to support the new drug's safety and efficacy is sufficient.

As opposed to traditional trial designs, in sequential trials the decision to continue or terminate the trial is made at different interim analysis points, based on patient response data collected thus far (Bassler et al. 2008; Whitehead 2004). Trial termination or continuation decisions in sequential trials commonly involve a statistic or primary measure of efficacy—representing the difference between the experimental treatment and a control—which is compared against pre-specified critical values (i.e., thresholds) at a series of interim analysis points throughout the

trial (Todd et al. 2001). The critical values lead to a boundary or stopping rule, whose precise form is determined based on statistical power and significance level, as well as the desired treatment advantage with respect to the primary measure of efficacy (Bassler et al. 2008). If the absolute value of the test statistic is higher than the critical value, an appropriate conclusion is drawn (e.g., rejection of the hypothesis of no difference between therapies), and if the statistic remains within the test boundary, the evidence is insufficient to reach to a conclusion and further analysis is performed at the next interim point (Todd et al. 2001). As an example, in Sect. 14.2.2, we discuss a commonly used stopping rule for adaptive trials, referred to as the "triangular test."

Research suggests that a trial with only two interim analyses considerably reduces the average number of patients exposed to an inferior treatment when compared to a fixed sample size trial (Pocock 1982). Further, the literature indicates that although, in theory, it is possible to conduct an interim analysis after every patient, conducting six or more interim analyses is unlikely to generate statistical, practical, or ethical benefits, unless an extremely large difference between the experimental treatment and the control is expected (McPherson 1982; Pocock 1977). According to Todd et al. (2001), the more frequently one analyzes the accumulating data, the more likely one is to draw incorrect conclusions from the trial.

Most of the existing literature focuses on adaptive trial design and implementation issues. In particular, numerous researchers have built upon the foundational works by Berry and colleagues (e.g., Berry (2005)), which utilize Bayesian analyses and bandit problem formulations to develop new adaptive clinical trial designs (see Chow and Chang (2008) for a review of adaptive design methods and Berry and Fristedt (1985) for a review of the bandit problem). For instance, Ahuja and Birge (2016) propose a novel adaptive design that uses Bayesian learning to adjust patient allocation to treatments. Other researchers analyze the exploration vs. exploitation trade-off within the context of the classical bandit problem (e.g., Macready and Wolpert (1998); Madani et al. (2004)). In addition, several researchers analyze the impact of interim analyses from clinical, statistical, and ethical standpoints (e.g., Jitlal et al. (2012); Zannad et al. (2012); Bassler et al. (2008); Mueller et al. (2007)). However, existing research does not evaluate how interim analyses impact the misclassification probability of the new drug, nor how the firm should adjust its resource allocation decision (patient enrollment plan) for the remainder of the trial at each interim analysis point; these questions are the focus of Sects. 14.3 and 14.4 of this chapter.

In the next section, we describe a commonly used stopping rule for adaptive trials, known as the triangular test, in the context of sequential adaptive trials with binary response—the subject of our analysis. We utilize the triangular test subsequently in this chapter.

14.2.2 Stopping Rules for Sequential Trials with Binary Response Under the Triangular Test

In the following, we provide an overview of various stopping rules used for parallel group sequential trials with binary response, and discuss the methodological details of the triangular test, as described in Whitehead (1997).

As discussed above, in sequential clinical trials, a decision on the continuation or termination of the trial is made periodically (i.e., at each interim analysis point), based on the treatment difference, θ, observed thus far between the experimental drug and the control. In particular, this termination or continuation decision is informed by a set of stopping rules. Different types of stopping rules allow decision makers to determine if the experimental drug is: (1) clearly better than the control (i.e., stopping for benefit), (2) clearly worse than the control (i.e., stopping for futility), or (3) clearly not going to be proven to be better than the control (i.e., stopping for futility) (Whitehead 2004). Some designs (e.g., O'Brien and Fleming (1979), Gordon Lan et al. (1982)) incorporate stopping rules that fulfill the first and second objectives: stop the trial if the experimental drug is clearly better or clearly worse than the control, while some other designs (e.g., the triangular test (Crowley and Hoering 2012)) satisfy the first and third objectives: stop the trial once it is evident that the experimental drug is either superior to the control or if it cannot be shown that it is better than the control.

The vast majority of stopping rules developed for sequential trials belong to one of the two main categories: the repeated significance testing approach and the boundaries approach. In the repeated significance testing approach, a series of conventional statistical analyses are performed, with significance levels adjusted depending on the patient sample size analyzed up to each interim point. These methods are based solely on the p-value, that is, the probability of finding the observed, or a larger, treatment difference when the null hypothesis (commonly, $H_0 : \theta = 0$) is true, and do not consider the treatment difference between the experimental and control therapies. On the other hand, in the boundaries approach, the null hypothesis, $H_0 : \theta = 0$, is rejected or retained with a desired statistical significance level α, and a desired statistical power $1 - \beta$. This approach utilizes two test statistics: Z, which measures the cumulative advantage of the experimental drug over the control (i.e., the efficient score for θ under the null hypothesis), and V, which indicates the amount of information about θ contained in Z (i.e., Fischer's information). The test statistics Z and V are plotted against each other until the stopping boundaries are crossed. Specifically, if the Z statistic's value crosses one of the boundaries, then the trial stops and the null hypothesis, of no difference between the two therapies, is rejected; if the Z statistic remains within the area delimited by the boundaries, then the evidence is not sufficient to arrive at a conclusion, the trial continues to the next interim analysis point (Todd et al. 2001), and the same decision process is repeated.

The triangular test, which belongs to the boundaries approach, is one of the most commonly used stopping methodologies (Rogers et al. 2005; Whitehead 2002; Todd

et al. 2001). Therefore, in what follows, we describe the features of the triangular test in the context of parallel group trials with binary response, where the possible responses are "success" and "failure," and each recruited patient is randomly assigned to either the control or the experimental group, hence, is administered only one of these two therapies. The stopping boundaries of the triangular test incorporate the expected treatment difference, and maintain both statistical significance and power levels. Unlike most repeated significance testing methodologies, which only contemplate stopping for benefit, the triangular test includes an upper boundary (to stop for benefit) and a lower boundary (to stop for futility). The triangular test has been compared with other sequential stopping methodologies, and shown to offer the largest reduction in sample size (Sebille and Bellissant 2000), as well as an early determination of the treatment effect (Hulot et al. 2003).

To describe the triangular test, we follow the notation utilized in Whitehead (1997). It is assumed that patient responses are binary and independent from each other, and that treatment on each patient in the experimental and control groups will be successful with probabilities p_E and p_C, respectively. Given that p_E is unknown, the decision maker has an estimate thereof, $\hat{p}_E$. In contrast, p_C is assumed to be known with certainty, because the control is a placebo or standard therapy, whose treatment effect has already been studied. Let N denote the total number of patients recruited, with N_E patients in the experimental group and N_C patients in the control group, i.e., $N = N_E + N_C$. Also let S_k and F_k, respectively, denote the number of successful and unsuccessful responses in N_k patients, for $k \in \{E, C\}$, with $S = S_E + S_C$ and $F = F_E + F_C$.

The desired treatment difference between the control and the experimental therapies is measured by the log-odds ratio $\theta_R = \ln\left\{\dfrac{\hat{p}_E(1-p_C)}{p_C(1-\hat{p}_E)}\right\}$. The corresponding Z and V statistics can then be obtained by the following expressions:

$$Z = \frac{(N_C S_E - N_E S_C)}{N}, \tag{14.1}$$

$$V = \frac{N_E N_C S F}{N^3}. \tag{14.2}$$

At the trial planning stage, it is common to assume that N is large, thus $S \approx N\bar{p}$ and $F \approx N(1-\bar{p})$, where $\bar{p} \equiv \dfrac{\hat{p}_E + p_C}{2}$. For the case of an equal split of the total number of patients among the experimental and control groups (i.e., $N_E = N_C = N/2$), the full sample size, N, can be derived as follows:

$$N = \frac{4\hat{V}}{\bar{p}(1-\bar{p})}, \tag{14.3}$$

where $\hat{V}$ is the ratio of the value of the information required by a fixed sample size test, satisfying the desired significance (α) and power ($1-\beta$) to θ_R^2 (i.e., the desired

treatment difference, measured by the log-odds ratio, squared). The value of the information required by fixed sample size tests for various values of α and $1-\beta$ are tabulated, see, e.g., Whitehead (1997).

Given the full sample size, N, the statistics Z and V are utilized to find the triangular test's upper and lower boundaries, which correspond to the lines defined by $Z = a + bV$ and $Z = -a + 3bV$, respectively, where the parameters a and b are given by the following expressions:

Case Where $\beta = \alpha$

$$a = \frac{2}{\theta_R} \ln\left(\frac{1}{2\alpha}\right), \tag{14.4}$$

$$b = \frac{1}{4}\theta_R. \tag{14.5}$$

Case Where $\beta \neq \alpha \cdot \theta_R$ in Eqs. (14.4) and (14.5) needs to be replaced with a corrected value, θ'_R, given below (Bellissant et al. 1990):

$$\theta'_R = \theta_R \left[\frac{2\Phi^{-1}(1-\alpha)}{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)}\right], \tag{14.6}$$

where $\Phi(x)$ denotes the standard normal cumulative distribution function.

When decision-making takes place on a continuous basis (i.e., after observing each patient), the stopping mechanism is fully defined by the linear upper and lower boundaries. In practice, however, decision-making occurs at discrete times (i.e., discrete monitoring), thus, the triangular test needs to be adjusted and the boundaries need to be modified. This results in "inner boundaries," often referred to as "Christmas tree boundaries" because of their shape; see Fig. 14.1 for linear boundaries and the corresponding inner boundaries.

In the discrete monitoring case, the stopping rule is as follows: at the ith interim analysis, the statistics Z_i and V_i are calculated using Eqs. (14.1) and (14.2) based on the data observed thus far, as well as an upper stopping limit, u_i, and a lower stopping limit, l_i, where $u_i > l_i$:

$$u_i = a + bV_i - 0.583\sqrt{(V_i - V_{i-1})}, \tag{14.7}$$

$$l_i = -a + 3bV_i + 0.583\sqrt{(V_i - V_{i-1})}, \tag{14.8}$$

where the term $0.583\sqrt{(V_i - V_{i-1})}$ is a correction factor for the discrete monitoring case, with $V_0 = 0$.

If $Z_i \geq u_i$, then the trial stops and the null hypothesis, of no difference between therapies, is rejected, with the conclusion that the experimental drug is more effective than the control. If $Z_i \leq l_i$, then the trial stops with two possible

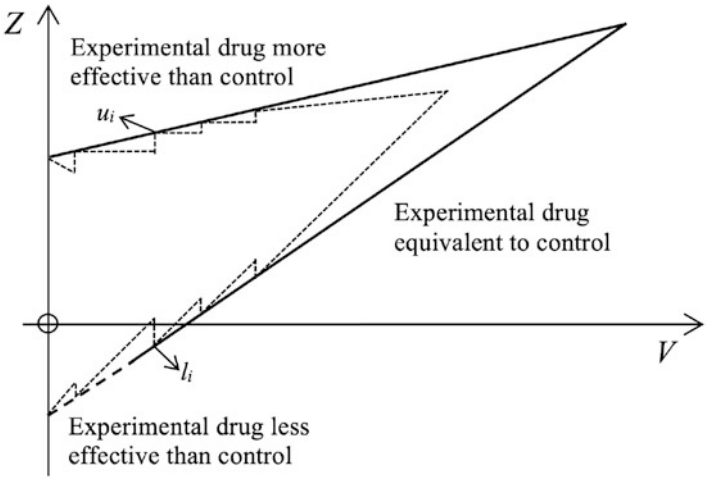


Fig. 14.1 Stopping boundaries for the triangular test under discrete decision-making (Adapted from Whitehead (1997))

conclusions, depending on the value of V_i: (1) that the experimental drug is less effective than the control, if the ith interim analysis takes place after a small patient sample has been observed (shown with the dashed part of the lower stopping boundary in Fig. 14.1), or (2) that the therapies are equivalent, thus, with the acceptance of the null hypothesis, if the ith interim analysis is performed after a larger patient sample has been observed (shown with the solid part of the lower stopping boundary in Fig. 14.1).

We apply the triangular test and its stopping boundaries in Sect. 14.3.1, with the objective of determining the likelihood of drug misclassification in sequential adaptive trials with binary response.

14.2.3 Overview of Research on Resource Allocation in the Pharmaceutical Industry

Resource allocation decisions have been studied in a variety of application areas in the pharmaceutical industry. In what follows, we provide some examples of optimization-based approaches applied to pharmaceutical R&D project portfolio construction and capacity planning.

Some examples of optimization-based approaches for pharmaceutical R&D project pipeline management include the works of Colvin and Maravelias (2010) and Christian and Cremaschi (2015). In the context of project portfolios, Solak et al. (2010) develop a multi-stage stochastic program for constructing an R&D portfolio under endogenous uncertainty, and Patel et al. (2013) develop an integer

programming model as the basis for a Bayesian analysis, with the objective of optimizing portfolios of Phase 3 trials. Other researchers use real options theory in R&D project portfolio optimization (e.g., Jacob and Kwak (2003); Enea and Lo Nigro (2011); Kouvelis et al. (2017)). In particular, Kouvelis et al. (2017) use real options theory to study Phase 3 trials under a stochastic revenue and a stochastic patient enrollment rate, and use a test statistic to make a decision (i.e., stop the trial for futility or efficacy), hence, they do not explicitly model the decision maker's learning process resulting from data updates.

Capacity planning in the pharmaceutical industry has also received attention in the literature. For example, Oh and Karimi (2004) present a mixed-integer linear programming formulation, including decisions to expand existing manufacturing facilities, build new capacity, or outsource. Rajapakse et al. (2005) consider similar decisions and develop a framework based on scenario analysis and discrete-event simulation. Some models address R&D portfolio management and manufacturing capacity planning problems simultaneously (e.g., Levis and Papageorgiou (2004)). Kaminsky and Yuen (2014) evaluate the impact of data updates on the capacity investment decision in the pharmaceutical industry, considering a fixed sample size clinical trial and a fixed and a priori determined patient recruitment rate. The firm needs to decide, under uncertainty on the outcome of the clinical trial, whether or not to invest in production capacity in each period: if sufficient capacity is not built by the end of the trial and the trial is successful, then the launch of the new drug will be delayed; if the trial is unsuccessful, then the resources invested on capacity will become a sunk cost.

Rojas-Cordova and Bish (2018) use a stochastic dynamic programming model to determine an optimal resource allocation decision, i.e. patient enrollment policy, for a group sequential adaptive trial that allows the decision maker to stop the trial at an interim analysis point for benefit or futility reasons, based on a periodically updated estimate of the drug's success probability. They consider that the potential revenue for a successful drug is time-decreasing, and assume that the firm is able to achieve the target enrollment rates. To the best of our knowledge, with the exception of the aforementioned work, the optimal resource allocation decision for adaptive clinical trials has not been explored, but the need to study it in a rigorous manner has been acknowledged by various researchers. For example, David et al. (2015) elaborate on the importance of evaluating the impact of different adaptive trial designs (such as sequential designs with interim analyses) on financial value, and argue that these designs may increase the firm's flexibility to balance risk, cost, and time in drug development, which are quantified in Rojas-Cordova and Bish (2018).

In the following sections, we describe the models that we have developed to analyze the current stopping practices in sequential adaptive clinical trials, and to address the corresponding patient enrollment and trial termination decisions under the flexibility that interim analysis points provide to decision-makers. We focus our analysis on three performance measures: drug misclassification risk, time-to-market, and the firm's profit from the commercialization of the experimental drug.

14.3 Drug Misclassification Risk and Optimal Resource Allocation in Sequential Adaptive Clinical Trials

In this section, we provide an overview of the various models that we have developed to: evaluate the likelihood of drug misclassification in trials conducted following the triangular test's stopping rules (Sect. 14.3.1), and determine an optimal resource allocation and trial termination policy (Sect. 14.3.2), in sequential adaptive clinical trials. We refer the interested reader to Rojas-Cordova and Hosseinichimeh (2018) and Rojas-Cordova and Bish (2018) for further details on each of these two models.

14.3.1 Quantifying the Likelihood of Drug Misclassification in Current Practices

Our objective in this first part of the study is to quantify the drug misclassification risk, or likelihood thereof, within the context of a sequential adaptive trial that is conducted following the triangular test design, and where the trial termination and continuation decisions follow the triangular test's stopping rules, see Sect. 14.2.2. Numerous researchers from the medical field have been skeptical towards sequential adaptive trials, and have affirmed that the ability to stop early for benefit could increase the risk of false positive classifications, that is, classifying an ineffective drug as effective (e.g., Whitehead (2004), Zannad et al. (2012), Pocock and White (1999), among others). The risk of false negatives (i.e., classifying an effective drug as ineffective) has not been analyzed to the same extent, although this type of drug misclassification is also detrimental to both the trial sponsors and patients. We fill this gap in the literature by analyzing the trial design parameters that lead to drug misclassification, and examining the magnitude of the "hot stove effect," a bias against risky and novel options that are not further explored because of initial bad experiences, preventing learning (Denrell and March 2001). In what follows, we summarize our model and findings, and refer the reader to Rojas-Cordova and Hosseinichimeh (2018) for more information.

In this model, we consider a candidate drug that needs to undergo Phase 3 testing within a sequential adaptive trial with a single interim analysis point. In line with FDA recommendations (US Food and Drug Administration 2010), the number of interim analysis points—one, in this case—cannot be modified once patients start being recruited to the trial. Thus, at the interim analysis point, the firm needs to decide whether to continue or stop the trial. The *full sample size* is also determined before the start of the trial, based on statistical power and significance. Similar to other models in literature (e.g., Patel et al. (2013) and the references therein), we assume that no safety concerns exist, and focus on efficacy.

We develop a system dynamics model to represent the Phase 3 testing process; see Fig. 14.2 for an overview of the model. In particular, each patient that is recruited is assigned to a treatment group during the *patient randomization* process, and receives the control or the experimental drug in the *drug testing process*. The

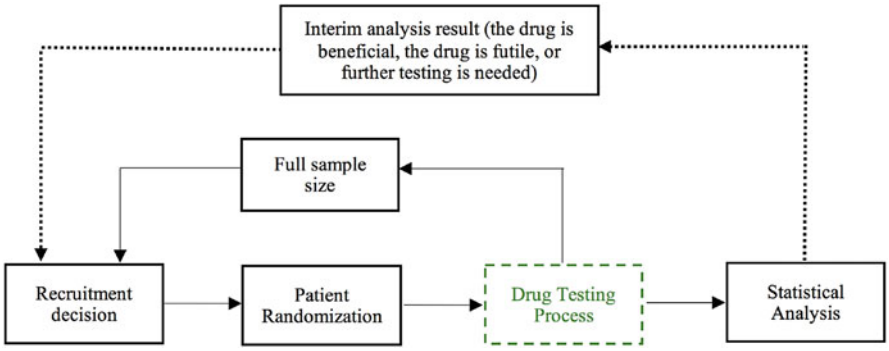


Fig. 14.2 Structure of the patient recruitment decision-making model for a single drug (Phase 3)

outcome of the *drug testing process* is subject to a *statistical analysis*, conducted at the interim analysis point under the triangular test's stopping rules, and corresponds to one of the following: (1) the drug is clearly beneficial, (2) the drug is clearly futile, or (3) the testing is inconclusive. Outcomes (1) and (2) will lead to the termination of the trial, whereas outcome (3) leads to a trial continuation decision, with more patients recruited.

We consider an experimental drug with an estimated success probability of 0.70, and a control with a true (and known) success probability of 0.50 to derive the triangular test's parameters. The full sample size N is derived utilizing Eq. (14.3), and is equal to 244 patients. We conduct a series of sensitivity analyses where we vary the number of patients analyzed at the interim analysis point and the true success probability of the experimental drug, in order to study the effect of imperfect information (i.e., optimistic and pessimistic prior expectations on the drug's efficacy), and the conditions under which the drug misclassification rate increases. We evaluate misclassification rates for drugs that are truly effective (i.e., with true success probabilities higher than the control's) and drugs that are truly ineffective (i.e., with true success probabilities lower than or equal to the control's). We simulate trials for drugs with success probabilities ranging from 0.10 to 0.90, and compute the percentage of simulation runs where the trial stops early for benefit, and the percentage of runs where the trial stops early for futility.

Across all interim analysis limits, we find that no more than 2.3% of the ineffective drugs tested are mistakenly classified as beneficial, with a larger proportion of false positives emerging with interim analysis limits equivalent to at least 55% of the sample size (Fig. 14.3a). On the other hand, the proportion of false negatives is significantly larger across the entire range of interim analysis limits we studied. As shown in Fig. 14.3b, the percentage of effective drugs that are misclassified as futile—as a result of early trial termination for futility—can be more than 10 times larger than the percentage of ineffective drugs misclassified as beneficial when evaluating the same sample size at the interim analysis point.

Larger interim analysis limits may involve larger drug misclassification rates. Our results also suggest that the overall likelihood of false negatives is significantly larger than the overall likelihood of false positives. We find that false negatives

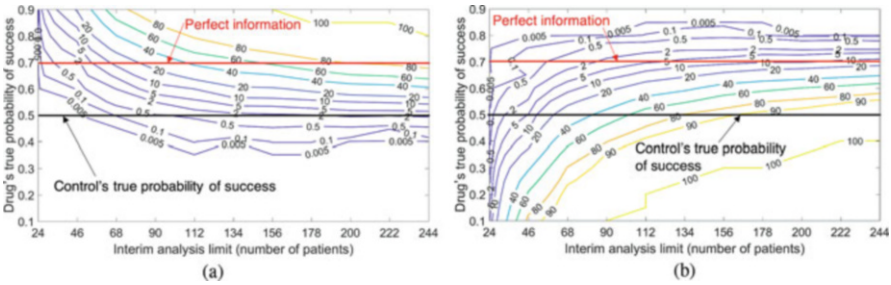


Fig. 14.3 Percentage of drugs whose trials were terminated early. (**a**) Termination for benefit. (**b**) Termination for futility

emerge with experimental drugs that are clearly better (i.e., with efficacies 0.05 to 0.20 higher) than the control, whereas false positives emerge with experimental drugs that are only slightly worse (i.e., with efficacies up to 0.02 lower) than the control. In other words, drugs that are clearly (and oftentimes, significantly) more effective than the control are misclassified as futile, whereas drugs that are only slightly (and not significantly) worse than the control are misclassified as beneficial. Our sensitivity analysis shows that drugs that are not as effective as expected, but still able to generate beneficial effects, can be misclassified as futile in as many as 92% of the simulation runs (for a drug with a true success probability of 0.55). Figures 14.3a and b also confirm that, regardless of the sample size analyzed at the interim analysis point, false negatives are more likely than false positives. In Rojas-Cordova and Hosseinichimeh (2018), we provide and discuss similar results in sequential adaptive trials with two and three interim analysis points.

In the next section, we study optimal resource allocation and trial termination policies for sequential adaptive clinical trials with one and two interim analysis points, as well as fixed sample size trials (used as a benchmark). We broaden our analysis to derive time-to-market and expected profit values in addition to drug misclassification risk.

14.3.2 Optimal Resource Allocation and Trial Termination Policies

Our objective in this section is to study the firm's optimal resource allocation (i.e., patient enrollment) and trial termination policy in a sequential trial with binary response. As discussed in Sect. 14.1, adaptive clinical trials promise to increase drug R&D productivity by reducing developmental costs and lead times, but pose new challenges to decision makers. As discussed in Sect. 14.2, the current literature offers a wide variety of statistical methodologies that lead to specific stopping rules

for determining whether to stop or continue the trial at different interim points throughout the trial's duration. However, the main objective of these methodologies is to preserve certain statistical qualities such as significance and power, without considering any financial factors, such as the budget available to the firm for the trial, or the potential revenue from the commercialization of the drug. Further, these statistical methodologies do not address the number of patients that need to be recruited in between two consecutive interim analysis points.

We develop a stochastic programming model to address the two main challenges presented by sequential adaptive clinical trials (i.e., the selection of stopping rules and the determination of the number of patients to enroll in between each pair of interim analysis points), while incorporating both statistical significance and financial considerations, as well as the fact that the efficacy of the candidate drug is highly uncertain at the outset of the Phase 3 trial.

Specifically, our model utilizes both the information available to the decision maker at the outset of the Phase 3 trial (e.g., expert opinions, past data on similar drugs, outcomes of Phase 1 and 2 trials) and the new information obtained up to each interim analysis point of the ongoing trial, within a Bayesian framework, to update the decision maker's beliefs on the efficacy of the candidate drug. We then use the updated probability of efficacy of the experimental drug, and determine the firm's optimal dynamic resource allocation policy throughout the course of the trial. That is, based on patient responses obtained up to an interim analysis point, the firm can either: (1) pursue a more aggressive investment policy (increasing the rate of enrollment) in the next period to expedite the submission of a new drug approval application to the FDA, (2) slow down enrollment to obtain more information on the drug before more resources are invested in the trial or (3) completely terminate the trial. In this section, we provide an overview of our modeling and results, and refer the interested reader to Rojas-Cordova and Bish (2018) for more details.

The firm determines the resource allocation, i.e., the number of patients to recruit in each period $t \in [1, \ldots, T]$, N_t, over the duration of the Phase 3 trial under uncertainty on the experimental drug's probability of efficacy, so as to maximize the expected profit from the commercialization of the drug. Let M denote the maximum number of observations (one from each patient) the firm is able to collect based on a budget of B (i.e., $M = \frac{B}{c}$, where c represents the recruitment and testing cost per patient). The experimental drug's success (i.e., deeming the drug effective in comparison with a placebo or a standard therapy) is assessed through an exogenously determined criterion, represented in terms of the minimum number of patients that need to be treated successfully with the experimental drug during the Phase 3 trial, $\gamma_c(M)$, which is a non-decreasing function of M, and is determined based on statistical significance and power. If this success criterion is met, then a time-dependent revenue is realized. The revenue for a successful experimental drug is modeled as a strictly decreasing function in time to represent the potential loss in market share, and the reduction in active patent life due to delays in commercialization.

A sequential adaptive Phase 3 design provides the firm with the flexibility to terminate the trial earlier than in a traditional setting of fixed sample size, due

to either proven benefit (i.e., the success criterion is achieved early) or perceived futility (i.e., initial observations from patients indicate a "low" likelihood of success). An "early" success enables the firm to market the drug earlier and realize a higher revenue, while an "early" termination due to perceived futility reduces the Phase 3 trial cost. An early termination option is not available for a fixed sample size clinical trial, which the firm must carry to completion (i.e., it must collect the number of observations initially determined to achieve the desired statistical significance and power), regardless of what the interim data may indicate, except for the situation in which the interim data indicate safety concerns—an aspect we do not model.

Therefore, the resource allocation decision in a sequential adaptive clinical trial is dictated by the exploitation vs. exploration trade-off, i.e., how should the resources be split between the efforts of obtaining information (hence better estimating the unknown success probability of the experimental drug), and performing the experiments, or in our context, recruiting patients so as to achieve the success criterion. In our setting, these efforts overlap, meaning that the firm can use the interim data both for information updating purposes and for achieving the success criterion. Thus, as discussed previously, the optimal decision needs to consider the trade-off between the incentive to reach the minimum number of successful observations as early as possible to realize the maximum revenue, and the benefit resulting from the collection of information and consequent updates to the decision maker's prior on the experimental drug's probability of success, without committing a large amount of budget early on.

We also model and analyze a fixed sample size Phase 3 trial, which we use as a benchmark. In the sequential adaptive trial, at the beginning of each period, the decision maker updates her estimate on the experimental drug's success probability, based on observations collected through the end of the previous period, and determines the number of observations to collect in the current period, constrained by the remaining number of observations that can be collected in the current and subsequent periods. In Rojas-Cordova and Bish (2018), we derive various key properties of the stochastic programming formulation, which allow us to characterize the structure of an optimal resource allocation and trial termination policy for both a sequential adaptive trial and a fixed sample size trial. Then, we perform a numerical study based on realistic data to quantify the impact of interim analyses and the proposed Bayesian updating scheme on drug misclassification risk, time-to-market, and expected profit.

In our numerical study, we use various financial parameters that are representative of the cost of a Phase 3 oncology trial, with an available budget of $10 million and a per observation (patient) cost of $50,000 (Pharmaceutical Research and Manufacturers of America (PhRMA) 2015). In such trials, the cost of patient enrollment and observation is typically very high, due to the level of professional resources and the physical infrastructure needed to support the commonly geographically dispersed Phase 3 trial activities, hence the value of optimization-based approaches can be significant. We consider a time-decreasing revenue with a reduction rate of 5% per period, and a final revenue of $1 billion. The revenue figures were chosen

based on global net sales per novel active substance (i.e., a molecular or biologic entity or combination product in which at least one element had not been previously approved by the FDA) in the 2005–2009 launch cohort (Berndt et al. 2015), and sales expectations for cancer drug Keytruda (Pembrolizumab) (Nissen 2016). The success criterion in Phase 3 is represented by a minimum number of successful observations of $\gamma_c(M) = 100$, corresponding to $M = 200$ maximum observations; this threshold exemplifies certain Phase 3 oncology studies (Nesse 2016; American Society of Clinical Oncology (ASCO) 2015; National Cancer Institute (NCI) 2015; Ye 2008) after which the drug can be submitted for FDA approval and subsequently commercialized.

We analyze how the firm's optimal resource allocation policy in two sequential adaptive trial variations (i.e., one and two interim analysis points) compares to that in the fixed sample size trial. We consider three different scenarios: undershooting the prior (pessimistic prior), perfect information, and overshooting the prior (optimistic prior). Our analysis shows that, although the revenue is time-decreasing, the optimal policy in the adaptive trials suggests to initially "test out" or "explore" the drug when the decision maker's prior is pessimistic. If the drug's performance in the first period is satisfactory, then the optimal solution consists in allocating a much larger budget to the trial in the second period, so as to "exploit" the drug's good performance. Thus, the optimal policy in the sequential adaptive trials is to act conservatively when the prior on the experimental drug's success probability is low. In the fixed sample size trial, however, in the absence of the exploration option, low priors automatically lead to zero budget allocation to the trial, i.e., the drug is immediately classified as futile without conducting a Phase 3 trial.

Figure 14.4 depicts the total number of observations for each trial in the three scenarios. The trade-off between the time-decreasing nature of the revenue and the flexibility offered by the multi-period nature of adaptive trials, to explore

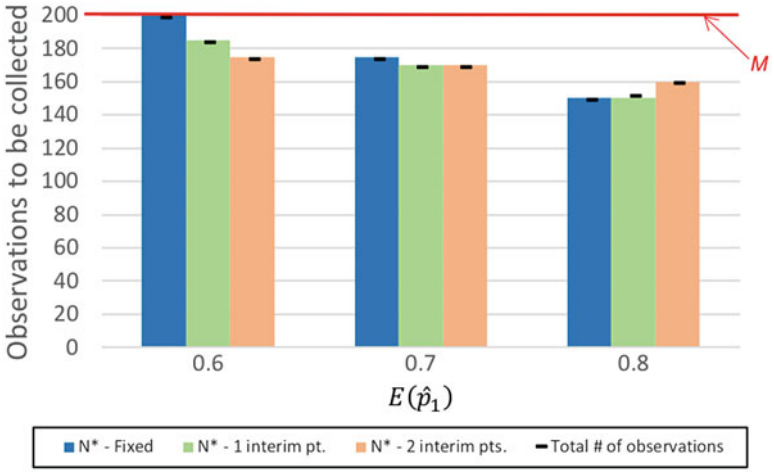


Fig. 14.4 Optimal resource allocation in the fixed sample size and sequential adaptive trial variations (one interim analysis and two interim analyses)

the performance of the experimental drug prior to committing a large amount of budget in the first period, explains why the total number of observations to be collected is the largest for the fixed sample size trial in the pessimistic and perfect information scenarios. On the other hand, when using an optimistic prior in the two interim analyses trial, the firm's incentive to realize the revenue early on exceeds the potential benefit of gathering some information about the drug, thus, the total number of observations in this sequential trial is larger than in the fixed sample size trial. As expected, in all cases, the optimal number of observations decreases as the prior increases. The total number of observations in the fixed sample size trial is the largest when using perfect information or a pessimistic prior, but the lowest (along with the one interim analysis trial) when using an optimistic prior. As Fig. 14.4 indicates, the adaptive trials stop at the first interim analysis point in five out of the six scenarios studied.

We next analyze the performance of the optimal resource allocation and trial termination policy, in terms of drug misclassification risk, expected time-to-market, expected profit, and the costs of optimism versus pessimism in each trial and for each scenario; see Tables 14.1, 14.2 and 14.3.

Table 14.1 Drug misclassification risk (%) for fixed sample size and sequential adaptive trials

	$E(\hat{p}_1)$								
Setting	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Fixed sample size	100.00	100.00	100.00	100.00	0.00	0.00	0.00	14.87	95.30
1 interim analysis	41.10	41.10	41.10	41.10	0.00	0.00	0.00	0.00	0.13
2 interim analyses	8.40	8.40	8.40	8.40	0.00	0.00	0.00	0.00	0.00

Table 14.2 Expected time-to-market (in number of periods) for fixed sample size and sequential adaptive trials

	$E(\hat{p}_1)$								
Setting	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Fixed sample size	NA	NA	NA	NA	3.00	3.00	3.00	3.00	3.00
1 interim analysis	3.00	3.00	3.00	3.00	1.50	1.50	1.51	1.74	2.99
2 interim analyses	2.28	2.28	2.28	2.28	1.00	1.00	1.00	1.01	1.01

NA not applicable, as the drug is misclassified as futile

Table 14.3 Expected profit and costs of pessimism vs. optimism for fixed sample size and sequential adaptive trials

	E [Profit] ($ million)			Cost ($ million)	
Setting	$E(\hat{p}_1) = 0.6$	$E(\hat{p}_1) = 0.7$	$E(\hat{p}_1) = 0.8$	Pessimism	Optimism
Fixed sample size	990.00	991.25	843.83	1.25	147.42
1 interim analysis	1071.08	1071.83	1059.90	0.75	11.93
2 interim analyses	1099.28	1099.48	1099.25	0.19	0.23

Not surprisingly, the fixed sample size trial leads to the highest misclassification risk overall: by not allocating any resources in the scenarios with pessimistic priors, the decision maker does not collect any observation and automatically classifies the drug as futile. When overshooting the prior, the fixed sample size trial still leads to a very high misclassification risk, in contrast to the virtually null misclassification risks in the sequential adaptive trials (Table 14.1). In terms of expected time-to-market (Table 14.2), in fixed sample size trials the drug is either: (1) not tested under pessimistic priors, or (2) submitted for FDA approval only after the trial's full duration (3 periods). Across all scenarios, both sequential adaptive trials translate into lower drug misclassification risks and shorter time-to-market.

Finally, as shown in Table 14.3, the cost of pessimism is only a fraction of the cost of optimism, because optimistic priors lead the decision maker to collect fewer observations and either misclassify the experimental drug as futile (in the fixed sample size trial) or delay the product commercialization (in sequential adaptive trials). Further, the cost of optimism in the fixed sample size trial is the largest due to the high misclassification risk caused by overshooting the prior (Table 14.1). In terms of expected profit, the fixed sample size trial performs the worst, and the single interim analysis trial offers the largest improvement because of its lower misclassification risk and shorter time-to-market.

In the next section, we discuss the implications of the analysis performed in Sects. 14.3.1 and 14.3.2 from the perspectives of different stakeholders. We also provide concluding remarks and various suggestions for expanding this line of research.

14.4 Conclusions

New medications play a vital role in public health, and have a large impact on the quality of human life. The current development lead times in drug R&D prevent patients from having prompt access to novel therapies, which can potentially cure or alleviate the symptoms of severe diseases. Research that can alleviate these issues can have a positive impact on the society, by providing decision makers with a better understanding of optimal resource allocation policies (patient enrollment and stopping rules) in adaptive clinical trials, thus making it possible for patients to have expedited access to effective drugs and therapies.

Adaptive clinical trials promise to increase drug R&D productivity by reducing development costs and lead times, but this new trial design poses new challenges to decision makers. The research described in Sect. 14.3 serves as a proof of concept that optimal resource allocation policies can enable pharmaceutical firms to optimally allocate their limited funds to promising drugs, and promptly withdraw funding from treatments that do not show sufficient evidence of efficacy. This not only entails higher financial gains to the firms, but also represents important benefits to the trial's participants, and patients in general: in case the drug under development is ineffective, trial participants will promptly stop receiving this treatment, and if

the experimental drug is effective, it will be submitted for regulatory approval much sooner than under a non-adaptive setting. Upon approval, this novel and effective drug will be made available to patients.

The model described in Sect. 14.3.1 captures the current practices of testing and analysis in a sequential adaptive trial, and allows us to quantify the likelihood of drug misclassification, in the form of false negatives and false positives, under the triangular test. The model discussed in Sect. 14.3.2 represents a framework in which the newly acquired information from the ongoing trial can be used to update the decision maker's beliefs about the experimental drug's efficacy, and shows the value of using this information within an optimization model.

While the research described in this chapter sheds light on the benefits of sequential adaptive clinical trials, and offers an optimal resource allocation policy within this new setting, there is much work to be done, especially in the application of various operations research and simulation methodologies, such as stochastic dynamic programming and system dynamics—discussed here—which can be utilized to model and study decision problems arising in sequential adaptive trials, and in pharmaceutical R&D in general. The overall goal is to enable pharmaceutical firms to further explore and apply adaptive designs, and regulatory bodies to define drug evaluation and approval mechanisms that contemplate the new flexibilities and potential of these adaptive designs.

Future research can expand the models discussed in this chapter in various directions. For example, in certain clinical trials, patient recruitment and observation can be highly stochastic due to participants dropping out, or therapies not showing evaluable results, therefore, optimal resource allocation schemes need to be devised considering stochastic patient recruitment and stochastic response collection rates. Given the complexities of the pharmaceutical market, competitors' actions, and changes in regulatory requirements, the firm's revenue from a successful drug can also be stochastic. Within the sequential adaptive clinical trial context, stochastic dynamic programming can be applied to decisions other than resource allocation; for instance, to the allocation of patients with different characteristics to different trial arms, with each arm corresponding to a different therapy, or a different dosage of a given therapy. In these problems, the reward may not be financial, but rather measured in terms of a specific health outcome.

Determining optimal resource allocation policies for trials following other types of adaptive designs (e.g., Phase II/III hybrid designs) represents a promising area of research. The shift towards these novel drug testing designs will require researchers to examine the nature and complexity of the decisions that pharmaceutical firms, clinical investigators, and policy makers face. This area would certainly benefit from interdisciplinary research efforts that take into account the different perspectives and needs of the stakeholders involved, in order to perform a holistic analysis of the different decision problems. If successful, such research efforts are likely to make a substantial contribution to the scientific literature, as well as to the pharmaceutical R&D decision-making and regulatory practices.

Acknowledgements We would like to thank Mr. Keith Gardner, Senior Director of Decision Science at AstraZeneca Pharmaceuticals, for many valuable discussions that improved our understanding of the drug R&D process. This research was supported in part by the Seth Bonder Foundation.

References

Ahuja V, Birge JR (2016) Response-adaptive designs for clinical trials: simultaneous learning from multiple patients. Eur J Oper Res 248(2):619–633

American Society of Clinical Oncology (ASCO) (2015) Phases of Clinical Trials. http://www.cancer.net/navigating-cancer-care/how-cancer-treated/clinical-trials/phases-clinical-trials. Accessed 28 Sept 2016

Barker A, Sigman C, Kelloff G, Hylton N, Berry D, Esserman L (2009) I-SPY 2: An Adaptive Breast Cancer Trial Design in the Setting of Neoadjuvant Chemotherapy. Clin Pharmacol Ther 86(1):97–100

Bassler D, Montori VM, Briel M, Glasziou P, Guyatt G (2008) Early stopping of randomized clinical trials for overt efficacy is problematic. J Clin Epidemiol 61(3):241–246

Bellissant E, Benichou J, Chastang C (1990) Application of the triangular test to phase II cancer clinical trials. Stat. Med 9(8):907–917

Berndt ER, Nass D, Kleinrock M, Aitken M (2015) Decline in economic returns from new drugs raises questions about sustaining innovations. Health Aff 34(2):245–252

Berry DA (2005) Introduction to Bayesian methods III: use and interpretation of Bayesian tools in design and analysis. Clin. Trials 2(4):295–300

Berry DA (2011) Adaptive clinical trials: the promise and the caution. J Clin Oncol 29(6):606–609

Berry DA (2012) Adaptive clinical trials in oncology. Nat Rev Clin Oncol 9(4):199–207

Berry DA, Fristedt B (1985) Bandit problems: sequential allocation of experiments (Monographs on Statistics and Applied Probability). Springer, Dordrecht

Berry SM, Carlin BP, Lee JJ, Muller P (2010) Bayesian adaptive methods for clinical trials. CRC Press, Boca Raton

Chow SC, Chang M (2008) Adaptive design methods in clinical trials–a review. Orphanet J Rare Dis 3(1):1

Christian B, Cremaschi S (2015) Heuristic solution approaches to the pharmaceutical R&D pipeline management problem. Comput Chem Eng 74:34–47

Colvin M, Maravelias CT (2010) Modeling methods and a branch and cut algorithm for pharmaceutical clinical trial planning using stochastic programming. Eur J Oper Res 203(1):205–215

Crowley J, Hoering A (2012) Handbook of statistics in clinical oncology. CRC Press, Boca Raton

David E, Tramontin T, Zemmel R (2009) Pharmaceutical R&D: the road to positive returns. Nat Rev Drug Discov 8(8):609–610

David FS, Bobulsky S, Schulz K, Patel N (2015) Creating value with financially adaptive clinical trials. Nat Rev Drug Discov 14(8):523–524

Denrell J, March JG (2001) Adaptation as information restriction: The hot stove effect. Organ Sci 12(5):523–538

DiMasi JA, Hansen RW, Grabowski HG (2003) The price of innovation: new estimates of drug development costs. J. Health Econ 22(2):151–185, URL http://www.sciencedirect.com/science/article/pii/S0167629602001261

DiMasi JA, Grabowski HG, Hansen RW (2016) Innovation in the pharmaceutical industry: new estimates of R&D costs. J Health Econ 47:20–33

Ding M, Eliashberg J, Stremersch S (2013) Innovation and marketing in the pharmaceutical industry: emerging practices, research, and policies. Springer, New York

Enea G, Lo Nigro G (2011) A real options based model to select a balanced R&D portfolio. In: 15th Annual International Conference on Real Options, Turku, Finland, Realoptions.org

Gordon Lan K, Simon R, Halperin M (1982) Stochastically curtailed tests in long term clinical trials. Seq Anal 1(3):207–219

Halliday RG, Drasdo AL, Lumley CE, Walker SR (1997) The allocation of resources for R&D in the world's leading pharmaceutical companies. R D Manag 27(1):63–77

Hulot JS, Cucherat M, Charlesworth A, Van Veldhuisen DJ, Corvol JC, Mallet A, Boissel JP, Hampton J, Lechat P (2003) Planning and monitoring of placebo-controlled survival trials: comparison of the triangular test with usual interim analyses methods. Br J Clin Pharmacol 55(3):299–306

Jacob WF, Kwak YH (2003) In search of innovative techniques to evaluate pharmaceutical R&D projects. Technovation 23(4):291–296

Jitlal M, Khan I, Lee S, Hackshaw A (2012) Stopping clinical trials early for futility: retrospective analysis of several randomised clinical studies. Br J Cancer 107(6):910–917

Kaminsky P, Yuen M (2014) Production capacity investment with data updates. IIE Trans 46(7):664–682

Kouvelis P, Milner J, Tian Z (2017) Clinical trials for new drug development: Optimal investment and application. Manuf Serv Oper Manag 19(3):437–452

Lacey MJ, Hanna GJ, Miller JD, Foster TS, Russell MW (2014) Impact of pharmaceutical innovation in HIV/AIDS treatment during the highly active antiretroviral therapy (HAART) era in the US, 1987–2010: an epidemiologic and cost-impact modeling case study. http://truvenhealth.com/Portals/0/Assets/Life-Sciences/White-Papers/pharma-innovation-hiv-aids-treatment.pdf. Accessed 20 Jan 2018

Levis AA, Papageorgiou LG (2004) A hierarchical solution approach for multi-site capacity planning under uncertainty in the pharmaceutical industry. Comput Chem Eng 28(5):707–725

Macready WG, Wolpert DH (1998) Bandit problems and the exploration/exploitation tradeoff. IEEE Trans Evol Comput 2(1):2–22

Madani O, Lizotte DJ, Greiner R (2004) The budgeted multi-armed bandit problem. In: International conference on computational learning theory, Springer, Berlin pp 643–645

McPherson K (1982) On choosing the number of interim analyses in clinical trials. Stat Med 1(1):25–36

Mueller PS, Montori VM, Bassler D, Koenig BA, Guyatt GH (2007) Ethical issues in stopping randomized trials early because of apparent benefit. Ann Intern Med 146(12):878–881

National Cancer Institute (2014) Surveillance, epidemiology, and end results program. https://seer.cancer.gov/statfacts/html/ld/all.html. Accessed 02 Apr 2017

National Cancer Institute (NCI) (2015) NCI dictionary of cancer terms. https://www.cancer.gov/publications/dictionaries/cancer-terms?cdrid=45833. Accessed 28 Sept 2016

National Center for Health Statistics (2015) Health, United States, 2014: with special feature on adults aged 55–64. https://www.cdc.gov/nchs/data/hus/hus14.pdf. Accessed 21 May 2017

Nesse E (2016) Clinical trial design. https://ccrod.cancer.gov/confluence/download/attachments/71041052/Clinical_Trial_Design.pdf. Accessed 28 Sept 2016

Nissen M (2016) Read the label on pharma's new drug sales. https://www.bloomberg.com/gadfly/articles/2016-08-16/big-pharma-new-drug-sales-tell-only-part-of-the-story. Accessed 29 Sept 2016

O'Brien PC, Fleming TR (1979) A multiple testing procedure for clinical trials. Biometrics 35(3):549–556

Oh HC, Karimi I (2004) Regulatory factors and capacity-expansion planning in global chemical supply chains. Ind Eng Chem Res 43(13):3364–3380

Orloff J, Douglas F, Pinheiro J, Levinson S, Branson M, Chaturvedi P, Ette E, Gallo P, Hirsch G, Mehta C, et al (2009) The future of drug development: advancing clinical trial design. Nat Rev Drug Discov 8(12):949–957

Patel NR, Ankolekar S, Antonijevic Z, Rajicic N (2013) A mathematical model for maximizing the value of phase 3 drug development portfolios incorporating budget constraints and risk. Stat Med 32(10):1763–1777

Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL (2010) How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov 9(3):203–214

Pharmaceutical Research and Manufacturers of America (PhRMA) (2015) Biopharmaceutical industry-sponsored clinical trials: impact on state economies. http://www.phrma.org/sites/default/files/pdf/biopharmaceutical-industry-sponsored-clinical-trials-impact-on-state-economies.pdf. Accessed 28 Sept 2016

Pharmaceutical Research and Manufacturers of America (2016a) 2016 Biopharmaceutical research industry profile. http://www.phrma.org/report/industry-profile-2016. Accessed 21 May 2017

Pharmaceutical Research and Manufacturers of America (2016b) A decade of innovation in chronic diseases. http://www.phrma.org/report/a-decade-of-innovation-in-chronic-diseases. Accessed 21 May 2017

Pocock SJ (1977) Group sequential methods in the design and analysis of clinical trials. Biometrika 64(2):191–199

Pocock SJ (1982) Interim analyses for randomized clinical trials: the group sequential approach. Biometrics 38(1):153–162

Pocock S, White I (1999) Trials stopped early: too good to be true? The Lancet 353(9157):943–944

President's Council of Advisors on Science and Technology (2012) Report to the President on propelling innovation in drug discovery, development, and evaluation. https://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-fda-final.pdf. Accessed 12 Oct 2016

Rajapakse A, Titchener-Hooker NJ, Farid SS (2005) Modeling of the biopharmaceutical drug development pathway and portfolio management. Comput Chem Eng 29(6):1357–1368

Rogers MS, Chang AM, Todd S (2005) Using group-sequential analysis to achieve the optimal sample size. BJOG Int J Obstet Gynaecol 112(5):529–533

Rojas-Cordova A, Bish EK (2018) Optimal patient enrollment in sequential adaptive clinical trials with binary response, Working paper

Rojas-Cordova A, Hosseinichimeh N (2018) Trial termination and drug misclassification in sequential adaptive clinical trials. Service Science 10(3):354–377

Sebille V, Bellissant E (2000) Comparison of four sequential methods allowing for early stopping of comparative clinical trials. Clin Sci 98(5):569–578

Solak S, Clarke JPB, Johnson EL, Barnes ER (2010) Optimization of R&D project portfolios under endogenous uncertainty. Eur J Oper Res 207(1):420–433

Sun E, Lakdawalla D, Reyes C, Goldman D, Philipson T, Jena A (2008) The determinants of recent gains in cancer survival: an analysis of the Surveillance, Epidemiology, and End Results (SEER) database. J Clin Oncol 26(15):6616–6616

Todd S, Whitehead A, Stallard N, Whitehead J (2001) Interim analyses and sequential designs in phase III studies. Br J Clin Pharmacol 51(5):394–399

US Congress (2016) H.R.34 - 21st century cures act. https://www.congress.gov/bill/114th-congress/house-bill/34/. Accessed 21 Jan 2018

US Food and Drug Administration (2010) Guidance for industry: adaptive design clinical trials for drugs and biologics. Food and Drug Administration, Washington DC

Whitehead J (1997) The design and analysis of sequential clinical trials. Wiley, Chichester

Whitehead J (2002) Sequential methods in clinical trials. Seq Anal 21(4):285–308

Whitehead J (2004) Stopping clinical trials by design. Nat Rev Drug Discov 3(11):973–977

Ye F (2008) Design and analysis of phase III clinical trials. https://medschool.vanderbilt.edu/cqs/files/cqs/media/2008Jun2008Fei.pdf. Accessed 28 Sept 2016

Zannad F, Stough WG, McMurray JJ, Remme WJ, Pitt B, Borer JS, Geller NL, Pocock SJ (2012) When to stop a clinical trial early for benefit: lessons learned and future approaches. Circ Heart Fail 5(2):294–302

Alba C. Rojas-Cordova is an Assistant Professor in the Department of Engineering Management, Information, and Systems at Southern Methodist University. She earned her Ph.D. and master's degrees in Industrial and Systems Engineering from Virginia Tech in 2017 and 2011, respectively, and her bachelor's degree in Industrial Engineering from the Bolivian Catholic University (Universidad Catolica Boliviana) in La Paz, Bolivia in 2007. Between 2011 and 2013, she worked as a production planner in a manufacturing company.

Alba decided to pursue studies in Engineering because of her strong interest in mathematics and physics, and chose Industrial Engineering for her undergraduate and graduate degrees because of the versatility of the field. She worked in the textile, food, and manufacturing industries in her home country, Bolivia, and in the United States. Her passion for research and teaching took her back to graduate school to complete a Ph.D. degree, which she obtained under the supervision of Dr. Ebru Bish and Dr. Niyousha Hosseinichimeh. These two remarkable women made a very big difference in Alba's life and career, by providing her with guidance, support, and encouragement to continue looking for solutions to impactful and complex problems arising in healthcare and public policy.

Dr. Ebru K. Bish is an associate professor of Industrial and Systems Engineering and an associate professor of Health Sciences at Virginia Tech. Dr. Bish's research interests lie in stochastic modeling, optimization, and decision-making under uncertainty, with applications to public health policy and health implementation science. Her specific research focuses on public health screening and surveillance of infectious diseases and genetic disorders; and on improving the safety of health care delivery. She is the recipient of the INFORMS Pierskalla Award for the Best Paper in Healthcare (first prize winner, runner-up, and a finalist), INFORMS JFIG Best Paper Award, and IIE Transactions Best Applications Paper Award. She is also the 2018–2019 Vice President/President-Elect of the INFORMS Health Applications Society.

Ebru's keen interest in mathematics and probability inspired her to choose engineering as her career path. She chose industrial engineering because of the strong emphasis the discipline places on probability and optimization, and due to the ability to work in non-traditional engineering domains, including healthcare and public policy decision-making. Ebru is especially passionate about utilizing Operations Research modeling and methodologies for the benefit of the society, and in particular, for improving the fairness, quality, efficiency, and effectiveness of healthcare delivery through improving healthcare systems and public policy.

Dr. Niyousha Hosseinichimeh is an assistant professor at the Department of Industrial and Systems Engineering at Virginia Tech. Her research focuses on systems modeling of complex public health problems. The applications of her studies include major depressive disorder and infant mortality. She uses system dynamics and statistical approaches in her studies. Her research has been funded by the National Institute of Health and the Ohio State Department of Health.

She was born in Iran and her inspiration to rebuild the country after the war, as well as her interest in mathematics and physics inspired her to choose mechanical engineering at Sharif University of Technology. After graduation, Niyousha joined a major consulting engineering company in Iran and contributed to two national dam construction projects. Her 3 years of experience as a mechanical engineer in the company and observing unintended consequences of dam constructions on Iran's environment showed her the importance of using systems perspective in designing public policies. So she came to the US to continue her education in system dynamics at the State University of New York at Albany. From then, she has focused on applying systems science on health issues with the goal of reducing the burden of diseases.

Chapter 15
Calibration Uncertainty and Model-Based Analyses with Applications to Ovarian Cancer Modeling

Jing Voon Chen and Julia L. Higle

Contents

15.1 Introduction 347
15.2 Model-Based Analysis for Ovarian Cancer 348
15.2.1 A History of Ovarian Cancer Models 349
15.2.2 Summary 351
15.3 Natural History Model for Ovarian Cancer 351
15.3.1 Data: Sources and Characteristics 352
15.3.2 Model Structure 352
15.3.3 Modeled Outcomes 355
15.3.4 Validity Conditions 356
15.3.5 Model-Based Calibration for Ovarian Cancer 356
15.4 Results 358
15.5 Conclusions 361
Appendix 1: Ovarian Cancer Data 362
Appendix 2: Validity Conditions 364
References 366

15.1 Introduction

Medical decision making (MDM) studies typically include comparative analyses of disease screening or treatment options in order to understand the relative costs and benefits of various strategies and programs under consideration. Randomized control trials (RCTs) are considered the gold standard for such purposes because the actual effects of different medical interventions across a representative segment of the population are observed and analyzed. However, RCTs are typically cost- and time-intensive, and the number of interventions that can be tested within a trial is

J. V. Chen · J. L. Higle (✉)
Daniel J. Epstein Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, CA, USA
e-mail: higle@usc.edu

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_15

limited. As a result, model-based analyses for MDM can be appealing as surrogates for RCTs.

Model-based analysis for MDM relies heavily upon a natural history (NH) model, which represents disease progression and regression in the absence of interventions. An NH model requires the specification of various model parameters, some of which may not be observable, and generates modeled outcomes that lead to a representation of the costs and benefits of interventions. The NH model can also be used to calculate modeled outcomes that can be compared to observed data from epidemiological and clinical studies. Because it serves as a replacement for direct observation, it is essential to ensure that an NH model is consistent with such data.

The process of selecting model parameters that provide consistency between modeled outcomes and observed data is known as calibration. In order to assess interventions such as screening, vaccination, etc., an NH model will necessarily incorporate phases of a disease that occur prior to diagnosis, and data associated with these phases are typically not available. This gives rise to a phenomenon known as "calibration uncertainty," where distinct sets of model parameters offer consistency with observed data while providing decidedly different representations of unobservable phases. Because model parameters influence comparative analyses, insufficient examination of the impact of calibration uncertainty on model recommendations can lead to misleading interpretations of the conclusions drawn.

In this chapter, we present an approach to model-based analysis for MDM that systematically examines the breadth of models that can plausibly represent the disease. We illustrate the approach within the context of ovarian cancer, with a particular emphasis on the corresponding variability of modeled outcomes that might impact typical comparative analyses of methods for early detection.

15.2 Model-Based Analysis for Ovarian Cancer

Ovarian cancer is the fifth deadliest cancer for females, with 22,500 estimated new cases and 14,000 estimated deaths in 2017 (American Cancer Society 2017). While a small fraction (15%) of the cases are diagnosed at the localized stage, where the 5-year survival rate is 92%, the majority of the cases (60%) are diagnosed at the distant stage, where the 5-year survival rate is only 29% (American Cancer Society 2017). Although early detection often increases treatment options and improves survival outlook for cancer patients, the U.S. Preventive Services Task Force (USPSTF) does not recommend screening for ovarian cancer (Moyer 2012).

The strongest empirical evidence of the effectiveness of cancer screening strategies often comes from RCTs. Two large RCTs involving ovarian cancer appear in the literature: the Prostate, Lung, Colorectal and Ovarian study (PLCO) (Buys et al. 2011) and the United Kingdom Collaborative Trial of Ovarian Cancer Screening study (UKCTOCS) (Menon et al. 2009). PLCO is a United States based RCT targeting women from age 55 to 74, which considers screening based

on either a biomarker (CA 125) or imaging (transvaginal ultrasound, TVS). The study concludes that the screening regimens examined did not reduce mortality from ovarian cancer. UKCTOCS is a United Kingdom based RCT targeting post-menopausal women from age 50 to 74, which concludes that screening that combines CA 125 and TVS may reduce ovarian cancer mortality (Jacobs et al. 2016).

15.2.1 A History of Ovarian Cancer Models

The literature also includes a very small number of models of the natural history of ovarian cancer. Skates and Singer (1991) present a stochastic simulation model designed to evaluate the potential benefit of using CA 125 radioimmunoassay to screen for ovarian cancer. Given the focus on a screening program, a representation of the disease prior to diagnosis is necessary. Thus, their model consists of four components:

- a model of the natural history of ovarian cancer,
- a model of the time of clinical detection of ovarian cancer,
- a model of the survival probability for each cancer stage, and
- a model of the screening program.

The natural history of ovarian cancer is modeled using the standard four-stage cancer staging system. The joint log-normal distribution of the durations of the four stages is represented as a function of two variables: the mean duration in stage I and the coefficient of variation of the duration of each stage, which is assumed to be constant across all stages. The survival distributions are differentiated by stage at diagnosis, and are derived via maximum likelihood estimations based on data from the Massachusetts General Hospital tumor registry. The resulting model is used to assess a comparison of post-diagnosis survival with, and without, a screening strategy based on CA 125 serum levels.

Urban et al. (1997) extend the model developed by Skates and Singer (1991) and evaluate six ovarian cancer screening strategies in terms of their efficacy and cost-effectiveness. Their natural history model is based on the four-stage log-normal model of Skates and Singer (1991). Their model for clinical detection assigns each case an age and stage at diagnosis according to the age- and stage-specific distributions given by the Surveillance, Epidemiology, and End Results Program (SEER) of the National Cancer Institute (NCI). Departing from Skates and Singer, the age- and stage-specific survival distributions are estimated by applying the Kaplan-Meier method to the SEER data. Six screening protocols involving TVS and/or CA 125 assay are evaluated.

Drescher et al. (2012) refine the model of Urban et al. (1997) by introducing hypothetical biomarker and imaging tests. Parameter estimates for the natural history model are derived from the SEER database, the U.S. Vital Statistics Report, and literature (Yabroff et al. 2008; Anderson et al. 2010; Katsube et al. 1982;

Havrilesky et al. 2011; Partridge et al. 2009), except that the malignant durations are point estimates provided by gynecologic oncologists. A hypothetical cohort of 1 million women are screened annually from ages 45 to 85 using a multimodal screening test with the first-line test being either CA 125 or a hypothetical biomarker assay, and the second-line test being either TVS or a hypothetical imaging test. The survival component assigns a time of death for all clinically diagnosed cases and screen detected cases, derived from the age-, stage-, histology-, and grade-specific survival data obtained from SEER.

Schapira et al. (1993) conduct an analysis of the effectiveness of a one-time ovarian cancer screening with CA 125 and TVS in a cohort of 40-year-old women in the United States. Unlike the approaches that developed from Skates and Singer (1991), Schapira et al. (1993) use a simplified decision tree to compare "no screening" to one-time screening with CA 125 and TVS at age 40. The disease model involves only:

- ovarian cancer prevalence,
- proportion of prevalent cases that are in the early stages at the time of screening, and
- probability of early-stage disease diagnosed clinically in the absence of screening.

The prevalence of ovarian cancer in 40-year-old women is estimated as the product of age-specific incidence (based on Cutler and Young (1975)) and the average duration of the preclinical disease phase (assumed to be 2 years). The time it takes to progress from early- to late-stage disease before clinical diagnosis is assumed to be 1 year. Associated with each branch of the decision tree is a terminal node where appropriate remaining life expectancy is assigned to each terminal node.

Also departing from the models that developed from Skates and Singer (1991), Havrilesky et al. (2008) develop a discrete-time Markov chain model of the natural history of ovarian cancer and study the cost-effectiveness of potential screening strategies. The model consists of 13 health states, including well, undiagnosed stage I–IV ovarian cancer, diagnosed stage I–IV ovarian cancer, benign oophorectomy (surgical removal of ovaries without a priori evidence of disease), ovarian cancer survivor, death from ovarian cancer, and death from other causes. All patients enter the model at age 20 and die by age 100. Data obtained from SEER that aids calibration includes the lifetime probability of developing ovarian cancer, stage distribution at diagnosis, and lifetime probability of death from ovarian cancer. Age-specific U.S. Life Tables are used to estimate age-specific probability of death from other causes. Stationary transition probabilities between pairs of cancer stages that are consistent with the SEER data are obtained by searching over clinically justifiable values. Age-specific probabilities of benign oophorectomy and mortality from oophorectomy are estimated from literature data (Keshavarz et al. 2002; Merrill 2006; Wingo et al. 1985) following model calibration. Imposed on their natural history model are several hypothetical screening strategies, including no screening and screening at intervals of 3–36

months. Two scenarios are examined in their study: screening within the general population and within a high-risk population (simulated based on the prevalence of having a risk factor and the relative risk of ovarian cancer). The study concludes that annual screening is potentially cost effective, especially in a high-risk population.

Havrilesky et al. (2011) extend their original model (Havrilesky et al. 2008) to include two distinct ovarian cancer phenotypes, referred to as "aggressive" and "indolent." Model parameters are estimated using methods described in Havrilesky et al. (2008). Following the selection of model parameters, age-specific probabilities of benign oophorectomy and mortality from oophorectomy are estimated from literature data (Keshavarz et al. 2002; Merrill 2006; Wingo et al. 1985). Screening strategies evaluated in this study include no screening and the use of a hypothetical screening test at intervals of 3–36 months.

In an effort to estimate the duration of the preclinical phase of serous ovarian cancer, Brown and Palmer (2009) examine prophylactic salpingo-oophorectomy specimens. A broad range of techniques are used to create a tumor-growth model for early- and advanced-stage tumors. Simulated tumors are then subjected to hypothetical screening tests in order to assess a relationship between tumor size at diagnosis and the benefits of early detection.

15.2.2 Summary

Because screening is undertaken in an effort to detect latent cancers, accurate descriptions of the latent period of the natural history are key to designing effective tools to study screening protocols. However, the lack of longitudinal data from large cohort studies such as RCTs renders the parameter estimation of the natural history model of ovarian cancer difficult. The lack of longitudinal data also challenges the estimation of the window of opportunity for early detection. In the remainder of this chapter, we discuss model-based analyses that investigate the screening potential for ovarian cancer. Our modeling approach aims to characterize the set of models that can be considered to provide "plausible" representations of ovarian cancer. By examining the breadth of such models, we can explore the range of potential unobservable disease characteristics, as well as the range of potential outcomes for various interventions.

15.3 Natural History Model for Ovarian Cancer

Natural history model development requires identification of data sources and model structure, specification of validity conditions, and calibrated model parameters that satisfy validity conditions and yield consistency with available data. In this section, we discuss each of these in turn.

15.3.1 Data: Sources and Characteristics

The SEER program routinely collects data from, and provides cancer statistics for, the United States. Its database currently covers approximately 30% of the U.S. population, and is a common source of data for NH cancer models. SEER data includes patient demographics, tumor characteristics, stage at diagnosis, initial course of treatment, and post-diagnosis follow-up for vital status. SEER*Stat is statistical software that has been developed under the SEER program to serve as an interface for retrieval and analysis of the SEER data. Unless stated otherwise, all data used in this study were obtained via SEER*Stat.

Ovarian cancer data describing annual age- and stage-specific incidence for 2000 through 2014 was analyzed using SEER*Stat for all ages up to 85 and 3 stages: (I) localized, (II) regional, and (III) distant. A cursory examination of this data indicates that the distribution of patient age and cancer stage at diagnosis (i.e., incidence) exhibits minimal variation over this 15-year period (see Fig. 15.4 in Appendix 1). Accordingly, we developed our model using the aggregation of these data.

An analysis of data describing the 5-year survival data following diagnosis indicates that these survival probabilities are a function of the age and stage at diagnosis. That is, for a given stage at diagnosis, an older patient is less likely to survive the given number of years post diagnosis than a younger patient, as illustrated in Fig. 15.5 in Appendix 1.

Disease mortality data for all ages up to 95 are obtained from DevCan, another statistical software developed under the SEER program. All-cause mortality data is obtained via the National Center for Health Statistics (NCHS) of the Centers for Disease Control and Prevention (CDC) for all ages up to 100 (Arias et al. 2016). Both disease mortality and all-cause mortality data are age-dependent, which suggests that competing-risk mortality is as well.

SEER provides a type of prevalence data called the limited-duration prevalence, which is the prevalence of ovarian cancer cases that were diagnosed since 1975. This indicates that ovarian cancer patients who were diagnosed prior to 1975 and are still alive currently are excluded from the limited-duration prevalence data. To estimate the complete prevalence regardless of when the diagnosis occurs, SEER employs a statistical model called the completeness index that is based on the limited-duration prevalence, incidence, and survival data (Capocaccia and De Angelis 1997; Merrill et al. 2000). Because the complete prevalence data are modeled values, we decided against the use of both complete and limited-duration prevalence data. Hence, our analysis excluded ovarian cancer prevalence data.

15.3.2 Model Structure

We use a discrete-time Markov chain to represent the natural history of ovarian cancer based on nine health states, as follows:

- healthy (H), indicates that the patient is free from ovarian cancer;
- stage 1 ("localized"), 2 ("regional"), 3 ("distant") which may be
  - undiagnosed ($1U, 2U, 3U$), or
  - diagnosed ($1D, 2D, 3D$), and
- mortality due to
  - disease (DD), or
  - other causes (DO).

Given the lack of evidence that suggests otherwise, we model ovarian cancer as a strictly progressive disease. That is, our model excludes transitions from a disease state to healthy or to a less advanced disease state. In addition, a person in a diagnosed disease state remains within the diagnosed states until death. Recognizing that SEER data is likely to have the impact of post-diagnosis treatments embedded within the data, our model permits transition probabilities from undiagnosed disease stages to differ from those for diagnosed stages.

Our Markovian model of the natural history of ovarian cancer is structurally decomposed into three discrete-time components:

- a component that represents disease activation as an age-dependent process,
- a component that represents the process of disease progression following activation as a stationary discrete-time Markov chain (DTMC), and
- a component that represents competing-risk mortality as an age-dependent process.

The cycle lengths for these component processes are 1 year. From one cycle to the next, a healthy female may remain healthy, transition to $1U$, or die from other causes. The transition from H to $1U$ is referred to as disease activation. Following activation, a female progresses according to the DTMC, or dies from competing risks. Patients enter the model in the healthy state at age 20.

Notation

- $\mathcal{S} = \{H, 1U, 2U, 3U, 1D, 2D, 3D, DD, DO\}$, the set of all health states.
- $\mathcal{D} = \{1D, 2D, 3D\}$, the set of all health states corresponding to a diagnosis of ovarian cancer.
- $\mathcal{U} = \{H, 1U, 2U, 3U\}$, the set of all undiagnosed health states.
- $\mathcal{A} = \{20, \ldots, 85\}$, the set of all ages for which patients are modeled.
- $\mathbb{P}(a) = [P_{i,j}(a)]$ = the transition probability matrix for a patient at age a. That is, $P_{i,j}(a)$ denotes the probability that a female in state $i \in \mathcal{S}$ at age $a \in \mathcal{A}$ will be in state $j \in \mathcal{S}$ at age $a + 1$.
- $\mathbb{P} = \{\mathbb{P}(a),\ a \in \mathcal{A}\}$.
- $p_{DO}(a)$ = the probability that a female at age a succumbs to competing risks by age $a + 1$.

Figure 15.1 illustrates the set of possible transitions. Since transitions into DO can occur from all states, these transitions are excluded from Fig. 15.1 in an effort to enhance visual clarity.

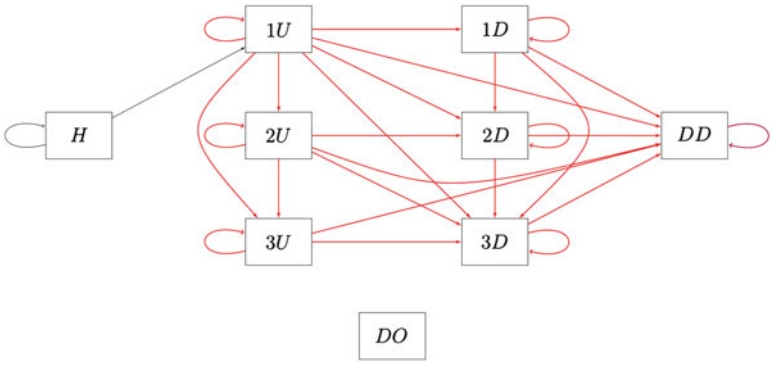


Fig. 15.1 Possible transitions for the ovarian cancer model (transitions to DO are not shown)

The nonstationary transition probability matrix, $\mathbb{P}(a)$, is represented as

$$\mathbb{P}(a) = \begin{array}{c} \\ H \\ 1U \\ 2U \\ 3U \\ 1D \\ 2D \\ 3D \\ DD \\ DO \end{array} \begin{array}{c} \begin{array}{ccccccccc} H & 1U & 2U & 3U & 1D & 2D & 3D & DD & DO \end{array} \\ \begin{bmatrix} P_{H,H}(a) & P_{H,1U}(a) & 0 & 0 & 0 & 0 & 0 & 0 & p_{DO}(a) \\ 0 & P_{1U,1U}(a) & P_{1U,2U}(a) & P_{1U,3U}(a) & P_{1U,1D}(a) & P_{1U,2D}(a) & P_{1U,3D}(a) & P_{1U,DD}(a) & p_{DO}(a) \\ 0 & 0 & P_{2U,2U}(a) & P_{2U,3U}(a) & 0 & P_{2U,2D}(a) & P_{2U,3D}(a) & P_{2U,DD}(a) & p_{DO}(a) \\ 0 & 0 & 0 & P_{3U,3U}(a) & 0 & 0 & P_{3U,3D}(a) & P_{3U,DD}(a) & p_{DO}(a) \\ 0 & 0 & 0 & 0 & P_{1D,1D}(a) & P_{1D,2D}(a) & P_{1D,3D}(a) & P_{1D,DD}(a) & p_{DO}(a) \\ 0 & 0 & 0 & 0 & 0 & P_{2D,2D}(a) & P_{2D,3D}(a) & P_{2D,DD}(a) & p_{DO}(a) \\ 0 & 0 & 0 & 0 & 0 & 0 & P_{3D,3D}(a) & P_{3D,DD}(a) & p_{DO}(a) \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array} \tag{15.1}$$

Unknown parameter values are:

- $p_{DO}(a), a \in \mathcal{A}$
- $P_{HH}(a), a \in \mathcal{A}$, and
- $P_{ij}(a), i \notin \{H, DD, DO\}, j \notin \{H, DO\}, a \in \mathcal{A}$.

There are (24+2)*65=1690 parameters in $\mathbb{P}$ whose values must be determined.

We incorporate three simplifications that serve to streamline the effort required to identify plausible model parameters. We represent the process of disease progression following activation as a stationary process. Accordingly,

$$P_{i,j}(a) = (1 - p_{DO}(a))\, P_{i,j} \quad i \notin \{H, DD, DO\},\ j \notin \{H, DO\}. \tag{15.2}$$

We model disease activation, $P_{H,1U}(a)$, as a piecewise linear function of age:

$$P_{H,1U}(a) = \begin{cases} 0 & \text{if } a \leq 30 \\ \beta(a-30) & \text{if } a \in (30, 75] \\ 45\beta & \text{if } a \in (75, 85]. \end{cases} \tag{15.3}$$

Finally, we define competing-risk mortality, $p_{DO}(\cdot)$, as the difference between the observed disease mortality and all-cause mortality. As a result of (15.2) and (15.3), the set of unknown parameter values is reduced to

- β
- $P_{ij}, i \notin \{H, DD, DO\}, j \notin \{H, DO\}$.

Note that with these simplifications, we have 24+1 = 25 variables, a significant reduction from 1690 variables in the original presentation of the transition matrices.

15.3.3 Modeled Outcomes

Transitions among health states over time result in modeled trends and tendencies governed by the laws of probability. These modeled values can be compared to available data when determining whether or not a given set of model parameters, $\mathbb{P}$, "fits" the available data. In modeling incidence by age and stage, we define

- $\alpha_i(a; \mathbb{P})$ = the probability that a patient receives an initial diagnosis in state $i \in \mathcal{D} \cup \{DD\}$ at age $a \in \mathcal{A}$.

Then

$$\alpha_i(a+1; \mathbb{P}) = \sum_{k \in \mathcal{U}} \pi_k(a; \mathbb{P}) P_{k,i}(a) \qquad \forall i \in \mathcal{D}, a \in \mathcal{A} \tag{15.4}$$

$$\alpha_{DD}(a+1; \mathbb{P}) = \sum_{k \in \mathcal{U} \cup \mathcal{D}} \pi_k(a; \mathbb{P}) P_{k,DD}(a) \quad \forall a \in \mathcal{A}, \tag{15.5}$$

where vectors $\{\pi(a; \mathbb{P}), a \in \mathcal{A}\}$ are calculated as

- $\pi_i(20; \mathbb{P}) = 1$ if $i = H$, and 0 otherwise
- $\pi(a+1; \mathbb{P}) = \pi(20; \mathbb{P}) \prod_{t=20}^{a} \mathbb{P}(t) \quad \forall a \in \mathcal{A}$.

Equations (15.4) and (15.5) consider incidence within a cohort, and provide a basis for comparing outcomes derived from the model to observations from data sources described in Sect. 15.3.1.

Finally, the n-year survival probability following diagnosis at age $a \in \mathcal{A}$ and state $i \in \mathcal{D}$, $S(i, a, n; \mathbb{P})$, is readily calculated as

$$S(i, a, n; \mathbb{P}) = \sum_{j \neq DD, DO} \Big[\prod_{t=a}^{a+n-1} \mathbb{P}(t) \Big]_{i,j}. \tag{15.6}$$

15.3.4 Validity Conditions

We impose limitations on model parameters, including transition probabilities and the slope parameter in the activation process, so that models that do not satisfy these restrictions are eliminated from consideration. For example, transition probabilities must follow basic probability laws; they are nonnegative and each row of each transition matrix must sum to one. In addition, we impose constraints such that prior to diagnosis, progression is more likely than after diagnosis. This accounts for the effect of treatments that are contained within the post-diagnosis data. Other types of constraints include:

- The likelihood of a diagnosis increases with the severity of the health state;
- Progression is more likely when an individual is in a more severe health state than in a less severe one;
- The likelihood of progression to a health state decreases with the severity of the state;
- The slope defining the activation process is confined to be strictly positive with an upper bound that ensures that basic probability laws are satisfied.

The full statement of the validity conditions for our ovarian cancer study appears in Appendix 2.

15.3.5 Model-Based Calibration for Ovarian Cancer

We present the task of identifying valid and well-calibrated parameter sets as an optimization problem. Similar to the modeling framework discussed in Chen et al. (2018), we define an objective function that represents the deviation between modeled outcomes and calibration targets obtained from SEER. We differentiate modeled outcomes from calibration targets by using a "hat" for the targets. Thus, while $\alpha_i(a; \mathbb{P})$ denotes the modeled value for incidence for state $i \in \mathcal{D}$ at age $a \in \mathcal{A}$, $\hat{\alpha}_i(a)$ represents the corresponding target value. Our computational work in this chapter uses the total weighted absolute deviation between modeled outcomes and calibration targets, as a function of the set of model parameter values. Our objective function is the weighted sum of the total deviations associated with the three types of targets used: incidence, disease mortality, and post-diagnosis survival. For a given set of parameter values, $\mathbb{P}$, the objective value is given by

$$\begin{aligned}\text{GOF}(\mathbb{P}) &= \sum_{t\in\{\text{inc},\ DD,\ S\}} w_t \text{GOF}_t(\mathbb{P}) \\ &= w_{\text{inc}} \sum_{a\in\mathcal{A}}\sum_{i\in\mathcal{D}} \left|\alpha_i(a;\mathbb{P}) - \hat{\alpha}_i(a)\right| + w_{DD} \sum_{a\in\mathcal{A}} \left|\alpha_{DD}(a;\mathbb{P}) - \hat{\alpha}_{DD}(a)\right| \\ &\quad + w_S \sum_{a\in\mathcal{A}}\sum_{i\in\mathcal{D}}\sum_{n} \left|S(i,a,n;\mathbb{P}) - \hat{S}(i,a,n)\right|\end{aligned}$$

where α and S are calculated as in (15.4)–(15.6). The coefficients, w_{inc}, w_{DD}, and w_S represent weights corresponding to the three types of calibration targets: incidence, disease mortality, and survival, respectively. These sets of target values have significant differences in magnitude. Thus, our computational work in this chapter uses weights that place the various types of targets on similar footing by using weights that are equal to the reciprocal of the mean of the target values.

We define the set $\mathbb{V}$ as the set of models, $\mathbb{P}$, that satisfy known validity conditions described in Sect. 15.3.4. Our calibration problem for ovarian cancer may now be stated as:

$$\begin{aligned}&\text{minimize GOF}(\mathbb{P}) \\ &\text{subject to } \mathbb{P} \in \mathbb{V}.\end{aligned} \tag{CP}$$

Note that $\mathbb{V}$ is a polyhedral set, while the objective function, GOF($\mathbb{P}$) is nonconvex, and requires numerous multi-step calculations. We use the numerical method of Nelder and Mead (1965) to identify solutions to (CP). This is a heuristic search method for nonlinear optimization that is widely used, but is not guaranteed to find an optimal solution to (CP). Accordingly, we solve (CP) repeatedly using randomized initializations. That is, when the Nelder-Mead method stabilizes at a particular set of model parameters, the parameters are retained and the search is restarted in a random fashion. The process is repeated several times, after which we examine the quality of fit to each type of target. A set of model parameters is considered to yield a plausible representation of the natural history of ovarian cancer when the modeled outcomes provide an acceptable fit for each type of target. Specifically, if T_i represents the sum of the targets of type i, plausible models satisfy

$$\frac{\text{GOF}_i(\mathbb{P})}{T_i} \leq c_1 \tag{15.7}$$

for each $i \in \{\text{inc}, DD, S\}$, and for at least two target types,

$$\frac{\text{GOF}_i(\mathbb{P})}{T_i} \leq c_2 \tag{15.8}$$

where $c_2 < c_1$. In combination, (15.7) and (15.8) ensure that model parameters that are considered to be plausible fit the modeled outcomes for all three target types

well, with at least two types being fit with a tighter tolerance. Within this study, we set $c_1 = 0.25$ and $c_2 = 0.2$.

15.4 Results

Using the approach described in Sect. 15.3.5 we identify approximately 3500 sets of model parameters, of which 150 are considered to yield plausible models of the natural history of ovarian cancer. With this collection of plausible models, we can develop a plausible range of modeled outcomes relevant to possible medical interventions. In this way, a sense of the broad set of projected outcomes is available.

For example, early detection programs can be most effective when they detect the presence of cancers during their preclinical phase. By definition, the preclinical phase is not observable. Its duration is unknowable, and the nature of this window of opportunity is subject to speculation. The collection of plausible models can shed some light on this. For each plausible model, we calculate the expected duration of the preclinical phase, which is defined as the expected duration from disease onset to a diagnosed disease state (i.e., the first passage time from $1U$ to $\{1D, 2D, 3D, DD\}$). This yields 150 plausible values of the expected duration of the preclinical phase. This set of values is summarized in Table 15.1. Note that 80% of the plausible mean durations fall between 4.1 years and 9.6 years, while the interquartile range is from 5.8 years to 8.3 years. In addition to the expected duration of the preclinical duration, there may be interest in understanding plausible ranges for various distributional characteristics. For example, associated with each plausible model is a cumulative distribution function (cdf) of the duration of the preclinical phase, from which ranges of percentiles can be obtained. Table 15.2 contains information regarding various percentiles that are typically of interest.

Despite the absence of recommended screening for ovarian cancer by the U.S. Preventive Services Task Force (USPSTF) (Moyer 2012), our results suggest a high likelihood of a multi-year window of opportunity for early detection. Using our plausible models, we assess the impact that a screening program might yield. We

Table 15.1 Mean duration of preclinical phase, percentiles from plausible models

10th Percentile	25th Percentile	Median	75th Percentile	90th Percentile
4.1 yrs	5.8 yrs	7.4 yrs	8.3 yrs	9.6 yrs

Table 15.2 Percentiles of the duration of the preclinical phase

Percentiles	Interquartile range from plausible models
25th	2.3–3.3 years
Median	4.1–6.1 years
75th	7.1–10.5 years

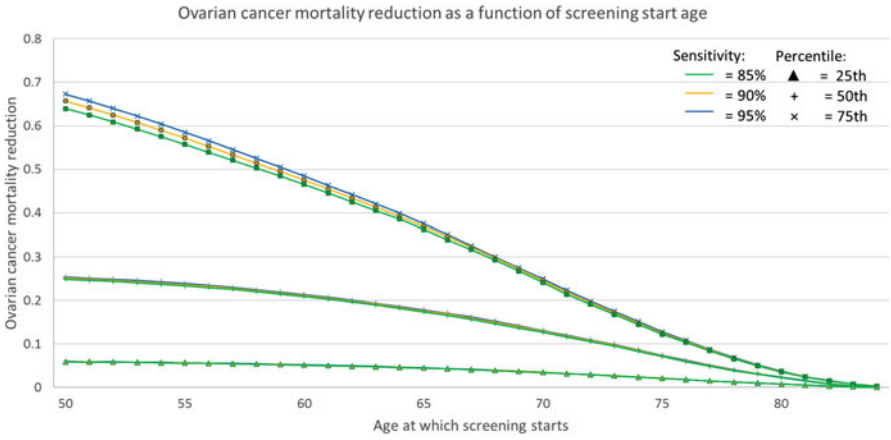


Fig. 15.2 Plausible mortality reduction as a function of the age at which screening is initiated

examine the opportunity for reductions in disease mortality as a function of the sensitivity of a hypothetical screening test and the age at which annual screening is initiated. The hypothetical screening program is two-phased, with a first-line screening test and a second-line diagnostic test. An individual with a positive screen result undertakes the second-line diagnostic test, after which her health state is known with certainty and she continues to transition according to $\mathbb{P}$. That is, a healthy individual with a false positive screen result will be correctly classified as healthy, while an individual in state iU who tests positive will be correctly diagnosed, placing her in state iD. This movement to state iD, which corresponds to an early diagnosis, causes the patient to progress and receive treatment as a diagnosed patient. Mortality reduction is defined as the reduction in the probability of dying from ovarian cancer by age 85 that is attributed to the screening program. The results for mortality reduction are illustrated in Fig. 15.2, where we see that the age at which annual screening is initiated exhibits a more strongly differentiated impact than the sensitivity of the screening test. This is especially true for the median and upper quartile obtained from the collection of plausible models than for the lower quartile. Not surprisingly, the effect on mortality reduction diminishes as the age at which screening is initiated increases.

Figure 15.2 illustrates mortality reduction as a result of early diagnosis, based in part on post-diagnosis outcomes embedded within SEER data. Figure 15.3 illustrates the manner in which the age and stage of cancers, including cancers that are currently latent, would be diagnosed with the hypothetical screening program. Using two plausible models selected from the set of 150 such models, we illustrate the changes in diagnostic states at various ages as a function of the age at which screening is initiated. Close inspection of the side-by-side graphs reveals differences between the two models when screening is introduced. Overall, we see a stark

Fig. 15.3 Evidence of early detection using age and stage at diagnosis (two plausible models)

increase in diagnoses with early introduction of screening, especially in the early stages. Information such as this would be useful in assessing the potential range of impact when new treatment protocols are under consideration.

15.5 Conclusions

We present a Markov model for ovarian cancer that is decomposed into separate processes representing disease activation, disease progression, and competing-risk mortality. Exploitation of our structural decomposition of the model of the natural history of ovarian cancer leads to a significant reduction in the number of parameters involved in the calibration process, thereby streamlining the computational challenges associated with identifying solutions to (CP). Additionally, we model disease activation as a piecewise linear function, which further reduces the number of parameters required to model disease activation. While a fully nonstationary model might appear to be more realistic and desirable, such a model will overwhelm search methods and exacerbate the magnitude of calibration uncertainty. The net result would likely render the analysis unreliable.

We illustrate how model-based analysis can help shed some light on our understanding of the duration of the preclinical phase for ovarian cancer, an unknowable value that has a significant impact on the potential efficacy of a screening program. We estimate that the mean duration is likely to fall between 4.1 years and 9.6 years. This range includes the estimate from Brown and Palmer (2009) (i.e., 5.1 years), excludes that of Skates and Singer (1991) (i.e., 1.3 years), and excludes the assumption of Cutler and Young (1975) (i.e., 1 year).

Using our plausible models to assess the efficacy of a hypothetical annual screening program initiated at age 50, we suggest a median of approximately 25% in mortality reduction across all first-line test sensitivities when the second-line diagnostic test is used. Among our models, the 25th percentile of mortality reduction is approximately 6%, whereas the 75th percentile is approximately 65%. Havrilesky et al. (2008) use model-based analysis to conclude that an annual screening initiated at age 50, without a second-line diagnostic test, would result in 43% reduction in ovarian cancer mortality. This corresponds to the 65th percentile among values observed from our plausible models, although the differences in the diagnostics test make a direct comparison difficult.

Our work is not without limitations. The disease activation process is modeled as a piecewise linear function that eliminates activation prior to age 30. Although the activation process is inherently age-dependent, the assumption that ovarian cancer activates according to a linear fashion after age 30 might be overly rigid and simplistic. Nevertheless, this process is unknowable and ovarian cancer is rarely diagnosed before age 40 (Boyd et al. 2000; Permuth-Wey and Sellers 2009). As a future direction, we will modify the piecewise linear function to examine the impact of this assumption on modeled outcomes.

Another limitation is that in modeling the general population, we did not create a risk-differentiated model. In addition to age, genetic/hereditary conditions are major risk factors (Chambers and Hess 2008; Schorge et al. 2010), as activation is influenced by these factors. Our future research includes an adaptation of our model to incorporate risk-differentiated activation, which will allow us to examine the potential of risk-differentiated early detection strategies.

Appendix 1: Ovarian Cancer Data

Figures 15.4 illustrates the boxplots of the age and stage at diagnosis from years 2000–2014, where the combined data for all 15 years are included at the last positions on the subplots. Figure 15.5 illustrates survival distributions as a function of the age and stage at diagnosis. These are clearly age-dependent, for all stages.

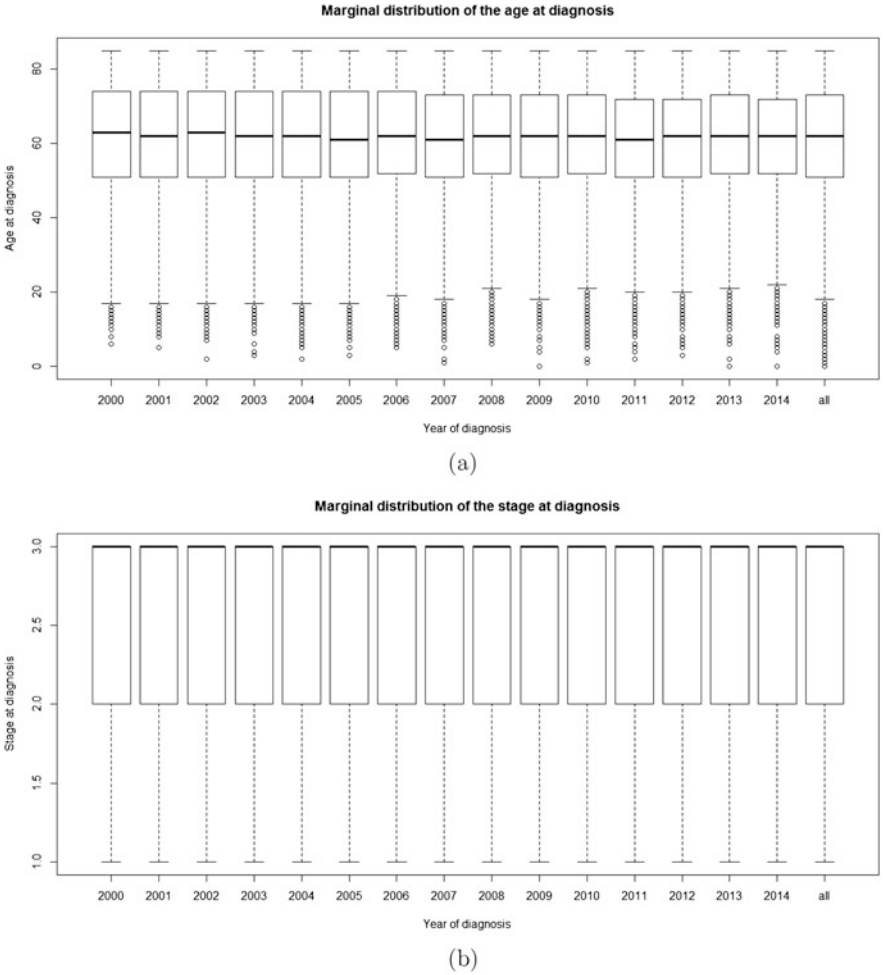


Fig. 15.4 The boxplots of the age and stage at diagnosis from years 2000–2014. (**a**) The boxplot of the age at diagnosis across years 2000–2014. (**b**) The boxplot of the stage at diagnosis across years 2000–2014. A cursory examination of these marginal distributions indicates minimal variation over the 15-year period. Accordingly, the models in this paper were developed using the aggregated data, which appears as "all"

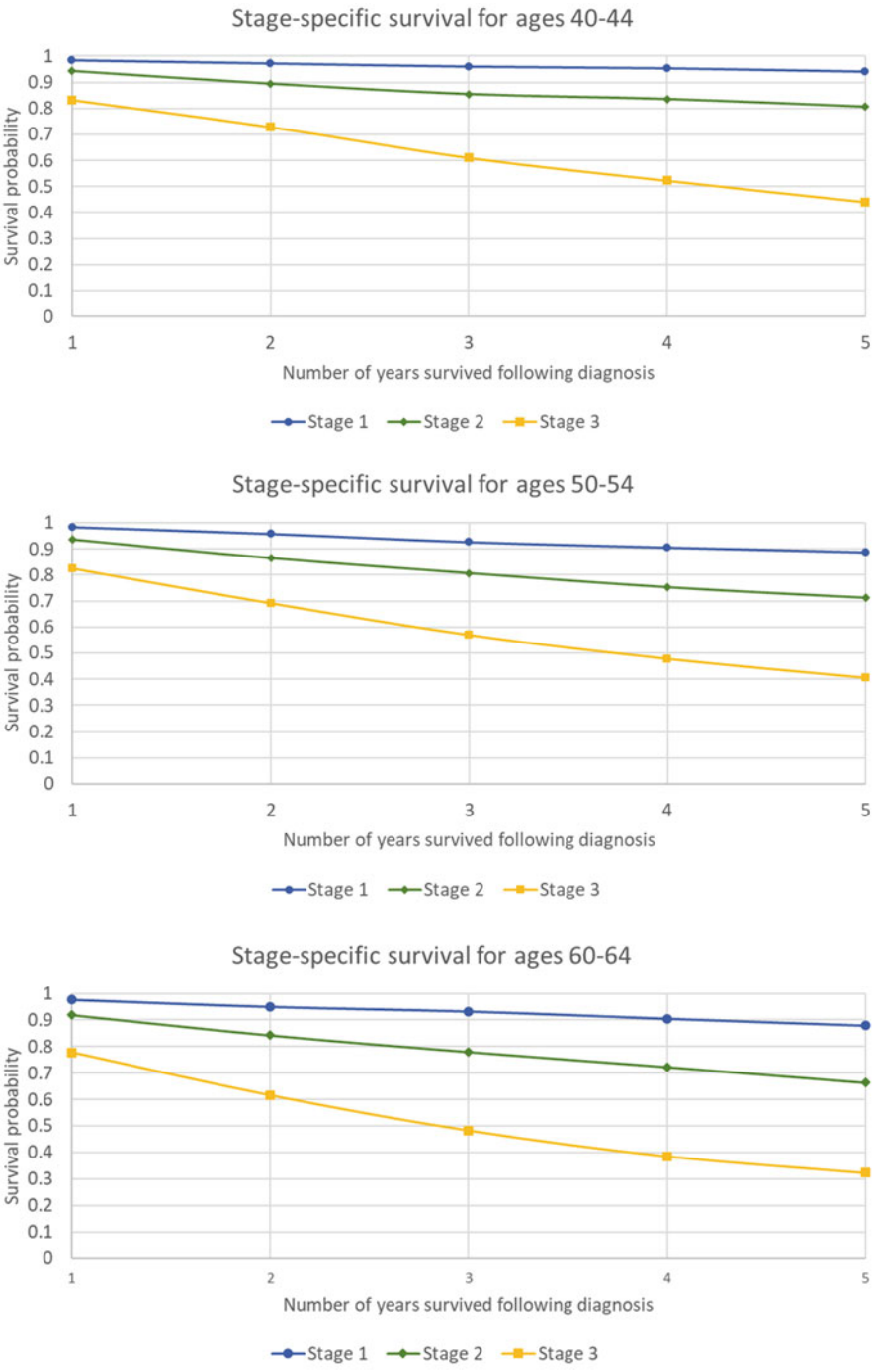


Fig. 15.5 Nonstationarity in post-diagnosis survival

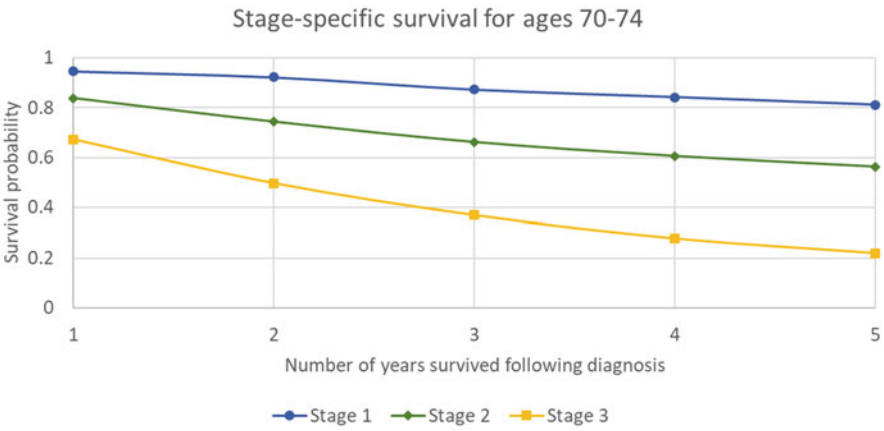


Fig. 15.5 (continued)

Appendix 2: Validity Conditions

$$P_{1U,2U} \geq P_{1D,2D} \tag{15.9}$$

$$P_{1U,3U} \geq P_{1D,3D} \tag{15.10}$$

$$P_{1U,DD} \geq P_{1D,DD} \tag{15.11}$$

$$P_{2U,3U} \geq P_{2D,3D} \tag{15.12}$$

$$P_{2U,DD} \geq P_{2D,DD} \tag{15.13}$$

$$P_{3U,DD} \geq P_{3D,DD} \tag{15.14}$$

$$P_{2U,3U} \geq P_{1U,2U} \tag{15.15}$$

$$P_{3U,DD} \geq P_{2U,3U} \tag{15.16}$$

$$P_{2D,3D} \geq P_{1D,2D} \tag{15.17}$$

$$P_{3D,DD} \geq P_{2D,3D} \tag{15.18}$$

$$P_{1U,2U} \geq P_{1U,3U} \tag{15.19}$$

$$P_{1U,3U} \geq P_{1U,DD} \tag{15.20}$$

$$P_{1D,2D} \geq P_{1D,3D} \tag{15.21}$$

$$P_{1D,3D} \geq P_{1D,DD} \tag{15.22}$$

$$P_{2U,3U} \geq P_{2U,DD} \tag{15.23}$$

$$P_{2D,3D} \geq P_{2D,DD} \tag{15.24}$$

$$P_{1U,1D} \geq P_{1U,2D} \tag{15.25}$$

$$P_{1U,2D} \geq P_{1U,3D} \tag{15.26}$$

$$P_{1U,3D} \geq P_{1U,DD} \tag{15.27}$$

$$P_{2U,2D} \geq P_{2U,3D} \tag{15.28}$$

$$P_{2U,3D} \geq P_{2U,DD} \tag{15.29}$$

$$P_{2U,DD} \geq P_{1U,DD} \tag{15.30}$$

$$P_{3U,DD} \geq P_{2U,DD} \tag{15.31}$$

$$P_{2D,DD} \geq P_{1D,DD} \tag{15.32}$$

$$P_{3D,DD} \geq P_{2D,DD} \tag{15.33}$$

$$P_{2U,3D} \geq P_{1U,3D} \tag{15.34}$$

$$P_{3U,3D} \geq P_{2U,3D} \tag{15.35}$$

$$P_{2U,2D} \geq P_{1U,2D} \tag{15.36}$$

$$P_{3U,3D} \geq P_{3U,DD} \tag{15.37}$$

$$P_{2U,2D} \geq P_{1U,1D} \tag{15.38}$$

$$P_{3U,3D} \geq P_{2U,2D} \tag{15.39}$$

$$p_{DO}(a) = \hat{p}_{DO}(a) \qquad \forall a \in \mathcal{A} \tag{15.40}$$

$$\sum_{j \in \mathcal{S}} P_{i,j}(a) = 1 \qquad \forall i \in \mathcal{S}, a \in \mathcal{A} \tag{15.41}$$

$$0 \leq P_{i,j}(a) \leq 1 \qquad \forall i, j \in \mathcal{S}, a \in \mathcal{A} \tag{15.42}$$

$$0 \leq \beta \leq \frac{1 - p_{DO}(85)}{45} \tag{15.43}$$

Equations (15.9)–(15.14) state that treatment helps slow down progression. Equations (15.15)–(15.18) impose that progression is more likely when the person is in a more severe health state. Equations (15.19)–(15.39) require that progressing to a more severe health state is less likely than to a less severe health state. Equation (15.40) sets $p_{DO}(a)$ as the corresponding observed value. Equations (15.41)–(15.42) refer to the basic probability laws, whereas (15.43) ensures that the slope defining the disease activation process is strictly positive and is at most $\frac{1 - p_{DO}(85)}{45}$. This is derived from the fact that $\hat{p}_{DO}(a)$ is a monotonically decreasing function of a, and that (15.41) must be satisfied.

References

American Cancer Society (2017) Cancer facts & figures 2017. American Cancer Society, Atlanta

Anderson GL, McIntosh M, Wu L, Barnett M, Goodman G, Thorpe JD, Bergan L, Thornquist MD, Scholler N, Kim N et al (2010) Assessing lead time of selected ovarian cancer biomarkers: a nested case–control study. J Natl Cancer Inst 102(1):26–38

Arias E, Heron M, Xu J (2016) United states life tables, 2012. Natl Vital Stat Rep 65(8):14–15

Boyd J, Sonoda Y, Federici MG, Bogomolniy F, Rhei E, Maresco DL, Saigo PE, Almadrones LA, Barakat RR, Brown CL et al (2000) Clinicopathologic features of BRCA-linked and sporadic ovarian cancer. JAMA 283(17):2260–2265

Brown PO, Palmer C (2009) The preclinical natural history of serous ovarian cancer: defining the target for early detection. PLoS Med 6(7):1–14

Buys SS, Partridge E, Black A, Johnson CC, Lamerato L, Isaacs C, Reding DJ, Greenlee RT, Yokochi LA, Kessel B, et al (2011) Effect of screening on ovarian cancer mortality: the prostate, lung, colorectal and ovarian (PLCO) cancer screening randomized controlled trial. JAMA 305(22):2295–2303

Capocaccia R, De Angelis R (1997) Estimating the completeness of prevalence based on cancer registry data. Stat Med 16(4):425–440

Chambers SK, Hess LM (2008) Ovarian cancer prevention. In: Alberts DS, Hess LM (eds) Fundamentals of cancer prevention, Chap. 17. Springer, Berlin, pp 447–473

Chen JV, Higle JL, Hintlian M (2018) A Systematic approach for examining the impact of calibration uncertainty in disease modeling. Comput Manage Sci 15:541–561

Cutler SJ, Young JL (eds) (1975) Third national cancer survey: incidence data, national cancer institute monograph 41, DHEW Publication No. (NIH) 75-787. U.S. Government Printing Office, Washington, DC

Drescher CW, Hawley S, Thorpe JD, Marticke S, McIntosh M, Gambhir SS, Urban N (2012) Impact of screening test performance and cost on mortality reduction and cost-effectiveness of multimodal ovarian cancer screening. Cancer Prev Res 5(8):1015–1024

Havrilesky LJ, Sanders GD, Kulasingam S, Myers ER (2008) Reducing ovarian cancer mortality through screening: is it possible, and can we afford it? Gynecol Oncol 111(2):179–187

Havrilesky LJ, Sanders GD, Kulasingam S, Chino JP, Berchuck A, Marks JR, Myers ER (2011) Development of an ovarian cancer screening decision model that incorporates disease heterogeneity. Cancer 117(3):545–553

Jacobs IJ, Menon U, Ryan A, Gentry-Maharaj A, Burnell M, Kalsi JK, Amso NN, Apostolidou S, Benjamin E, Cruickshank D et al (2016) Ovarian cancer screening and mortality in the UK collaborative trial of ovarian cancer screening (UKCTOCS): a randomised controlled trial. Lancet 387(10022):945–956

Katsube Y, Berg J, Silverberg S (1982) Epidemiologic pathology of ovarian tumors: a histopathologic review of primary ovarian neoplasms diagnosed in the Denver Standard Metropolitan Statistical Area, 1 July–31 December 1969 and 1 July–31 December 1979. Int J Gynecol Pathol 1(1):3–16

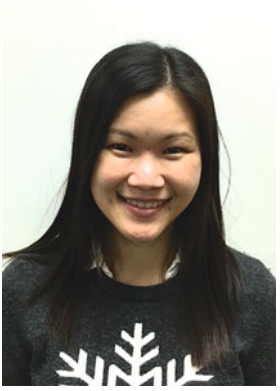
Keshavarz H, Hillis SD, Kieke BA, Marchbanks PA (2002) Hysterectomy surveillance—United States, 1994–1999. MMWR CDC Surveill Summ 51:1–8

Menon U, Gentry-Maharaj A, Hallett R, Ryan A, Burnell M, Sharma A, Lewis S, Davies S, Philpott S, Lopes A et al (2009) Sensitivity and specificity of multimodal and ultrasound screening for ovarian cancer, and stage distribution of detected cancers: results of the prevalence screen of the UK collaborative trial of ovarian cancer screening (UKCTOCS). Lancet Oncol 10(4):327–340

Merrill RM (2006) Impact of hysterectomy and bilateral oophorectomy on race-specific rates of corpus, cervical, and ovarian cancers in the United States. Ann Epidemiol 16(12):880–887

Merrill RM, Capocaccia R, Feuer EJ, Mariotto A (2000) Cancer prevalence estimates based on tumour registry data in the surveillance, epidemiology, and end results (SEER) program. Int J Epidemiol 29(2):197–207

Moyer VA (2012) Screening for ovarian cancer: U.S. preventive services task force reaffirmation recommendation statement. Ann Intern Med 157(12):900–904
Nelder JA, Mead R (1965) A simplex method for function minimization. Comput J 7(4):308–313
Partridge E, Greenlee RT, Xu JL, Kreimer AR, Williams C, Riley T, Reding DJ, Church TR, Kessel B, Johnson CC et al (2009) Results from four rounds of ovarian cancer screening in a randomized trial. Obstet Gynecol 113(4):775–782
Permuth-Wey J, Sellers TA (2009) Epidemiology of ovarian cancer. Humana Press, Totowa, pp 413–437
Schapira MM, Matchar DB, Young MJ (1993) The effectiveness of ovarian cancer screening: a decision analysis model. Ann Intern Med 118(11):838–843
Schorge JO, Modesitt SC, Coleman RL, Cohn DE, Kauff ND, Duska LR, Herzog TJ (2010) SGO white paper on ovarian cancer: etiology, screening and surveillance. Gynecol Oncol 119(1): 7–17
Skates SJ, Singer DE (1991) Quantifying the potential benefit of CA 125 screening for ovarian cancer. J Clin Epidemiol 44(4):365–380
Urban N, Drescher C, Etzioni R, Colby C (1997) Use of a stochastic simulation model to identify an efficient protocol for ovarian cancer screening. Control Clin Trials 18(3):251–270
Wingo PA, Huezo CM, Rubin GL, Ory HW, Peterson HB (1985) The mortality risk associated with hysterectomy. Am J Obstet Gynecol 152(7):803–808
Yabroff KR, Lamont EB, Mariotto A, Warren JL, Topor M, Meekins A, Brown ML (2008) Cost of care for elderly cancer patients in the United States. J Natl Cancer Inst 100(9):630–641



Jing Voon Chen holds a PhD in Industrial and Systems Engineering from the University of Southern California. Her dissertation focuses on medical decision making for ovarian cancer and a robust examination of the impact of calibration uncertainty on modeled outcomes. Her research interests include operations research in healthcare and health policy. She received a BS in mathematics from the University of Central Arkansas and an MS in mathematics from Texas A&M University. She is a member of INFORMS (Institute of Operations Research and the Management Science). She chose a career in STEM because she enjoys applying mathematics and engineering tools to solve real-world problems.

Julia L. Higle, Professor and Chair, Daniel J. Epstein Department of Industrial and Systems Engineering. Julie Higle joined the University of Southern California as Professor and Chair in 2012. Prior to that, she served as Professor and Chair of the Department of Integrated Systems Engineering at The Ohio State University, and as Professor of Systems and Industrial Engineering at the University of Arizona.

Julie's research interests are in the development of models and methods for decision making under uncertainty, stochastic programming models and algorithmic methods for their solution, and stochastic modeling for health-care applications, with a particular focus on the development of cancer screening strategies. She has served as the chair of the Council of

Industrial Engineering Academic Department Head. She has served the Institute of Industrial Engineering as the Senior Vice President for Academics, and Institute for Operations Research and the Management Sciences in a variety of positions. She received her PhD in Industrial and Operations Engineering from the University of Michigan.

Part V
Logistics

Chapter 16
Contributions to Humanitarian and Non-profit Operations: Equity Impacts on Modeling and Solution Approaches

Burcu Balcik and Karen Smilowitz

Contents

16.1 Introduction 371
16.2 Humanitarian Operations 373
16.2.1 Characterizing and Modeling Equity 374
16.2.2 Trade-Offs Between Equity and Other Metrics/Objectives 377
16.2.3 Methodological Advances 379
16.3 Non-profit Operations 381
16.3.1 Food Distribution Services 382
16.3.2 Provision of Healthcare Services 383
16.4 Conclusions 385
16.5 Dedication 387
References 387

16.1 Introduction

In 2010, we wrote a book chapter on the impact of equity in humanitarian and non-profit routing and distribution problems with our collaborator Seyed Iravani (Balcik et al. 2010). The chapter explored the ways in which broader operations research application areas, such as community services (e.g., public library operations, postal services), emergency and disaster relief, and various health and social services, in nonprofit and public sector operations research have led to interesting new problems and solution methods. Of the 29 papers cited in that chapter, just under 30% had at least one female co-author and two papers were written entirely by women. Almost

B. Balcik
Industrial Engineering, Ozyegin University, Istanbul, Turkey
e-mail: burcu.balcik@ozyegin.edu.tr

K. Smilowitz (✉)
Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL, USA
e-mail: ksmilowitz@northwestern.edu

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_16

a decade later, we see how women have been instrumental in the significant growth of the literature in humanitarian and non-profit applications in IE/OR.

In this chapter, we highlight methodological and practical contributions related to humanitarian and non-profit operations made by women in the fields of IE/OR in the 10 years since we wrote our initial study. The chapter explores applications of IE/OR to improve operations of humanitarian and non-profit organizations, focusing primarily on the role of equity considerations in terms of modeling and solution approaches. We review several key papers by women in the field and highlight how equity considerations necessitate new advances in IE/OR methodologies.

Equity is not a new concept in the IE/OR literature. Many studies that focus on public sector applications have stressed the importance of providing equitable access to users when designing and managing a variety of operations. In particular, there is a rich facility location literature that focuses on determining the locations of public facilities, including public parks, libraries, emergency facilities (e.g., ambulance, police and fire stations), and obnoxious facilities (e.g., landfills). As reviewed in Balcik et al. (2010), a variety of equity metrics are employed in these studies (see Marsh and Schilling (1994) for a comprehensive analysis of equity metrics). In many of these applications, equity is defined in terms of the distance (or travel time) between users and facilities. Additionally, there are studies that involve routing decisions in designing services; for instance, in transporting hazardous materials across the populated areas, it could be important to identify solutions that minimize the exposure of populations to the dangerous materials.

Recent work focused on operational problems in humanitarian and non-profit settings presents new equity metrics and modeling and solutions approaches. In particular, allocation of scarce resources, achieving quick response, prioritization of service, and providing access to vulnerable populations have been overarching themes in humanitarian and non-profit operations that require innovative interpretations of equity. Therefore, existing equity metrics from the well-established IE/OR literature focusing on public sector applications and other related disciplines such as economics have been adapted to address the specific needs of the corresponding humanitarian and non-profit settings.

In this chapter, we review 34 papers in total, which have been co-authored by 38 women researchers. These papers have been published between years 2008 and 2018. We used Google Scholar and Web of Science to search for papers that consider equity as an important concern in modeling humanitarian and/or non-profit operations. Our discussion at the end of the chapter includes feedback from a survey of several of the authors cited in this chapter, related to their views of the challenges of including equity in operations planning and open problems to be addressed in future work.

The remainder of this chapter is organized as follows. In Sect. 16.2, we discuss the studies that focus on modeling and solving problems that involve equity considerations in the context of humanitarian operations. In Sect. 16.3, we extend our review and discussion to non-profit operations. In Sect. 16.4, we conclude our chapter and discuss future work.

16.2 Humanitarian Operations

While equity has been widely studied within the scope of public sector operations, it has been introduced as an important concept within the scope of humanitarian operations relatively recently. There are a few earlier studies that focus on disaster management, but the literature that focus on implementing IE/OR techniques to improve humanitarian operations has grown extensively over the past 15 years. In particular, the scale of relief efforts carried by international humanitarian organizations beginning with the 2004 Asian Tsunami, and followed by Hurricane Katrina in 2005 and the 2010 earthquake in Haiti, has brought to light the complexity and challenges of humanitarian operations; increasing numbers of researchers since then have devoted efforts to improve humanitarian supply chains and logistics operations. The first study that discusses the differences between humanitarian supply chains and the traditional commercial supply chains was published as a conference proceeding in the same year (Beamon 2004).

With the increasing attention on humanitarian logistics field over the past decade, equity has been acknowledged as an important concern in designing and managing humanitarian operations. Given the significant demand for relief supplies created by a disaster and the scarcity of resources (such as supplies, vehicles, equipment), it is inevitable that some needs will be satisfied later than others, and effective prioritization of delivery is crucial in humanitarian operations. Different than commercial operations, where efficiency concerns are the main driver of the decisions related to whom to serve first, humanitarian organizations are challenged with the need to deliver resources in an equitable way to increase the chances of survival of people.

The importance of providing equitable service to the people affected by disasters has been particularly considered by IE/OR studies that focuses on last mile stage of humanitarian operations. Moreover, some disaster preparedness studies have incorporated the post-disaster last mile decisions into pre-planning, and considered achieving equity after a disaster as an important concern while making strategic decisions. The literature has focused on two main problems in the last mile stage: network design and relief distribution. Last mile *network design* problems focus on the establishment of local distribution networks in regions affected by a disaster and mainly involve decisions related to determining locations of local distribution points, assignment of beneficiaries to facilities, allocation of resources (relief supplies) among facilities and/or beneficiaries. Last mile *distribution* problems focus on delivery of supplies through the network from local distribution points to beneficiary locations over the relief horizon and mainly involve vehicle routing and resource allocation decisions.

At the time of writing (Balcik et al. 2010), there were only two studies that explored equity modeling in last mile distribution (Campbell et al. 2008; Balcik et al. 2008). Important contributions have been made by women related to equity in last mile relief networks over the past decade. In the following subsections, we review key papers written by women, which have advanced the literature

in characterizing and modeling equity in humanitarian relief and exploring the methodological implications of equity.

16.2.1 Characterizing and Modeling Equity

As acknowledged in much of the related work, there is no universal definition of equity. Defining and characterizing equity in humanitarian settings can be affected by a number of factors, including the type of the problem addressed, the planning horizon, the decisions, the services, the users, etc. However, similar to other public sector applications, the general criteria in designing humanitarian operations is to ensure equitable access to the services. Additionally, achieving equity in terms of response time and supply allocation are important considerations in many studies. In this subsection, we discuss studies that present different ways to define and model equity in humanitarian logistics problems.

One of the first papers to define and model equity in a last mile distribution setting was Balcik et al. (2008). The authors define a last mile distribution problem that focuses on distributing supplies from a local distribution center to a set of demand ("beneficiary") points by a fleet of vehicles over a relief horizon. The main objective focuses on achieving equity in supply allocation in this study, along with a secondary efficiency objectives to minimize total transportation costs. More specifically, equity is modelled by *minimizing the maximum proportion of unsatisfied demand* each day among the demand locations. This objective function aims to ensure *equity in supply allocation and response time*. Subsequent work has focused on characterizing last mile distribution problems in different settings and defining alternative equity metrics and objectives, and also developing efficient solution methods. For instance, Vitoriano et al. (2011) present a multi-objective model and goal programming approach to support distribution of aid in disaster relief operations. Besides ensuring equitable service, the authors consider other performance measures such as time of response, and reliability and security of the operation routes. Equity is achieved by minimizing the maximum proportion of unsatisfied demand over the affected locations.

Noyan et al. (2015) focus on characterizing and modeling equity in designing last mile distribution networks, which are set up after a disaster occurs. The authors focus on a post-disaster setting, which involves a single local distribution center (LDC), points of distribution (POD), and beneficiary (demand) locations; the beneficiaries must access PODs to receive relief supplies. Given the locations of the LDC and beneficiary locations, the problem is to determine the locations and capacities of the PODs such that high levels of accessibility and equity are achieved in the network. Since available information on demands and network conditions are usually rough and incomplete at this stage, demands and travel times are assumed to be uncertain. The authors develop metrics to characterize and measure accessibility and equity in this setting. Since beneficiaries may have different characteristics that would affect their mobility after a disaster (e.g., gender, age, vehicle ownership),

demographical and socio-economical aspects of the beneficiaries are considered in addition to the network travel times to calculate an accessibility score for each demand point. Then achieving *equity in accessibility* is ensured by assigning each demand location to a POD that satisfies a minimum accessibility requirement; that is, coverage constraints are imposed to limit the worst accessibility. The authors consider two policies to allocate the scarce supplies at the LDC among the PODs equitably; *equity in supply allocation* is measured based on the *maximum proportion of unsatisfied demand*. Specifically, the first equitable allocation policy divides supplies among the PODs according to a proportional allocation policy, while the second allocation policy allocates supplies by limiting the shortages at the PODs by a specific common proportion of the corresponding total demand. The approach presented in Noyan et al. (2015) is illustrated through a case study, which is based on real world data based on an earthquake that occurred in 2011 in Van province of Turkey.

Noyan and Kahvecioğlu (2017) extend the last mile network design problem in Noyan et al. (2015) to a three-echelon setting in which there are multiple LDCs in the first echelon and the additional decisions involve determining which LDCs to operate in the last mile. Furthermore, some of these LDCs may already exist in the region before the disaster and they may contain some pre-positioned supplies. Therefore, reallocation of existing supplies among the LDCs can also be an important decision in the last mile, as effective redistribution of supplies among the facilities may highly affect equity in supply distribution and accessibility. To ensure equitable accessibility in a three-echelon post-disaster network, Noyan and Kahvecioğlu (2017) control the worst accessibility scores between the LDCs and PODs, and also PODs and demand points separately by imposing coverage constraints.

Muggy and Stamm (2017) address locating temporary health-care facilities in a post-disaster environment for providing treatment and/or preventive services during infectious disease outbreaks. Depending on the dynamic needs and resource levels, humanitarian organizations may wish to open and close facilities to serve beneficiaries. The authors consider uncertainties in demands (in terms of beneficiaries' ability to travel to facilities as a function of distance) and supplies (in terms of facility capacities). The main objective in making location decisions in this network is maximizing accessibility to the services. The authors present an approach to measure accessibility, which does not necessitate assigning beneficiaries to facilities, nor does it assume that each individual visits the nearest facility. Specifically, they define the demand in the catchment zones for the potential facilities and measure accessibility through a weighted capacity-to-demand ratio for each facility. Then accessibility for each population location is calculated by summing the capacity-to-demand ratios of the facilities within its catchment zone. Equity among demand locations is ensured by providing sufficient accessibility to all demand locations, where access is deemed sufficient if it is greater than or equal to a pre-determined access score. The authors present a model that minimizes the maximum regret in demand-weighted access and maximizes the number of people that receive sufficient access.

While the contributions discussed above focus on the disaster response stage, there are also studies that consider equity during the disaster preparedness stage. These studies mostly address prepositioning relief supplies in a humanitarian supply network in advance of a disaster while considering distribution of these prepositioned supplies to the demand points after a disaster. That is, the location and amount of inventories prepositioned in the network affect both the response time and amount of demand that can be satisfied, and therefore equity can be an important metric to consider for prepositioning decisions as well. We discuss some examples next, which incorporate equity in disaster preparedness.

Davis et al. (2013) focus on inventory allocation for hurricane preparedness. Given a set of local facilities in a region with prepositioned supplies, the problem is to relocate the existing inventory among facilities based on short-term forecasts for an approaching hurricane. The local facilities are owned by different agencies, which coordinate by allocating their inventory among the facilities to better respond to the hurricane; that is, depending on the track of the hurricane, supplies are moved to safer locations. The authors present a mathematical model that minimizes costs associated with prepositioning and distribution of supplies. The proposed model incorporates constraints that enforce that each demand location is served from the stocks proportionally. They also ensure that each demand point is within the desired response time. That is, relocation of supplies are made by considering equity both in response time and allocation.

Battarra et al. (2018) focus on related inventory allocation problems for disaster preparedness, motivated by the earthquake preparedness efforts of the Turkish Red Crescent. The authors adapts a methodology from earthquake engineering to estimate the people affected by disasters by using historical earthquake data. Given the estimated demands and the total amount of inventory in the network, the inventory at each existing warehouses is determined by using a resource allocation model. In the model, equitable allocation of supplies among the warehouses is ensured by maximizing the minimum coverage level achieved across potential demand locations. The coverage level is measured in terms of the fraction of the emergency supply demand that can be satisfied within the desired response time, which is set as 2 h by the Turkish Red Crescent.

Noham and Tzur (2018) incorporate post-disaster decisions into pre-disaster planning by considering a three-echelon network. The warehouses at the first echelon are located in the preparedness stage while considering the implications of post-disaster decisions, which involve locating LDCs, assignment of LDCs to the warehouses, and assignment of demand points to the LDCs. That is, facilities are located both in pre- and post-disaster stages. To model equitable supply allocation, the authors define a service level gap, which is the maximal ratio allowed between the proportions of the satisfied demand at all demand points. They allow some difference in demand satisfaction because an equal allocation can be quite limiting. By setting the maximum service gap parameter, the planners can control the level of equity and determine the best allocation that satisfy this equity level for the given total available supply. However, the service level gap is enforced to be zero among all the demand points served by the same LDC.

Equity has also been considered in post-disaster response and recovery operations to restore the damaged infrastructure. Aksu and Ozdamar (2014) focus on a road restoration planning problem to allocate limited equipment for clearing blocked roads by removing debris as soon as possible. The primary objective is to maximize network accessibility, which is ensured by maximizing the total weighted earliness of all paths' restoration completion times, where the weight assigned to a path represents its criticality. The earliness of a path is defined as the difference between the operation due date and its restoration completion time. For this problem, the authors present and compare two models with and without equity concerns; the model with equity additionally minimizes the maximum difference between the earliness values of any two districts.

As with most of the studies reviewed in this subsection, decisions in disaster preparedness and response usually require considering multiple performance metrics. Some studies have explored the tradeoffs between equity and other performance metrics analytically, numerically, or qualitatively, which are reviewed next.

16.2.2 Trade-Offs Between Equity and Other Metrics/Objectives

Distribution of supplies in humanitarian settings often involves multiple performance metrics and/or objectives. Gralla et al. (2014) specify three main groups of performance criteria in humanitarian relief: (1) efficiency, (2) effectiveness, and (3) equity. Efficiency metrics capture cost of operations; depending on the setting, efficiency can be represented in terms of traveling costs, inventory costs, or number of opened facilities. Effectiveness metrics capture the extent to which operations achieve target goals; effectiveness can be formulated in terms of total amount of demand satisfied and average response time. In an earlier work, Beamon and Balcik (2008) propose three types of performance metrics for humanitarian relief operations: resource (relates to efficiency), output (relates to effectiveness), and flexibility. In that study, equity is classified as an output metric. The performance measurement framework in Beamon and Balcik (2008) is adapted from the seminal study on performance measurement in supply chains in Beamon (1999). Therein, the output metrics are defined to be related to customer responsiveness, quality, and the quantity of final product; therefore, equitable supply distribution can also be considered as an output (effectiveness) indicator. However, given the increased interest in modeling equity and studying its methodological implications in humanitarian operations, it is now being acknowledged as a primary criteria, rather than being one of the dimensions of effectiveness.

Many papers in humanitarian operations consider a combination of metrics. Some studies present multi-objective models, while most optimize one of this criteria in the objective function and incorporate the others through the constraints. For instance, Noham and Tzur (2018) express effectiveness in the objective function,

while efficiency and equity criteria are handled through constraints. Balcik et al. (2008) propose a multi-objective model where the primary objective focuses on ensuring equity, while the secondary objective focuses on efficiency. Vitoriano et al. (2011) propose a goal programming approach for handling multiple objectives of costs, response time, equity, security, and reliability. In Aksu and Ozdamar (2014), the primary objective is effectiveness, while equity is addressed in an alternative model through constraints. Noyan et al. (2015) maximize total expected accessibility in the objective function, while ensure equity in accessibility and supply allocation through constraints. Muggy and Stamm (2017) balance efficiency and equity with a weighted objective function.

Some work in the last decade has considered the implications of equity analytically. The first study to explore the impact of equity in a humanitarian routing problem is Campbell et al. (2008). The authors address a last mile routing problem and present an analytical study that compares the effect of using different objectives on response time, and show that the choice of objective significantly affects how aid is distributed. More specifically, the authors consider alternative objectives of minimizing the last arrival time and minimizing the sum of arrival times and show that superior response times can be achieved compared to those resulting from a traditional VRP objective (i.e., minimizing total travel times). Campbell et al. (2008)'s highly cited study has an important role in the literature to stress not only that equity is an important concern in humanitarian relief efforts, but also the choice of objective function can have a significant effect on performance of relief efforts.

In Campbell et al. (2008), each demand location is visited exactly once and full satisfaction of demand is ensured; consequently, the authors focus on response time equity, and supply equity is not considered. Huang et al. (2012) extend the work of Campbell et al. (2008) by considering demands at each recipient which necessitates formulating equity metrics for supply allocation. Huang et al. (2012) measure efficiency in terms of total travel time of the constructed routes, and efficacy is formulated by the time-weighted unsatisfied demand. Additionally, three alternative equity objectives are proposed, which aim to ensure that each node in the network gets equitable service in terms of *demand-weighted arrival time*. Specifically, the first equity objective minimizes the maximum pairwise difference in service across nodes, and the second one minimizes the standard deviation of the service level. The third equity objective weighs units of demand differently and minimizes the disutility-weighted arrival time; more specifically, more utility can be gained by demands satisfied earlier, which is formulated by using a piecewise linear disutility function. Huang et al. (2012) show that the model variations that incorporate different metrics can yield significantly different solutions. The paper explicitly discusses the impact on route structure.

The impact of different equitable supplies policies on accessibility and equity metrics is explored in Noyan et al. (2015). The authors show that a proportional allocation policy, which imposes strict requirements on dividing supplies among demand locations may decrease the expected total accessibility in the network. Accessibility is measured by incorporating mobility scores of different population groups into travel times; so accessibility in this study can be considered as a

measure of effectiveness. The authors propose a new allocation policy, which relaxes proportional allocation requirements but limits the maximum proportion of unsatisfied demand, and show that the proposed policy can achieve high levels of equity and accessibility simultaneously.

Gralla et al. (2014) conduct a survey of experienced humanitarian logisticians, and evaluate their preferences over different important performance attributes in relief distribution. Specifically, they focus on the amount of cargo delivered, the prioritization of aid by commodity type, the prioritization of aid by delivery location, the speed of delivery, and the operational costs. Based on the survey results, the importance of each attribute is quantified. Noyan et al. (2015)'s study provides important empirical evidence about the performance metrics used in real-world last mile operations, and validates that equity is an important factor considered in making last mile distribution decisions in practice.

16.2.3 Methodological Advances

The studies that focus on humanitarian operations with equity concerns have also contributed to the literature by presenting new methodologies to model and solve the proposed problems. A variety of approaches are used; primarily, the studies that involve last mile routing/distribution decisions mostly present heuristics, and the studies that address the uncertainties in different pre- and post-disaster problem settings mostly develop stochastic programming models, which are solved by exact and heuristic methods.

The last mile distribution problems involve vehicle routing decisions and are difficult to solve optimally for realistic size problem instances. Equity concerns may bring additional complexity in solving these problems, as modeling equity may require additional constraints or different objective functions than the traditional VRP objectives, as reviewed in the prior subsections. It is frequently acknowledged in the literature that last mile distribution decisions must be made quickly in chaotic post-disaster environments, where humanitarian organizations may not have access to advanced computational resources; therefore, there has been an increasing attention on developing practical solution methods, which can provide good solutions quickly and without requiring special software. To solve the last mile routing problems with alternative objectives, Campbell et al. (2008) present a constructive insertion heuristic, and Huang et al. (2012) develop a greedy random adaptive search procedure (GRASP).

As discussed earlier, in humanitarian supply chains, both pre- and post-disaster decisions have to be often made under significant uncertainty; before a disaster occurs, there are uncertainties regarding the location, time and scale of the next disaster, while after a disaster, there are uncertainties regarding the location, number and conditions of the affected people, available resources (supplies, vehicles, funding/budget, etc.), and network conditions. Consequently, there has been a vast interest in the literature to develop methods to support decision making in

humanitarian settings that involve uncertainty. In particular, two-stage stochastic programming has been one of the most popular modeling approaches; see the literature review by Grass and Fischer (2016).

Noyan and her co-authors presented several studies that develop stochastic programming formulations and exact solution approaches (branch-and-cut algorithms based on Benders decomposition) for humanitarian network design problems involving uncertainties in demands and transportation network conditions. Noyan et al. (2015) and Noyan and Kahvecioğlu (2017) present two-stage stochastic programming models for designing post-disaster last mile networks in which achieving equity in accessibility and supply allocation is critical. Noyan and her co-authors also present different modeling approaches for pre-disaster network design problems for disaster preparedness; primarily, these studies focus on incorporating risk into optimization models to counteract the unfavorable effects of potentially high level of random variability inherent in chaotic disaster relief systems. Their modeling approaches mainly involve risk measures and chance constraints, and lead to novel and computationally challenging risk-averse stochastic optimization models. For instance, Hong et al. (2015) use probabilistic constraints to ensure that the demand for relief supplies across all demand nodes in the network is satisfied with a high probability. Their modeling approach appears to be computationally effective, since it ensures the feasibility of the second-stage problem without introducing recourse decisions; specifically, Gale–Hoffman inequalities are used to represent the conditions on the existence of a feasible network flow in the second stage.

Elçi and Noyan (2018) present a chance-constrained two-stage mean-risk stochastic programming model; more specifically, conditional value-at-risk (CVaR) is considered as the risk measure in a mean-risk objective, and a joint probabilistic constraint is used to ensure the feasibility of the second-stage problem concerned with distributing the relief supplies to the affected areas. In fact, they develop a hybrid-type risk-averse model, which controls the supply shortages by utilizing a quantitative approach via incorporating the shortage cost into the mean-risk objective function and a qualitative approach via a joint chance constraint. The authors also enforce a common supply shortage penalty and a common coverage threshold for all demand points to ensure equity in supply allocation and response times, respectively. Noyan et al. (2017) introduce an elaborate risk-averse two-stage modeling approach, which enforces a stochastic benchmarking preference relation between vectors of performance measures. In particular, in the first-stage, they enforce multivariate risk constraints based on CVaR and the multiple performance measures (responsiveness and equitable supply allocation) associated with the second-stage decisions, while minimizing the expected values of efficiency (costs) and efficacy (shortage) related metrics. It is important to highlight that these studies provide general and flexible ways of modeling decision makers' risk preferences based on multiple stochastic performance criteria. Indeed, these risk-averse modeling approaches are not limited to the pre-disaster network design problems; they are general enough to accommodate other logistics and transportation problems

modeled as a two-stage stochastic program, where the second-stage problem is a linear program.

There are also heuristic methods developed to solve two-stage stochastic programming formulations developed for humanitarian network design problems; for instance, Noham and Tzur (2018), which was described before, present a tabu search heuristic to solve the humanitarian network design problem that addresses both pre-and post-disaster phases. One difficulty of stochastic programming methods in humanitarian logistics problems is the need for developing probabilistic scenarios to represent uncertainties. That is, probabilistic representations may be difficult to get in humanitarian networks, where data is usually scarce or inaccurate. As an alternative approach to generating probabilistic demand scenarios, Battarra et al. (2018) present a methodology that utilizes forecasting methods from the earthquake engineering literature to estimate the number of affected people by earthquakes. Moreover, Muggy and Stamm (2017) present a scenario-based robust optimization method to address uncertainties in demands and supplies. Although robust optimization methods have been applied less frequently compared to stochastic optimization methods, there is a growing interest in using robust optimization to model humanitarian logistics problems primarily due to the challenges associated with describing uncertainties in humanitarian relief environments probabilistically.

16.3 Non-profit Operations

As with humanitarian logistics, equity plays a significant role in non-profit operations. A recent review of modeling approaches and metrics to evaluate non-profit operations (Berenguer 2016) cites a range of papers analyzing equity metrics, including the work in Campbell et al. (2008) in disaster response, discussed in the prior section and the work in Lien et al. (2014) in food distribution, discussed next. Around the time of our earlier book chapter on equity, Leclerc et al. (2012) reviewed equity in the allocation of public resources, a topic that both Laura Albert and Maria Mayorga have studied in their related work (Chanta et al. 2011; Toro-Díaz et al. 2015; McLay and Mayorga 2013). Ayer et al. (2014) reviewed public health research related to efficient, effective, and equitable outcomes. In this section, we consider works cited in the above references, as well as other recent studies, on equity in non-profit operations. In the 2010 book chapter, we discussed two papers that explicitly modeled equitable distribution of scarce resources in non-profit operations; one focused on distribution of food to the homebound (Johnson et al. 2005) and one focused on the operations of a regional food bank (Lien et al. 2014). The work by Lien et al. (2014) has led to several new studies of food bank operations, many of which include some modeling of equity. In what follows, we discuss those new papers, as well as papers in the provision of health services, that represent the growing literature on the complexity of non-profit operations, particularly when equity is considered. As discussed in the prior section on humanitarian logistics,

these operations often involve limited resources; thus, decisions of whom to serve and to what extent are critical.

16.3.1 Food Distribution Services

Balcik et al. (2010) include a review of work in equitable distribution of food to food-insecure populations, with a particular focus on Lien et al. (2014). In that paper, the authors introduce a min-max function to ensure equitable distribution along a route distributing food sequentially to various agencies under limited supply and stochastic demands. As the authors discuss, there is no perfect metric for equity that can both simultaneously maximize equity, minimize wasted resources and remove the bias of position along a route. Thus, this led to continued exploration of equity metrics, and much of these studies were conducted by leading women researchers in the area of non-profit operations.

Balcik et al. (2014) extend the work of Lien et al. (2014) to consider multiple vehicles. In the multi-vehicle setting, one must consider how the clustering of food donors and recipients into vehicle routes impacts equity, since the nature of this partitioning limits the extent to which one can ensure equitable and effective allocation with the allocation policy of Lien et al. (2014). Therefore, Balcik et al. (2014) formulate the equity-maximizing multi-vehicle sequential resource allocation problem and examine how maximizing equity influences the assignment of donors and recipients to routes, as compared with waste-minimizing solutions. As with the single vehicle case, it is shown that finding equitable solutions in a strategic manner can lead to minimal waste. Insights from initial studies of demand and supply variations lead to a scalable decomposition-based heuristic to solve the multi-vehicle problem for larger instances.

Julie Ivy and Lauren Davis, working with a number of their students and colleagues, have also studied equity in food distribution. As the work in humanitarian logistics discussed in Sect. 16.2.2 has shown, equity objectives can often be in conflict with other metrics. Orgut et al. (2016) present work performed in collaboration with the Food Bank of Central and Eastern North Carolina aimed at policies for both equitable and effective distribution of food donations. The authors define an equitable distribution as allocation that is proportional to county's demand (based on population in poverty)—rather than striving for perfect equity, the authors constrain waste-minimizing solutions by equity across counties. That is, there is a limit set on the absolute deviation between the percent of food allocated to a county and the relative need of that county. In Sengul Orgut et al. (2017), the authors extend this work to the case with stochastic capacities of the recipients. Fianu and Davis (2018) consider uncertainty not in recipient capacity, but in donations. Their model considers equitable allocation policies (using metrics similar to the above studies) as supply inventory levels change over time with stochastic donations.

In Davis et al. (2014), the authors consider an alternate collection and delivery strategy from the distribution routes studied in earlier work to use resources more

efficiently while providing equitable access. In particular, the paper introduces the concept of "food delivery points" (FDP) which are satellite locations from which recipients can pick up donations. The authors propose a two-phase approach to determining FDP locations and routing and scheduling vehicles. In the first phase, FDP locations are determined with a set covering formulation and in the second phase, routes are generated and scheduled with a periodic pickup and delivery problem. Equity is considered with the models through constraints in both phases: the set covering model enforces a maximum travel distance for recipients and the periodic routing model ensures demand satisfaction. The authors show that the new strategy can improve access to food while using resources more efficiently. The paper includes a comparison equity achieved with the new strategy versus the existing strategy. Using a deviation equity metric, calculating the difference between the best and worst travel time effects on agencies grouped by county, the authors show that the new strategy is more equitable.

Michal Tzur and her colleagues have been collaborating with other food banks on studies related to equity in food distribution. In particular, her work with her student Ohad Eisenhandler explores joint routing and allocation decisions with equity considerations (Eisenhander and Tzur 2018). In particular, her work with her student Ohad Eisenhandler explores joint routing and allocation decisions with equity considerations. As with the earlier studies cited, they look to balance objectives of effectiveness and equity. The effectiveness objective is measured as the total supply distributed (F) and the equity objective is measured as 1 minus the Gini coefficient (G). Their routing and allocation model maximizes a combined objective function $Z = F(1 - G)$, since effectiveness increases as the distribution of supply increases and equity increases as the Gini coefficient decreases representing a broader distribution of wealth. As discussed in the paper, the function Z can be easily linearized and the structure of the objective function suggests natural decompositions of the problem to be used in constructing solution approaches. Further, the objective has three desirable properties: it is component-wise increasing, satisfies the principle of transfers, and finds a balance between objectives. The proposed function is evaluated relative to other objectives that also satisfy these properties to some extent.

16.3.2 Provision of Healthcare Services

In their tutorial on public health research, Ayer, Keskinocak, and Swann discuss the challenges of balancing objectives related to efficient, effective, and equitable outcomes (Ayer et al. 2014). These challenges are often particularly pronounced when healthcare is provided by non-profit or nongovernmental organizations. Several recent studies have explored these challenges, focusing on complexities introduced when both uncertainty and equity are jointly considered.

A stream of work by Smilowitz and collaborators has explored a collaboration with a non-profit organization in Chicago that screens and treats pediatric asthma

(Deo et al. 2013). In this setting, the health care provider serves a broad catchment of asthmatic children in a public school system with limited access to health care for their chronic asthma. Thus, in this setting, rather than scheduling appointments according to cost efficiency or patient satisfaction measures, the objective function considers aggregate health outcomes of the target community population and equity of access. Uncertainty in this setting comes from uncertainty in disease progression and patient response to treatment. Thus, the authors develop scheduling models, given limited appointment capacity, to ensure that appointments are allocated equitably and provide access to those in the lowest health state, thus improving the aggregate health outcomes. In follow-up work, Smilowitz and Savelsbergh extend the study of equitable appointment allocation to consider patient no-shows (Savelsbergh and Smilowitz 2016). As no-shows are often the result of the inability of parents and caregivers to attend, the model considers allocation given time-dependent preference for slots outside of work hours.

Two recent collaborations with nongovernmental organizations in Africa have considered the challenges of equitable provision of healthcare. Julie Swann and Ozlem Ergun and collaborators worked with a nongovernmental organization in South Africa that coordinates a supply chain in South Africa to distribute donated breastmilk (Cao et al. 2016). Their collaboration focuses on supply chain design for equitable distribution given uncertainty in donations and demand. As the others note, consistent with the discussion in this chapter, "In addition to the trade-off between equity and efficiency, modeling for equity involves two important challenges. First, there is no single universal equity measure that works well in all possible settings, therefore definition of an appropriate equity objective is context-dependent. Second, even when such a measure can be defined, finding a solution to the corresponding model is generally computationally more challenging than its efficiency-based counterpart." The authors study three measures for equity, defined by the target population level for equity (referred to often in the equity literature as "equity to whom"). At the national level, the equity metric is essentially an efficiency metric as the authors maximize the total percent of demand met. At the regional and local levels, the authors maximize the minimum percent of demand met. The national level metric is easier to solve, while the regional and local maxmin objectives require heuristics to solve. Notably, in order to decrease inequity at the local level caused by transportation costs, the authors propose multi-stop routes (rather than existing direct routes). Thus, the problem becomes a location routing problem. This is an interesting example of how the underlying optimization problem itself changes to improve equitable access.

Location routing problems are also considered in VonAchen et al. (2016) for a nongovernmental organization in Liberia providing community based health care to remote communities in Liberia. Rather than optimizing equity metrics as in Cao et al. (2016), equitable access is modeled through coverage constraints that enforce a maximum distance from a health care provider and the community served by the health care provider.

16.4 Conclusions

The work cited in this chapter represents key advances in the ways in which we model and solve distribution problems for non-profits with equity considerations. The papers also show growing collaborations with non-profit agencies and non-governmental organizations, expanding the reach of industrial engineering and operations research. Further, many of the women faculty cited above are mentoring woman graduate students and post-doctoral researchers in this work, who will likely lead the next way of research in equitable operations in the non-profit sector.

In the process of writing this chapter, we reached out to some of the authors cited and asked the following questions related to how considerations of equity has influenced their work: (1) What challenges did the consideration of equity present in your work on humanitarian and/or non-profit logistics, either in modeling a problem or developing solution methods (or in other ways)? and (2) What do you see as open questions/next steps in the study of equity in humanitarian and/or non-profit logistics?

The authors referenced many of the challenges related to defining equity that are discussed in this chapter. As Lauren Davis commented *The challenge for me was how to accurately measure equity/fairness based on the problem setting (there are several metrics in the literature to choose from) and then how to incorporate that into the model.... Should I stick with traditional cost-based metrics that you see in many inventory problems and make equity a constraint (like in Davis et al. (2013))? or Should equity be the driving and primary objective influencing the decision (like in Fianu and Davis (2018)). When considering equity, there is the acceptance though that all demand is not going to be satisfied which is really not an ideal situation in the humanitarian logistics setting.* Of note, Ann Campbell reflected on the particular challenges associated with equity in delivery of relief items in disaster response *When it comes to delivery times, someone is always first on a route and someone is last. How to enforce an equity to delivery times was not straightforward.*

Michal Tzur offered some insights on how to choose among the multiple ways to define equity: *You don't want your results to be sensitive to parameters whose values may seem arbitrary. It is unclear how much we are willing to 'sacrifice' from the supposedly 'real' objective, to achieve equity. While aiming to represent equity as accurately as possible, this may lead to concave or otherwise difficult to handle functions within the framework of a mathematical program.*

Nilay Noyan stressed the need for developing effective modeling approaches to incorporate equity and other metrics: *Incorporating the concept of equity in decision-making problems calls for optimization models, which address multiple and possibly conflicting performance criteria such as the total acquisition cost, equity in terms of relief distribution, and accessibility of relief supplies. In addition, it is of crucial importance to take into the consideration the uncertain aspects of the system of interest. Ultimately, we are challenged to develop elaborate multicriteria stochastic optimization models and effective solution methods to address problems in humanitarian logistics with equity considerations.*

In terms of future work to be addressed, the responses covered open challenges in implementation, modeling and solution approaches. Ann Campbell noted that *It seems like so much needs to be done on the side of better communication of need and cooperation among different groups, more so than on the optimization side.* Although the importance for having coordinated planning and response systems in humanitarian supply chains has been widely highlighted in the literature, there are indeed only a few analytical studies that develop methods to assist collaborative planning/response and study implications of equity in collaborative settings; in our review, only Davis et al. (2013) and Muggy and Stamm (2017) explore the implications of a coordinated response with multiple agencies. In these studies, providing equitable service to beneficiaries is the main concern. A multi-agency setting study that uses cooperative game theory to "fairly" allocate costs and benefits of coordination among multiple agencies is presented by Ergun et al. (2014). Specifically, the authors focus on using information technology (IT) tools collaboratively in managing multiple camps that serve internally displaced people after the Haiti earthquake in 2010, and present a game theory framework to allocate the costs and benefits associated with these IT investments among the partner agencies. Future work that analytically explore humanitarian settings with multiple agencies and propose methods for fair resource sharing among agencies to provide equitable service to beneficiaries would be valuable.

Both Lauren Davis and Michal Tzur commented on equity across multiple dimensions. *One of the things I am looking at is equity in multiple dimensions (or tiered equity) in a problem and the impact of that. An example from the food distribution problem is what if I have to ensure that overall I am distributing food equitably into multiple counties (from a total poundage perspective), but I also want to ensure that the type of food that is being distributed is also equitable.* (Davis) Michal Tzur also pointed to the challenges related to solution approaches: *Developing methodologies to achieve equity in a dynamic environment, i.e., when new information arrives over time.* Nilay Noyan commented that *The study of equity can benefit from recent progress in stochastic operations research, which provides a large variety of alternative flexible and tractable modeling approaches to incorporate various definitions of equity into optimization models. Making these approaches more accessible to practitioners has a great potential to better serve the society.*

The focus of this paper is on reviewing the contribution of women researchers in studying equity in humanitarian/non-profit operations by using IE/OR methods, which has been increasingly acknowledged as an important concern over the past decade. However, needless to say, the contributions of women researchers in studying the humanitarian/non-profit operations has been far larger and more significant than studying the equity aspect. The role of women for the establishment and advancement of these fields has been immense not only in publishing a large number of high quality papers but also in starting and running special chapters/clusters in professional societies; serving as the members in the editorial boards and the editorial teams of the field-related journals; and establishing research

centers and organizing conferences. We hope that the contributions of women in the advancement of the field will continue to grow.

16.5 Dedication

As we recognize the work of women in the field, the authors dedicate this chapter to the memory of Benita Beamon, whose work has been foundational to the fields of humanitarian and non-profit operations research. We are grateful for her inspiration both personally and professionally.

References

Aksu DT, Ozdamar L (2014) A mathematical model for post-disaster road restoration: enabling accessibility and evacuation. Transp Res E Log Transp Rev 61:56–67

Ayer T, Keskinocak P, Swann J (2014) Research in public health for efficient, effective, and equitable outcomes. In: Bridging data and decisions. Informs, Catonsville, pp 216–239. Chapter 9

Balcik B, Beamon BM, Smilowitz K (2008) Last mile distribution in humanitarian relief. J Intell Transp Syst 12(2):51–63

Balcik B, Iravani SM, Smilowitz K (2010) A review of equity in nonprofit and public sector: a vehicle routing perspective. Wiley Encyclopedia of Operations Research and Management Science. Wiley, Chichester

Balcik B, Iravani S, Smilowitz K (2014) Multi-vehicle sequential resource allocation for a nonprofit distribution system. IIE Trans 46(12):1279–1297

Battarra M, Balcik B, Xu H (2018) Disaster preparedness using risk-assessment methods from earthquake engineering. Eur J Oper Res 269:423–435

Beamon BM (1999) Measuring supply chain performance. Int J Oper Prod Manag 19(3):275–292

Beamon BM (2004) Humanitarian relief chains: issues and challenges. In: Proceedings of the 34th international conference on computers and industrial engineering San Francisco, CA, November 14–16, vol 34, pp 77–82

Beamon BM, Balcik B (2008) Performance measurement in humanitarian relief chains. Int J Public Sect Manag 21(1):4–25

Berenguer G (2016) Modeling approaches and metrics to evaluate nonprofit operations. In: Advances in managing humanitarian operations. Springer, Cham, pp 9–31

Campbell AM, Vandenbussche D, Hermann W (2008) Routing for relief efforts. Transp Sci 42(2):127–145

Cao W, Çelik M, Ergun Ö, Swann J, Viljoen N (2016) Challenges in service network expansion: an application in donated breastmilk banking in South Africa. Socio-Econ Plan Sci 53:33–48

Chanta S, Mayorga ME, Kurz ME, McLay LA (2011) The minimum p-envy location problem: a new model for equitable distribution of emergency resources. IIE Trans Healthc Syst Eng 1(2):101–115

Davis LB, Samanlioglu F, Qu X, Root S (2013) Inventory planning and coordination in disaster relief efforts. Int J Prod Econ 141(2):561–573

Davis LB, Sengul I, Ivy JS, Brock LG, Miles L (2014) Scheduling food bank collections and deliveries to ensure food safety and improve access. Socio-Econ Plan Sci 48(3):175–188

Deo S, Iravani S, Jiang T, Smilowitz K, Samuelson S (2013) Improving health outcomes through better capacity allocation in a community-based chronic care model. Oper Res 61(6):1277–1294

Eisenhander O, Tzur M (2018) The humanitarian pickup and distribution problem. Oper Res 67(1):10–32

Elçi Ö, Noyan N (2018) A chance-constrained two-stage stochastic programming model for humanitarian relief network design. Transp Res B Methodol 108:55–83

Ergun Ö, Gui L, Heier Stamm JL, Keskinocak P, Swann J (2014) Improving humanitarian operations through technology-enabled collaboration. Prod Oper Manag 23(6):1002–1014

Fianu S, Davis LB (2018) A Markov decision process model for equitable distribution of supplies under uncertainty. Eur J Oper Res 264(3):1101–1115

Gralla E, Goentzel J, Fine C (2014) Assessing trade-offs among multiple objectives for humanitarian aid delivery using expert preferences. Prod Oper Manag 23(6):978–989

Grass E, Fischer K (2016) Two-stage stochastic programming in disaster management: a literature survey. Surv Oper Res Manag Sci 21(2):85–100

Hong X, Lejeune MA, Noyan N (2015) Stochastic network design for disaster preparedness. IIE Trans 47(4):329–357

Huang M, Smilowitz K, Balcik B (2012) Models for relief routing: equity, efficiency and efficacy. Transp Res E Logist Transp Rev 48(1):2–18

Johnson MP, Gorr WL, Roehrig S (2005) Location of service facilities for the elderly. Ann Oper Res 136(1):329–349

Leclerc PD, McLay LA, Mayorga ME (2012) Modeling equity for allocating public resources. In: Community-based operations research. Springer, New York, pp 97–118

Lien RW, Iravani SM, Smilowitz KR (2014) Sequential resource allocation for nonprofit operations. Oper Res 62(2):301–317

Marsh MT, Schilling DA (1994) Equity measurement in facility location analysis: a review and framework. Eur J Oper Res 74(1):1–17

McLay LA, Mayorga ME (2013) A dispatching model for server-to-customer systems that balances efficiency and equity. Manuf Serv Oper Manag 15(2):205–220

Muggy L, Stamm JLH (2017) Dynamic, robust models to quantify the impact of decentralization in post-disaster health care facility location decisions. Oper Res Health Care 12:43–59

Noham R, Tzur M (2018) Designing humanitarian supply chains by incorporating actual post-disaster decisions. Eur J Oper Res 265(3):1064–1077

Noyan N, Kahvecioğlu G (2017) Stochastic last mile relief network design with resource reallocation. OR Spectr. 40(1):187–231

Noyan N, Balcik B, Atakan S (2015) A stochastic optimization model for designing last mile relief networks. Transp Sci 50(3):1092–1113

Noyan N, Merakli M, Kucukyavuz S (2017) Two-stage stochastic programming under multivariate risk constraints with an application to humanitarian relief network design. arXiv preprint arXiv:1701.06096

Orgut IS, Ivy J, Uzsoy R, Wilson JR (2016) Modeling for the equitable and effective distribution of donated food under capacity constraints. IIE Trans 48(3):252–266

Savelsbergh M, Smilowitz K (2016) Stratified patient appointment scheduling for mobile community-based chronic disease management programs. IIE Trans Healthc Syst Eng 6(2):65–78

Sengul Orgut I, Ivy J, Uzsoy R (2017) Modeling for the equitable and effective distribution of food donations under stochastic receiving capacities. IISE Trans 49(6):567–578

Toro-Díaz H, Mayorga ME, McLay LA, Rajagopalan HK, Saydam C (2015) Reducing disparities in large-scale emergency medical service systems. J Oper Res Soc 66(7):1169–1181

Vitoriano B, Ortuño MT, Tirado G, Montero J (2011) A multi-criteria optimization model for humanitarian aid distribution. J Glob Optim 51(2):189–208

VonAchen P, Smilowitz K, Raghavan M, Feehan R (2016) Optimizing community healthcare coverage in remote Liberia. J Humanit Logist Supply Chain Manag 6(3):352–371

Burcu Balcik is an Associate Professor in Industrial Engineering Department at Ozyegin University, Istanbul, Turkey. She received her Ph.D. in Industrial and Systems Engineering from University of Washington, and her M.S. and B.S. degrees in Industrial Engineering from Middle East Technical University. Dr Balcik has also been a Postdoctoral researcher in Industrial Engineering and Management Sciences Department at Northwestern University, a visiting researcher at HUMLOG Institute at Hanken School of Economics, and a visiting professor at HEC Montreal and Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT). Dr Balcik is a recipient of the Turkish Science Academy's Young Scientist Award. She recently served as President of the Public Sector Operations Research Section INFORMS. She is a member of the executive board of the EURO Working Group on Humanitarian Operations. She is on the Editorial Boards of the Journal of Humanitarian Logistics and Supply Chain Management, Productions and Operations Management, and Journal of Operations Management.

Dr Balcik's research focuses on the design and management of humanitarian supply chain networks and logistics operations. Since her PhD, she has undertaken numerous projects addressing humanitarian operations, which has contributed to the advancement of this field. She particularly enjoys working with humanitarian agencies to develop strategies and methods that support their strategic and operational decision making processes.

She was born in Turkey, and attended a STEM specialized high school in Istanbul, in which only a quarter of students were female. She particularly liked studying mathematics in high school and chose to study industrial engineering at the Middle East Technical University. She chose industrial engineering due to the flexibility it offers. She appreciates that her choice has eventually led to an academic career focusing on the exciting field of humanitarian logistics, and that her work could make a positive impact on the lives of populations living under disaster risk.

Karen Smilowitz, Charles Deering McCormick Professor of Teaching Excellence, is a Professor in the Department of Industrial Engineering and Management Sciences at Northwestern University. Dr Smilowitz is a leading expert in modeling and solution approaches for logistics and transportation systems in both commercial and non-profit applications, working with transportation providers, logistics specialists, and a range of non-profit organizations. Dr Smilowitz received a CAREER award from the National Science Foundation and a Sloan Industry Studies Fellowship. Dr Smilowitz is the founder of the Northwestern Initiative on Humanitarian and Non-Profit Logistics. She has been instrumental in promoting the use of operations research within the humanitarian and non-profit sectors through the Woodrow Wilson International Center for

Scholars, the American Association for the Advancement of Science, and the National Academy of Engineering, as well as various media outlets. Dr Smilowitz recently served as President of the Transportation Science and Logistics Society within INFORMS. Dr Smilowitz is an Associate Editor for Transportation Science and Operations Research. In 2016, Dr Smilowitz received the Award for the Advancement of Women in OR/MS from the Women in OR/MS Forum of INFORMS.

Dr Smilowitz began her career in industrial engineering and operations research through the study of transportation systems; however, her career evolved over the past twenty years to include a significant focus on humanitarian and non-profit logistics. Work in this area allows her to combine her interests in design and optimization of systems with her involvement with various non-profit organizations. Dr Smilowitz truly enjoys her work with female researchers at the undergraduate, graduate, and post-doctoral level. Writing a chapter for this series has been a wonderful opportunity for her to work once again with her former post-doctoral advisee, Dr Burcu Balcik.

Chapter 17
Simulation-Based Approach to Evaluate the Effects of Food Supply Chain Mitigation and Compliance Strategies on Consumer Behavior and Risk Communication Methods

Jessye Talley and Lauren B. Davis

Contents

17.1 Introduction 392
17.1.1 Chapter Outline 393
17.2 Consumer Behavior and Food Supply Chain Safety 394
17.2.1 Communication with the Consumer 394
17.2.2 Consumer Compliance 395
17.2.3 Intervention Strategies 396
17.2.4 Opportunities for Food Supply Chain Modeling 397
17.3 Literature Review 397
17.3.1 Food Supply Chain Models 397
17.3.2 Consequence Assessment Models 398
17.3.3 Consumer Behavior 398
17.3.4 Interventions 399
17.3.5 Research Contribution 399
17.4 New Methods to Model Food Supply Chains 400
17.4.1 Consequence Assessment Compartment Model 400
17.4.2 Agent-Based Simulation Model 407
17.5 Future Research Opportunities for Food Supply Chain Contamination, Mitigation, and Risk Communication 412
References 413

J. Talley
Morgan State University, Baltimore, MD, USA

L. B. Davis (✉)
North Carolina A&T University, Greensboro, NC, USA
e-mail: lbdavis@ncat.edu

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_17

17.1 Introduction

Approximately 351,000 deaths occur annually on a global scale due to food poisoning (Sifferlin 2015). Most cases of food poisoning can be traced back to the presence of physical, biological, or chemical contaminants introduced at some point in the food supply chain. According to the Centers for Disease Control (CDC), there are over 250 foodborne diseases caused by the presence of contaminants introduced accidentally or intentionally. This has led to food safety concerns on a national and global scale.

In the United States, several key stakeholders play a vital role in protecting the food supply against intentional and accidental contamination incidents. The primary organizations within the federal government are the United States Department of Agriculture (USDA), The Food and Drug Administration (FDA), and the Centers for Disease Control and Prevention (CDC). The USDA Food Safety and Inspection Services (FSIS) is responsible for ensuring meat, poultry, and processed egg products are safe. FSIS creates regulations for food safety, inspects and enforces the regulations at food processing facilities, and works collaboratively with other agencies to respond to contamination incidents (USDA 2014). The FDA is also responsible for ensuring the safety of the food supply. However, their specific focus is on products not addressed by the USDA (American Bar Association 2014). Some of their responsibilities include creating regulations that food producing companies must follow, ensuring firms are compliant with food safety regulations, performing inspections and product sampling, conducting research into new technologies for inspection, and creating food defense tools that can be used by companies to prevent and respond to food safety incidents (FDA 2017). They also work with the CDC to track cases of foodborne illness. The CDC tracks information about foodborne illness to assist with risk mitigation and response efforts.

Developing optimal risk mitigation strategies requires an understanding of the prevalence of certain contaminants and food type combinations as well as any trends in consumer behavior. In addition, Batz et al. (2011) define three components to use a risk-based approach for food safety: (1) gather information on food that causes significant risk from a public health perspective, (2) prioritize resources and information to reduce these risks based on their effectiveness and cost associated with particular interventions, and (3) develop the optimal intervention and allocate the necessary resources. Batz et al. (2011) address the first component as part of their research and note that there are four high-risk food groups: produce, baked goods, dairy, and meat (Batz et al. 2011). These food groups are associated with the nutrients needed for daily intake. Besides these high-risk foods there are other trends associated with consumer behavior that impact food safety and the increase in foodborne illness. Consumers are dining at restaurants that grow their own produce or herbs outside of the establishment (Kramer and Fasone 2016). This causes regulatory concern with the food being grown in an environment susceptible to contaminants (Kramer and Fasone 2016). More consumers are seeking locally grown produce foods which when used in other products causes

challenges for tracing the product back to the origin (Kramer and Fasone 2016). Consumers are also more focused on healthy eating. In 2016, there was an increase in the amount of recalls for frozen fruits and vegetables used in other products and smoothies (Kramer and Fasone 2016). New food services are available that deliver both cooked and uncooked meals to homes based on consumer demand for convenience foods (see for example companies likes Blue Apron and Hello Fresh). However, there is no way to identify if the food has been stored at the proper temperature for consumption. Lastly, consumer sensitivity to certain food products coupled with mass processing of convenience foods has contributed to several recalls due to cross-contamination or mislabeling. For example, food companies are mandated to label foods that are considered gluten free (GF) according to the standards given by the FDA (Kramer and Fasone 2016). Many restaurants have stated that they are not asked to validate GF during food safety inspections nor given the proper tools to determine the concentration of gluten (Kramer and Fasone 2016).

In this chapter, we will explicitly examine this relationship between consumer behavior and food safety. We use purchasing and consumption behavior as a proxy for estimating consumer preferences for certain convenience foods. We also incorporate a consumer compliance component that represents the absence or delay associated with receiving and responding to recall information.

17.1.1 Chapter Outline

This chapter discusses the operational challenges and latest research in food supply chain risk mitigation (FSCRM). We discuss some novel approaches to investigate food supply chain risk management bringing special attention to those models within the operations research/management science discipline. We also discuss new approaches to inform decision-making in FSCRM based on epidemic models and agent-based simulation. The epidemic model is based on the well-known SEIR (susceptible-exposed-infective-recovered) model. We adapt this model to determine the number of people that can become ill based on their purchasing and consumption behavior. We also present an agent-based modeling methodology that incorporates three interacting components within the food supply chain: the consumer, the manufacturer, and public health. The first component considers the human population as agents that can progress through various stages of health. Their health status is updated based on their behavior (product purchasing/consumption). The second component incorporates the recall process issued by the manufacturer. We specifically consider a recall intervention because it is the most common strategy used by companies to reduce the spread of the foodborne illness caused by a contaminated food product. The third component consists of assessing different messaging techniques during these food contamination events. A numerical study is conducted to understand the impact of consumer behavior and messaging on the spread of foodborne illness.

The remainder of this chapter is organized as follows. The second section presents the related challenges and opportunities associated within the food supply chain. The third section discusses the literature for food supply chain contamination, consumer behavior, and messaging. The fourth section presents two new approaches and corresponding results for modeling food supply chain contamination incidents. The conclusions and suggestions for future work are presented in Sect. 17.5.

17.2 Consumer Behavior and Food Supply Chain Safety

This section discusses some of the challenges associated with food safety and defense and the role of consumer behavior along three main themes: (1) communicating with the consumer, (2) measuring the consumer's level of compliance, and (3) applying various intervention strategies. Tracking contamination events is one of the first lines of defense for reducing the risk of illness to the consumer. However, communicating with the consumer, under-diagnosis and under-reporting of illness, and new food and toxin combinations make tracking contamination events a challenging process.

17.2.1 Communication with the Consumer

Many consumers do not receive information regarding contamination events until a recall is issued or they are contacted by a distribution channel with the information (Hallman and Cutie 2009). Currently, the easiest way for a retail distribution channel to inform the consumers of this event is by telephone or email since the store can track consumer food purchases through their rewards program. However, all consumers do not necessarily use this program. Furthermore, certain segments of the population (e.g., minorities and underserved populations) might not receive or have access to food contamination and recall information. According to the survey done by Steelfisher et al. (2010), African Americans (86%) and Hispanics (81%) were less likely to remember a recall during the time of their survey compared to whites (94%). While government websites contain information about recalls, approximately 15% of American consumers that used those sites indicated the sites were not organized in a way to make it easy to find the food products involved in a recall.

The National Research Service reported that 49% of managers from the FDA felt that communication and coordination would help them to improve their ability to carry out the mission. Once a recall ends, most consumers are unaware unless a food agency shares the information on their website (Government Accountability Office 2012).

According to a Rutgers survey on improving communication during recall events (Hallman and Cutie 2009), less than 60% of American consumers have checked

their home for a recalled food product. Even with information from the media about contaminated food, 12% of consumers stated that they ate a food product that was recalled. Out of this 12%, 9 respondents from the study thought they became ill from consuming the food product, with only 4 respondents taking action to go the doctor for the symptoms. Under-diagnosis and reporting also contributes to the lack of timely and accurate tracking of foodborne outbreaks. The Public Health Department in each state is responsible for reporting instances of foodborne outbreaks to the CDC for tracking and reporting purposes. Since, most consumers do not go to their healthcare provider to report symptoms, this leads to under-diagnosis and reporting. As a result, there is uncertainty in the amount of people affected by the contamination event.

17.2.2 Consumer Compliance

Although food agencies give notice to the public through various avenues, consumers do not heed the warnings. Many of these same consumers still eat food that is unsafe because they do not deem it to be a risk or they have not received any information on a recall being issued.

The FDA has experienced challenges with coordinating with other agencies to issue recalls appropriately to give consumers accurate information (Government Accountability Office 2012). In 2006, the FDA issued a warning about spinach that resulted in companies losing a large amount of money. In addition, many agencies do not give out information until they receive it which may leave consumers confused on what steps to take to discard food (Government Accountability Office 2012).

Many consumers, after purchasing and consuming food, are often unaware that some of their food puts them at risk from contaminants. For this reason, recalls occur. During the recall process, some consumers may choose to follow the guidelines to discard food or to not comply.

Steelfisher et al. (2010) did telephone surveys with regard to two major food recalls. This survey evaluated the consumers' thoughts on food recalls and how it impacts their actions for the future.

The findings from this survey shows (1) better communication is needed on the actual products that are a part of a recall; (2) it is helpful to use media outlets such as television or radio to send out information to consumers, and (3) the need to reach out to minority populations about recalls and foodborne illness. Saulo and Moskowitz (2011) develop a set of food safety messages, which were shown to a population of respondents through the Internet. These respondents numerically answered how they would react to the various messages. From this data, conclusions were drawn on how specific messages can influence various demographics and behavior of consumers. Freberg (2012) uses a consumer panel survey to understand the effects of compliance, based on messages received from an organization versus those generated by a user. The results show that more people comply with messages received from organizations.

17.2.3 Intervention Strategies

The food protection plan lists interventions as one of its core elements (FDA 2007). An intervention is defined as targeted inspections and testing that verify that preventive controls are working and that resources are applied to high priority areas (FDA 2007). Interventions can be proactive, prior to food distribution, or reactive in response to a foodborne outbreak. Some of the common proactive intervention strategies are inspections, sampling, and surveillance within the supply chain. The most well-known reactive strategy is a product recall. Recalls can be initiated by the product manufacturer or by the FDA. The Federal Drug Administration (FDA) has developed standards to help determine if a food contamination event should result in a recall; however, the FDA is not given sufficient data in order to issue a recall in a timely manner (Government Accountability Office 2012). For example, their internal procedures do not help them effectively conclude that there is enough evidence to proceed with a recall.

Many recalls are issued by food companies that are possibly at fault. However, this results in false alarms, loss of revenue and customer support. There are mechanisms in place to help companies recover from the negative effects associated with a recall or advisory that might have been false; however, they are not always beneficial.

If they are falsely accused, companies are able to move through the judicial process to take the FDA to court. Companies may also be eligible to participate in government purchasing to improve their image with the public. For example, the Federal government provided one-time funding for tomato growers in South Carolina after the 2008 outbreak was incorrectly linked to their farms (Government Accountability Office 2012). Companies may also take out loans or insurance for these cases. This gives cause to consider the guidelines and protocol associated with timing of implementing a recall or other type of intervention.

Since there are a diverse set of contaminants that can enter the food supply chain, it is challenging to know the best time to initiate the appropriate intervention strategy.

To address these challenges, food companies need to test a variety of intervention strategies on contamination events with different levels of severity to improve food safety and defense. By utilizing information from past events, it is possible to develop baseline rates that will help food companies and organizations to respond to these types of events. It is also important to incorporate inspection and detection technology to help with removal of tainted food from the production phase of the food supply chain.

The third challenge focuses on the emerging pool of food and toxin combinations. The CDC maintains databases that track and use surveillance data through the National Outbreak Reporting Systems in order to link contamination information over states, agents, or food responsible for illness to the FOOD Tool. Discovery of these new agents can lead the way to create new technologies, techniques, and studies. They can provide more insight into how toxins evolve and track their

growth. Public health officials will also need to develop new guidelines and courses of action to handle new cases.

17.2.4 Opportunities for Food Supply Chain Modeling

The prior sections illustrate the role of the consumer, beyond just someone who purchases the product. Reducing foodborne illness requires a coordinated effort to communicate with the consumer effectively about recalls, get the consumer to recognize and report foodborne illness to improve tracking of these events, and also getting the consumer to comply. Each of these challenges will be addressed through the agent-based simulation model.

17.3 Literature Review

Food supply chain contamination is a growing concern due to the rise in consumer illnesses and recalls. Within the literature there are various models to assess consequences, recall strategies, as well as individual information about a consumer and risk communication. However, there is limited research that utilizes consumer behavior data in these models.

17.3.1 Food Supply Chain Models

The food supply chain models focus on three main areas: production, distribution, and intervention strategies. Vlijac et al. (2012) characterize a set of supply chain disturbances and vulnerabilities which are used to develop redesign principles for supply chains. This framework is tested on a meat supply chain. Rong and Grunow (2010) develop a model that sends food from a distribution facility in batches to the retailer. They consider the effects of dispersion by reducing the amount of batches that are sent to specific retailers in order to prevent a recall. Akkerman et al. (2010) gives a summary of the literature related to models that solve diverse problems with food quality, distribution, and planning. Chen et al. (2013) create a model to evaluate quality control methods in the Chinese dairy industry under two supply chain structures. For each structure, they calculated the cost of sampling strategies from the retailer and supplier perspective. Buchanan and Appel (2010) discuss the integration of analysis and mathematical models to enhance information used by risk managers to meet regulatory requirements of food safety and defense. Manizini and Accorsi (2013) introduce a framework for an integrated food supply chain which considers the following characteristics such as quality, safety, efficiency, food products, and food processes. Chebolu-Subramanian and Gaukler (2015) propose an analytical model which is validated through simulation to study the origin and mode

of detection for contamination on various food supply chain designs where illness has already occurred.

There have been some studies based on agent-based simulation modeling. Knowles-McPhee (2015) presents an agent-based model that tests three inspection strategies at a retailer in order to understand the interaction of consumers, retailers, and inspectors. Chaturvedi et al. (2014) develop an agent-based model that can be used for food defense training. The model incorporates the total food supply chain to help companies understand and prepare for a food contamination event. Crooks and Hailegiorgis (2014) create an agent-based model to show the interactions between humans and their environment by using a Susceptible-Exposed-Infected-Recovered (SEIR) model as the underlying structure. This model considers the spread of cholera from unclean water resources and is used to understand the humanitarian relief perspective. Zechmann (2011) uses a multiagent-based approach to determine the optimal mitigation strategies after a water contamination event.

17.3.2 Consequence Assessment Models

There is a limited number of food consequence assessment models, but they focus on various areas to capture information to determine which mitigation strategies to implement during some event. One of the main characteristics of these models is to determine the number of people that will be affected by an event. Many models capture this by counting the total number of casualties of a food contamination event (Liu and Wein 2005, 2008; Hartnett et al. 2009). Another characteristic is the effect of contaminants that enter into the supply chain through various modes such as production facilities, distribution, and retailers (Jaine 2005; Liu and Wein 2005, 2008). After food contamination takes place, some models consider the impact of public health response directed at consumers (Hartnett et al. 2009). Lastly, none of these models consider the effects of consumer behavior on their risk of illness and compliance.

17.3.3 Consumer Behavior

Various studies and surveys are used to capture information on consumers' behavior in relation to food choices, purchases, consumption, food handling, and compliance. Erongul (2013) reports on a survey to evaluate the relationship between consumer food safety and consumption patterns. The results from this study show that more consumers need education in best practices for handling food in order to prevent illness. Grunert (2002) presents a Total Food Quality Model framework. This framework is used to understand the consumers' perception of food quality, food technology, and changes in behavior due to a contamination event. Caraballo-Martinez and Burt (2011) use a survey to collect data on the changes in a consumer's

choice in grocery or household products. This behavior can assist retailers in determining which products to sell and whether to open new stores.

17.3.4 Interventions

Fendyur (2011) provides an overview of the operations research techniques used to handle outbreak management of infectious diseases. Dasakalis et al. (2012) provide a review of models that focus on epidemic control and logistics with many applications. Many researchers have developed models that focus on control measures related to public health measures for diseases and food.

Liu and Wein (2005) develop a differential equation model that implements two antibiotic residue testing of trucks after an attack occurs. With testing, the number of people affected reduced to half for one strategy. The other strategy when used in isolation prevents even more people from becoming poisoned. Liu and Wein (2008) consider detection after a certain amount of consumers develop symptoms from eating a contaminated food product. After this threshold is reached, all food consumption is stopped. Hartnett et al. (2009) create a discrete event simulation model. One component of the simulation considers the number of reported cases. Once this number reaches a threshold, then an advisory is sent out to the public. They issue a delay to determine the point of contamination. After this is confirmed, exposures can still occur based on the way consumers comply to the advisory and all products are moved out of the supply chain. Jaine (2005) develops a simulation that has an interface for food safety officials to use to pick the type of intervention based on the outbreak. The model uses information from literature, various websites, and data sets that were provided by certain companies. Chang et al. (2015b) develops a partially observable Markov game which allows two agents to interact based on little information to make decisions on how to affect the overall system. A food supply chain is used to illustrate this model. Chang et al. (2015a) uses a partially observable multi-objective game, to help a production manager determine the best way to maximize productivity of a supply chain and minimize the number of products that leave the production facility contaminated. Chebolu-Subramanian and Gaukler (2015) use sampling during the production process to detect contamination in food. Chen et al. (2013) uses an analytical model to do an acceptance sampling test of food products to make sure they conform to specifications required by the retailer. Based on the results from this test, they can choose to accept or reject the production lot.

17.3.5 Research Contribution

While agent-based models have been used to study some contamination problems, there are very few for the food supply chain. The papers discussed in the literature

Table 17.1 Paper by characteristics

Author	Agents	Applications	Consumer behavior	Compliance
Knowles-McPhee (2015)	Consumer, Retailer, Inspector	Food	No	No
Chaturvedi et al. (2014)	Individual Refugees, Leaders, Citizens, Organizations, Rebels, Institutions, Infrastructure	Refugees	Yes	No
Crooks and Hailegiorgis (2014)	Refugees	Non-linear	Yes	Yes
Zechmann (2011)	Consumer and Utility Operator	Water Distributions Systems	Yes	Yes
Talley (2016)	Consumers and Food	Food	Yes	Yes

review, related to food consequence assessment models, do not focus on the consumers at the individual level. In this research, we show the progression of illness by using the following parameters: the rate of purchase, the rate of consumption, and the time until illness and recovery. The time it takes to purchase and consume the contaminated food products is considered stochastic. The literature presented does not consider illness from the perspective of the consumer. This research merges the characteristics of the consequence assessment models and food supply chain models with consumer behavior, specifically purchasing, consumption, and compliance (Table 17.1).

17.4 New Methods to Model Food Supply Chains

Two new methods are presented to model food supply chains. The first method is based on epidemic and compartmental models. The second method is an agent-based simulation model.

17.4.1 Consequence Assessment Compartment Model

Methodology

Problem Overview

This research considers a food supply chain with the following structure: (1) The producer is a farm or company that produces an original product; (2) Processors

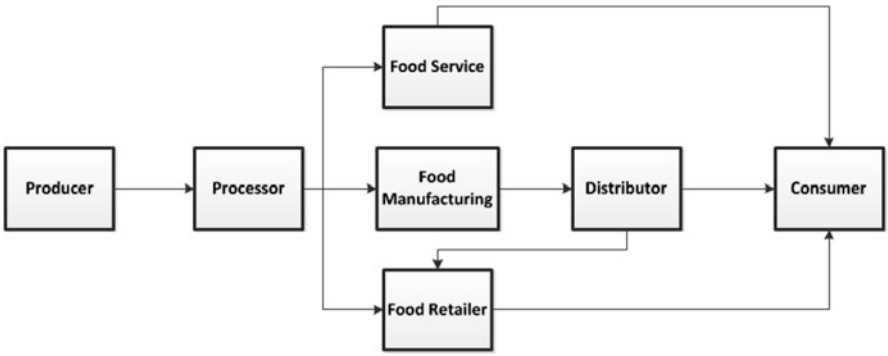


Fig. 17.1 The food supply chain

are farms or companies that change the original product into a different form; (3) Distributors supply food products to different companies and industries; (4) Food reaches a consumer from one of three different distribution channels: (1) Food Retail, (2) Food Manufacturing, and (3) Food Service (Fig. 17.1).

A chemical or biological agent can be released at different stages of the food supply chain.

Distribution channels receive impure food, which consumers purchase and ingest. From this attack the amount of casualties are determined based on symptoms that affect consumers from the agent. A deterministic SEIR model is presented to show the purchase and consumption behavior of consumers assuming a non-constant population.

Model Description and Assumptions

A deterministic ordinary differential equation progression model is developed to illustrate food contamination within consumer populations. The model shows the progression of a consumer from a healthy to unhealthy state because of eating contaminated food. The phase of progression is as follows: (1) Susceptible (S) population represents the people who purchase contaminated food products from a distribution channel but have not yet consumed the product; (2) Exposed (E) population represents the people that consume a contaminated product but show no signs of illness; (3) Infected (I) population represents the people that show signs of illness after consuming a contaminated product; (4) Recovered (R) population represents the people that are no longer ill after a single outbreak. This model is different from the original SEIR model because sickness is a result of contact with a particular food product and not a person. Therefore, there is no interaction term. Based on the flow of each stage we construct a system of differential equations (Fig. 17.2).

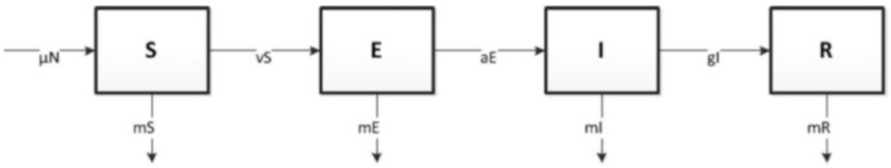


Fig. 17.2 SEIR model flow

Table 17.2 Model notation

Symbol	Description
$S(t)$	The number of customers who have purchased a contaminated product from a distribution channel (DC) at time t
$E(t)$	The number of people that consume a contaminated product but show no sign of illness at time t
$I(t)$	The number of people who become ill after consuming a contaminated food product at time t
$R(t)$	The number of people that are no longer sick after consuming a contaminated food product at time t
N	Total number of consumers in the population
m	The rate which consumers leave any class due to natural cause of death
μ	The rate which consumers purchase food products
ν	The rate a consumer is exposed to a contaminated food product
a	The rate at which a consumer exposed to a contaminated food product becomes ill
g	The rate at which an ill consumer recovers
t	Number of days in the food outbreak

Table 17.2 displays the notation used for the variables and parameters in this model.

Using the notation from (Table 17.2) the deterministic SEIR model for a food contamination event is formulated as follows:

$$\frac{\mathrm{d}S}{dt} = \mu N - mS(T) - \nu S(t) \tag{17.1}$$

$$\frac{dE}{dt} = \nu S(t) - mE(t) - aE(t) \tag{17.2}$$

$$\frac{dI}{dt} = aE(t) - mI(t) - gI(t) \tag{17.3}$$

$$\frac{dR}{dt} = gI(t) - mR(t) \tag{17.4}$$

The total population N becomes susceptible by purchasing tainted food products with rate μ. They can exit this compartment (S) due to natural death with rate m or from exposure to a contaminant in food with rate n (Eq. 17.1). A consumer can move to the exposed compartment with rate v only after they purchase a contaminated product. They exit the compartment with rate m because of natural causes or if they become ill with rate a (Eq. 17.2). The consumers that become ill with rate a can eventually move to recovery with rate g. They can also exit compartment (I) due to natural causes with rate m (Eq. 17.3). Consumers receive some medical care in order to go to the recovered class. They can exit with a natural death with rate m (Eq. 17.4). Consumers can progress through the five stages consecutively to denote their place during a contamination event (i.e., susceptible, exposed, infective, recovered). This model considers a non-constant homogeneous population of consumers that can purchase food items that are always in stock. Contaminated food products are sent to a distribution channel (DC) (i.e., food retail, food service) for purchase and consumption. The consumer illness is based on the interval of time between the consumer purchasing and ingesting a contaminated food product. This model represents a worst case where food continuously enters the market undetected by the producer.

Experimental Design

The purpose of this research is to understand the effects of various population parameters on the number of people affected by food contamination. The parameters considered in this study are population size, consumer consumption behavior, consumer purchasing behavior, and discarding policy. A numerical example is developed to answer the research questions in Table 17.3.

Each model case uses data from a food and consumer behavior survey which includes time until symptoms occur from contaminants and the food shelf life (Watkins 2015). There were 83 responses to the survey, which capture various demographic, consumer purchasing and consumption behavior information. The limitations of this study were the lack of diversity and size of the sample population.

Table 17.3 Research questions

Number	Question
1	What percentage of consumers found in each compartment (S,E,I,R) change as consumption and purchasing increase?
2	How do consumer stocking policies affect the risk of exposure to impurities in food products?

Table 17.4 Notation for parameter computation

Symbol	Description
v^{-1}	The average time until initial consumption for each food type f
c_{fr}	The time until consumption for food type f and respondent r
μ_{f}	The average purchasing rate per day for each food type f
j_{fr}	The purchase frequency (in days) for food type f and respondent r over 30 days
m_{f}	The shelf life rate for food type f
h_{f}	The shelf life of food type f
M	The total number of respondents

However, two questions were used from this survey to determine purchase and consumption rates.

Question 1. In the last 30 days, how many times did you buy the following food from the grocery store?

Question 2. How many days after purchase do you store the food (in the refrigerator or pantry) before you first eat it?

Stocking policies were created for each food type. Dairy, vegetables, bread, and eggs have a short shelf life, which is five days or fewer. This could result in people stocking these products more often because they are consumed faster. Baked goods have a medium shelf life which is two to three weeks with meat having the longest shelf life of over one month. Using the data from Question 2, a range was created for each food type to represent the time until consumption. The data from Question 1 was used to develop the purchasing rates.

Table 17.4 summarizes the notation used to develop these rates. Equation 17.5 calculates the consumption rate by using the average time until consumption in days for all food types and responses. Equation 17.6 calculates the purchasing rate using the average frequencies that a respondent shops for each food type. Equation 17.7 calculates the shelf life rate using the shelf life of the various food types. The shelf life data is obtained from publicly available sources (Tasty 2015).

$$v^{-1} = \frac{1}{M} \sum_{r=1}^{M} c_{\text{fr}} \tag{17.5}$$

$$\mu_f = \frac{1}{M} \sum_{r=1}^{M} \frac{j_{\text{fr}}}{30} \tag{17.6}$$

$$m_{\text{f}} = \frac{1}{h_{\text{f}}} \tag{17.7}$$

Results

Single parameter sensitivity analysis

The SEIR model captures the amount of consumers that purchase, ingest, become ill and recover due to contaminated food products. Figure 17.3 shows the general behavior of the SEIR Model for the case where there are no interventions introduced into the model. The total population of 100 consumers all purchase food products that are contaminated. The long-term behavior shows that the amount of people purchasing contaminated food products decreases. This is a result of consumption and consumer illness. The number of consumers exposed to the contaminants peaks at around 3 days. The long-term behavior shows that eventually the amount of people exposed reaches a steady state. The number of ill consumers increases based on the rate of consumption. Over time all of those consumers receive some treatment that allows them to recover from their illness.

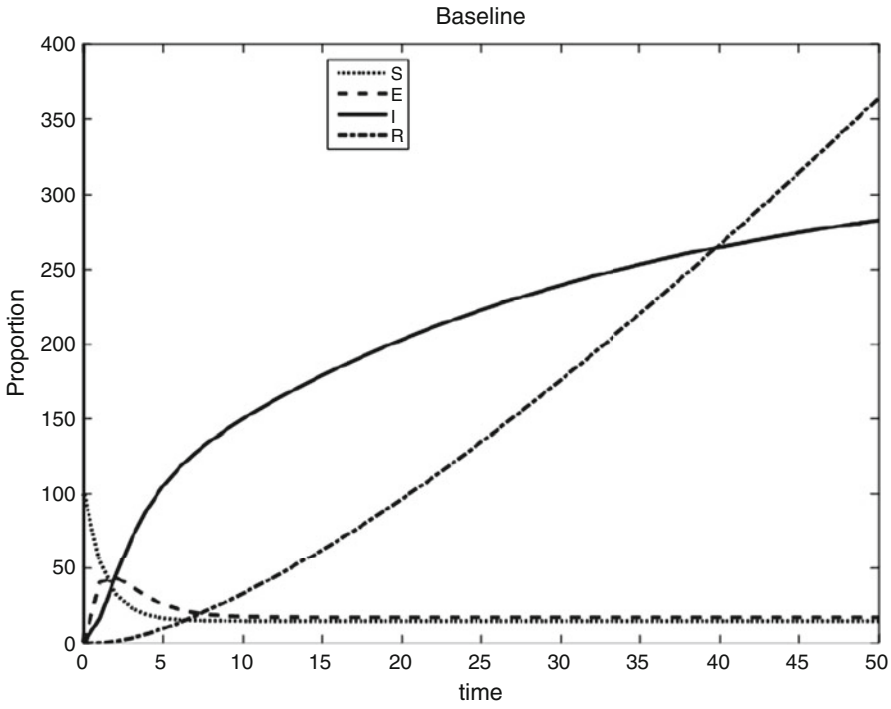


Fig. 17.3 Baseline results

Purchasing Behavior

Figure 17.4 shows the amount of people exposed to contaminated food based on consumption and purchasing behavior. In general, the exposure percentage increases as the purchasing rate increases. In addition, a 50% increase in the purchasing rate results in a 50% increase in the exposure percentage. The relationship between the purchasing and consumption rate is intuitive and serves to validate the model. Also, an increase of population size increases the number of people that become ill.

Vegetables are a short shelf life product which consumers purchase at a rate of 4.55 every 30 days; this results in 29.87% of consumers being exposed to salmonella. Baked goods have a medium shelf life which consumers purchase at a slower rate of 1.24 every 30 days; this results in 8.12% of the total population being exposed to contaminants.

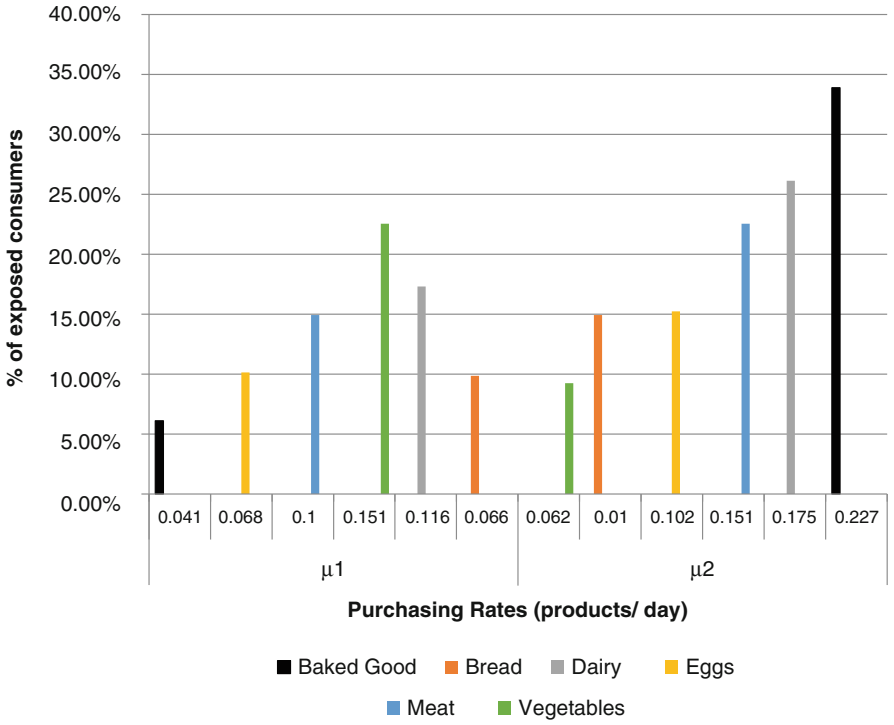


Fig. 17.4 The percentage of consumers that are exposed to contaminated food products based on purchasing rate

17.4.2 Agent-Based Simulation Model

Methodology

This section describes the agent-based simulation model based on the Overview, Design Concepts, and Details ODD framework (Grimm et al. 2006).

Purpose

The purpose of this study is to model the effects of three types of consumer behavior: purchasing, consumption, and compliance, on (1) the food supply chain, (2) risk of illness, and (3) recall.

State Variables and Scales

The intervention simulation will consist of two agents: the consumers and food (Figs. 17.5 and 17.6). The consumer population moves into the purchase state where they can buy food products from two different distribution channels (i.e., food service and food retail). After purchasing products, consumption takes place and people can become ill. Based on the threshold of consumers that can become ill, a recall or some other intervention can occur. If the recall occurs, all stakeholders are notified and a warning given so that contaminated food products are removed from the shelf. Also, at this time a message is sent to the consumer to give warning that the food they purchased has been contaminated. Lastly, a consumer can recover from illness. The food agent models the recall process. Table 17.5 shows the notation used for this model.

Process Overview and Scheduling

A consumer progresses through different compartments based on their health status. Based on the rates associated with the agent, some may progress faster than others.

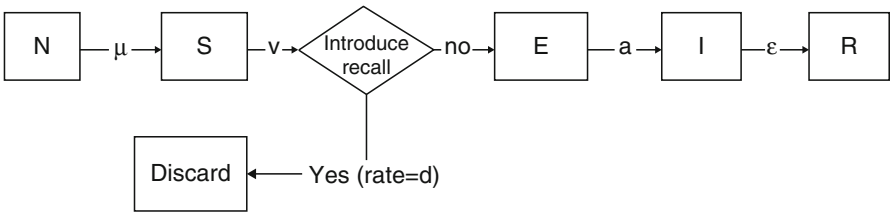


Fig. 17.5 Consumer statechart

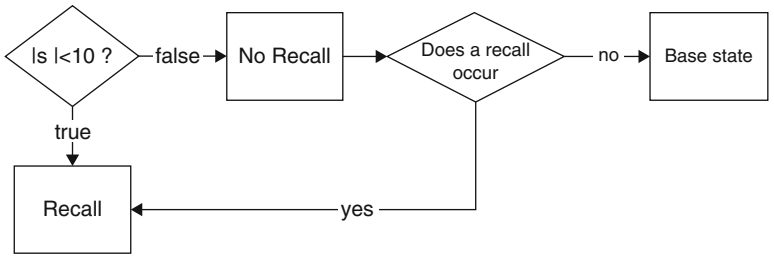


Fig. 17.6 Food statechart

Table 17.5 Food simulation model notation

Notation	Description
Susceptible (S)	The number of consumers who have purchased a contaminated product from a DC at time t
Exposed (E)	The number of people that consume a contaminated product but show no sign of illness at time t
Ill (I)	The number of consumers who are symptomatic after eating a contaminated food product at time t
Recovered (R)	The number of people that are no longer sick after consuming a contaminated food product at time t
Totalpop (N)	Total number of consumers in the population
Discard	The number of products discarded after a recall is issued
Recall	An intervention strategy to remove contaminated food products
No Recall	No products are removed
isRecall	The probability of a recall occurring
FoodRecall (Introduce Recall)	Signifies to consumers to discard food
computer (Introduce Recall)	The number of messages sent out to the consumer during the recall process
badfood	Signifies that food is contaminated
μ	The rate which consumers purchase food products
ν	The rate at which a consumer exposed to a contaminated food product
a	The rate at which a consumer exposed to a contaminated food product becomes ill
ε	The rate in which an ill consumer recovers
t	Number of days in a food outbreak

Once the number of ill consumers is greater than the threshold, the recall occurs. Based on this recall, the consumer compliance behavior is tested to see if they discard or keep products.

Design Concepts

For this simulation, the general food supply chain is considered; however, all parts are not modeled explicitly: (1) producer, (2) distributor, (3) retailer, and (4) consumer. For the recall process, we use a threshold value to signify that a food contamination outbreak has occurred; at the same time as testing of products is initiated. Lastly, we utilize information of consumer behavior to model purchasing, consumption, and compliance.

Emergence

The results from this model will show (1) the number of people that are affected in relation to compliance measures and (2) the number of products that were discarded as a result of the recall process.

Adaptation/Learning

Consumers update their compliance behavior based on information regarding a recall. The recall process is updated based on consumer purchasing and consumption behavior.

Objectives

The objectives of this research are (1) to understand how consumer compliance reduces illness and (2) the effects of uncertainty in consumer purchasing and consumption behavior.

Results

Table 17.6 displays the number of people that are affected from a food contamination event when compliance is introduced. For each compliance rate there are fewer people that are exposed and ill at the end of the time period. However, most consumers eventually progress to the recovered class at the end of the time period. Overall the results show that with a lower discard rate, less consumers (748.08) comply.

Table 17.6 Number of consumers affected based on compliance

Compliance rate	E	I	R	EIR
0.25	0.04	0.24	747.8	748.08
0.75	0.04	0.14	251.4	251.58

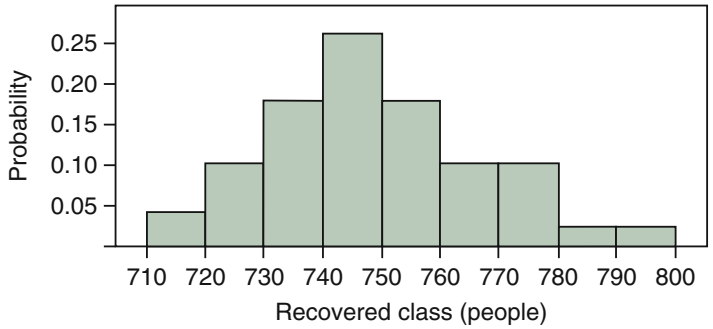


Fig. 17.7 Recovered class with compliance rate = 0.25

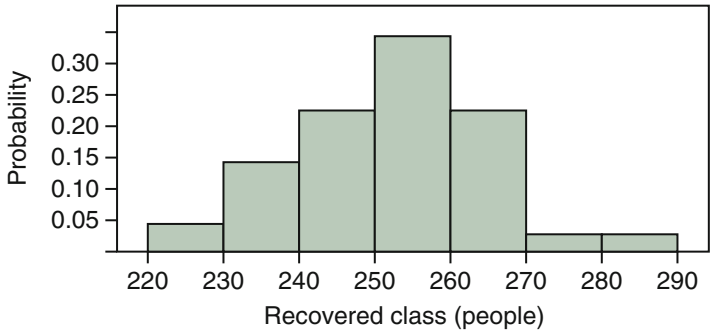


Fig. 17.8 Recovered class with compliance rate = 0.75

Table 17.7 Statistics for recovered class

Compliance rate	Mean	Std. dev.	Min	Max
0.25	747.8	17.2	714.0	790.0
0.75	251.4	12.9	223.0	284.0

Table 17.8 Statistics for discard class

Compliance rate	Mean	Std. dev.	Min	Max
0.25	251.84	17.23	210.00	286.00
0.75	748.26	12.82	719.00	777.00

Figures 17.7 and 17.8 display graphs of the distribution of the recovered and discard class, which has the most impact on the number of people affected by contamination based on compliance.

Table 17.7 and 17.8 provide descriptive statistics for each class. At a lower discard rate, the recovered class shows more variability (17.2) in the number of consumers that are affected compared to a higher discard rate (12.9) (Figs. 17.7 and 17.8, Table 17.7). This represents people that receive information about a contaminated food product but ignore the warnings. It also represents people that may have already consumed the product before they were aware of any information about contamination.

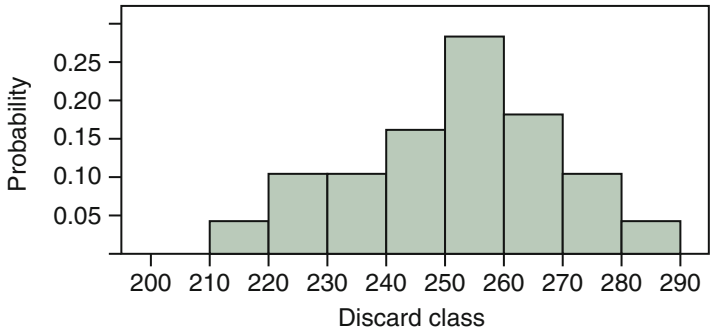


Fig. 17.9 Discard class with compliance rate = 0.25

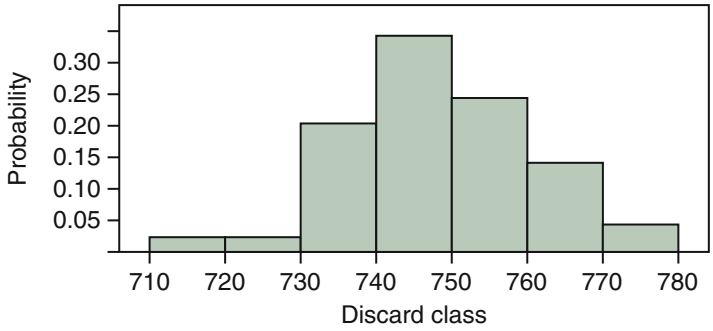


Fig. 17.10 Discard class with compliance rate = 0.75

Table 17.9 Statistics for 0.25 discard rate with messaging

Compartment	Mean	Std. dev.	Min.	Max
E	179.32	12.03	154.00	210.00
I	50.57	6.25	37.00	62.00
R	237.73	12.51	208.00	270.00
Discard	152.38	11.165	125.00	183.00

Table 17.10 Statistics for 0.25 discard rate without messaging

Compartment	Mean	Std. dev.	Min.	Max
R	747.80	17.20	714.00	790.00
Discard	251.84	17.23	210.00	286.00

Similarly, with the discard class, the lower discard rate also shows more variability (17.23) than the higher discard rate (12.82) (Figs. 17.9 and 17.10, Table 17.8). However, a higher discard rate results in more consumers heeding the information of food safety officials, especially for major food outbreaks (Fig. 17.10). The information presented to consumers from food safety organizations can promote better prevention of illness.

Tables 17.9 and 17.10 display the descriptive statistics for the main compartments associated with a person coming into contact with a contaminated food product with

Table 17.11 Relative increase and decrease of compliance of 0.25

Compartment	Relative value
R	0.6820
Discard	0.3938

using messaging and without messaging. Table 17.9 shows that most consumers become exposed to a contaminated food product however they are not remaining sick for long. Also, we can see the number of products has decreased based on introducing messaging to the consumers. This is due to the low compliance policy of 0.25 or consumers behaving contrary to the guidelines (Table 17.9). Overall, without messaging we have more consumers coming through the system without any warning which can keep contaminated food products in homes. Although all consumers do not comply, it is important to observe this behavior to develop the best policies to prevent further spread of contaminated food. Table 17.11 shows us the relative increase or decrease in the recovered and discard class given the messaging and non-messaging case. During the messaging case between 400 and 600 messages were sent to consumers to warn them the food they purchased was contaminated. However, comparing the two cases there was a decrease in the amount of people that recover which shows that more people are receiving the information about the contaminated food products even if they do not necessarily discard the food. The consumers might choose not to eat the food product anymore which shows that over time they may develop new behavior based on warnings.

17.5 Future Research Opportunities for Food Supply Chain Contamination, Mitigation, and Risk Communication

This chapter focuses on challenges and opportunities that arise from intentional contamination in the food supply chain such as (1) tracking contamination events, (2) intervention strategies, and (3) consumer compliance. It addresses each of the challenges by using various approaches. Two models were developed to understand the number of consumers that become ill, the number of products linked to illness, and the effects of a simple intervention strategy. An agent-based simulation model was presented to understand how a simple intervention can affect the number of illnesses and consumer compliance.

The first set of models showed that compartment models can be used to understand population data and progression of symptoms in consumers. It allowed for flexibility in the number of characteristics to show for the population. Results for these models show that even by changing all the population sizes the percentage of people affected from contaminated food remains the same. Risk increases as the consumer purchasing frequency increases for all food types this is due to exposure to contaminants. Depending on the purchasing and consumption behavior some food products result in higher risk. The third model presents an agent-based model to

consider consumer compliance. The results show that as we increase information about compliance, fewer illnesses occur. This modeling approach can be used to analyze consumer behavior at the individual level.

As food contamination events continue to rise, more research is needed to address how to prevent illness from spreading to help food safety officials. The future work will consider the following areas: (1) Consumer characteristics, (2) Data collection, (3) Cost analysis, (4) Intervention and mitigation strategies, (5) Compliance, (6) Public health, and (7) Traceability. The current models only account for a limited number of characteristics however, introducing more into the model could add extra insight especially if the consumer may be in a vulnerable population. More data is needed on consumer behavior as well an information about the recall process and various contamination events to use in the models. This can be used to develop real-world scenarios to test and validate the models. Currently there are a limited amount of studies done to show the impact of cost related to intervention strategies and cost of illness in a food contamination event. This can help manufacturers to develop more robust food defense plans. More research is needed to evaluate various timelines of food contamination and the implementation of intervention strategies. This will allow the FDA and other food organizations to create a baseline for responding to different levels of severity for food contamination events. Although messaging is being used to warn consumers about contaminated food, there is still low compliance. All messaging needs to be able to reach all populations. There needs to be better guidelines in place to instruct consumers on how to dispose or return contaminated food so it can be tested for the agent considered. Another aspect of this problem that needs more development is public health. Many consumers do not report their symptoms or may not think they are symptomatic from eating contaminated food. Better ways are needed to track these consumers so they can receive the proper care and understand the effects of the spread of illness in various regions especially for multi-state outbreaks. This can help identify food outbreaks earlier. Lastly, traceability is becoming harder as the global food supply chain expands. Some areas for development are gaining understanding of how producers and manufacturers work together when using certain food in other products and developing policies to continue traceability throughout the whole supply chain process.

Based on these developments in our food systems it is clear that transparency and safety is key to keep our food safe and consumers supportive because they have healthy food choices.

References

Akkerman R, Poorya F, Grunow M (2010) Quality, safety and sustainability in food distribution: a review of quantitative operations management approaches and challenges. OR Spectr 32(2):863–904

American Bar Association (2014) The role of the government in regulating food: an overview. https://www.americanbar.org/publications/aba_health_esource/2013-14/february/the_role_of_the_government.html. Accessed 7 Feb 2018

Batz M, Hoffmann S, Morris J (2011) Ranking the risks: the 10 pathogen-food combinations with the greatest burden on public health. https://epi.ufl.edu/blog/keeping-americas-food-supply-safe/

Buchanan R, Appel B (2010) Combining analysis tools and mathematical modeling to enhance and harmonize food safety and food defense regulatory requirements. Int J Food Microbiol 139:S48–S56

Caraballo-Martinez K, Burt S (2011) Determining how and why consumer purchasing of grocery and household products varies. Afr J Bus Manag 5(16):6917–6926

Chang Y, Erera A, White C (2015a) A leader-follower partially observed, multiobjective Markov game. Ann Oper Res 235(1)

Chang Y, Erera A, White C (2015b) Value of information for a leader-follower partially observed Markov game. Ann Oper Res 235(1):129–153

Chaturvedi A, Armstrong B, Chaturvedi R (2014) Securing the food supply chain: understanding complex interdependence through agent-based simulation. Health Technol 4:159–169

Chebolu-Subramanian V, Gaukler G (2015) Product contamination in a multi-stage food supply chain. Eur J Oper Res 244:164–175

Chen C, Zhang J, Delaurentis T (2013) Quality control in food supply chain management: An analytical model and case study of the adulterated milk incident in china. Int J Prod Econ 152:188–199

Crooks A, Hailegiorgis A (2014) An agent-based modeling approach applied to the spread of cholera. Environ Model Simul 62:164–177

Dasakalis T, Pappis C, Rachaniotis N (2012) Epidemics control and logistic operations: A review. J Prod Econ 139:393–410

Erongul B (2013) Consumer awareness and perception to food safety: a consumer analysis. Food Control 32:461–471

FDA (2007) Food protection plan

FDA (2017) Where did the FDA come from, and what does it do? https://www.smithsonianmag.com/science-nature/origins-FDA-what-does-it-do-180962054/. Accessed 7 Feb 2018

Fendyur A (2011) Applications of operations research/statistics in infection outbreak management. Int Business Econ Res J 10(2)

Freberg K (2012) Intention to comply with crisis messages communicated via social media. Public Relat Rev 38:416–421

Government Accountability Office (2012) FDA's food advisory and recall process needs strengthening

Grimm V, Berger U, Bastiansen F, Eliassen S, Ginot V, Giske J, Goss-Custard J, Grand T, Heinz SK, Huse G, Huth A, Jepsen JU, Jorgensen C, Mooij WM, Müller B, Pe'er G, Piou C, Railsback SF, Robbins AM, Robbins MM, Rossmanith E, Rüger N, Strand E, Souissi S, Stillman RA, Vabe R, Visser U, DeAngelis DL (2006) A standard protocol for describing individual-based and agent-based models. Ecol Model 198:115–126

Grunert K (2002) Current issues in understanding of consumer food choice. Trends Food Sci Technol 13:275–285

Hallman W, Cutie C (2009) Food recalls and the American public: improving communications. Technical report. Rutgers University

Hartnett E, Paoli GM, Schaffner DW (2009) Modeling the public health system response to a terrorist event in the food supply. Risk Anal 29(11):1506–1520

Jaine A (2005) A predictive modeling and decision-making tool to facilitate government and industry response to an intentional contamination of the food supply. In: The Institute of Food Technologists' First Annual Food Protection and Defense Conference

Knowles-McPhee S (2015) Growing food safety from the bottom up: an agent based model for food safety inspections. J Artif Soc Social Simul 18(2):1–11

Kramer GJ, Fasone V (2016) Consumer food trends create food safety challenges for the food service industry. https://www.foodsafetymagazine.com/magazine-archive1/junejuly-2017/consumer-food-trends-create-food-safety-challenges-for-the-foodservice-industry/. Accessed 21 Jan 2018

Liu Y, Wein LM (2005) Analyzing a bioterror attack on the food supply: the case of botulinum toxin in milk. Proc Natl Acad Sci U S A 102(8):9984–9989

Liu Y, Wein LM (2008) Mathematically assessing the consequences of food terrorism scenarios. J Food Sci 73(7):M346–M353

Manizini R, Accorsi R (2013) The new conceptual framework for food supply chain assessment. J Food Eng 115:251–263

Rong A, Grunow M (2010) A methodology for controlling dispersion in food production and distribution. OR Spectr 32:957–978

Saulo A, Moskowitz H (2011) Uncovering the mind-sets of consumers towards food safety messages. Food Qual Prefer 22:422–432

Sifferlin A (2015) 351,000 people die of food poisoning globally every year. http://time.com/3768003/351000-people-die-of-food-poisoning-globally-every-year/

Steelfisher G, Weldon K, Benson J, Blendon R (2010) Public perceptions of food recalls and production safety: two survey's of the American public. J Food Saf 30:848

Talley JB (2016) Modeling individual consumer food contamination progression with interventions. Dissertation. North Carolina Agricultural and Technical State University, Greensboro

Tasty S (2015) Your ultimate shelf life guide. stilltasty.com

USDA (2014) Food safety and inspection service protecting public health and preventing foodborne illness. https://www.fsis.usda.gov/wps/wcm/connect/7a35776b-4717-43b5-b0ce-aeec64489fbd/mission-book.pdf?MOD=AJPERES. Accessed 3 Feb 2018

Watkins ME (2015) Modeling consumer behavior for high risk foods. Master's thesis. North Carolina Agricultural and Technical State University, Greensboro

Zechmann E (2011) Agent-based modeling to simulate contamination events and evaluate threat management strategies in water distribution systems. Risk Anal 31(5):758–772

Dr. Jessye Talley attended North Carolina Agricultural & Technical State University, where she received all her degrees in Industrial & Systems Engineering with a concentration in Operations Research. Her research expertise includes stochastic and deterministic modeling of supply chains using stochastic programming, Markov chains, differential equations, linear programming, and queueing theory. Dr. Talley's current research interest consists of applications in humanitarian relief, emergency preparedness and response to address ports, healthcare, and food supply chain safety and defense. Dr. Talley was exposed to information about STEM at an early age and participated in many programs to fuel her interest in those fields. She was given the opportunity through the Science and Engineering Apprenticeship Program (SEAP) to work with a famous Astrophysicist, Dr. George Carruthers at the Naval Research Lab and she also presented research she was learning through a program called Joint Education Facilities. During this program she developed her skills in research, presenting, and writing during her high school years. As a result, she decided to choose the field of Industrial Engineering with a concentration in Operations Research because of the diverse set of problems and techniques available to provide unique solutions. Lastly, based on these experiences Dr. Talley loves to share with the next generation of engineers and STEM professionals. She also

has a strong interest in encouraging other girls and woman to obtain degrees in STEM fields. Dr. Talley is currently an Assistant Professor at Morgan State University in the Industrial & Systems Engineering department.

Lauren B. Davis is an Associate Professor at North Carolina A&T State University in the Industrial & Systems Engineering department. Her research addresses the management of for-profit and not-for-profit (humanitarian) supply chains with specific focus on performance in dynamic environments. Characteristics of the dynamic environment include one or more of the following: (1) constrained supply; (2) uncertain supply (3) uncertain demand; (4) demand surges due to extreme events, like natural disasters; (5) disruptions in operations caused by extreme events. Primary emphasis is on decision-making related to effective planning, allocation, and distribution of inventory. The goal is to identify policies that are optimal from a financial (cost) and/or service (response time, need) perspective. Dr. Davis has always had an interest in STEM and received strong encouragement by her mother and teachers to pursue computer science and mathematics given her affinity for the topic. She also has a strong interest in encouraging younger students to pursue STEM fields. As a member of Delta Sigma Theta Sorority Inc. (Durham NC Chapter) and The Durham Chapter of The Links Incorporated, Dr. Davis is able to channel her passion of paying it forward to the next generation by participating in science enrichment opportunities for students in the Durham community.

Chapter 18
Contributions of Women to Multimodal Transportation Systems

Heather Nachtmann

Contents

18.1 As a Girl 418
18.2 As a Female Engineering Student 419
18.3 As a Female Assistant Professor 420
18.4 As a Female Associate Professor 422
18.5 As a Female Professor 424
18.6 A Look into the Future 428
References 431

Multimodal transportation systems are critical infrastructure components that are essential to promoting and preserving economic health and general societal welfare. These assets facilitate efficient movement of people, goods, and services, and their operations are highly interconnected with numerous other infrastructure systems including communications, emergency response, energy, water supply, agricultural production, and manufacturing. Having 17 years of experience working within the multimodal transportation community, I am continually impressed with the female academics and practitioners who are working to improve our multimodal transportation system. This chapter will focus on major contributions I and other women have made in multimodal transportation research and the impacts these women have had on my career.

The freight transportation system is heavily utilized due to increasing economic activities among/within countries as the result of product specialization and globalization. The United States has one of the world's largest transportation networks including 25,000 miles of navigable waterways, four million miles of public roads, 140,000 miles of railways, and considerable transportation infrastructure (U.S. Department of Transportation (USDOT) 2017). Only 11 ports account for 85% of the United States' containerized international trade (Federal Maritime Commission

H. Nachtmann (✉)
Department of Industrial Engineering, University of Arkansas, Fayetteville, AR, USA
e-mail: hln@uark.edu

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_18

Bureau of Trade Analysis 2015). This dependence and vast infrastructure makes our freight transportation system vulnerable to disruptions and delays due to natural disasters and security incidents. The high demand and frequency of cargo carried by the multimodal transportation system suggests significant impacts will result from future disrupted freight movement. It was reported that the major bridge collapse in Minneapolis influenced approximately 140,000 daily vehicle trips and led to $400,000 daily cost to the commercial vehicles and road users for rerouting (Zhu and Levinson 2012). A series of events including gate failure and inspection closed the Ohio River at Hannibal Locks and Dam for 5 days and resulted in a conservative estimated cost of $5.1 million according to the USACE (2006). Other recent events such as Hurricane Katrina, Superstorm Sandy, and recent flooding and drought events highlight disruption impacts on system performance. Multimodal transportation systems under emergency conditions including evacuation, response, and recovery require planning, design, management, operation, and preservation of transportation systems to economically, efficiently, and safely respond to changing conditions and demands that may occur. Under emergency conditions, the amount and timing of the travel demand often quickly and overwhelmingly exceeds the ability of the transportation system to serve it.

18.1 As a Girl

I grew up in Pittsburgh, Pennsylvania along the Allegheny River. Pittsburgh, known as the Steel City, is home to more than 400 bridges that connect communities across the three rivers that surround the city. Pittsburgh is home to the head of the Ohio River which is formed by the confluence of the Allegheny and Monongahela Rivers. I remember watching coils of steel floating through the locks and dams down the Ohio River to unknown designations. I often wondered why there were only one or two rolls of steel on each “boat” and where they were traveling. These questions would not be answered until much later as I built my expertise in inland water freight transportation and realized the “boats” were actually barge tows and began to understand the commodities and freight movements along the United States’ inland waterways and across the globe. Ranked 31st based on short tons transported in 2016, the Port of Pittsburgh remains one of the busiest inland ports in the United States (U.S. Army Corps of Engineers (USACE) 2017).

It was not until I was grown when I realized the greatest influence on my career was my mother, Ms. Lauren Nachtmann. As I seek to find balance between my career, my family, and my personal well-being, I remember back to my time as a girl observing my mother as she worked full time as a nurse administrator at the University of Pittsburgh Medical Center, earned her B.S. and M.S. in Nursing, raised two children, and managed our household. She is my inspiration and source of hope as I forge my own career as a working mother and navigate through the sleepless nights, daycare stress, homework, extracurricular activities, illnesses, laundry, and other demands that come along with the joy and pleasures of being a parent. Even

today while she is happily retired and filling her days with volunteer work, golf, and friends, she gives me strength and courage to pursue my professional goals.

While there are numerous formal mentoring programs for female engineers across the United States, I believe women in engineering find themselves being mentors to other women without even knowing they are doing so. Sometimes simply demonstrating that navigating a career in engineering and raising a family is possible is enough to give other women hope for a happy and successful future as a professional engineer and mother. In this chapter, I will describe my research in multimodal transportation and how this has been shaped by other female industrial and systems engineers. I will discuss how this research has established my career and recognize the women who have impacted me along the way.

18.2 As a Female Engineering Student

I earned my B.S., M.S., and Ph.D. from the Department of Industrial Engineering at the University of Pittsburgh. In high school, I debated between engineering and architecture sparked by my interests and abilities in math and science with an inclination towards engineering because it appeared to be a broad major with more career options. My decision to study engineering was solidified when my high school Physics teacher (an old white guy) told me that I would never be successful in engineering. Nothing like an ignorant sexist to motivate you towards the most challenging major you can find!

My time as an undergraduate engineering student was shaped by my friends and the faculty, and in particular, Dr. Cynthia Atman who employed me as an undergraduate researcher (Atman et al. 1994). Through this work experience, I discovered my love for conducting research. Cindy, professor of human centered design and engineering and Mitchell T. and Lella Blanche Bowie Endowed Chair at the University of Washington, gave me the opportunity to return to graduate school with the offer of a research and teaching doctoral assistantship while I was a new mother of a 1 year son. To this day, I appreciate the gift of having a strong engineering professor who could lift me up while I was initiating my career as a new mother. I also have Cindy to thank for my most referenced article with more than 500 citations (Atman et al. 1999).

Two other female industrial and systems engineering faculty were critical to the completion of my Ph.D. degree and shaped my research interests for years to come. Dr. Alice Smith, now Joe W. Forehand/Accenture Distinguished Professor of industrial and systems engineering at Auburn University and the editor of this book, introduced me to rigorous ways to analyze data and solve mathematically challenging decision problems that benefit real world stakeholders. Alice also introduced me to *The Engineering Economist* journal, which led me to my dissertation advisor, Dr. Kim LaScola Needy, and thus my future research in this area.

My dissertation research contributed methods for handling estimation imprecision and uncertainty in activity-based costing systems (Nachtmann and Needy 2001,

2003). This work expanded my ability to quantitatively analyze data and model engineering economic decisions and led me to my interest in solving real world problems through applied research. Kim currently serves as the dean of the Graduate School and International Education at the University of Arkansas and has been a constant role model for me throughout my career. She is a dedicated servant leader having served as president of the American Society for Engineering Management and the Institute of Industrial and Systems Engineers and a devoted mother of two sons. Kim makes balancing career and family look easy as she navigates her way through a successful academic career as an engineering leader. She taught me that being extremely well organized and in control of your calendar are critical to maintaining a work–life balance.

18.3 As a Female Assistant Professor

I began my career as an assistant professor of industrial engineering at the University of Arkansas (U of A) in 2000. My decision to join this department had as much to do with the community we would live in as the professional opportunities it would provide. As a single mother of a 6-year-old son, navigating an urban environment while living in a one bedroom apartment or living in a very small town with no after school care was not a desirable option for me. I was fortunate to have the opportunity to take a position that offered a tenure track career in a livable community. It is critical for female engineers to influence the factors they can control during their careers such as choosing the best professional opportunity while controlling for your quality of life or choosing who your co-investigators are with the criteria that they understand that you do not typically work from 5 to 7 pm because you spend time with your family. Although I was the only female tenure-track faculty member in our department, I was fortunate to have an office across the hall from Dr. Carol Gattis, now Associate Dean Emeritus of the U of A Honors College, who has been a mentor and a friend throughout my career at U of A. Carol provided a powerful voice and sympathetic ear as I navigated my way through our male-dominated work environment.

I owe the start of my research into multimodal freight transportation to Dr. Melissa Tooley, director of the Southwest Region University Transportation Center at the Texas A&M Transportation Institute. At the time I started my career as an assistant professor at U of A, Melissa served as the director of the Mack-Blackwell Transportation Center and a member of the civil engineering faculty at U of A. Melissa sought me out because of my background in engineering economic analysis and offered me a $7000 grant to complete a research project left unfinished by the departure of another faculty member. This project, *Economic Evaluation of the Impact of Waterways on the State of Arkansas* (Nachtmann 2002), was my introduction to maritime and multimodal transportation and the jumpstart of my career in this field. To be clear, $7000 was a small grant even in 2001, and I use this experience as a lesson to new faculty and my graduate students. You need

to carefully evaluate opportunities that come your way. Although the grant was small, I was conducting the project as a sole principal investigator allowing me to demonstrate my ability to lead and conduct research on my own and the work would introduce me to our state transportation agencies. This project even gave me an opportunity to testify before the Arkansas House Committee on Public Transportation during my first year at U of A. At the time, I had no idea this project would be the first of many research projects in maritime and multimodal transportation. There were very few female transportation engineers working in the state of Arkansas. Having Melissa as a strong and powerful role model who was willing to invest in me and my career was critical to my success as an assistant professor and my research in this field.

Under Melissa's center direction, I received additional grants from the Mack-Blackwell Transportation Center that expanded my technical knowledge in engineering economic analysis and multimodal transportation. Through the development of WebShipCost (Li et al. 2003, 2004, 2006), we developed a geographic information system (GIS) integrated, web-based application that provides cost, time, and uncertainty analysis of multimodal freight transportation based on up-to-date network information. The valuable attributes of our WebShipCost tool include providing users with more convenient and efficient data management methods to support the decision-making in terms of route planning and data visualization of results to industry end users.

As one of the few female faculty in our college of engineering, I understood I was a role model for our female students and perhaps for our male students who benefited from learning from a female engineer. Our college has made great strides of recruiting female students to engineering over the past 15 years, and industrial and systems engineering is a discipline that generally appeals to female engineers. While supporting and building a diverse and inclusive environment was personally important to me, I was careful to protect my time from related services activities that would pull my time and effort away from the critical scholarly activities necessary to reach promotion and tenure. My personal view on this is that I could not be of full service to my current and future students if I was not successfully promoted and tenured as a faculty member in our college.

As I reflect on my time as an assistant professor, I think of two other strong women who I benefited from knowing as I built my career as an engineering professor, Dr. Alisha Youngblood and Dr. Kellie Schneider. Alisha (also a single mom at the time) and Kellie were both students in our department, brilliant engineers, and young women who I could related to. Alisha graduated as my first doctoral student and currently serves as associate dean of the Harrison College of Business & Computing at Southeast Missouri State University, and Kellie is currently an assistant professor of management systems at University of Dayton. These women exemplify that you find mentors in the most unexpected places, and while faculty are typically thought of as mentoring students, these women were a great support to me during my early career. Towards the end of my time as an assistant professor, I remarried and my second child was born. In 2005, I was tenured and promoted to associate professor.

18.4 As a Female Associate Professor

As previously mentioned, I was careful to protect my time as an assistant professor. After I was tenured, I chose to serve as the faculty advisor to our college's Society of Women Engineers chapter and our Women in Engineering mentoring program. This gave me an opportunity to work with and support our female engineering students in a more concerted way. Being promoted also brought leadership opportunities my way. In 2006, I was asked to serve as the associate director of the Mack-Blackwell Transportation Center and then was promoted to center director in 2007. These positions helped me develop my leadership skills and expand my research into other areas of multimodal transportation. I continued my work in economic impact analysis of inland waterway transportation and expanded into a new research area that focused on the role of transportation in homeland security (Nachtmann and Pohl 2011).

Given that we knew emergency transportation planning requires systematic contingency preparation, my initial work in this area developed the Transportation Readiness Assessment and Valuation for Emergency Logistics (TRAVEL) scorecard which enables emergency operations planners to assess the quality of their emergency operation plans (EOPs) from a transportation perspective (Nachtmann and Pohl 2013b). Our scorecard tool is easily implemented in a spreadsheet software and does not require a high level of analytical ability to provide a user-friendly implementation. TRAVEL adopts an all-hazards approach to emergency operations planning and was developed using a value-focused thinking (Keeney 1992) framework.

Our next contribution provided a decision tool for emergency management teams to design effective and efficient inland waterway-based emergency response systems and improve their emergency operations planning through the development of a Waterway Emergency Services (WES) Index (Nachtmann and Pohl 2013a). Our review of the literature found that transportation plays a key role in emergency operations planning (Federal Emergency Management Agency 2010) but no work existed that explored the role that inland waterways could play in emergency response planning. With 41 states having access to the inland waterways, the potential for these waterways to provide multimodal transportation support during an emergency response was clear. Our first contribution in this area was the development of the Waterway Emergency Medical Service (WEMS) index based on six measureable factors including *Accessibility to Navigable Inland Waterway*, *Proximity to Barge Origin*, *Population Demands*, *Social Vulnerability*, *Risk of Disaster*, and *Limited Access to Medical Services*. With support from Dr. Leily Farrokhvar, then graduate student researcher and now assistant professor of industrial and management systems engineering at West Virginia University, we expanded the WEMS index to the WES Index which added two additional factors, *Limited Access to Resources* and *Limited Access to Transportation Modes* (Nachtmann et al. 2012). Our case analysis demonstrated that more than 73% of counties with public port access along the Mississippi River have medium to high potential to benefit

from incorporating emergency services via inland waterways into their emergency operations planning. The expanded WES index was incorporated into a framework to identify the optimal starting locations of emergency response barges. This multi-objective Emergency Response Barge System integrates a model to determine the minimum number of barges required to provide a pre-defined level of emergency response coverage and a model to determine the optimal locations to stage response barges in order to provide maximum response coverage. Upon graduation with her M.S. in Industrial Engineering from U of A, Leily went on to complete her doctoral degree at Virginia Tech and focused her dissertation on distribution issues faced by the industrial gas industry, a critical research area in multimodal transportation (Farrokhvar 2016).

In 2012, my network of multimodal transportation experts expanded to include the U.S. Army Engineer Research and Development Center, and I began working on expanding the application of systems optimization within the USACE (Nachtmann and Mitchell 2012). In support of the USACE's navigation mission to "provide safe, reliable, efficient, effective, and environmentally sustainable waterborne transportation systems for movement of commerce, national security needs, and recreation" (USACE 2018) and with recognition of ongoing budget constraints, we explored prior research in system optimizations and how this work could be applied to three key navigation decisions including sediment management in coastal systems and across watersheds, lock and dam operations and maintenance, and dredge scheduling and sequencing. We went on to make specific contributions in dredge scheduling and sequencing through the development of a systems-based approach based on constraint programming to achieve increased efficiencies for annual USACE operations and maintenance (O&M) dredging of navigation projects (Nachtmann et al. 2014; Gedik et al. 2016). The USACE annually oversees hundreds of dredging projects in support of its navigation mission. Our work systematically allocates dredge vessels to projects under a variety of constraints including environmental restrictions on when dredging can take place, dredge vessel availability, and equipment productivity rates. This research contributed software that was implemented by the USACE to develop more efficient detailed schedules of dredge resources under current operational restrictions. Additional benefits include guidance into where to direct future research efforts towards the environmental restrictions that are having the most significant impacts on overall dredge program efficiency and insights into the needs for the next-generation dredge fleet.

Interestingly as my professional network expanded to new collaborators, I was faced with more pros (invited to speak at conference because an agency official wanted "that woman from Arkansas who knows a lot about the waterways") to cons (a project sponsor questioning my ability to complete a project due to my advanced pregnancy—this was right after I flew out of state to lead a successful project progress meeting in my 35th week of pregnancy) of being a female researcher in the field of multimodal transportation. As my research career expanded, my family was also expanding with the births of my third and fourth children, respectively, in 2007 and 2010. Adding to my family did delay my progress towards promotion to professor as I was in maintenance mode instead of expansion mode for about a

year after each of my children were born. Since promotion to professor does not have a strict timeclock, this slower path was one that I was willing and able to take. Balancing work and family became even more critical as I raised four children and managed a large research program. Did I execute this perfectly? Absolutely not. Did I take conference calls while bouncing a sleeping baby? Did I have to pump breastmilk in convention center restrooms during multiple conferences? Did I wear two different colors of the same shoe to graduation? Yes, yes, and yes (hey, they were black and navy and it was dark in the room when I tiptoed out). But were my children happy, healthy, and left the house wearing clean clothes? Did I start every class on time? Did I get my research proposals submitted prior their deadlines? Also yes, yes, and yes. Balancing work and family requires choosing your battles, creativity, dedication, and saying no (a lot!). If you choose to have a life partner, I believe the most important career success factor is not related to your actual job, it's who you choose as that partner. This person must care as much about your professional success as you do. In my case, my ability to balance work and family is entirely due to my husband who is 100% supportive of my career, believes our careers are equally important, and invests significant effort into our children and household. In 2013, I was successfully promoted to full professor.

18.5 As a Female Professor

As a professor, additional leadership opportunities came my way. In 2013, I led a multi-institution, multi-disciplinary proposal to form a USDOT University Transportation Center, the Maritime Transportation Research and Education Center (martrec.uark.edu), which was successfully competed and awarded twice from 2013 to 2022. This gave me great opportunity to expand my research partners and in particular my network of women in multimodal transportation including Dr. Bethany Stich and Ms. Carol Short of the University of New Orleans Transportation Institute and Dr. Hiba Baroud, assistant professor of civil and environmental engineering at Vanderbilt University, who are all members of our center's leadership team. Bethany serves as department chair and associate professor of planning and urban studies and is an expert in freight transportation within the Gulf Coast region (Stich 2014) and policy impacts on transportation (Stich and Griffith 2014). Carol has significant experience working with the maritime industry and frequently serves on the organizing committees for conferences and seminars on transportation and women's leadership issues. Hiba's research investigates data analytics and statistical methods to analyze risk, reliability, and resilience in critical infrastructure systems. Hiba is already making significant contributions to inland waterway network resilience (Baroud et al. 2014a, b). I also had the opportunity to join the executive committee of the American Society for Engineering Management and served as the society president from 2016 to 2017.

Today's multimodal transportation network has evolved into a set of complex systems due to globalization and decentralization. These systems depend on the

configuration of their primary components (suppliers, warehouses, service centers, staging areas, ports of debarkation, and transportation modes). The location, transportation mode selection, and supply chain partner identification constituting these components are strategic decisions with major cost implications that have a fundamental impact on freight flow patterns of national and regional transportation systems. These decisions must create robust, reliable, and resilient multimodal supply chains while not compromising financial goals. The USDOT recognizes that investing in multimodal infrastructure can reduce roadway congestion and pollution impacts, and there is a critical need for maritime infrastructure planning that considers increasing throughput to avoid future shipment delays, intermodal transfer point congestion, and increased costs (USDOT 2017). There is limited information available on the economic impacts of waterborne commerce at national, regional, and local levels and how waterway disruptions affect multimodal freight flows and economic outcomes. Better information about waterborne freight can inform private and federal investment in port development and infrastructure improvements which in turn can increase competitive advantages without negatively affecting social and environmental outcomes.

Throughout my career, I have continued to conduct research into the economic impacts of the inland waterways. Mostly recently I partnered with the Arkansas Department of Transportation, University of Arkansas at Little Rock, and Oklahoma Department of Transportation (ODOT) to complete a regional economic impact study of the McClellan-Kerr Arkansas River Navigation System (MKARNS) (Nachtmann et al. 2015). Our findings show the MKARNS contributes total impacts of $8.5 billion in sales, $4.3 billion in gross domestic product (GDP), and 55,872 jobs to the national economy. In this project, I was able to work with Ms. Deidre Smith who was the ODOT Waterways Manager and is now the Executive Director for the Arkansas Waterways Commission. Dede has made tremendous contributions to the inland waterways and the multimodal transportation network through her work as a transportation manager and her participation on the American Association of State Highway and Transportation Officials Standing Committee on Water Transportation, executive and membership committees of the National Waterways Conference, and executive committee of the Arkansas River Historical Society. There are many other talented and dedicated women working in the transportation industry, and it is a privilege to work among them.

We expanded our economic impact work to systematically explore the impact of disruptions to the multimodal transportation system. Predicting the economic impacts of inland waterway disruption response enables multimodal stakeholders to increase their preparedness and potentially reduce economic losses. Our simulation-based economic impact disruption decision model is based on publicly available data, which allows our approach to be easily adopted (Oztanriseven and Nachtmann 2017). Our simulation model examines scenarios when the MKARNS is closed down due to a disruptive event for short-term, medium-term, and long-term disruptions and predicts the economic losses due to these potential disruptions. Our findings indicate that the expected disruption duration determines whether decision makers are better off waiting for the waterway system to reopen or switching to an

alternative mode of transportation, and as expected, improved estimation accuracy of disruption duration can help reduce negative economic impacts caused by the disruptive event.

We wanted to explore better methods for studying the operational behavior and economic impacts of the MKARNS and other multimodal transportation systems and partnered with Dr. Suzanna Long, department chair and professor of engineering management at Missouri University of Science and Technology, to do so. Suzie's research interests include critical infrastructure systems and supply chain and transportation. Her recent work in multimodal transportation ranges from sustainable rail infrastructure systems (Rangarajan et al. 2013) to inland freight hub location (Long and Grasmann 2012) to electric vehicles (Egbue et al. 2017). We identified system dynamics as an appropriate approach to study the maritime transportation system and its integration with multimodal transportation. We then conducted a literature review of system dynamics applications in the maritime transportation system and found that the existing body of knowledge primarily focused on ports with very little vessel research (Oztanriseven et al. 2014). Our review indicated that many researchers integrated their system dynamics model with other models and conducted sensitivity analysis and scenario analysis for model validation. In addition, we found that researchers primarily faced data-related and complexity-related modeling challenges. This project also gave me the opportunity to work with then graduate student and now assistant professor of supply chain at the University of North Carolina Wilmington, Dr. Lizzette Pérez-Lespier. Liz, advised by Suzie Long, focused her thesis research on examining the efficiency of multimodal transportation systems through a system dynamics approach in order to provide decision support that will improve customer satisfaction and system performance (Pérez-Lespier 2013).

Our capabilities in modeling the operational behavior and economic impacts of the MKARNS expanded with the development of the maritime transportation simulator (MarTranS) that integrates agent-based modeling, discrete-event simulation, and system dynamics along with a multiregional input-output model (Oztanriseven and Nachtmann 2018). With MarTranS, we are able to study the interactions between inland waterway transportation system components and economic impact factors which can help stakeholders make informed inland waterway infrastructure investment decisions to improve economic impacts. Our simulation results illustrate that the economic performance of the MKARNS is not sustainable without future infrastructure investments as economic impacts and commodity flow will drop to only 10% of their current values in approximately two decades if infrastructure improvements are not made. Our results also show that multiple ports and lock and dam systems are reaching utilization rates over 80% creating increased transportation delays and costs.

To further explore emergency operations planning in multimodal transportation, we developed the cargo prioritization and terminal allocation problem (CPTAP) which integrates assignment and scheduling of disrupted barges to terminals in the event of inland waterway disruptions where the cargo is offloaded to be transported by an alternative transportation mode with the aim of minimizing the total value

loss (Tong and Nachtmann 2017). This work gave me the opportunity to work with Dr. Jingjing Tong, my former doctoral student and now an assistant professor of polytechnic studies at Southeast Missouri State University. Jingjing is a gifted researcher whose work focuses on applied operation research in transportation. The goal of CPTAP is to minimize the total value loss of cargo disruptions on the inland waterways. The total value loss objective depends on a multi-attribute, commodity-specific parameter, the value decreasing rate. To allow stakeholders to assess the value decreasing rate beyond simply economic value, we developed a value-focused thinking (Keeney 1992) approach to determine value decreasing rates for disrupted cargo based on emergency need, response need, community need, military need, local priorities, public health, environmental security, market value, and perishability (Tong et al. 2015). We employed a genetic algorithm to solve realistically sized problem instances efficiently (in terms of CPU solution time) and effectively (in terms of consistency with assumptions and optimality on small problems) as compared to representative experimental instances and other decision approaches. In addition to its emergency response application, CPTAP can be employed in emergency planning by assessing the resiliency of the inland waterway transportation system to handle potentially disrupted cargo based on the existing commodity capacity of the offload terminals and alternative modes of land-based transportation. One of the greatest challenges in recovering cargo from barges is the large amounts of cargo transported and how to prepare the land-based transportation recovery system in the event of a disruption. We are currently working to expand the original CPTAP to handle uncertainty. There are multiple uncertain terms in the structure of the CPTAP decision including barge handling time, value decreasing rate, water transportation time, and land transportation time. We are exploring a simulation-optimization approach that will allow us to systematically incorporate this uncertainty into our decision-making process to enable more realistic information about the multimodal transportation system behavior under disruption to the inland waterways.

I am also expanding CPTAP through my ongoing research with Dr. Liliana Delgado Hidalgo, assistant professor of industrial engineering at Universidad del Valle in Cali, Colombia and my former doctoral student. Liliana utilized the Analytic Hierarchical Process (AHP) (Saaty 1980) to develop a multi-attribute decision approach to CPTAP (Tong and Nachtmann 2013) and effectively divided CPTAP into two decision components, assignment of the disrupted barges to the terminals and scheduling of the barges assigned to each particular terminal (Delgado Hidalgo et al. 2015). The results of our sequential AHP approach, which is easier to implement, do not differ significantly from the more complex, non-linear CPTAP approach. Liliana's dissertation research investigates how to improve inland waterway post-disaster outcomes including reducing cargo value loss and response time by developing new decision support techniques to redirect disrupted barges and prioritize offloading at accessible terminals during disruption response. Our current research objective is to reduce cargo value loss during inland waterway disruption response by developing a pure mathematical approach that integrates barge-terminal assignment and scheduling decisions. We have formulated CPTAP

as a mixed integer linear program model which is improved through the addition of valid inequalities and are working to implement exact methods to obtain improved solutions of the CPTAP.

I am happy to report that my personal life is stabilizing as well. My oldest son graduated from U of A and is gainfully employed as a mechanical engineer. One down and three to go! My other children grow more self-sufficient each day. We even took the three youngest to China for a summer vacation. Of course it took me about a month to pack after work and on weekends but we made it there and back and had an amazing experience. I have started reading for pleasure again (a favorite pastime) and even get to spend some time with friends. My personal advice is to take time to enjoy your friends and family. When I look back at my life, I am proud of my professional accomplishments but my happiest moments are with the ones I love.

18.6 A Look into the Future

In Table 18.1, I summarize the contributions of the women highlighted in this chapter. These women have made important contributions to our field and are advancing the role of women in industrial and systems engineering.

As I look forward, I see a bright future for women in multimodal transportation. Comprehensively described in the Beyond Traffic 2045 report (USDOT 2017), it is critical for the United States to improve the operational efficiency and infrastructure sustainability of our multimodal transportation system to meet the future demands of our population and freight. This remains a rich research area for engineers as data analytics and new technologies and materials are critical to support our next-generation transportation system. Personally I was recently appointed to the Earl J. and Lillian P. Dyess Endowed Chair in Engineering, and my home department now has five female tenured and tenure-track faculty members including two in campus leadership roles. This is a 500% increase from when I was the lone such professor in our department 17 years ago. In 2014, I became associate dean for research for our college of engineering. This gives me the opportunity to work directly with our engineering faculty and observe firsthand the contributions our female faculty are making on the transportation system of the future. Dr. Sarah Hernandez, assistant professor of civil engineering, focuses on freight transportation planning challenges for long range planning in the public sector (Hernandez et al. 2016). Dr. Ashlea Milburn, associate professor of industrial engineering at the University of Arkansas, is investigating the application of information gathered from social media in support of disaster response planning (Kirac et al. 2015). Contributions by another assistant professor of industrial engineering at the University of Arkansas, Dr. Sarah Nurre, have advanced network optimization in infrastructure restoration (Nurre et al. 2012) and electric vehicle charging station placement (Nurre et al. 2014). Dr. Shengfan Zhang, also an assistant professor of industrial engineering at the University of Arkansas, has explored dynamic decision-making during inland

Table 18.1 Main contributions of women cited in this chapter

Female contributors	Main contributions as cited in this chapter
Dr. Cynthia Atman, Professor of Human Centered Design and Engineering and Mitchell T. and Lella Blanche Bowie Endowed Chair, University of Washington	Atman CJ, Nair I, Nachtmann H (1994) Do engineers and humanities majors perceive STS issues differently? American Society for Engineering Education Annual Conference Proceedings, June 1994 Atman CJ, Chimka JR, Bursic KM, Nachtmann H (1999) A comparison of freshman and senior engineering design processes. Design studies 20 (2):131–152
Dr. Hiba Baroud, Assistant Professor of Civil and Environmental Engineering, Vanderbilt University	Baroud H, Ramirez-Marquez JE, Barker K, Rocco CM (2014a) Stochastic measures of network resilience: applications to waterway commodity flows. Risk Analysis 34:1317–1335 Baroud H, Barker K, Ramirez-Marquez JE (2014b) Importance measures for inland waterway network resilience. Transportation Research Part E: Logistics and Transportation Review 62:55–67
Dr. Liliana Delgado Hidalgo, Assistant Professor of Industrial Engineering, Universidad del Valle in Cali, Colombia	Delgado Hidalgo L, Nachtmann H, Tong J (2015) Analytic hierarchy approach to inland waterway cargo prioritization and terminal allocation. American Society for Engineering Management Conference Proceedings, October 2015
Dr. Suzanna Long, Department Chair and Professor of Engineering Management, Missouri University of Science and Technology	Egbue O, Long S, Samaranayake VA (2017) Mass deployment of sustainable transportation: Evaluation of factors that influence electric vehicle adoption. Clean Technologies and Environmental Policy 19(7):1927–1939 Long S, Grasmann S (2012) Strategic decision model for evaluating inland freight hub locations. Research in Transportation Business and Management (5):92–98 Rangarajan K, Long S, Tobias A, Keister M (2013) The role of stakeholder engagement in the development of sustainable rail infrastructure systems. Research in Transportation Business and Management 7:106–113
Dr. Leily Farrokhvar, Assistant Professor of Industrial and Management Systems Engineering, West Virginia University	Farrokhvar L (2016) Strategic planning models and approaches to improve distribution planning in the industrial gas industry. Dissertation, Virginia Tech
Dr. Sarah Hernandez, Assistant Professor of Civil Engineering, University of Arkansas	Hernandez S, Tok A, Ritchie SG (2016) Multiple-classifier systems for truck body classification at WIM sites with inductive signature data. Transportation Research Part C: Emerging Technologies 68:1–21
Dr. Ashlea Milburn, Assistant Professor of Industrial Engineering, University of Arkansas	Kirac E, Milburn AB, Wardell CL (2015) The traveling salesman problem with imperfect information with application in disaster relief tour planning. IIE Transactions 47(8):783–799
Dr. Shengfan Zhang, Assistant Professor of Industrial Engineering, University of Arkansas	Madadi M, Holmer R, Zhang S, Nachtmann H (2016) Dynamic decision modeling for inland waterway disruptions. Proceedings of the 2016 Industrial and Systems Engineering Research Conference

(continued)

Table 18.1 (continued)

Female contributors	Main contributions as cited in this chapter
Dr. Heather Nachtmann, Earl J. and Lillian P. Dyess Endowed Chair in Engineering, Professor of Industrial Engineering, and Associate Dean for Research, University of Arkansas	Nachtmann H, Mitchell KN, Rainwater CE, Gedik R, Pohl EA (2014) Optimal dredge fleet scheduling with environmental work windows. Transportation Research Record 2426:11–19 Nachtmann H, Needy KL (2001) Fuzzy activity based costing: A methodology for handling uncertainty in activity based costing systems. The Engineering Economist 46(4):245–273 Nachtmann H, Needy KL (2003) Methods for handling uncertainty in activity based costing. The Engineering Economist 48(3):259–282 Nachtmann H, Pohl EA (2013a) Emergency medical services via waterways. Risk Management 15(4):225–249 Nachtmann H, Pohl EA (2013b) Transportation readiness assessment and valuation for emergency logistics. International Journal of Emergency Management 9(1):18–36 Nachtmann H, Pohl EA, Farrokhvar L (2012) Decision support for inland waterways emergency response. Engineering Management Journal 24(4):3–14 Oztanriseven F, Nachtmann H (2017) Economic impact analysis of inland waterway disruption response. The Engineering Economist 62(1):73–89
Dr. Sarah Nurre, Assistant Professor of Industrial Engineering, University of Arkansas	Nurre SG, Bent R, Pan F, Sharkey TC (2014) Managing operations of plug-in hybrid electric vehicle (PHEV) exchange stations for use with a smart grid. Energy Policy 67:364–377 Nurre SG, Cavdaroglu B, Mitchell JE, Sharkey TC, Wallace WA (2012) Restoring infrastructure systems: An integrated network design and scheduling problem. European Journal of Operational Research 223(3):794–806
Dr. Lizzette Pérez-Lespier. Assistant Professor of Supply Chain, University of North Carolina Wilmington	Pérez-Lespier L (2013) Examining the efficiency of multimodal transportation systems: A systems dynamics approach. Thesis, Missouri University of Science and Technology
Dr. Bethany Stich, Department Chair and Associate Professor of Planning and Urban Studies, University of New Orleans	Stich B (2014) Intermodal transportation disruption and reroute simulation framework. Transportation Research Record 2410:150–159 Stich B, Griffith K (2014) Ending transportation neglect in America. Public Works Management & Policy 19(4)
Dr. Jingjing Tong, Assistant Professor of Polytechnic Studies, Southeast Missouri State University	Tong J, Nachtmann H (2017) Cargo prioritization and terminal allocation problem for inland waterway disruptions. Maritime Economics and Logistics 19(3):403–427 Tong J, Nachtmann H (2013) Multi-attribute decision model for cargo prioritization within inland waterway transportation. Proceedings of the 2013 Industrial Engineering Research Conference, May–June 2013 Tong J, Nachtmann H, Pohl EA (2015) Value-focused assessment of cargo value decreasing rate. Engineering Management Journal 27(2):73–84

waterway disruptions (Madadi et al. 2016). You can see the impacts women in industrial and systems engineering are having on the multimodal transportation system just by taking a quick glance at these researchers in my home department. As female faculty in engineering, it is critical that we lift up other women and support their careers in terms of productivity and success and in achieving work–life balance. In order to address the multi-dimensional and complex challenges our transportation system will face in the future, it is critical that we nurture and support gender diversity in our research community and provide an inclusive research culture for these women.

References

Atman CJ, Nair I, Nachtmann H (1994) Do engineers and humanities majors perceive STS issues differently? In: American Society for Engineering Education annual conference proceedings, June 1994

Atman CJ, Chimka JR, Bursic KM, Nachtmann H (1999) A comparison of freshman and senior engineering design processes. Des Stud 20(2):131–152

Baroud H, Ramirez-Marquez JE, Barker K, Rocco CM (2014a) Stochastic measures of network resilience: applications to waterway commodity flows. Risk Anal 34:1317–1335

Baroud H, Barker K, Ramirez-Marquez JE (2014b) Importance measures for inland waterway network resilience. Transport Res E-Log 62:55–67

Delgado Hidalgo L, Nachtmann H, Tong J (2015) Analytic hierarchy approach to inland waterway cargo prioritization and terminal allocation. In: American Society for Engineering Management conference proceedings, October 2015

Egbue O, Long S, Samaranayake VA (2017) Mass deployment of sustainable transportation: evaluation of factors that influence electric vehicle adoption. Clean Techn Environ Policy 19(7):1927–1939

Farrokhvar L (2016) Strategic planning models and approaches to improve distribution planning in the industrial gas industry. Dissertation, Virginia Tech

Federal Emergency Management Agency (2010) Developing and maintaining emergency operations plans: comprehensive preparedness guide. Report FEMA CPG 101 V2, U.S. Department of Homeland Security, Washington, DC

Federal Maritime Commission, Bureau of Trade Analysis (2015) U.S. container port congestion & related international supply chain issues: causes, consequences & challenges

Gedik R, Rainwater C, Nachtmann H, Pohl E (2016) Analysis of a parallel machine scheduling problem with sequence dependent setup times and job availability intervals. Eur J Oper Res 251(2):640–650

Hernandez S, Tok A, Ritchie SG (2016) Multiple-classifier systems for truck body classification at WIM sites with inductive signature data. Transport Res C Emerg Technol 68:1–21

Keeney RL (1992) Value-focused thinking: a path to creative decision making. Harvard University Press, Cambridge

Kirac E, Milburn AB, Wardell CL (2015) The traveling salesman problem with imperfect information with application in disaster relief tour planning. IIE Trans 47(8):783–799

Li Z, Rossetti MD, Nachtmann H (2003) WebShipCost—an intermodal transportation web-based application. In: Proceedings of the 2003 industrial engineering research conference, May 2003

Li Z, Nachtmann H, Rossetti MD (2004) WebShipCost—quantifying risk in intermodal transportation. In: Proceedings of the 2004 industrial engineering research conference, May 2004

Li Z, Rossetti MD, Nachtmann H (2006) WebShipCost—risk analysis with a geographical information system. In: Proceedings of the 2006 industrial engineering research conference, May 2006

Long S, Grasmann S (2012) Strategic decision model for evaluating inland freight hub locations. Res Transp Bus Manag 5:92–98

Madadi M, Holmer R, Zhang S, Nachtmann H (2016) Dynamic decision modeling for inland waterway disruptions. In: Proceedings of the 2016 industrial and systems engineering research conference

Nachtmann H (2002) Economic evaluation of the impact of waterways on the state of Arkansas. In: Mack-Blackwell Transportation Center final report

Nachtmann H, Mitchell KN (2012) Systems optimization within the U.S. Army Corps of Engineers' navigation program. In: Proceedings of the 2012 industrial engineering research conference, May 2012

Nachtmann H, Needy KL (2001) Fuzzy activity based costing: a methodology for handling uncertainty in activity based costing systems. Eng Econ 46(4):245–273

Nachtmann H, Needy KL (2003) Methods for handling uncertainty in activity based costing. Eng Econ 48(3):259–282

Nachtmann H, Pohl EA (2011) The inland waterway transportation systems' role in response and recovery. J Homel Secur

Nachtmann H, Pohl EA (2013a) Emergency medical services via waterways. Risk Manage 15(4):225–249

Nachtmann H, Pohl EA (2013b) Transportation readiness assessment and valuation for emergency logistics. Int J Emerg Manag 9(1):18–36

Nachtmann H, Pohl EA, Farrokhvar L (2012) Decision support for inland waterways emergency response. Eng Manag J 24(4):3–14

Nachtmann H, Mitchell KN, Rainwater CE, Gedik R, Pohl EA (2014) Optimal dredge fleet scheduling with environmental work windows. Transp Res Rec 2426:11–19

Nachtmann H, Boudham O, Oztanriseven F (2015) Regional economic impact study of the McClellan-Kerr Arkansas River Navigation System. In: Arkansas Department of Transportation final report

Nurre SG, Cavdaroglu B, Mitchell JE, Sharkey TC, Wallace WA (2012) Restoring infrastructure systems: an integrated network design and scheduling problem. Eur J Oper Res 223(3):794–806

Nurre SG, Bent R, Pan F, Sharkey TC (2014) Managing operations of plug-in hybrid electric vehicle (PHEV) exchange stations for use with a smart grid. Energy Policy 67:364–377

Oztanriseven F, Nachtmann H (2017) Economic impact analysis of inland waterway disruption response. Eng Econ 62(1):73–89

Oztanriseven F, Nachtmann H (2018) Modeling dynamic behavior of navigable inland waterways. Working Paper

Oztanriseven F, Perez-Lespier L, Long S, Nachtmann H (2014) A review of system dynamics in maritime transportation. In: Proceedings of the 2014 industrial engineering research conference, May–June 2014

Pérez-Lespier L (2013) Examining the efficiency of multimodal transportation systems: a systems dynamics approach. Thesis, Missouri University of Science and Technology

Rangarajan K, Long S, Tobias A, Keister M (2013) The role of stakeholder engagement in the development of sustainable rail infrastructure systems. Res Transp Bus Manag 7:106–113

Saaty T (1980) The analytic hierarchy process: planning, priority setting, resource allocation. McGraw-Hill, Columbus

Stich B (2014) Intermodal transportation disruption and reroute simulation framework. Transp Res Rec 2410:150–159

Stich B, Griffith K (2014) Ending transportation neglect in America. Public Works Manag Policy 19(4):328–333

Tong J, Nachtmann H (2013) Multi-attribute decision model for cargo prioritization within inland waterway transportation. In: Proceedings of the 2013 industrial engineering research conference, May–June 2013

Tong J, Nachtmann H (2017) Cargo prioritization and terminal allocation problem for inland waterway disruptions. Marit Econ Log 19(3):403–427
Tong J, Nachtmann H, Pohl EA (2015) Value-focused assessment of cargo value decreasing rate. Eng Manag J 27(2):73–84
U.S. Army Corps of Engineers (2006) Hannibal locks and dam: causes and consequences of lock closures 21 October to 16 November 2005. http://www.corpsnets.us/docs/EventHannibal/06-NETS-R-05.pdf
U.S. Army Corps of Engineers (2018) Navigation. http://www.usace.army.mil/Missions/Civil-Works/Navigation/. Accessed 28 Jan 2018
U.S. Army Corps of Engineers, Institute for Water Resources (2017) Waterborne commerce of the United States calendar year—2016: part 5, national summaries. CEIWR-WCUS-16-5
U.S. Department of Transportation (2017) Beyond Traffic 2045
Zhu S, Levinson DM (2012) Disruptions to transportation networks: a review. In: Network reliability in practice. Springer, New York, p 5–20

Heather Nachtmann is the Associate Dean for Research in the College of Engineering, Earl J. and Lillian P. Dyess Endowed Chair in Engineering, and Professor of Industrial Engineering at the University of Arkansas. Dr. Nachtmann serves as director of the Maritime Transportation Research and Education Center (MarTREC), a U.S. Department of Transportation University Transportation Center, and the Mack-Blackwell Transportation Center. Her research includes economic and operations analysis of inland waterways, intermodal network optimization, and vulnerability assessment and emergency preparedness for transportation networks. Dr. Nachtmann has published more than eighty peer-reviewed publications and led over ten million dollars in research grants as principal investigator. Dr. Nachtmann teaches in the areas of operations research, engineering economics, and cost analysis. Dr. Nachtmann is a Fellow of the American Society for Engineering Management and Institute of Industrial and Systems Engineers, a member of the Arkansas State Highway and Transportation Department Research Advisory Council, and co-editor-in-chief of the *Engineering Management Journal* and an associate editor for *The Engineering Economist*. She received her Ph.D. from the University of Pittsburgh, where she made the decision to become an industrial engineer because it integrates her mathematical and scientific skills with her interests in business and working with people. She dedicate this chapter to all of the women in her life.

Chapter 19
Combining Exact Methods to Construct Effective Hybrid Approaches to Vehicle Routing

Rym M'Hallah

Contents

19.1 Introduction 435
19.2 Literature Review 436
19.3 Problem Description and Formulation 438
19.3.1 Problem Definition 438
19.3.2 Mixed Integer Program 438
19.3.3 Constraint Program 441
19.3.4 Enhancing MIP and CP 442
19.4 Solution Approach 442
19.4.1 Stage 1 443
19.4.2 Stage 2 448
19.5 Results 448
19.6 Practical Considerations 451
19.6.1 Single Versus Multiple Depots 451
19.6.2 Distances and Travel Times 451
19.6.3 Time Windows 452
19.6.4 Collection Versus Distribution 452
19.6.5 Packing Constraints 452
19.6.6 Stochastic Travel Times and Time Windows 453
19.7 Conclusions 454
References 454

19.1 Introduction

Globalization has fostered a prosperous climate for industrial growth. However, this growth is coupled with new challenges: a fiercer competition, broadened distribution networks, diversified supply chains, tighter profit margins, and more exigent clients. These latter are demanding lower prices, requiring higher quality

R. M'Hallah (✉)
Department of Statistics and Operations Research, College of Science, Kuwait University, Kuwait City, Kuwait
e-mail: rym.mhallah@ku.edu.kw

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_19

levels, and imposing stricter deadlines. Consequently, the manufacturing, service, and transportation industries that are affected by this new order have to manage their activities judiciously while dealing with more complex, larger-scale, real-life systems. Managing these activities is equivalent to solving several inter-related industrial engineering problems. Because of the new global environment, industries can no longer address some abstract form of these problems or decompose them into independent subproblems. They have to model the problems as closely as possible to their true real-life context. Such models are important for three reasons. First, they offer a better reflection of the problem. Second, they account for more constraints, take into account the true nature of the variables, and build in intricate interdependencies. Third and last, they enhance the chances of direct real-life implementation of the resulting solutions without any modification and at no supplementary cost.

The optimization of the new complex problems that have emerged is difficult not only because of their compounded nature but most importantly because their components are themselves difficult to solve. This chapter considers one such problem: vehicle routing with multiple time windows (VRPMTW), a problem of substantial impact. Its good management alleviates environmental concerns caused by pollution, reduces costs, and offers better work conditions for drivers and higher satisfaction for clients. In VRPMTW, every client specifies more than one availability period for receiving delivery.

Unlike existing methods, which consider the routing and scheduling aspects of VRPMTW independently, this chapter addresses the problem as a single entity. It approximately solves VRPMTW in reasonable run times using an efficient search technique that takes advantage of the advances of computing technologies. Larger computer memories and faster processors allow the (near-)exact resolution of more realistic problems, an unimaginable phenomenon few years ago.

The proposed approach uses exact techniques as approximate ones. It substitutes difficult constraints with easier ones, and limits the search space to neighborhoods that contain the optimum. Most importantly, it explores the respective strengths of mixed integer programming, constraint programming, and search methods.

Section 19.2 reviews the literature. Section 19.3 defines and models the problem. Section 19.4 details the proposed approach. Section 19.5 presents some results. Section 19.6 discusses some practical considerations. Finally, Sect. 19.7 summarizes this research and presents future extensions.

19.2 Literature Review

Because they occur in most services and industries, vehicle routing problems (VRP) are continuously drawing the attention of researchers (Vidal et al. 2014). This is clearly reflected by the extensive literature and numerous review papers (Adewumi and Adeleke 2018; Braekers et al. 2016; Eksioglu et al. 2009, Gendreau et al. 2008, Koç et al. 2016; Vidal et al. 2013). These surveys classify the literature according

to the problem constraints, mathematical models, and solution techniques. Most of them (e.g., Baldacci et al. 2012) single out cases where loads (Alba and Dorronsoro 2008; Gendreau et al. 1994; Minocha and Tripathi 2013; Nagata and Bräysy 2009; Toth and Vigo 2002) and delivery times (Cordeau et al. 2001; Favaretto et al. 2007; Fu et al. 2007; Figliozzi 2010; Moccia et al. 2012; Nagata et al. 2010; Nazif and Lee 2012; Solomon 1987) are constrained. These constraints fathom large parts of the search space and thus eliminate a large number of potential nodes of any tree-based procedure. However, they make finding feasible solutions much harder.

Hashimoto et al. (2006) and Ibaraki et al. (2005), to cite a few, consider VRP with general time windows (TW). They associate to each client more than one delivery window, but penalize the violation of these soft or flexible TWs, with penalties that are not necessarily linear. Beheshti et al. (2015) consider a variant where the multiple soft TWs are not specified by the clients but by the delivery company. Clients rank the TWs in decreasing order of their preference. This variant, which applies in rapid post deliveries, does not guarantee the suitability of the alternatives with the clients' constraints.

Belhaiza et al. (2013) address VRPMTW via a tabu search variable neighborhood search (TSVNS) that minimizes the weighted sum of the trip duration and the penalties associated to the TWs and capacity constraints. Belhaiza and M'Hallah (2016) consider the multiple objective VRPMTW where the criteria are: drivers' utility, total cost, and costumers' utility. They speed their search by fathoming all dominated solutions and investigate the space of solutions that satisfy the Nash equilibrium conditions of a non-cooperative multiple-agent game. Finally, Belhaiza et al. (2017) extend the single objective VRPMTW to the multiple-depot heterogeneous fleet case with the objective of minimizing either the total travel time or the total traveled duration. They address the problem using a multiple-start hybrid genetic variable neighborhood search.

Unlike the techniques surveyed by Archetti and Speranza (2014), the proposed method combines random search with mathematical/constraint programming. This line of research is successfully applied in cutting and packing where the constraints are non-linear (M'Hallah et al. 2013; Al-Mudhahka et al. 2011) and in timetabling where the problem size is large and the planning horizon is long (M'Hallah and Alkhabbaz 2013). Similarly, it is applied in scheduling where a steepest descent heuristic (Laalaoui and M'Hallah 2016) calls iteratively a mixed integer program to insert a job on a machine (M'Hallah 2007; M'Hallah and Al-Khamis 2012) or to enhance an existing solution and assess its optimality gap (M'Hallah 2014) or to choose jobs that will be exchanged among agents (Polyakovskiy and M'Hallah 2014). Recently, it is extended to bin packing where mixed integer/constraint programs apply a look-ahead strategy that reserves areas in the bins for unpacked items (Polyakoskiy and M'Hallah 2018). In all aforementioned applications, the models are augmented with feasibility constraints and with bounds on the objective function values forcing the solvers to only investigate feasible improving directions. They are subject to runtime limits because in most cases, the incumbent is identified at an early stage of the search while most of the computational effort is devoted to proving optimality. Finally, they are fed with partial or complete initial solutions.

19.3 Problem Description and Formulation

Section 19.3.1 defines VRPMTW. Sections 19.3.2 and 19.3.3 model it as a mixed integer and as a constraint program. Finally, Sect. 19.3.4 discusses ways to enhance the performance of both models.

19.3.1 Problem Definition

Consider a set $I = \{1, \ldots, n\}$ of n clients that are served, from a single depot d, by a set $K = \{1, \ldots, m\}$ of m vehicles. The depot d, denoted also as client 0, is located in position $\left(o_0^h, o_0^v\right)$ of the Cartesian coordinate system. It is open during the time interval $[\ell_0, u_0]$; i.e., no vehicle leaves d prior to ℓ_0 or returns to d later than u_0. A vehicle may wait at the depot or at a client's site. Let $\delta_{i,j}$ denote the travel times between i and j, $i \in I^*, j \in I^*, i \neq j$, where $I^* = I \cup \{0\}$.

Client i, $i \in I$, is characterized by its Cartesian coordinate system's position $\left(o_i^h, o_i^v\right)$, positive demand q_i delivered by a single vehicle, known positive service time s_i, and ordered set $W_i = \left\{w_{i,1}, \ldots, w_{i,\overline{p}_i}\right\}$ of $\overline{p}_i$, $\overline{p}_i \in \mathbb{N}^*$, non-overlapping availability TWs. TW p, $p = 1, \ldots, \overline{p}_i$, of client i, $i \in I$, is $w_{i,p} = [\ell_{i,p}, u_{i,p}]$, with $0 < \ell_{i,1} < u_{i,1} < \cdots < \ell_{i,p} < u_{i,p} < \cdots < \ell_{i,\overline{p}_i} < u_{i,\overline{p}_i} < \infty$.

Let $I_k \subset I$ be the ordered subset of clients assigned to k. The subsets are mutually exclusive; i.e., $\cup_{k \in K} I_k = I$, $\cup_{k \in K} I_k = \varnothing$. Thus, partial delivery is prohibited. Vehicle k, $k \in K$, has a limited capacity $Q_k > 0$. Its time out of d, including its waiting time, should not exceed $\overline{D}_k > 0$. Its travel duration D_k is the sum of its travel and service time. It excludes any waiting time.

19.3.2 Mixed Integer Program

Herein, VRPMTW assigns the clients to the vehicles and schedules their respective deliveries with the objective of minimizing the total travel time over all vehicles. It can be formulated as a mixed integer program. The model is inspired from the mixed integer program (MIP) of Belhaiza et al. (2013). It uses three types of positive decision variables and three types of binary variables.

The first positive variable $\omega_{i,k}$ denotes the waiting time of vehicle k, $k \in K$, at client i, $i \in I^*$. A vehicle may have to wait prior to starting delivery to a specific client and may choose to wait at d in order to reduce its total time out of d. The second positive variable $t_{i,k}$ indicates the time k, $k \in K$, reaches i, $i \in I^*$. Consequently, k starts delivery to i, $i \in I$, at $t_{i,k} + \omega_{i,k}$. Finally, the third positive variable $c_{i,k}$ is the departure time of k, $k \in K$, from i, $i \in I^*$. It follows that $c_{0,k}$ is the departure time of k from d and $t_{0,k}$ its return time to d.

The first binary variable $z_{i,k} = 1$ if client i, $i \in I$, is assigned to vehicle k, $k \in K$, and 0 otherwise. The second binary variable $x_{i,j,k} = 1$ if i, $i \in I^*$, immediately precedes j, $j \in I^*$, $j \neq i$, and both j and i are served by vehicle k, $k \in K$, and 0 otherwise. Finally, the binary variable $v_{i,p} = 1$ if client i, $i \in I$, is served during $w_{i,p}$, $p = 1, \ldots, \overline{p}_i$, and 0 otherwise.

Then, the VRPMTW using the above six types of decision variables follows.

$$\text{Min} \sum_{k \in K} \sum_{(i,j) \in I^* \times I^*} \delta_{i,j} x_{i,j,k} \tag{19.1}$$

$$s.t. \quad \sum_{k \in K} z_{i,k} = 1, \quad i \in I \tag{19.2}$$

$$\sum_{j \in I^*} x_{i,j,k} - \sum_{j \in I^*} x_{j,i,k} = 0, \quad i \in I^*, k \in K \tag{19.3}$$

$$2x_{i,j,k} - z_{i,k} - z_{j,k} \leq 0, \quad i \in I, j \in I, i \neq j, k \in K \tag{19.4}$$

$$\sum_{k \in K} \sum_{i \in I^*} x_{i,j,k} = 1, \quad j \in I \tag{19.5}$$

$$\sum_{k \in K} \sum_{j \in I^*} x_{i,j,k} = 1, \quad i \in I \tag{19.6}$$

$$\sum_{i \in I} q_i z_{i,k} = Q_k, \quad k \in K \tag{19.7}$$

$$c_{i,k} - t_{i,k} - \omega_{i,k} - s_i + M\left(1 - z_{i,k}\right) \geq 0, \quad i \in I, k \in K \tag{19.8}$$

$$t_{j,k} - c_{i,k} - \delta_{i,j} - M\left(1 - x_{i,j,k}\right) \leq 0, \quad i \in I^*, j \in I^*, i \neq j, k \in K \tag{19.9}$$

$$t_{j,k} - c_{i,k} - \delta_{i,j} + M\left(1 - x_{i,j,k}\right) \geq 0, \quad i \in I^*, j \in I^*, i \neq j, k \in K \tag{19.9'}$$

$$t_{i,k} + \omega_{i,k} - \ell_{i,p} + M\left(1 - z_{i,k}\right) + M\left(1 - v_{i,p}\right) \geq 0, \quad i \in I, p \in \left\{1, \ldots, \overline{p}_i\right\}, k \in K \tag{19.10}$$

$$t_{i,k} + \omega_{i,k} - u_{i,p} - M\left(1 - z_{i,k}\right) - M\left(1 - v_{i,p}\right) \leq 0, \quad i \in I, p \in \left\{1, \ldots, \overline{p}_i\right\}, k \in K \tag{19.11}$$

$$\sum_{p=1}^{\overline{p}_i} v_{i,p} = 1, \quad i \in I \tag{19.12}$$

$$c_{0,k} \geq \ell_0, \quad k \in K \tag{19.13}$$

$$t_{0,k} \leq u_0, \quad k \in K \tag{19.14}$$

$$t_{0,k} - c_{0,k} \leq \overline{D}_k, \quad k \in K \tag{19.15}$$

$$\omega_{i,k} \geq 0, z_{i,k} \in \{0, 1\}, \quad i \in I, k \in K \tag{19.16}$$

$$c_{i,k} \geq 0, t_{i,k} \geq 0, \quad i \in I^*, k \in K \tag{19.17}$$

$$v_{i,p} \in \{0, 1\}, \quad i \in I, p \in \left\{1, \ldots, \overline{p}_i\right\} \tag{19.18}$$

$$x_{i,j,k} \in \{0, 1\}, \quad i \in I^*, j \in I^*, i \neq j, k \in K \tag{19.19}$$

where M is a large positive number. Equation (19.1) defines the objective, which minimizes the total travel time. Constraint (19.2) ensures that each client is assigned to a single vehicle. Constraint (19.3) is the classical flow conservation. Constraint (19.4) guarantees that client i may immediately precede client j if and only if both are assigned to the same vehicle k. Constraints (19.5) and (19.6) ensure that each client i is visited once: i has exactly one immediate predecessor and one immediate successor. Constraint (19.7) guarantees that the quantity delivered by a vehicle does not exceed the vehicle's capacity. Constraint (19.8) defines the departure time of vehicle k from client i as the sum of its arrival, service, and waiting time. This equation holds when i is assigned to k, and is redundant otherwise. Constraints (19.9) and (19.9′) compute the arrival time of k to client j as the sum of the departure time of k from the immediately preceding client i and the travel time between i and j. Constraints (19.10) and (19.11) impose that vehicle k starts delivery to client i during TW p of i if and only if i is assigned to k and is served during p. They are redundant otherwise. Constraint (19.12) imposes that delivery to client i occurs during exactly one of the TWs of i. Constraints (19.13) and (19.14) impose that a vehicle k leaves and returns to d during the open time of d, whereas constraint (19.15) limits the time of k out of the depot to the maximal permissible threshold. Finally, constraints (19.16)–(19.19) define the variable types. Solving the above MIP is a challenge.

CPLEX fails to identify feasible solutions for instances with more than 20 clients when allocated 3 h of runtime.

19.3.3 Constraint Program

Alternatively, VRPMTW can be modeled as a constraint program. Let $\zeta_{i,k}$ be an optional interval variable indicating the activity of visiting client i using vehicle k. Let the interval variable h_i represent the activity of visiting client i. In addition, let $\boldsymbol{x_k}$ denote the route of k; that is, $\boldsymbol{x_k}$ is the sequence of interval variables $\zeta_{i,k}$. Then, the CP model follows.

$$\text{Min} \sum_{k \in K} \sum_{i \in I^*} \delta_{ij} : \text{next}\left(\boldsymbol{x}_k, i\right) = j \tag{19.20}$$

$$\text{s.t. } u_{i,1} \geq \text{startOf}\left(h_i\right) \geq \ell_{i,1} \parallel \ldots \parallel u_{i,\overline{p}_i} \geq \text{startOf}\left(h_i\right) \geq \ell_{i,\overline{p}_i}, \quad i \in I \tag{19.21}$$

$$\text{startOf}\left(\zeta_{0,k}\right) \geq l_0 \&\& u_0 \geq \text{startOf}\left(\zeta_{n+1,k}\right), \quad k \in K \tag{19.22}$$

$$\text{noOverlap}\left(\boldsymbol{x_k}, \boldsymbol{\delta}\right), \quad k \in K \tag{19.23}$$

$$\text{first}\left(\boldsymbol{x_k}, \zeta_{0,k}\right), \quad k \in K \tag{19.24}$$

$$\text{last}\left(\boldsymbol{x_k}, \zeta_{n+1,k}\right), \quad k \in K \tag{19.25}$$

$$\sum_{i \in I} q_i \ \text{presenceOf}\left(b_i^k\right) \leq Q_k, \quad k \in K \tag{19.26}$$

$$\text{endOf}\left(\zeta_{n+1,k}\right) - \text{startOf}\left(\zeta_{0,k}\right) \leq \overline{D}_k \quad k \in K \tag{19.27}$$

$$\text{alternative}\left(h_i, \zeta_{i,k}\right), \quad k \in K, i \in I \tag{19.28}$$

where || denotes the exclusive "or" operator, && the "and" operator, $\boldsymbol{\delta}$ the travel matrix expressed in time units, and $n + 1$ the depot d. Equation (19.20) minimizes the total travel time. Constraint (19.21) ensures that service to client i starts within exactly one of the TWs of i while constraint (19.22) guarantees that a vehicle leaves and returns to d when d is open. Equation (19.23) makes the subsets of clients served by different vehicles mutually exclusive. Equations (19.24) and (19.25) force each

vehicle to start and end its tour at d. Equation (19.26) imposes that the load of a vehicle be less than or equal to the vehicle's capacity. Equation (19.27) limits the time out of d of vehicle k. Finally, Eq. (19.28) is the alternative constraint, which assigns every client to a single vehicle. CP obtains near-global optima for some benchmark instances, in particular, when the TWs are wide.

19.3.4 Enhancing MIP and CP

We use the above models as stand-alone heuristics by allocating fixed runtimes to their commercial solvers. To enhance their solvability, we couple them with some strategies.

- First, we reduce the size of the problem by imposing logical assumptions that truck drivers adopt when constructing their solutions. A client j can't immediately follow a client i on the route of vehicle k if the travel time δ_{ij} is too large. Both MIP and CP formulations are augmented by the constraint $x_{ijk} = 0$ if $\left|o_i^h - o_j^h\right| > \frac{\overline{o}^h}{m}$ or $\left|o_i^v - o_j^v\right| > \frac{\overline{o}^v}{m}$ or $\delta_{ij} > \overline{\delta}$, where $\overline{o}^h = \max_{(i,j)\in I^* \times I^*} \left\{o_i^h - o_j^h\right\}$, $\overline{o}^v = \max_{(i,j)\in I^* \times I^*} \left\{o_i^v - o_j^v\right\}$, and $\overline{\delta}$ is a maximal threshold time.
- Second, we start the solvers from several feasible solutions. The multitude of starting points guards against the stagnation in local optima that are far from the global optimum.
- Third, we iteratively fix clients to vehicles, deliveries to TWs, and solve the resulting problems. This avoids the need for disjunctive constraints and the sensitivity of solvers to the choice of M. At each iteration, we augment the model with the tightest known bounds. We obtain the lower bound by removing all TW-related constraints, thus reducing the problem to a series of single machine scheduling problems with sequence-dependent setup times and due date windows. The upper bound is updated every time a VRPMTW feasible solution is obtained.
- Fourth, we iteratively divide the clients and vehicles into two or more disjoint subsets, solve the corresponding subproblems, merge their solutions, and compute the total travel time.
- Fifth, we solve MIP with emphasis on constraint satisfaction rather than on optimization. This yields better quality solutions while tackling larger problems.

19.4 Solution Approach

To apply the above enhancements in a more systematic manner, we approximately solve VRPMTW via a two-stage approach. Stage 1 constructs a pool of initial solutions, where a solution assigns each client to a vehicle and sequences the clients

of every vehicle. Stage 1 obtains solutions by one of nine constructive heuristics. These solutions are inspired from the way truck drivers construct their routes. They divide clients into sectors and choose the clients of a single vehicle from adjacent sectors. They construct the routes logically along the grid, a feature that enhances their chance of real-life implementation.

Stage 2 enhances the solutions of Stage 1 by perturbing the incumbent using a steepest descent neighborhood search. The search uses MIP to reassign clients to vehicles and CP to search for a feasible route of a vehicle. It embeds a look-ahead strategy that guards against infeasible/non-improving search directions. Sections 19.4.1 and 19.4.2 detail these two stages.

19.4.1 Stage 1

Because of the different distributions of the clients around the depot and of the varying number of vehicles and clients, it is difficult for a single heuristic to consistently generate good initial solutions. Therefore, we propose a panoply of constructive heuristics CH1, ..., CH9. They all impose a maximal travel time $\overline{\delta}$ between any (i', i^+) when appending location i^+ to the current location i' of a vehicle. In addition to constraining the search space, this conforms to the way drivers plan their routes. Drivers prefer balanced routes; they would rather have equal stretches than a long haul followed/preceded by a series of small hauls. The constructive heuristics differ in the way they assign clients to trucks and how they route them. For example, CH1, CH7, and CH8 start from the closest client to the depot while CH2, ..., CH6 select the first client randomly.

Some of the constructive heuristics (e.g., CH3 and CH4) use the notion of a sector. When the number of vehicles is large, the grid is divided into a series of radial sectors, where sector k, $k = 1, \ldots, m$, is delimited by the angles $\frac{2(k-1)\pi}{m}$ and $\frac{2k\pi}{m}$, with sectors $m+1$ and $m+2$ coinciding with sectors $k=1$ and $k=2$, respectively. Client i, $i \in I$, is part of sector k if its polar coordinate $\theta_i = \tan^{-1}\left(\frac{o_i^v - o_0^v}{o_i^h - o_0^h}\right)$ falls in the sector of k. When the number of vehicles is small, the grid is split into m guillotine rectangular adjacent bands, with sector k corresponding to the band delimited by $o^h = \frac{(k-1)(\overline{o}^h - \underline{o}^h)}{m}$ and $o^h = \frac{k(\overline{o}^h - \underline{o}^h)}{m}$, where $\underline{o}^h = \min_{(i,j)\in I^* \times I^*} \left\{ o_i^h - o_j^h \right\}$.

Let F denote the set of free clients. Unless differently stated, F is initially set to I. In addition, let $I_k = \{[1], \ldots, [n_k]\}$ denote the ordered set of $n_k = |I_k|$ clients of k where $I_k^* = I_k \cup \{d\}$ with $[n_k + 1] = [0] = d$ signaling the end and the beginning of the tour at the depot. Using these definitions, we detail below each of the nine constructive heuristics.

CH1 considers the vehicles sequentially starting with $k = 1$. For the current vehicle k, it initializes the vehicle's current capacity $Q_k^{\text{current}} = 0$ and the index of its current location $i' = d$. Iteratively, it assigns to k a free client $i^+ \in F$ such that

C1. i^+ is the closest client to i',
C2. $Q_k^{\text{current}} + q_{i^+} \leq Q_k$ —i.e., appending the demand of i^+ to the current load of k does not violate the load capacity of k,
C3. delivery to i^+ may start during one of the TWs of i^+, and upon completing delivery to i^+, vehicle k may reach the depot
C4. on-time (prior to u_0),
C5. without exceeding its own maximal time out of the depot $\overline{D}_k$.

Consequently, CH1 updates the vehicle's current load: $Q_k^{\text{current}} = Q_k^{\text{current}} + q_{i^+}$, current location: $i' = i^+$ and the set of free clients: $F = F \setminus \{i^+\}$. When CH1 fails to append any of the free clients to the current route of k, it considers the next vehicle: it sets $k = k + 1$.

Generally, CH1's total travel distance is smaller than those of all other constructive heuristics; however, it may induce large waiting times. It is greedy and myopic by nature.

CH2 assigns all clients of a given sector to a same vehicle. Then, for each vehicle k, it applies CP to search for a feasible route; i.e., it applies Eqs. (19.20)–(19.25) and (19.27) of CP for $K = \{k\}$. When CP fails to identify a feasible route, we relax Eq. (19.25).

CH2 identifies good quality solutions when the sectors concord with the geographical clusters. CH2 offers balanced routes with every vehicle having to eventually serve those far away clients of its sector.

CH3 mimics CH2 except that it considers the clients of a sector sequentially. It assigns each vehicle $k \in K$ to sector k. It then considers the vehicles sequentially starting with $k = 1$. For the current vehicle k, it initializes $Q_k^{\text{current}} = 0$ and the index of its current location $i' = d$. Iteratively, it assigns to vehicle k a free client i^+ from sector k such that $\delta_{i'i^+} < \overline{\delta}$, and constraints (C1)–(C5) hold. Consequently, it updates Q_k^{current}, i', and F.

When it fails to heuristically append any of the free clients of sector k to the current route, CH3 tries to insert a free client using an exact model. Let $f \in F$ denote the free client to be inserted in the route of k. Let the binary decision variable $\chi_{[i]} = 1$, $[i] \in I_k^*$, if client f is inserted immediately before client $[i]$ on the route of k, and 0 otherwise. Let the positive decision variables $t_{[i]}$, $\omega_{[i]}$, and $c_{[i]}$ denote, respectively, the time k reaches $[i]$, waits before delivery, and ends delivery to $[i]$, $[i] \in I_k$, and (c_0, t_0) the time k leaves and returns to d. Finally, let parameter $\Delta_{[i]} = \delta_{[i-1]f} + \delta_{f[i]} - \delta_{[i-1][i]}$ denote the net travel time change caused by inserting client f immediately before client $[i]$. In fact, k no longer travels between $[i-1]$ and $[i]$. Instead, it travels from $[i-1]$ to f to $[i]$. Using the above parameters and decision variables, MIP_f tries to insert iteratively every free client $f, f \in F$, that satisfies C3 to the route of k using the following mixed integer program.

$$\text{Min} \sum_{i \in I_k} \Delta_{[i]} \chi_{[i]} \tag{19.29}$$

$$\text{s.t.} \sum_{i \in I_k} \chi_{[i]} = 1 \tag{19.30}$$

$$c_{[i]} - t_{[i]} - \omega_{[i]} - s_{[i]} = 0, \quad i \in I_k \tag{19.31}$$

$$t_{[i+1]} - c_{[i]} - \delta_{[i][i+1]} \left(1 - \chi_{[i+1]}\right) + M \chi_{[i+1]} \geq 0, \quad i \in I_k \tag{19.32}$$

$$t_{[i+1]} - c_f - \delta_{f[i+1]} \chi_{[i+1]} - M \left(1 - \chi_{[i+1]}\right) \geq 0, \quad i \in I_k \tag{19.32'}$$

$$t_{[i]} - c_f - \delta_{f[i]} \chi_{[i]} + M \left(1 - \chi_{[i]}\right) \geq 0, \quad i \in I_k \tag{19.33}$$

$$t_{[i]} + \omega_{[i]} - \ell_{[i]p} + M \left(1 - v_{[i]p}\right) \geq 0, \quad i \in I_k, p \in \left\{1, \ldots, \overline{p}_{[i]}\right\} \tag{19.34}$$

$$t_{[i]} + \omega_{[i]} - u_{[i]p} - M \left(1 - v_{[i]p}\right) \leq 0, \quad i \in I_k, p \in \left\{1, \ldots, \overline{p}_{[i]}\right\} \tag{19.35}$$

$$\sum_{p=1}^{\overline{p}_{[i]}} v_{[i]p} = 1, \quad i \in I_k \tag{19.36}$$

$$c_0 \geq \ell_0 \tag{19.37}$$

$$t_0 \leq u_0 \tag{19.38}$$

$$t_0 - c_0 \leq \overline{D}_k \tag{19.39}$$

$$\omega_{[i]} \geq 0, \quad c_{[i]} \geq 0, \quad t_{[i]} \geq 0, \quad [i] \in I_k \tag{19.40}$$

$$v_{[i]p} \in \{0, 1\} \quad [i] \in I_k, p \in \left\{1, \ldots, \overline{p}_{[i]}\right\} \tag{19.41}$$

$$\chi_{[i]} \in \{0, 1\} \quad i \in I_k^*. \tag{19.42}$$

Equation (19.29) chooses the position of the free client f on the route of vehicle k. It minimizes the additional net travel time. Equation (19.30) imposes that f be inserted in exactly one position. It is possible that f can't be inserted without violating the TWs of the already routed clients. In such a case, MIP$_f$ is infeasible. Equation (19.31) relates the arrival, waiting, service and departure times for client $[i]$. Equations (19.32) and (19.33) compute the arrival time to a client relative to the departure time of the vehicle from the preceding client. Constraints (19.34)–(19.36) ensure that any client's service starts within exactly one of the client's TWs. Equations (19.37)–(19.39) restrict the vehicle's departure from, return to, and time out of d. Finally, constraints (19.40)–(19.42) define the variable types.

MIP$_f$ is much easier to solve than MIP. First, the clients' sequence is fixed. Second, the number of alternative positions for inserting f is small. Third and last, for every client $[i]$, $[i] \in I_k$, many TWs can be removed from the model by preprocessing the data. For $p \in \left\{1, \ldots, \overline{p}_{[i]}\right\}$, $v_{[i]p} = 0$, if

$$u_{[i]p} \leq \sum_{[j]=1}^{[i]} \delta_{[j-1][j]} + \sum_{[j]=1}^{[i-1]} s_{[j]}. \tag{19.43}$$

When the insertion of a free client of the current sector into the route of k deems impossible (i.e., MIP$_f$ is infeasible), CH3 appends it to sector $k + 1$.

CH3 is less sensitive to the physical limits imposed by the sectors while it yields good results when the clients are clustered. It maintains low traveled times per vehicle and makes the evolution of the route progress logically.

CH4 proceeds as CH3 does except that it considers clients from larger areas. It assigns to vehicle k a free client i^+ from sectors $k - 1$, k, and $k + 1$ such that $\delta_{i'i^+} < \overline{\delta}$ and C1–C5 hold. It generally obtains as good routes as CH3 or better, but may require more calls to MIP$_f$ (depending on the distribution of the clients on the grid and of the TWs) as it considers more free clients for each vehicle k, $k \in K$. CH4 is particularly useful in the presence of a mixture of clustered and randomly distributed clients.

CH5 and **CH6** are similar to CH3 and CH4 except that they build a route from both ends: assigning simultaneously two clients to the route: the client that is closest to the previous position of the truck and a client that is closest to the next position in the return path of the truck.

CH7 assigns each client i, $i \in I$, to a vehicle k, $k \in K$, where k is randomly selected from the discrete uniform $[1, m]$. It then applies Eqs. (19.20)–(19.25) of CP to sequence the clients of each vehicle. CH7 is a blind random search. As such, it can't be competitive when the clients are clustered. It can be used as part of a diversification strategy with a population-based meta-heuristic. Alternatively, it can be applied a fixed large number of times, and its best feasible solution is retained for further use. Obviously, it is useful when the clients are uniformly distributed.

CH8 balances the number of clients per truck. It considers the m trucks sequentially, assigning to each truck k, $k \in K$, the client that is closest to the previous location of k. It reiterates through the m trucks until all clients are assigned. When $\frac{n}{m}$ is integer, all trucks have an equal number of clients; otherwise, the numbers of clients for any two trucks do not differ by more than one. It then reduces the violations of the TWs of the clients of each route by applying CP.

CH9 is inspired from CH8 in that it assigns an equal number of clients to each truck except that it only considers clients from the sector of vehicle k, $k \in K$. When it exhausts the clients of its sector, vehicle k chooses clients of sector $k + 1$. Any unassigned clients of a sector k are automatically appended to sector $k + 1$. CH9 uses MIP_f to insert each client to the route of k.

CH3–CH9 apply a look-ahead search strategy that prohibits the search in infeasible directions. They abort constructing a solution if the demand of the free clients can't be assigned to the residual capacities of the vehicles or the residual travel time can't accommodate visiting them. This is guaranteed by solving a binary multiple choice multiple knapsack problem (MCKP) whose decision variables $y_{f,k} = 1$ if the demand of free client $f, f \in F$, can be loaded to vehicle k, $k \in K$, during one of the next iterations. MCKP is also used to choose the subset of free clients that may potentially be appended to the current route of the vehicle k when applied with $K = \{k\}$. Let c_k^{current} be the current total travel time of k and $\underline{\delta}_i = \min_{i' \in I^* \{i\}} \{\delta_{ii'}\}$ be the smallest travel time from i, $i \in I^*$, to any other client $i^{'}$, $i^{'} \in I^* \{i\}$. MCKP follows.

$$\text{Max} \sum_{k \in K} \sum_{f \in F} y_{f,k} \tag{19.44}$$

$$\text{s.t.} \quad \sum_{k \in K} q_f y_{f,k} \leq Q_k - Q_k^{\text{current}}, \quad k \in K \tag{19.45}$$

$$c_k^{\text{current}} + \sum_{f \in F} \left(\underline{\delta}_f + s_f \right) y_{f,k} + \underline{\delta}_0 - t_0 \leq \overline{D}_k, \quad k \in K \tag{19.46}$$

$$\sum_{k \in K} y_{f,k} \leq 1, \quad f \in F \tag{19.47}$$

$$y_{f,k} \text{ Binary}, \quad f \in F, k \in K \tag{19.48}$$

Equation (19.44) provides an upper bound to the number of free clients that can be inserted. Equation (19.45) ensures that the demand of the clients that could be assigned to a vehicle will not exceed the vehicle's residual capacity. Equation (19.46) guarantees that the additional travel time caused by the assignment of free client to a vehicle will not exceed that vehicle's residual travel time. Equation (19.47) reinforces that a free client is served by at most a single vehicle. Finally, Equation (19.48) defines the binary nature of the assignment variable; that is, no

partial delivery to a client. When the objective function equals the number of free clients, it may be possible to find a feasible solution to VRPMTW, without any guarantee that such a solution exists. On the other hand, a zero upper bound signals an infeasible direction of the search and halts the construction process. As this problem is solved a large number of times, its integer programming form is only considered when its linear relaxation succeeds in packing all free items.

19.4.2 Stage 2

Stage 2 enhances each solution of Stage 1 by perturbing the incumbent using a steepest descent neighborhood search and using CP to define both feasible and infeasible directions of the search. Three perturbation mechanisms are used:

- Merge I_k and $I_{k'}$ for k and $k^{'}$ serving adjacent sectors and apply CP for $|K| = 2$ and $I = I_k \cup I_{k'}$,
- Remove every client i, $i \in I_k$, and insert i into $I_{k'}$, $k \neq k^{'}$, using MIP$_f$
- Remove i_1 and i_2 from I_k and insert them, respectively, into $I_{k'}$ and $I_{k''}$ using MIP$_f$

Inserting client i into the route of vehicle $k^{'}$ is tested if (1) the residual capacity $Q_{k'} - Q_{k'}^{\text{current}}$ of $k^{'}$ exceeds demand q_i, (2) the slack time of $k^{'}$ is less than the minimal additional travel and service times caused by the insertion of i, and (3) the minimal variation Δ of the travel time caused by the swap is negative. That is, CP and MIP$_f$ are only applied when the perturbation may lead to an improving solution. A neighbor is improving if it (1) reduces the fitness of the incumbent or (2) is feasible while the incumbent is not.

19.5 Results

We applied the resulting heuristic on benchmark instances, which have randomly distributed (rm) clients, clustered (cm) clients, randomly distributed clusters of clients (rcm), tight (1) and wide (2) TWs. The tighter the TWs, the larger the number of vehicles is. The heuristic was run on a laptop with a 2.90 GHz Intel Core i7 processor and a 16.0 GB RAM. It was allocated a maximal runtime of 9 min per initial solution per instance. All calls to CP and MIP were allocated 3 and 10 s of runtime with all travel times rounded to their nearest integer.

Table 19.1 illustrates the variation of the solution values obtained by the constructive heuristics for 24 benchmark instances. Columns 1–5 display the label of the instance, the minimal and maximal number of TWs $\underline{p}$ and $\overline{p}$, the minimal and maximal time gap $\underline{\tau}$ and $\overline{\tau}$ between two consecutive TWs, the minimal and maximal ranges $\underline{w}$ and $\overline{w}$ of the TWs, and the number of vehicles m. Columns 6 and 7 display

Table 19.1 Variation of the solution values of Stage 1

	$(\underline{p},\overline{p})$	$(\underline{\tau},\overline{\tau})$	$(\underline{w},\overline{w})$	m	z^*	σ	Δ (%)	CH
rm101	(5,9)	(10,30)	(10,30)	10	1013.81	39.18	4.49	6
rm102	(5,7)	(10,30)	(10,30)	9	969.87	25.93	4.81	4
rm103	(4,7)	(10,50)	(10,30)	9	918.82	36.88	4.23	5
rm104	(3,6)	(10,70)	(10,30)	9	907.67	28.41	1.54	4
rm105	(2,6)	(10,100)	(10,30)	9	890.13	29.57	0.33	9
rm106	(2,3)	(50,100)	(30,50)	9	912.34	35.19	1.27	1
rm107	(1,3)	(50,150)	(30,50)	9	900.26	33.57	1.72	1
rm108	(1,2)	(100,200)	(50,100)	9	960.05	26.23	4.80	2
Average							2.90	
rcm101	(5,10)	(10,30)	(10,30)	10	1097.26	49.27	3.17	4
rcm102	(5,7)	(10,30)	(10,50)	10	1186.83	48.02	5.24	3
rcm103	(3,7)	(10,50)	(10,50)	10	1174.08	59.79	3.73	5
rcm104	(3,5)	(10,50)	(10,50)	10	1208.73	37.17	8.72	3
rcm105	(2,5)	(10,70)	(10,70)	10	1253.75	38.26	5.72	8
rcm106	(2,4)	(30,70)	(30,70)	10	1281.41	48.09	7.38	7
rcm107	(1,3)	(30,100)	(30,70)	11	1370.62	72.74	6.44	1
rcm108	(1,3)	(30,100)	(30,100)	11	1406.69	74.68	5.56	9
Average							5.75	
cm201	(5,10)	(100,150)	(50,100)	5	937.26	29.27	3.79	3
cm202	(5,7)	(100,200)	(50,100)	6	816.80	36.56	3.29	9
cm203	(3,7)	(100,300)	(50,100)	5	962.22	33.03	1.64	1
cm204	(3,5)	(100,500)	(50,100)	5	852.61	40.65	2.45	3
cm205	(2,5)	(200,500)	(100,200)	4	1058.42	32.04	3.86	1
cm206	(2,4)	(200,700)	(100,200)	4	941.89	38.84	3.58	7
cm207	(1,3)	(200,1000)	(100,300)	4	1173.93	27.34	3.60	5
cm208	(1,3)	(500,1000)	(100,500)	4	964.94	33.85	3.70	3
Average							3.24	

$\overline{z}$ and s, the average and standard deviation of CH1 to CH9 solution values. Column 7 reports $\Delta = \frac{(\overline{z}-z^u)}{z^u}100\%$, the percent deviation of $\overline{z}$ from the upper bound z^u obtained by TSVNS while column 8 indicates the index of the CH yielding the best solution value.

Table 19.1 suggests that none of the constructive heuristics obtains the best initial solution value consistently. Each of them is adapted to a particular class of TWs and tailored to a different distribution of clients on the grid. Table 19.1 further indicates that the initial solutions are reasonably good with a 3.96% average Δ from z^u.

Table 19.2 assesses the heuristic's performance. Columns 1–5 and 6–10 display the label of the instance, the number of vehicles m, the heuristic's best solution value z^*, the known bound z^u, and $\Delta^* = \frac{(z^*-z^u)}{z^u}100\%$the percent deviation of z^* from z^u.

None of the constructive heuristics makes stage 2 systematically obtain the best traveled time. This suggests that the nine local optima identified by stage 2 are not

Table 19.2 Heuristic's performance on benchmark instances

Instance	m	z^u	z^*	Δ^* (%)	Instance	m	z^u	z^*	Δ^* (%)
rm101	10	970.23	951	−2.02	rm201	3	686.42	688	0.23
rm102	9	925.34	921	−0.47	rm202	2	684.35	670	−2.14
rm103	9	881.49	883	0.17	rm203	2	674.01	657	−2.59
rm104	9	893.88	879	−1.69	rm204	2	664.88	657	−1.20
rm105	9	887.16	881	−0.70	rm205	2	651.30	648	−0.51
rm106	9	900.90	899	−0.21	rm206	2	672.80	659	−2.09
rm107	9	885.07	896	1.22	rm207	2	657.27	657	−0.04
rm108	9	916.04	912	−0.44	rm208	2	663.59	653	−1.62
Average				−0.52					−1.25
cm101	10	1101.16	1147	4.00	cm201	5	903.08	886	−1.93
cm102	11	1139.73	1190	4.22	cm202	6	790.81	825	4.14
cm103	12	1120.47	1152	2.74	cm203	5	946.67	979	3.30
cm104	14	1248.04	1251	0.24	cm204	5	832.21	834	0.21
cm105	10	1010.23	1096	7.83	cm205	4	1019.10	1069	4.67
cm106	10	982.22	999	1.68	cm206	4	909.37	930	2.22
cm107	11	1056.50	1059	0.24	cm207	4	1133.13	1179	3.89
cm108	10	967.33	968	0.07	cm208	4	930.54	980	5.05
Average				2.63					2.69
rcm101	10	1063.52	1069	0.51	rcm201	2	778.68	738	−5.51
rcm102	10	1127.68	1160	2.79	rcm202	2	815.90	757	−7.78
rcm103	10	1131.84	1124	−0.70	rcm203	2	721.97	695	−3.88
rcm104	10	1111.81	1128	1.44	rcm204	2	698.41	701	0.37
rcm105	10	1185.89	1184	−0.16	rcm205	2	754.51	711	−6.12
rcm106	10	1193.37	1196	0.22	rcm206	2	769.62	764	−0.74
rcm107	11	1287.67	1310	1.70	rcm207	3	749.78	756	0.82
rcm108	11	1332.57	1340	0.55	rcm208	2	742.70	681	−9.06
Average				0.79					−3.99

in the same vicinity. Indeed, their delivery plans differ in terms of routes and clients' sequencing. Thus the importance of the multiple restart of the heuristic.

The nine constructive heuristics are fundamental to building good initial starting points. Stage 2 obtains consistently lower traveled distances when initiated from these nine solutions than when initiated randomly. The comparison of the solution values obtained in Tables 19.1 and 19.2 highlights the importance of stage 2 in identifying a near-global optimum.

The proposed heuristic lowers the travel time for 22 out of 48 benchmark instances with the reduction reaching 9.06% for rcm208. All the time reductions were recorded for rm and rcm classes. The analysis of the results further showed that the best known solutions z^u correspond to solutions that are on the edge of feasibility. For instance, any slight increase of the travel time between pairs of clients may cause the violation of the TW constraints. This questions the validity of the best known solutions in a real-life context, as explained below.

19.6 Practical Considerations

Rincon-Garcia et al. (2018) discuss barriers to the implementation of VRP-related solutions in real-life contexts. Herein, we enumerate few such barriers and explain how the proposed model can handle them.

19.6.1 Single Versus Multiple Depots

Even though MIP and CP are given for a single depot, extending them to a multiple-depot case is straightforward. Two cases are possible: Subsets of vehicles are already allocated to each of the depots or the vehicles are allocated to the depots once an optimal solution to the problem is obtained. The former case is the more prevalent because vehicles are generally tagged to depots and goods are already stocked in specific depots. Such a case reduces to solving several single-depot problems. The latter case can be handled by defining I^* as the union of the set of depots and the set of clients. Then, Equations (19.9) and (19.9′) will apply to all $i \in I^*$ while (19.12)–(19.14) will be defined for each of the depots $d \in I^* \backslash I$. The same applies to Equations (19.22)–(19.24) of CP.

19.6.2 Distances and Travel Times

In MIP and CP, the time matrix $\boldsymbol{\delta}$ may be the result of a transformation into time units of the Euclidean distances that separate any pair of locations $(i, j) \in I^* \times I$. These distances are an accurate estimation of the true distances in the absence of natural or human made barriers or when locations are very far away, i.e., when the time needed to cover the travel distance is much larger than delays that may be caused by local traffic or the routing within the neighborhood of either i or j. In this case, transforming the distances into time units can be based on the average speed of the vehicle (on the highway). On the other hand, when the depot and clients are within the same city, a more relevant measure may be the actual distance the vehicles have to travel. For example, in newer cities, these distances could be estimated by the sum of the distances along the two axes because of the grid structure of the streets. In older cities, these distances should correspond to the true route that the vehicle will use. For instance, they can be approximated using Google Maps or similar software. Their transformation into time units should account for local traffic. The actual travel times are stochastic in nature because of traffic, weather conditions, road construction, unforeseen accidents, driver's state of mind, time of day, etc. Consequently, the time estimates corresponding to the traveled distances should be worst case estimates. This will avoid generating solutions that are on the edge of feasibility and that will cause tardy deliveries.

In the proposed models, travel times are not necessarily symmetric. Even when distances are equal, travel times may differ. In this sense, the above models are general. Travel time can always be expressed as a polynomial relationship of trip time.

19.6.3 Time Windows

The TWs are herein defined as hard; i.e., they must be satisfied for any solution to be feasible. This may be the case for vehicles in cities with travel bans or for deliveries in downtown areas or for private households. To ensure that their deliveries be on-time, constructors tend to account for possible traffic delays and reduce the range of their TWs. That is, their declared TWs are much tighter than their true TWs. For businesses, TWs are either soft or slightly flexible. Businesses tend to wait few extra minutes hoping to receive a delivery rather than to have it postponed. Similarly, households give conservative/tighter TWs than their true availability. They may resort to the help of a neighbor or a family member to receive the delivery. All these factors can be easily accommodated by altering Equations (19.9)–(19.14) and (19.21)–(19.22), e.g., by adding a flexibility factor φ and studying the sensitivity of the solutions to the size of φ.

19.6.4 Collection Versus Distribution

Both models apply whether the goods are being delivered or collected. However, for a distribution problem, items requested by a customer could be in different depots. In this case, the problem can be divided into two steps: Collecting all goods into a single depot (or moving from a national depot to a local one), then distributing the cumulated goods (i.e., delivery from a local depot to customers). Both steps can be perceived as VRPMTW with the clustering solved as in the generalized traveling salesman problem.

19.6.5 Packing Constraints

Equations (19.7) and (19.22) are loading constraints, expressed in terms of the capacity of the vehicle. They assume that the loads are shapeless volumes. However, unless the goods are liquid (e.g., petroleum for gas stations, water), these constraints are a relaxation of three-dimensional non-overlap and envelopment constraints. In some instances, the packing has to also account for the accessibility of the items

as a function of their order of delivery. Adding such constraints eliminates a large number of alternative solutions and reduces the search space.

19.6.6 Stochastic Travel Times and Time Windows

In MIP and CP, the travel times are deterministic. Yet, delays always occur. Travel times may vary according to time of day, traffic, weather conditions, etc. Even though delays are generally undesirable, they may be used in some circumstances to the advantage of the vehicle. When a vehicle is delayed on a route that doesn't include enough slack time to compensate for it, deliveries may have to be postponed to a later TW for a particular client but may allow an earlier delivery to another client. In such circumstances, it is judicious to apply a dynamic rerouting of the non-served clients with the objective of minimizing the travel time and the violation of TW penalties. This is of course only possible in small cities where deliveries are within the same vicinity.

Because travel times are non-deterministic, the above models assume that travel times and TWs correspond to worst case scenarios rather than average or modal values. In real life, these bounds are fuzzy. Vehicles reaching clients early may be able to start service immediately. Similarly, those reaching few minutes late may still get a chance to serve their clients during the intended TWs. Again, clients are generally willing to wait few extra minutes for the arrival of a delivery and may be available prior to the TW; i.e., the TWs' bounds are very conservative (i.e., tightest bounds).

When the average in lieu of the worst case performance is sought, a stochastic VRPMTW is at hand. It can be addressed via a stochastic program, which determines the routes that minimize the expected total travel time. The stochastic model can be tackled using a sample average approximation (Al-Khamis and M'Hallah, 2011) that iteratively determines the optimal routes of the vehicles for given samples of uncertain travel times and TWs. This method converges to the optimal travel time in the expected sense as the number of sampled scenarios increases. Each sample corresponds to a deterministic VRPMTW.

An alternative is to apply robust optimization (Mulvey et al. 1995). This method identifies solution robust (near-optimal) model robust (almost feasible) delivery plans. These plans are less sensitive to the uncertainty of the problem. They bound the infeasibility of the plans corresponding to the different scenarios and their distances from the optima by transforming the uncertain problem into a goal program. The weighted goal reflects the tradeoff between feasibility and optimality under all possible scenarios with the weights reflecting the likelihood of each scenario. Despite its attractiveness, robust optimization may become intractable for VRPMTW because of the very large number of possible scenarios, a number that increases with the number of clients, of TWs, and vehicles.

19.7 Conclusions

This chapter proposed a general framework for solving a complex compounded problem. This framework explores the complementary strengths of constraint and mixed integer programming. It builds and enhances solutions using constraint/mixed integer programming-based neighborhood search. The search is guided by feasibility constraints and look-ahead strategies that fathom large parts of the infeasible search space and focuses the search towards improving solutions. This framework was illustrated on the vehicle routing problem with multiple time windows. It can be extended to other compounded problems such as vehicle routing with packing constraints, bin packing with due dates, scheduling with cutting constraints or assembly limitations or resource unavailability.

This framework has several advantages. It requires no tuning and guarantees the reproducibility of the results. These two features can't be achieved by meta-heuristics that rely on seeds of random numbers. Such meta-heuristics include genetic algorithms, simulated annealing, ant colonies, swarm optimization, and bee colonies. Such techniques are generally assessed via their "average performance," a practice that is not justifiable in the service industry.

References

Adewumi AO, Adeleke OJ (2018) A survey of recent advances in vehicle routing problems. Int J Syst Assur Eng Manag 9(1):155–172

Alba E, Dorronsoro B (2008) A hybrid cellular genetic algorithm for the capacitated vehicle routing problem. In: Abraham A, Grosan C, Pedrycz W (eds) Engineering evolutionary intelligent systems. Springer, Heidelberg, pp 379–422

Al-Khamis T, M'Hallah R (2011) A two-stage stochastic programming model for the parallel machine scheduling problem with machine capacity. Comput Oper Res 38(12):1747–1759

Al-Mudahka I, Hifi M, M'Hallah R (2011) Packing circles in the smallest circle: An adaptive hybrid algorithm. J Oper Res Soc 62(11):1917–1930

Archetti C, Speranza MG (2014) A survey on matheuristics for routing problems. Eur J Comput Optim 2(4):223–246

Baldacci R, Mingozzi A, Roberti R (2012) Recent exact algorithms for solving the vehicle routing problem under capacity and time window constraints. Eur J Oper Res 218(1):1–6

Beheshti AK, Hejazi SR, Alinaghian M (2015) The vehicle routing problem with multiple prioritized time windows: a case study. Comput Ind Eng 90:402–413

Belhaiza S, M'Hallah R (2016) A pareto non-dominated solution approach for the vehicle routing problem with multiple time windows. In: IEEE Congress on Evolutionary Computation Proceedings, Vancouver, Canada, pp 3515–3524

Belhaiza S, Hansen P, Laporte G (2013) A hybrid variable neighborhood tabu search heuristic for the vehicle routing problem with multiple time windows. Comput Oper Res 52:269–281

Belhaiza S, M'Hallah R, Ben Brahim G (2017) A new hybrid genetic variable neighborhood search heuristic for the vehicle routing problem with multiple time windows. In: IEEE Congress on Evolutionary Computation, San Sebastian, Spain, pp 1319–1326

Braekers K, Ramaekers K, van Nieuwenhuyse I (2016) The vehicle routing problem: state of the art classification and review. Comput Ind Eng 99:300–313

Cordeau JF, Laporte G, Mercier A (2001) A unified tabu search heuristic for vehicle routing problems with time windows. J Oper Res Soc 52(8):928–936
Eksioglu B, Vural AV, Reisman A (2009) The vehicle routing problem: a taxonomic review. Comput Ind Eng 57(4):1472–1483
Favaretto D, Moretti E, Pellegrini P (2007) Ant colony system for a VRP with multiple time windows and multiple visits. J Interdisciplinary Math 10(2):263–284
Figliozzi MA (2010) An iterative route construction and improvement algorithm for the vehicle routing problem with soft time windows. Transp Res 18(5):668–679
Fu Z, Eglese R, Li LYO (2007) A unified tabu search algorithm for vehicle routing problems with soft time windows. J Oper Res Soc 59(5):663–673
Gendreau M, Hertz A, Laporte G (1994) A tabu search heuristic for the vehicle routing problem. Manag Sci 4(10):1276–1290
Gendreau M, Potvin JY, Bräysy O, Hasle G, Lokketangen A (2008) Metaheuristics for the vehicle routing problem and its extensions: a categorized bibliography. In: Golden B, Raghavan S, Wasil E (eds) The vehicle routing problem: latest advances and new challenges. Springer, New York, pp 143–169
Hashimoto H, Ibaraki T, Imahori S, Yagiura M (2006) The vehicle routing problem with flexible time windows and traveling times. Discret Appl Math 154(16):2271–2290
Ibaraki T, Imahori S, Kubo M, Masuda T, Uno T, Yagiura M (2005) Effective local search algorithms for routing and scheduling problems with general time-window constraints. Transp Sci 39(2):206–232
Koç Ç, Bektaş T, Jabali O, Laporte G (2016) Thirty years of heterogeneous vehicle routing. Eur J Oper Res 249(1):1–21
Laalaoui Y, M'Hallah R (2016) A binary multiple knapsack model for single machine scheduling with machine unavailability. Comput Oper Res 72(8):71–82
M'Hallah R (2007) Minimizing total earliness and tardiness on a single machine using a hybrid heuristic. Comput Oper Res 34(10):3126–3142
M'Hallah R (2014) Minimizing total earliness and tardiness on a permutation flow shop using VNS and MIP. Comput Ind Eng 75:142–156
M'Hallah R, Alkhabbaz A (2013) Scheduling of nurses: a case study of a Kuwaiti health care unit. Oper Res Healthcare 2(1–2):1–19
M'Hallah R, Al-Khamis T (2012) Minimizing total weighted earliness and tardiness on parallel machines using a hybrid heuristic. Int J Prod Res 50(10):2639–2664
M'Hallah R, Alkandari A, Mladenovic N (2013) Packing unit spheres into the smallest sphere using VNS and NLP. Comput Oper Res 40(2):603–615
Minocha B, Tripathi S (2013) Two phase algorithm for solving VRPTW problem. Int J Artif Intel Expert Syst 4:1–16
Moccia L, Cordeau JF, Laporte G (2012) An incremental tabu search heuristic for the generalized vehicle routing problem with time windows. J Oper Res Soc 63(2):232–244
Mulvey JM, Vanderbei RJ, Zenios SA (1995) Robust optimization of large-scale systems. Oper Res 43(2):264–281
Nagata Y, Bräysy O (2009) Edge assembly-based memetic algorithm for the capacitated vehicle routing problem. Networks 54(4):205–215
Nagata Y, Bräysy O, Dullaert W (2010) A penalty-based edge assembly memetic algorithm for the vehicle routing problem with time windows. Comput Oper Res 37(4):724–737
Nazif H, Lee L (2012) Optimized crossover genetic algorithm for capacitated vehicle routing problem. Appl Math Models 36(5):2110–2117
Polyakoskiy S, M'Hallah R (2018) A hybrid feasibility constraints-guided search to the two-dimensional bin packing problem with due dates. Eur J Oper Res 266:819–839
Polyakovskiy S, M'Hallah R (2014) A multi-agent system for the weighted earliness tardiness parallel machine problem. Comput Oper Res 44(4):15–136
Rincon-Garcia N, Waterson BJ, Cherrett TJ (2018) Requirements from vehicle routing software: perspectives from literature, developers and the freight industry. Transp Rev 38(1):117–138

Solomon M (1987) Algorithms for the vehicle routing and scheduling problems with time window constraints. Oper Res 35(2):254–265
Toth T, Vigo D (2002) Models, relaxations and exact approaches for the capacitated vehicle routing problem. Discret Appl Math 123:427–512
Vidal T, Crainic TG, Gendreau M, Prins C (2013) Heuristics for multi-attribute vehicle routing problems: a survey and synthesis. Eur J Oper Res 231(1):1–21
Vidal T, Crainic TG, Gendreau M, Prins C (2014) A unified solution framework for multi-attribute vehicle routing problems. Eur J Oper Res 234(3):658–673

Rym M'Hallah is currently a professor at the Department of Statistics and Operations Research at Kuwait University, and a professor of Quantitative Methods and Information System at the University of Sousse, Tunisia. She joined academia after industrial experience in Tunis, Tunisia. Dr. M'Hallah has earned degrees in industrial engineering and operations research, all from Pennsylvania State University.

Dr. M'Hallah's research focuses on modeling and optimizing large-scale systems using operations research techniques: mathematical programming, meta-heuristics, simulation, scheduling, and quality control. Applications include health care, manufacturing and transportation, scheduling, and cutting and packing. She has an extended international research collaboration network. Her current interest is enhancing health care systems and service industries using optimization techniques.

Dr. M'Hallah's choice of career was influenced at a very early stage by her grandmother and parents who persuaded her of the limitless benefits of a scientific career: notoriety along with a life style that involved discovering new worlds and making a difference in people's lives. Dr. M'Hallah was naturally gifted in Mathematics. Opting for engineering was a matter of luck: She got a scholarship to study in the USA. She was a natural optimizer, which made improving processes and enhancing designs a perfect match. She has been blessed with unconditionally supportive parents, inspiring women role models, enlightening scientific mentors, an encouraging husband and two loving wise kids. Her resilience and determination to succeed have garnered her more than 45 peer-reviewed journal papers. Today, she remains in the field of Science and Engineering as an educator, researcher, and mentor. She likes the challenges that real-life problems bring: the continuous/lifetime/independent learning, the personal growth, and the international collaborations with the friendships they build and the new experiences they offer. Last but not least, she appreciates the freedom of time management and of investigating her research interests.

Chapter 20
Modeling and Analysis of the Port Logistical Business Processes and Categorization of Main Logistics Costs

Carla Vairetti, Rosa G. González-Ramírez, Luisa Fernanda Spaggiari, and Alejandra Gómez Padilla

Contents

20.1 Introduction 457
20.2 Bibliographical Review 460
20.3 Research Methodology 461
20.3.1 Background of Foreign Trade in Chile 462
20.4 Port-Logistics Business Processes: Analysis and Categorization 463
20.5 Details of the Port-Logistics Business Processes Selected for the Analysis 467
20.5.1 Dispatching Full Direct Import Manifested to BW Business Process 470
20.5.2 Stacking and Consolidation Export from Bounded Warehouse Business Process 472
20.6 Determining and Estimating the Logistics Costs of the Selected Logistics Business Processes 476
20.7 Conclusions 480
References 482

20.1 Introduction

In the past few decades, the integration of economies in globalization has enabled a significant increase of foreign trade worldwide. Furthermore, the capacity and efficiency in the related operations of the global transport chain have also influenced this significant increase. The development of information systems, technology,

C. Vairetti · R. G. González-Ramírez (✉)
Facultad de Ingeniería y Ciencias Aplicadas, Universidad de los Andes, Mons Alvaro del Portillo, Santiago, Chile
e-mail: rgonzalez@uandes.cl

L. F. Spaggiari
Institución Universitaria Politécnico Grancolombiano, Bogotá, Colombia
Empresa Multimodal, Bogotá, Colombia

A. Gómez Padilla
Centro Universitario de ciencias Exactas e Ingeniería, Universidad de Guadalajara, San Pedro Tlaqupaque, Jalisco, Mexico

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_20

and transport infrastructure has an important impact on the competitiveness and efficiency of a nation's foreign trade (Ismail and Mahyideen 2015; Octavio and Sánchez Ricardo 2006).

Maritime transport mode has the most significant participation in foreign trade, with around 80% of the total volume as has been reported by UNCTAD from United Nations in the Review of Maritime Transport each year. As indicated by Song and Panayides (2008), ports are strategic nodes of the global transport chain and play a vital role in the competitiveness of foreign trade. The export and import processes imply a physical flow of the cargo and transportation vehicles, as well as an information flow that is exchanged among the different stakeholders involved. Information transmission can be paper-based or electronic and following international standards and protocols. Ports, as strategic nodes of this transport chain, need to foster more fluid procedures that may be coordinated among the different stakeholders, with visibility of the cargo throughout the port supply chain and efficient transmission of the related information and documents.

Digital innovation and the adoption of technologies at ports have been shown to lead to competitive advantages and the modernization of ports. To achieve such expected benefits, ports need to transform their intra- and inter-organizational business processes as indicated by Heilig et al. (2017). Ports have implemented inter-organizational information systems such as the port community systems (PCS) and related single window platforms to efficiently exchange information and enable fluid operations. In recent years, ports have been implementing advanced information systems and single windows-based systems (PCS) that enable electronic data interchange to further improve their performance and compete in the international port market. International competition demands huge investments in infrastructure and equipment to serve ever-growing ships that force the prices down due to economies of scale. Therefore, one of the key elements to maintain or improve their position in the global markets is to achieve more fluid operations for the import and export processes.

In the particular case of Chile, the country has a long coast of approximately 4200 kilometers. Foreign trade is a relevant economic sector of the country, as several of their main products, such as mining and forestry, are exported. Moreover, the country has not developed a strong manufacturing sector, and is mainly focused on services.

Hence, most of the consumption products from the retail are imported. In 2015, from the total volume, 92% of the imports and 96% of the exports were transferred by maritime ports, as reported in 2015 by National Customs of Chile. From all the volume, containerized cargo accounts to around 23% of the total imports and 26% of the total exports. Furthermore, this represents around 50% of the total cargo transferred by the main ports in the country that are the Ports of Valparaiso and San Antonio, as indicated by DIRECTEMAR for the volumes transferred in 2016 in Chile.

At present, there are several inter-organizational information systems being developed in Chile as a single window. A national single window (NSW) known as "SICEX" is currently being developed and implemented to provide a single

platform for the related business processes that involve public entities. In the Port of Valparaiso, a PCS referred as "SILOGPORT" is under implementation, providing functionalities for the coordination of vehicle flows between the pre-terminal (extension zone of logistics activities, ZEAL) and the port terminals (Terminal Pacifico Sur, TPS; Terminal Cerros de Valparaiso, TCVAL). It has also incorporated a tracking functionality of the status of the cargo at the different stages of the import and export processes.

There are several stakeholders involved in the operations of foreign trade, both public and private. This includes transport operators, stevedores, port authorities, public entities such as customs, freight forwarders, cargo agents, and logistic operators, customs agents, empty container depots, and bonded warehouses. A bonded warehouse is a facility that has been certified by customs as a primary zone. This means that import cargo can be manifested to this facility and remain up to 90 days prior to officially entering the country. This allows importers to postpone the payment of the corresponding fees. Additionally, it is possible for entities such as customs to perform cargo inspection, and provide other value-added services to the cargo such as stuffing/de-stuffing.

In this chapter, we model the logistics business processes with the participation of a bonded warehouse. These are referred to as the "Dispatching full direct import manifested to bonded warehouse" and the "Stacking and consolidation export from bonded warehouse." As a case study, we considered the Port of Valparaiso and one of the bonded warehouses located in the inter-port area. Furthermore, we propose a framework for the logistics cost analysis associated to the business process. The models and framework proposed have been validated with stakeholders of the port and bonded warehouse, based on semi-structured interviews and site visits. Accordingly, we provide recommendations to support the standardization of the related processes that are presented to port authorities and related stakeholders. Furthermore, we provide a comparative analysis of the logistics costs for the case of the Port of Valparaiso, the Port of Buenaventura in Colombia, and the Port of Buenos Aires in Argentina.

This chapter is structured as follows: Sect. 20.2 provides a bibliographical review where we present related work and state the contributions of this chapter. Section 20.3 provides a description of the research methodology and a context of the current situation of foreign trade in Chile. Section 20.4 presents the modeling of port-logistics business processes, providing a categorization. The documents and stakeholders involved are described. Section 20.5 presents the modeling for the two logistics businesses processes selected for the analysis, including a description of the required documents, and the interactions between stakeholders and processes. The estimation of logistics costs for the selected business processes being studied is provided in Sect. 20.6, where a comparative study is also presented for the Ports of Valparaiso, Buenaventura, and Buenos Aires. Finally, the main conclusions of this study, including recommendations for future developments, are provided in Sect. 20.7.

20.2 Bibliographical Review

The competitive advantage of a small country could be generated by improving the effectiveness rates and indexes of the activities of the port-centric logistics system as well as by the harmonization of strategic goals, tasks, and measurements (Sujeta and Navickas 2014). Integrated logistics support of operational and marketing strategies yields quality customer service, but its ability to support even broader and higher service strategies creates a sustainable and competitive advantage (Orlic Protega et al. 2011).

According to Jing and Jia-Wei (2010), the level of the logistics service is one of the most important determinants of port competitiveness. Accordingly, they propose an evaluation index system to measure the port-logistics competitiveness and evaluate the Port of Ningbo-Zhoushan as a case study. They identified 23 indicators divided into 5 criteria: size of the port, infrastructure, hinterland economy, development, and management level. Dou and Li (2015) propose a methodology to identify the key factors of port-logistics optimization; in their analysis that considers data from 9 years of operations at the Port of Hekou, they found that the influencing factors changed every year, but the key factors will not fluctuate significantly. Such traditional factors are: the support and political willingness, the import and export value of Honghe state, the import and export value of the Port of Hekoum, and the Honghe state's port information level.

Feng et al. (2012) compared port performances in port regions, finding that the most critical differences between the case ports are government support, proximity, speed of cargo handling, safety, and port technical infrastructure in descending order. All these previous factors have an impact also in the competitiveness of a port and a region, and a strong impact on logistics costs. On the other hand, Sánchez et al. (2003) analyze port efficiency and its impact on maritime transportation costs, while Avelar-Sosa et al. (2014) analyze the impact of regional infrastructure on services and logistics cost.

Martinez Zarzoso et al. (2011) and Martínez Zarzoso and Wilmsmeier (2010) study freight rates in Latin America, and Márquez-Ramos et al. (2007a,b) study the case of Spain. Pérez-Salas et al. (2015) provide a framework to measure the logistics costs with a focus on natural resources or agricultural supply chains. As a case study, they apply their proposed framework to estimate cost overrun in natural resources logistics chains of Bolivia. Nguyen et al. (2016) study the interaction between ports in a network and examine how a port sets its prices for infrastructure services given those of its competitors. Suárez-Alemán et al. (2018) present a study of the port regulation in LAC from a competition perspective in the region, indicating that the level of competition is low in the region. Their results indicate that Chile and Mexico have the best practices in the region, with better conditions to ensure competition in the market.

As observed, there are several contributions in the literature in which the importance of logistics systems, infrastructure, and service level is highlighted. Port performance strongly depends on the fluidity of the multiple supply chains

in which the port is a critical echelon of global transport chains. In this regard, port-logistics business processes and inter-organizational information systems (IOS) play a critical role on the performance of such global supply chains. Elbert et al. (2017) model key business processes in maritime transport chains and analyze how information is exchanged.

In this chapter, we extend previous analysis to consider a particular case study of a port in Chile, considering in particular, the business processes in which a bonded warehouse is participating. The concrete contributions of this work can be summarized as follows:

- A characterization of the port-logistics business processes in order to standardize and determine categories of the different business processes involved in the port for the foreign trade processes (import, export, and transit).
- As a case study, the Port of Valparaiso is considered for modeling two specific logistics business processes that involve the participation of a bonded warehouse in an import and export process.
- A comparative analysis of the related costs for the logistics business processes under consideration for the Port of Valparaiso related to the Port of Buenaventura, and the Port of Buenos Aires.
- Recommendations for port authorities and related stakeholders.

20.3 Research Methodology

The research methodology is summarized in Fig. 20.1. The first stage begins with the literature review and the state of the art, with the aim to provide a background on the topic. We reviewed the literature in which global supply chains and port-logistics business processes are reviewed, as well as the notation and methodologies of business process modeling and analysis (BPM and BPA). Furthermore, we analyze the current volumes of cargo transferred in Chile for foreign trade and, more specifically, in the port of study, the Port of Valparaiso. As a result of this stage, we provide a categorization of the main port-logistics business processes and its participation, to further indicate which are the logistics business processes that are the focus of this study. These correspond to those in which a bonded warehouse is participating in the processes.

The second stage considers the empirical work to modeling the processes and the detailed description of the stakeholders' interactions and documents required in the logistics business processes selected. The feature-based modeling approach considers UML Use Cases to represent the processes and interactions (Elbert et al. 2017). The models are segregated by stages and we also identify the related costs and the documents involved. For this, we organized interviews with personnel of a bonded warehouse that was participating in the study.

The third stage considers determining and estimating the corresponding costs to each of the logistics business processes under consideration. More specifically,

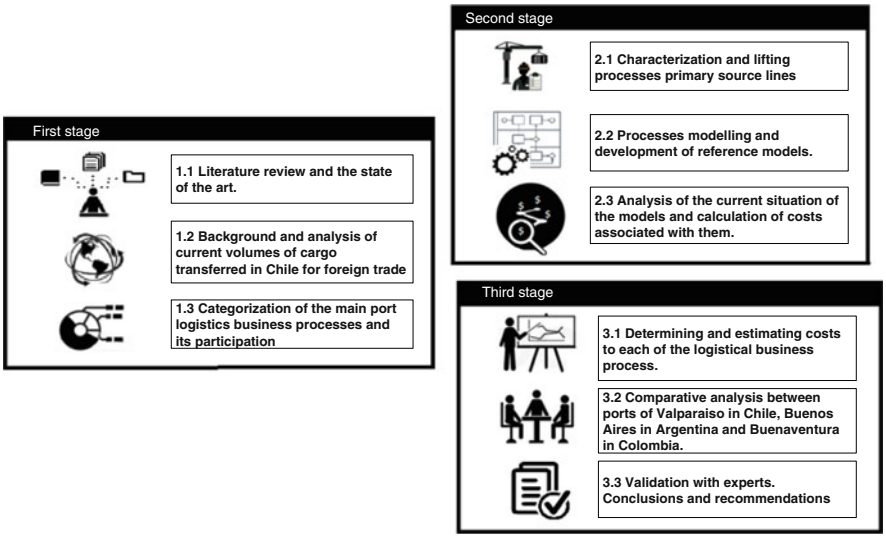


Fig. 20.1 Research methodology

we consider the operations of the Valparaiso South Pacific Port Terminal (TPS in Spanish), based on public records and information provided by the bonded warehouse. We present a comparative analysis with the Port of Buenos Aires in Argentina and the Port of Buenaventura in Colombia. We finalize this stage with the validation of the models and logistics cost analysis with personnel from the bonded warehouse. We discuss the results and provide recommendations as well as proposals for further extensions.

20.3.1 Background of Foreign Trade in Chile

As it was previously mentioned, maritime transport mode accounts for more than 80% of the total volume according to the Review of Maritime Transport published by the UNCTAD (UNCTAD 2017). UNCTAD estimates that seaborne trade is increasing by 2.8%, with total volumes reaching up to 10.6 billion tons. UNCTAD expects those volumes will expand across all segments with containerized cargo and major dry bulk commodities presenting the fastest growth. Particularly, in the case of developing economies such as those in the region of Latin America, the evolution of world seaborne trade volumes is expected to be 59% for exports and 64% for imports. Maritime ports are the nodes that support maritime and hinterland connectivity and containerized cargo is increasing its participation compared to other transport conditions. The introduction of containers has been a determining factor of the growing volumes of foreign trade. In the case of Chile, containerized

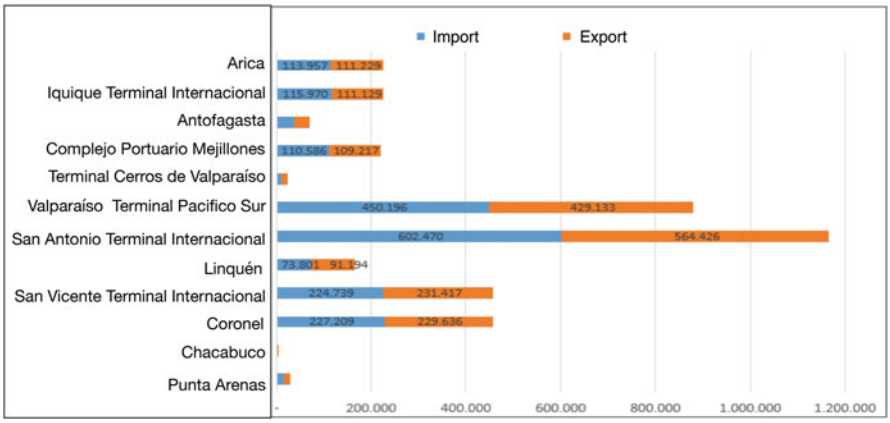


Fig. 20.2 Total flow of containers mobilized in national ports

cargo accounts to around 23% of the total imports and 26% of the total exports. Furthermore, this represents around 50% of the total cargo transferred by the Ports of Valparaiso and San Antonio.

In the particular case of the Port of Valparaiso, and particularly the main port terminal, TPS, 64% of the containerized cargo in 2015 corresponded to 40' containers and 36% to 20" containers for the exports, and 57% to 40" containers and 33% to 20" containers for the imports. On the other hand, regarding trade flows that correspond to bonded warehouses, 70% of the containers handled corresponded to 40" containers in 2015 (Fig. 20.2).

20.4 Port-Logistics Business Processes: Analysis and Categorization

In this section, we propose a reference model for the analysis of the port-logistics business processes, considering the operations of seven public Chilean ports, based on the information provided by the Ministry of Transport in Chile. Chile is a country characterized by a long coast with an open economy, a reason for which foreign trade is a significant sector in the economy of the country. Chile has 10 ports of public usage along its coast: Arica, Iquique, Antofagasta, Coquimbo, Valparaiso, San Antonio, Talcahuano-San Vicente, Puerto Montt, Chacabuco, and Austral. In terms of the volume of TEUs transferred, the Ports of Valparaiso and San Antonio are the country's principal ports, and they are both located around 100 km from the metropolitan area of Santiago, the country's capital and main origin and destination of the cargo. At each port, there is a port authority. The main stakeholders that interact in the port-logistics business processes are briefly described, based on Brodies Dictionary of Shipping Terms:

- **Port Terminal Operator (PTO)**: Entity that operates the area (terminal) designated for the handling, storage, and possibly loading or unloading of cargo into or out of containers, and where cargo can be picked up, dropped off, maintained, stored, or loaded or unloaded from one mode of transport to another (that is, vessel, truck, barge, or rail).
- **Shipping Line (ShL)**: Entities dedicated to the operation of the vessels, which may be of their property or not. They are responsible for maritime cargo transportation.
- **Shipping Agency (ShA)**: Entities that act in representation of the ShL in each country and provide services related to the cargo and the fulfillment of clearance and other procedures required at each country by the shipping line.
- **Transport Carrier (TC)**: An inland transportation company that hauls export or import traffic between ports and inland points. The inland carrier has a fleet of trucks for the transportation service of the cargo by road, both for the drayage of containers in the inter-port area and for the atomized transport service to/from the hinterland.
- **Freight Forwarder (FFW)**: Person or company who arranges for the carriage of goods and associated formalities on behalf of a shipper. The duties of a forwarder include booking space on a ship and providing all the necessary documentation.
- **Custom Agent or Broker (CB)**: Person or firm, licensed by the customs authority of their country when required, engaged in entering and clearing goods through customs for the shipper and consignee.
- **Shipper**: Cargo exporter.
- **Consignee (CC)**: Importer or the entity to whom the imported cargo is manifested.
- **Custom Officers**: Officers from National Customs who perform the inspection of cargo when required.
- **Empty Container Depot (DEPOT)**: Facility that provides the service of maintenance and storage of empty containers to the ShL.
- **Bonded Warehouse (BW)**: A warehouse authorized by customs authorities for storage of goods on which payment of duties is deferred until the goods are removed.
- **Container Freight Station (CFS)**: Dedicated port or container terminal area, usually consisting of one or more sheds or warehouses and uncovered storage areas where cargo is stuffed or de-stuffed into/from containers and may be temporarily stored in the sheds or warehouses.
- **Central Bank or Banks (B)**: Commercial bank or the Central Bank of the country.
- **Port of Origin (PoO)**: Port of origin of the cargo from which it is transported to its destination.

The main port-logistics business processes are characterized by the operations of cargo whose control and coordination processes in the port are similar by nature. Accordingly, we propose the following categorization (Fig. 20.3) based on the operations of the Chilean ports, focused on the containerized cargo.

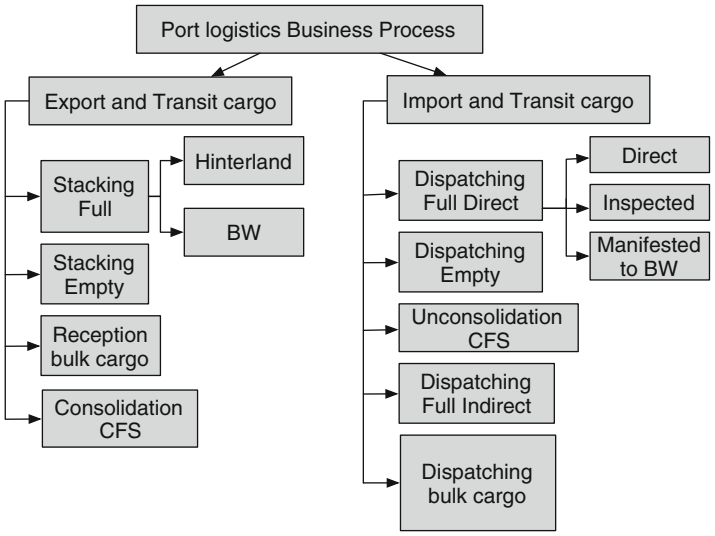


Fig. 20.3 Classification of the main port-logistics business processes

As observed in the previous figure, the port-logistics business processes related to the export business processes considered both export cargo and transit cargo. The latter refers to cargo that is exported from a different country (e.g., Argentina and Bolivia in the case of Chile), but transferred at a national port.

Accordingly, the export-logistics business processes are:

1. **Stacking Full-Hinterland**: This refers to the process of stacking full containers (dry and reefer) at the port terminal transported from the hinterland, considering dry, reefer, and transit containers, as well as transshipment which, in the case of Chile, is not significant.
2. **Stacking Full-BW**: It considers the operations related to stacking full containers that have previously been inspected or stored at a bonded warehouse (BW) that is located at the inter-port area. This type of movement may involve the transport of several containers (massive movement) and is also referred to as staging or bulk runs.
3. **Stacking Empty**: It considers the stacking of export-empty containers to be loaded in a vessel for the re-positioning operations of empty containers of the ShL, and is received mainly from an empty container depot or in some cases from the CC if the port terminal offers the service of maintenance and storage of empty containers to the ShL (situation that, in Chile, is common at the ports of the south region such as Talcahuano). When containers are received in batches called massive movements, the operation is also known as staging or bulk runs.
4. **Stuffing-CFS**: This refers to the operations related to the stuffing of general cargo at the CFS of the port terminal. Hence, cargo is received in bulk condition

and stuffed in the CFS. This type of operation is very typical at Chilean ports with the exception of the Port of Valparaiso, due to limitations of space.
5. **Reception Bulk Cargo**: This refers to the operations for the reception of bulk cargo that will be loaded on the vessel in this condition. In this business process we also consider fractional cargo (pallets).

On the other hand, the import-logistics business processes identified are:

1. **Dispatching Full Direct–Direct**: The direct dispatching procedure considers the operations in which the CC has complied with all the regulations and requirements to enter the cargo to the country and, in this case, considering full containers. The direct–direct corresponds to those containers that are manifested under this condition and fulfill all the requirements prior to the arrival of the vessel to the port, and are scheduled based on an appointment system in which trucks are given a specific schedule (appointment) to pick up cargo once the containers have been unloaded from the vessel. The appointments are defined according to the unloading sequence, so that there are no container reshuffles done by the yard crane. Trucks are required to arrive with anticipation to their appointments. If the truck is late, a fine is imposed.
2. **Dispatching Full Direct-Manifested to BW**: In this case, the full container is manifested to a bonded warehouse so that the clearance and regulations compliance are not performed at the port terminal. The CC has to manifest this condition prior to the arrival of the vessel and these containers may be segregated for their dispatching in a batch or massive movement of containers. This operation is also known as bulk runs or staging.
3. **Dispatching Full Direct-Inspected**: In this case, the full container was manifested as a direct container by the CC, but the container was selected for inspection reasons for which it cannot be dispatched until this condition is cleared. Hence, this business process considers all the required processes for coordinating the inspection of the container with the corresponding officers from customs and/or agricultural and quarantine.
4. **Dispatching Full Indirect**: This corresponds to the dispatching of full containers that need to be stored at the port at a certain time because the CC did not perform the required procedures for the authorization and clearance of cargo (supported by the custom broker) with anticipation. In this case, containers under this condition are typically stacked in different positions of the yard.
5. **Dispatching Empty**: This corresponds to the dispatching of empty containers that were imported for re-positioning operations of the ShL and will be also dispatched in batches or massive movements to the corresponding empty container depots. Except for the ports at the south of the country that also handle empty containers at the terminals, the ShL maintain their stock of empty containers mainly at empty container depots.
6. **De-stuffing CFS**: This operation corresponds to the process of de-stuffing cargo from the containers at the CFS.
7. **Dispatching Bulk Cargo**: This corresponds to the dispatching process of bulk cargo that is handled by the port under this condition.

20.5 Details of the Port-Logistics Business Processes Selected for the Analysis

In this section, we describe two business processes: The first one is related to the dispatching full direct import manifested to bounded warehouse (Sect. 20.5.1) and the second one is related to the stacking and consolidation export from bounded warehouse (Sect. 20.5.2).

Before describing the processes, we introduce the different documents that are generated and exchanged between the related stakeholders, indicating also for the specific case of the Port of Valparaiso how the documents are generated and transmitted (either paper-based or electronically, or by e-mail). It is important to mention that the ShL use the INTTRA Global Shipping Platform, which is an electronic platform for the business transactions between shipping partners.

- **National Import Declaration (DIN by its acronym in Chile)**: is a document by means of which a customs destination is formalized for the entry of import cargo to the country, indicating the value of the merchandise. This is generated by the CB and transmitted electronically to customs authorities.
- **National Export Declaration (DUS by its acronym in Chile)**: is a document that declares the cargo to be exported. This is generated by the CB and transmitted electronically to customs authorities. There are preliminary versions of this document (DUS-SAT) and the final and legalized version (DUS-LEG).
- **Request and Confirmation of De-stuffing Service**: These are documents exchanged among the CB or FFW and the BW.
- **Letter for the Opening of the Container**: This document is transmitted by the FFW to the ShA. In the Port of Valparaiso, it is generated physically.
- **Bill of Lading (B/L)**: This is one of the most important documents that is issued by a carrier (ShL in this case or the ShA) to acknowledge the receipt of cargo for shipment. During each sub-process, it is possible that some original and non-negotiable copies are generated, as well as the breaking down B/L or the payment receipt. Depending on the case, the document can be generated by the information system of the shipping line and transmitted by e-mail to the FFW or if it is on paper by courier or in a physical form. In turn, the B/L that the ShA generates and transmits to customs is generated electronically and transmitted among the systems just as when customs transmits the B/L to the BW.
- **DRES**: This is the document that is generated by the PTO at destination to confirm the receipt of cargo at the port terminal facilities. In the case of the Port of Valparaiso, this document is transmitted on paper.
- **DR**: This is very similar to the DRES but corresponds to the document generated by the BW and is also transmitted on paper.
- **Cargo Unloaded Record/PTO**: This document is the registry that the cargo has been unloaded from the vessel at the port terminal and is transmitted electronically from the PTO to the Shipping Line.

- **Record Gate-Out/PTO**: Document that registers the gate-out from the PTO and is elaborated electronically in the system of the PTO at destination of the cargo and sent to the information system of the shipping line.
- **Record Gate-In/BW (DEPOT)**: This document confirms the gate-in at the BW and is transmitted electronically to the shipping line. Similarly, the same document is generated when an empty container is received at a depot and is also transmitted electronically to the shipping line by the DEPOT. A related document that is also generated is the invoice for the gate-in at the DEPOT. The DEPOT transmits this document to the CB electronically, but the CB transmits this document to the PTO at destination, paper-based or by e-mail.
- **Release Order**: This document is generated by the ShA and transmitted to the PTO at destination by e-mail.
- **CAL**: This is the list of containers (empty and full) to be loaded onto a vessel, and this is generated by the shipping line and delivered to the PTO by e-mail.
- **De-stuffing Letter**: This document is generated by the PTO at destination to the FFW at destination physically.
- **Request and Confirmation of Stuffing Service**: These are documents exchanged among the CB or FFW and the BW. Depending on the port, they may be generated and transmitted by e-mail.
- **Temporary Container Admission Deed (TATC)**: This is the document that allows the entry of an empty container to a country considering it as a transport mode and not as an imported product. The document is generated by the ShA and transmitted to the PTO at destination electronically.
- **National Cargo Manifest**: This is a document that must be issued by vendors and service providers, when it has been decided to postpone the delivery of the invoice in a sale and when goods are transferred, regardless of whether this constitutes a sale. This is considered for national cargo (or import cargo that has been nationalized).
- **Truck Card**: This is a paper-based document that is delivered to the truck driver when they enter the BW to deliver the import container brought from the port terminal. Depending on the port digital level, this could be also a pin number.
- **EIR (Equipment Interchange Receipt)**: This is a document generated to transfer the container from one stakeholder to the other. For instance, from the PTO to the BW where a full container is received, or from the BW to the DEPOT where the empty container is delivered. The documents are generated physically in paper.
- **Damages Record and Repair Confirmation Letter**: These are documents generated by e-mail in which the DEPOT notifies the damages that the empty container has and requirements for repairing the container by the ShA. And the ShA notifies by a confirmation to the depot that it accepts that the DEPOT perform the repairs.
- **Booking Request and Confirmation**: The FFW or cargo agent generates the booking of space with the shipping line electronically on INTTRA, and receives the confirmation electronically in the same system.

- **Loading Instructive**: This is a document that is generated by the FFW or CB at the origin of the cargo and transmitted by e-mail to the BW by the shipper information system.
- **Shipping Instructive**: This is a paper-based document that is transmitted physically by the shipper to the transport carrier. The transport carrier delivers this document to the BW when it arrives with the cargo.
- **National Import Document (DIN by its acronym in Chile)**: is a document that is generated by the FFW or CB at the origin in the Customs Information Portal and in Chile is transmitted electronically to customs. In addition, the FFW or CB generates a physical copy which is transmitted paper-based to the BW.
- **Customs License**: This is a document that is generated by the BW in paper and exchanged with the Customs System.
- **Packing List**: This is an inventory of the incoming cargo. The BW generates this document and transmits it to the FFW at origin by e-mail.
- **Dispatching Instructive**: This is a document that is generated by the FFW at origin and transmitted by e-mail to the BW.
- **Dispatching Request and Order**: This is a document that the BW generates for picking up an empty container at the DEPOT and is transmitted by e-mail. In addition, the BW generates a dispatching order physically and provides it to the transport carrier.
- **Release Order—DEPOT**: The shipping line generates this order electronically for the pick-up of an empty container and transmits it to the DEPOT.
- **EIR DEPOT/BW and BW/PTO**: The EIR Depot/BW is generated by the DEPOT and delivered physically to the transport carrier. The EIR BW/PTO is generated by the BW and delivered physically to the transport carrier.
- **Record Gate-Out/DEPOT**: Document that registers the gate-out from the DEPOT and is elaborated electronically in the DEPOT system and sent to the information system of the shipping line.
- **Record Gate-In and Gate-Out/BW**: This document confirms the gate-in at the BW and is transmitted electronically to the shipping line. Similarly, the same document is generated for the gate-out of the full container from the BW and sent electronically to the shipping line.
- **Record Gate-In/PTO**: This document confirms the gate-in at the PTO at origin and is transmitted electronically to the shipping line.
- **Loading Record**: This document confirms that the cargo has been loaded onto the vessel and is generated electronically by the PTO and transmitted to the shipping line.

20.5.1 Dispatching Full Direct Import Manifested to BW Business Process

In this business process (BP) the container carries out its procedure in advance to be manifested to a bounded warehouse, granting it full responsibility over the container, and taking charge of the customs procedures for the admission and control of the corresponding payments (administrative role). Finally giving authorization to the direct delivery of the container on the land transport of the importer to be dispatched from the terminal to their warehouse and then be returned to the empty deposit. This BP contains 14 main processes that interact between actors and documents. Based on Fig. 20.4 we describe the business process separated in three interfaces:

1. **Maritime Stage:** In this case, the process begins with the negotiation between the CC and the FFW, establishing an import contract, and sending the corresponding documents for the cargo to the FFW. With this, the FFW performs the required procedures for the B/L with the shipping line, presenting a request for its emission and submission. With this document, the FFW completes the required data of the cargo and fleet in the offices of the shipping line. In this moment, the shipping line verifies that all the required data is correct and proceeds to legalize the B/L and inform the corresponding port of destination. Then, the FFW sends the documents, including the B/L, to the customs broker (CB). The CB proceeds to do the required procedures of the B/L with the BW and proceeds to internalize the cargo. At the same time, the ShL presents to customs authorities all the B/L exchanged that is referred to as the maritime manifest. If such documents are approved, then the ShL proceeds to submit the manifest to the PTO and inform the total number of containers that are expected to arrive so the PTO can plan the loading and unloading of cargo from the vessel. On the other hand, the CB performs the procedures for the warranty of the container along with the ShL, with the corresponding payment of the gate-in in order to obtain the temporary assignment of the container. A proof of payment is generated along with the TATC document. Once the container has been assigned, the CB proceeds with the

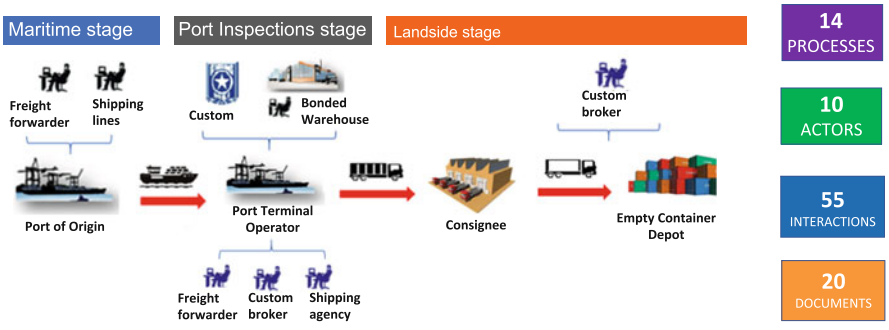


Fig. 20.4 Dispatching full direct import manifested to BW business process

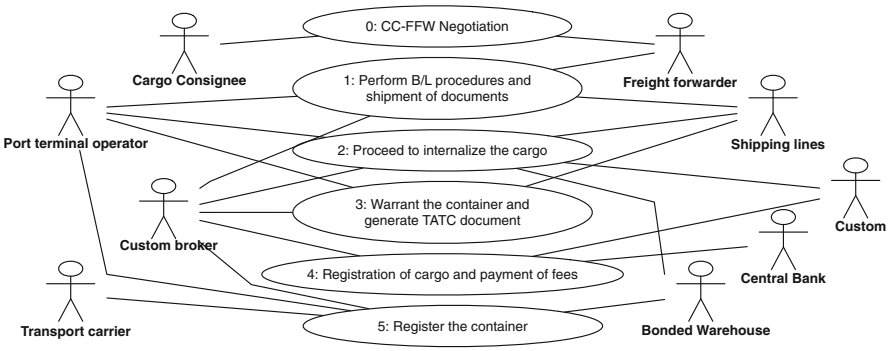


Fig. 20.5 Use cases of Maritime Interface

registration of the cargo by the entry declaration that is also presented to customs. If customs authorizes the cargo and it does not require inspections, then the CB makes the payment of the corresponding fees for the containers, reflecting the ad valorem taxes. Then, the CB registers the container for the sequence of the direct dispatching at the PTO system. The PTO publishes the final sequence indicating the time the truck should wait for the container.

The related process with this interface finalizes with the coordination of the transportation between the CB and the TC that will be in charge of picking up the container from the PTO. The six main associated processes, along with the stakeholders involved, are summarized in Fig. 20.5.

2. **Port Inspections Stage:** The corresponding processes of this stage begin with the arrival and mooring of the vessel at the PTO. Then, the PTO registers the documents of the containers and issues the corresponding receipt that is sent to the BW with the aim to notify the reception of the documentation. Then, the PTO validates the dispatching sequence of the container according to the condition that was manifested previously. This document is sent to the CB to fulfill the requirements so the entry of the cargo is complete. Next, the CB presents all the documentation at the BW to be validated. After that, the CB proceeds to submit all the required documentation to the customs authorities such as the TACT, the DIN, the corresponding payments, selection of inspection, the reception receipt at the BW, and the national cargo manifest. Once cargo has been nationalized, the BW sends the approved documentation to the CB in order to conclude with the payment of the services and confirm the transport carrier that will pick up the container. In parallel, the PTO approves the receipt of the container and proceeds to unload it from the vessel and stack it in the yard according to the dispatching sequence that was published.

   The main processes involved along with the stakeholders that participate in this interface are summarized in Fig. 20.6.

3. **Landside Stage:** The process begins with the coordination between the CB and the TC, when the CB sends the shipment guide and reception record. Then the CB

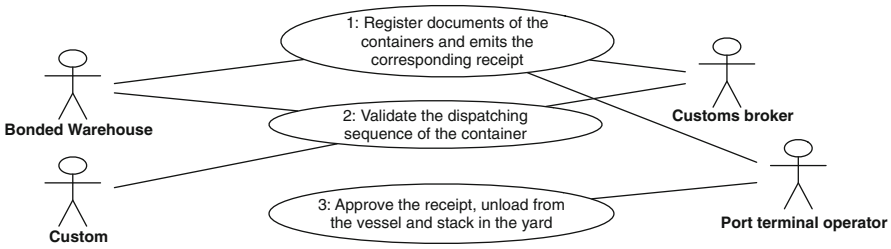


Fig. 20.6 Use cases of Port Inspections Interface

requests the dispatching of the cargo to the BW. Once the BW confirms with the PTO the cargo has been dispatched, they physically send the truck card to the TC in the pre-terminal. This document specifies the approval of the BW that assumes the responsibility of the cargo according to the manifest that will be picked up in the PTO. Once the TC receives the truck card with the schedule time, they arrive to the PTO gate at the time scheduled for the appointment. At the gate, the documents are verified, and the gate personnel proceed to register and approve the entry of the truck. The truck is directed to the stacking position in the yard where the container will be loaded onto the truck. The gate personnel proceeds to emit the exit request, preparing the corresponding documentation (DIN, TACT, and reception record), appending the EIR emitted by the PTO. This document indicates the date and physical status of the container that was dispatched and the potential damages that were observed when it was unloaded from the vessel. Once the truck exits the PTO, it transports the container to the BW where the cargo will be unconsolidated and further dispatched to the consignee facilities. When the container is received at the BW, the corresponding documentation is generated at the gate, and then the container is temporarily stacked while a confirmation for de-stuffing the cargo is received from the consignee. Once this is authorized, cargo is de-stuffed and the empty container is sent to the corresponding empty container depot, where it is checked both physically and documentarily. The process finishes with the closure of the TACT. The main processes involved along with the stakeholders that participate in this interface are summarized in Fig. 20.7.

20.5.2 Stacking and Consolidation Export from Bounded Warehouse Business Process

This logistics business process considers the stuffing of cargo into the container at the facilities of the BW to proceed with the document procedures and obtain the DUS that authorizes the cargo to exit the country. Then, the container is transported to the corresponding PTO in the stacking area where it will be

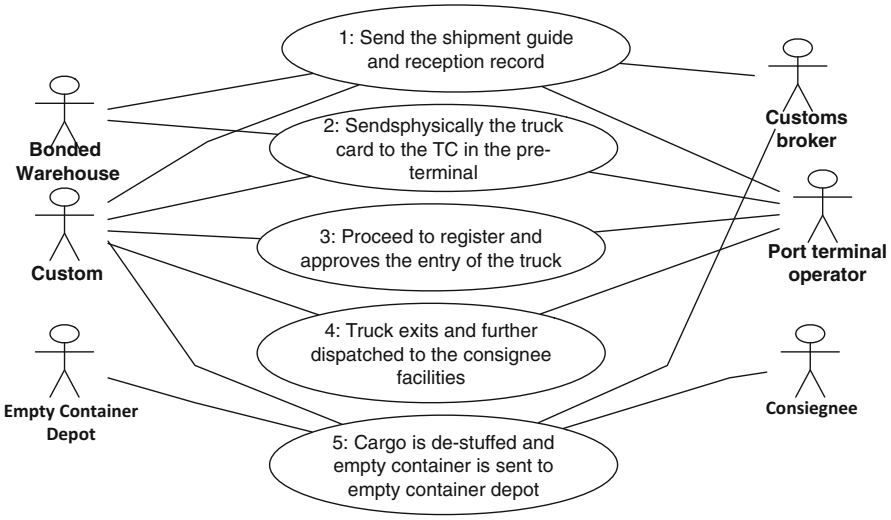


Fig. 20.7 Use cases of Landside Interface

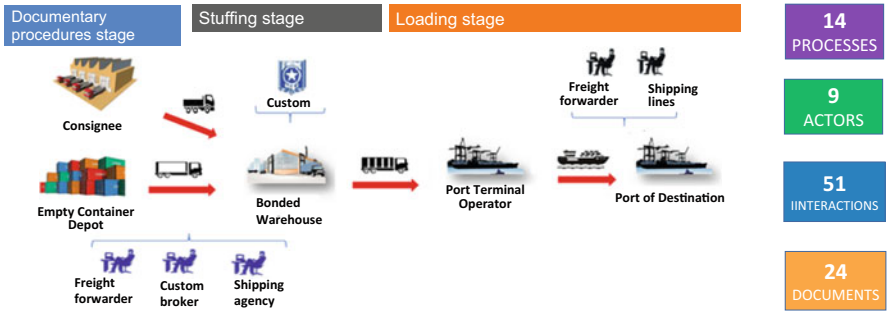


Fig. 20.8 Stacking and consolidation export from bounded warehouse business process

loaded into the corresponding vessel. In total, it considers 14 main processes that generate interactions among the stakeholders and documents involved. Figure 20.8 summarizes the business process separated in three interfaces:

1. **Document Procedures Stage:** This stage begins with the negotiation between the shipper and the FFW, establishing a service contract for the export, indicating the cargo and the port of destination. Then, the FFW begins the required procedures with the ShL with the aim to book a space in the vessel for the container. Once the booking request is confirmed, the ShL sends the corresponding confirmation to the FFW and elaborates the CAL, where all the bookings are indicated. This is sent to the OPT informing the cargo to be handled. In parallel, the FFW proceeds to submit the stuffing services request to the BW, appending

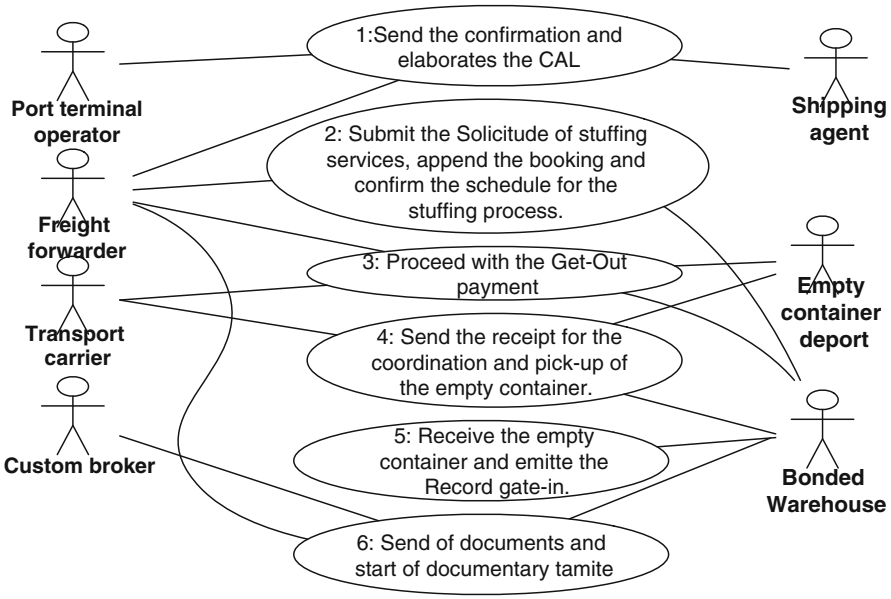


Fig. 20.9 Use cases of Document Procedure Stage

the booking and the cargo instructive. Once the BW accepts the request, they confirm the schedule for the stuffing process.

Then, the FFW proceeds with the get-out payment with the empty container depot, in order to coordinate with the TC the pick-up of the empty container. Once the service is paid, the FFW sends the receipt to the BW for the coordination of the BW with the TC for the pick-up of the empty container at the depot. In this case, the depot emits the gate-out record/depot document and the TC transports the container to the BW facilities. Once the empty container is received at the BW, the gate-in record/BW is emitted, where the physical status and potential damages of the empty container are registered. Then the FFW is notified by the BW of the receipt. The FFW, along with the shipper, coordinates the arrival time of the bulk cargo to the BW. Finally, the related documents of the cargo and its booking and loading instructive are sent to the CB. The CB proceeds to do all the required procedures to get the authorization for the cargo to exit the country (DUS). The main procedures and stakeholders involved are summarized in Fig. 20.9.

2. **Stuffing Stage:** This begins with the coordination of the FFW or shipper with the TC for the transportation of the bulk cargo to the BW. When the cargo arrives to the BW, the BW records the reception and emits the gate-in record/BW. This document is sent to the FFW and the FFW sends this document, along with the dispatching instructive, to the CB so that the required documentation for the DUS is complete. In this document, the CB indicates the destination, specifications

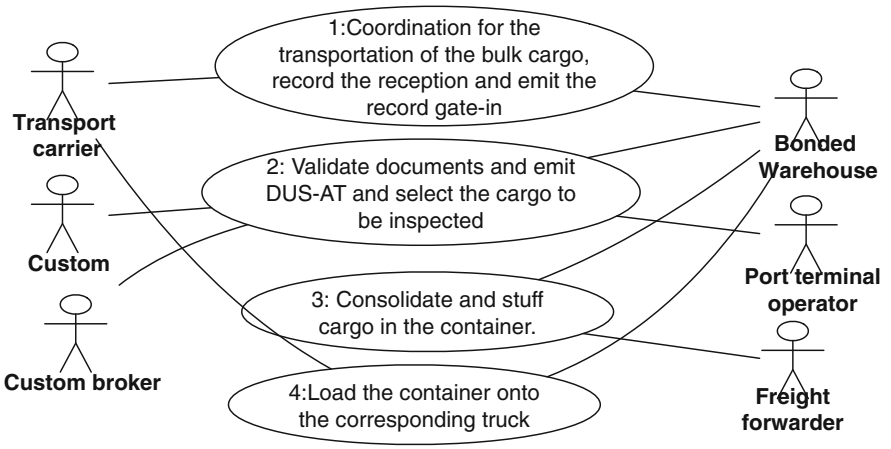


Fig. 20.10 Use cases of Stuffing Stage

of the cargo, and the booking of the vessels. Once the DUS is elaborated and the dispatching instructive is attached, the CB presents all the documents to the customs authorities. Customs proceeds to validate the documents and emits a preliminary DUS (DUS-AT) and the selection of the cargo to be inspected. If no cargo requires inspections, then the cargo is approved to be loaded into the vessel. Then the CB sends the documents required to the BW notifying the approval for the electronic validation with the terminal and initiates the process of stuffing cargo in the container. Once cargo has been stuffed in the container, according to the stuffing-instructive, the documents related to the packing list and the stuffing list are generated. Such documents indicate the status of the cargo that was consolidated and are sent to the FFE notifying that the process has finalized. After this, the BW proceeds to the load weighing and loading of the container onto the corresponding truck (considering that the truck arrives according to what was coordinated with the BW). Once the truck exits the BW, it is directed to the PTO to drop off the container. The main processes involved, and their participants, are illustrated in Fig. 20.10.

3. **Loading Stage:** This process begins with the arrival of the truck with the container to the PTO. At the gate, the container and related information are validated and authorized to enter the PTO. The truck driver presents the DUS-AT, the receipt of the electronic validation, the ticket of the weighing, and the selection for inspection. With this, the personnel at the gate emit the gate-in record/PTO. The PTO sends an approval request for the entering of the cargo to customs and receives the approval (if everything is correct). Once the truck enters the PTO, it is directed to the stacking area at the yard, where the container will be unloaded and temporarily stacked. Then the TC leaves the terminal and notifies the reception of the cargo to the FFW. With this information the FFW confirms the receipt of the cargo to the ShL and emits the gate-in record/PTO to the ShL.

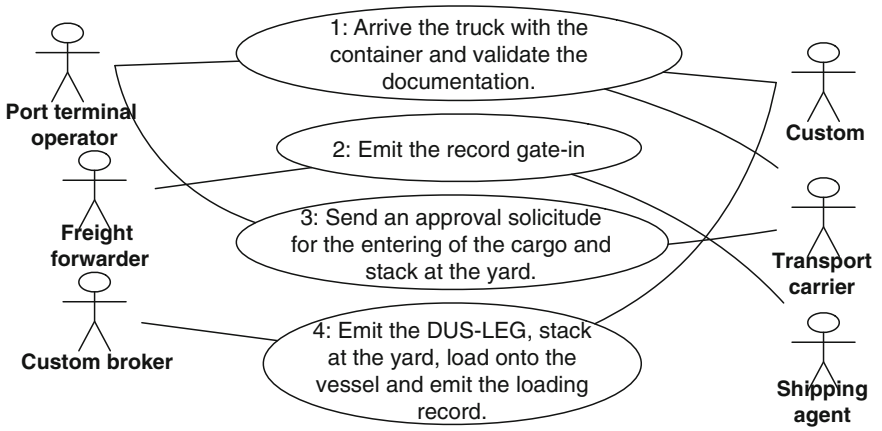


Fig. 20.11 Use cases of Loading Stage

Then the FFW sends these documents to the CB so that they issue the second DUS or what is referred to as the legalized DUS (DUS-LEG) to be presented to the customs authorities, as well as the B/L for its emission in the destination of the cargo. The container, on the other hand, is stacked at the yard while it is loaded onto the corresponding vessel. Once it is loaded onto the vessel, the PTO emits the loading record. The main processes involved, and their participants, are illustrated in Fig. 20.11.

20.6 Determining and Estimating the Logistics Costs of the Selected Logistics Business Processes

In this section, we present an analysis of the costs incurred, on average, at each of the logistics business processes under analysis. The items that compound the total costs were selected according to their frequency in the required documents that are exchanged in the port-logistics business processes under consideration. A number of samples of such documents were collected at the bonded warehouse facilities and estimations were also made based on expert opinions. With such information, it was possible to provide an average value of the costs paid by each item. There were some specific costs that were not considered due to their variability, such as the value of the cargo. This item is extremely diverse and is included in the FOB and CIF incoterms, but very difficult and irrelevant to consider. This is due to the fact that we aim to show the logistics costs rather than the value of the cargo. We also excluded from the analysis the cost of the maritime fleet (ocean freight) as this is very variable and depends on the final destination or port of origin and does not reflect the logistics operations that we aim to analyze.

Table 20.1 Associated charges for the direct and manifested BW-import process

Direct and manifested BW-Import		Valparaiso	Buenos Aires	Buenaventura
Category	Source	Rate		
Emission B/L	Public information (websites)	$ 30	$ 35	$ 40
Chartering B/L	Public information (websites)	$ 35	$ 35	$ 40
Correction B/L	Public information (websites)	$ 35	$ 35	$ 40
Terminal handling charges (THC)	Sampling	$ 80	$ –	$ 120
Insurance	Sampling	$ 75	$ 70	$ 75
Ground transport	Experts opinions	$ 400	$ 450	$ 500
Inspection	Public information (websites)	$ 190	$ 200	$ 250
Gate in—TATC	Sampling	$ 85	$ 95	$ 90
Drayage to bonded warehouse	Public information (websites)	$ 130	$ 120	$ 100
Handling	Public information (websites)	$ 60	$ 90	$ 75,00
Comodato charges	Sampling and experts opinions	$ 90	$ 100	$ 75
Documental procedures	Public information (websites)	$ 15	$ 20	$ 15
International documental procedures	Public information (websites)	$ 15	$ 20	$ 25
Reception of container	Public information (websites)	$ 50	$ 50	$ 50
Gate-in and Gate-out	Public information (websites)	$ 95	$ 95	$ 85
Fee custom broker	Experts opinions	$ 100	$ 120	$ 150
Fee freight forwarder	Experts opinions	$ 100	$ 120	$ 110

Tables 20.1 and 20.2 provide a brief overview of the main costs associated to each logistics business processes, respectively. Due to confidentiality of the data, the average values of such costs are not provided.

Charges presented in previous tables were classified according to its nature and characteristics, as well as according to the related interactions among stakeholders involved. Such categories are described below:

- **B/L Charges**: These charges are related to the emission and change of the B/L document of the import cargo, including such documents at the port of origin of the cargo. For the case of the export business process, such costs are related to the elaboration of the B/L after the cargo has been loaded onto the vessel, including the charges associated to the booking of the cargo and the documents that will be transmitted at the destination of the cargo, as well as the fees that correspond to the export of cargo.
- **Insurance and Transport Charges**: These involve those costs related to the insurance of the cargo to be transported, both import and export cargo, as well as the landside transportation costs to move the cargo from/to the hinterland and the port.
- **Charges Associated with the Container**: These are the costs that guarantee the permission for the temporal admission of the container for the case of import

Table 20.2 Associated charges for the direct and manifested BW-export process

Direct and Manifested BW-Export		Valparaiso	Buenos Aires	Buenaventura
Category	Source	Rate		
EMISSION OF THE EXPORT MATRIX	Public information (websites)	$ –	$ 25	$ 35
Booking fee	Public information (websites)	$ 10	$ 15	$ 15
Release documents	Public information (websites)	$ 30	$ 35	$ 40
Assistance Filing Ok	Public information (websites)	$ 30	$ 25	$ 20
Terminal handling charges (THC)	Sampling	$ 80	$ –	$ 95
Insurance	Sampling	$ 70	$ –	$ 90
Ground transport of bulk cargo	Experts opinions	$ 400	$ 450	$ 500
Inspection	Public information (websites)	$ 190	$ 200	$ 250
Gate-out	Sampling and Experts opinions	$ 70	$ 80	$ 90
Bulk cargo reception	Public information (websites)	$ 200	$ 220	$ 250
Warehousing	Public information (websites)	$ 110	$ 120	$ 130
Handling	Public information (websites)	$ 50	$ 90	$ 75
Consolidation/stuffing cargo	Public information (websites)	$ 350	$ 370	$ 400
Weighted of cargo	Public information (websites)	$ 60	$ 50	$ 40
Drayage to port terminal	Public information (websites)	$ 130	$ 120	$ 100
Fee custom broker	Experts opinions	$ 100	$ 120	$ 150
Fee freight forwarder	Experts opinions	$ 100	$ 120	$ 110

cargo. This is charged as the gate-in of the empty container at the corresponding empty container depot (once the cargo has been de-stuffed). In the case of exports, it considers the gate-out of the empty container from the depot for consolidating the cargo. These rates consider all the charges related to the empty container.

- **Customs Fees and Allowances**: These correspond to the customs fees and related allowances in which the import cargo incurs. For exports, there is no fee charged for the cargo.
- **Port Terminal and Bonded Warehouse Charges**: These correspond to the services charged by the port terminal operator and the BW for the administrative and document procedures of the cargo including, in some cases, the drayage service.
- **Custom Brokers and Freight Forwarders Fees**: These correspond to the fees paid for the services of the custom brokers and freight forwarders, for the administrative and document procedures for the cargo, as well as for ensuring the legal aspects of the cargo and the logistics coordination with the different stakeholders involved in the process.

The cost benchmark that is presented below approximates the cost item by item, based on expert opinions. This analysis aims to present the reader with the differences existing at each port, which are explained by the differences on the technical and regulation conditions of each port, as well as management models. The final costs and fees charged to port users will be determined by the technological and technical capabilities and the ports internal processes. The fees charged to port users are related to the main processes that the port performs on a daily basis, so the total charge to the user depends on the time and amount of the import and export procedures. When governments and regulators simplify the legal and taxing process, ports can simplify their processes and lower fees and rates to their users. The agility and costs that a port represents to the logistics chains in the different countries are decisive to the competitiveness of the country as a whole. Understanding these steps is the key to simplifying and improving the performance of the entire international trade system of a country.

Table 20.3 presents a benchmark considering the cases of the Port of Valparaiso, the Port of Buenaventura, and the Port of Buenos Aires for three types of charges: (1) Shipping line; (2) Port terminal operator and bonded warehouse; and (3) Shipping agents, FFW, and custom brokers. To do so, we considered a sample of data from each port, and then we presented the percentage that such costs represent in each port. Due to the significant differences that exist when cargo is inspected, we present the analysis for each case: with inspections and without inspections. In the table, we highlight the items in boldface that represent the highest percentage, noticing that when there are inspections, the costs incurred per port terminal operator and bonded warehouse are the highest percentage for both ports, while there are significant differences when there are no cargo inspections.

As observed in the previous table and considering the case of imports and the scenario of normal conditions, it is possible to observe that costs invoiced by the port terminal are more expensive (or have more participation) in the case of the Port of Buenos Aires as there are several charges for control operations, fees, and extra cargo handling. The case of the Port of Buenaventura is very similar and also represents a high percentage invoiced by the port terminal that, even so, is not the highest. In comparison, the Port of Valparaiso has a very low percentage of costs charged by the port terminal, and the most significant are related to the shipping companies, which are mainly represented by the distances to transport cargo.

In the case of an import operation with inspections, the most significant costs are represented by those charged by the port terminal. This is explained as the consignee has to pay for the related services of inspection which significantly increase the fees charged. However, we can notice that the Port of Buenos Aires has the highest percentage, followed by Buenaventura. Given that in the Port of Valparaiso there is a significant participation of the bonded warehouses in the import business processes, costs have been minimized in comparison to the other ports under analysis.

On the other hand, when we analyze the case of export operations and the normal conditions scenario, there are no significant differences observed in the three ports. However, the Port of Valparaiso has the more significant percentage of charges invoiced by the port terminal. In the case of operations with inspections, the most

Table 20.3 Benchmark of import and export cases (with and without inspections) between the Ports of Valparaiso, Buenos Aires, and Buenaventura

	Valparaíso	Buenos Aires	Buenaventura
Import normal			
Subtotal invoiced by shipping company	**0.57%**	0.37%	**0.43%**
Subtotal invoiced by port terminal	0.16%	**0.42%**	0.39%
Subtotal invoiced by shipping agents and bonded warehouse	0.26%	0.21%	0.18%
Export-normal			
Subtotal invoiced by shipping company	0.34%	**0.42%**	**0.37%**
Subtotal invoiced by port terminal	**0.45%**	0.31%	0.32%
Subtotal invoiced by shipping agents and bonded warehouse	0.22%	0.28%	0.31%
Impo with inspection			
Subtotal invoiced by shipping company	0.37%	0.24%	0.32%
Subtotal invoiced by port terminal	**0.46%**	**0.63%**	**0.54%**
Subtotal invoiced by shipping agents and bonded warehouse	0.17%	0.14%	0.14%
Export with inspection			
Subtotal invoiced by shipping company	0.24%	0.23%	0.26%
Subtotal invoiced by port terminal	**0.60%**	**0.62%**	**0.52%**
Subtotal invoiced by shipping agents and bonded warehouse	0.16%	0.15%	0.22%

significant charges are represented by the costs invoiced by the port terminal. This is very similar as the case of import operations.

As a final takeaway of the analysis presented herein, we can point out that port-logistics business processes, with the participation of bonded warehouses, provide a competitive advantage to the import operations, mainly of a port. In the case of the Ports of Buenaventura and Buenos Aires, bonded warehouses have a limited participation, while in the Port of Valparaiso, they can represent a significant percentage, which impacts the participation of total costs as it was categorized in this chapter.

20.7 Conclusions

We have presented a categorization of the different port-logistics business processes, considering how unique the operations in Chile are. This categorization can be adapted and extended for other ports in different countries. The most significant contribution relies on the application of a standard methodology such as the UML that uses cases to model the interactions among different stakeholders, as well as the detailed description of the documents required. The analysis was focused on two

business processes in which a bonded warehouse participates. Further research can consider the modeling and analysis of the rest of the business processes and proceed with the standardization of such procedures. As a result of the work presented herein, we could observe the importance of understanding the very complex business processes of the ports that operate under a multi-stakeholder environment, and the need to cooperate to deepen the understanding of such processes is evident. For the case presented in this chapter, during the process modeling and gathering of information, we observed a lot of opportunities to improve operations at ports. A collaborative scheme between the industry and the academy is clearly a way to achieve better understanding of such processes with the support of academics and students who may have the knowledge concerning modeling and analysis tools.

In this work, a second contribution was a categorization of the different logistics costs involved in such business processes, providing a detailed description of the different charges. In order to compare and determine which are the most significant categories of costs, we provide a benchmark (as a percentage) for the Ports of Valparaiso, Buenaventura, and Buenos Aires. Results show that the charges invoiced by the port terminal are higher in the case of the Ports of Valparaiso and Buenos Aires under the scenario of normal conditions (no inspections of cargo) for the import business processes. One important element to highlight is that, in this case, the bonded warehouses have a significant participation on the process, in comparison to the other ports. So, we can clearly observe significant benefits for the total logistics costs. Another important result is that total logistics costs are significantly higher when inspections are required, and under this scenario, the charges invoiced by the container terminal are very significant for the three ports. However, we observe that for the case of Valparaiso, these charges have the least participation.

In the case of exports operations, no significant differences were observed, especially as in these cases, few inspections operations are demanded and are mainly document inspections, and not physical, as in the case of imports.

One important element to highlight is that gathering information related to logistics costs is not an easy task, as information is confidential and subject to business and marketing conditions. So, the data provided in this chapter is based on estimates of expert opinions and public information available at each port. As a recommendation for governments and public policy decision makers, we believe that there is a need to have logistics observatories that collect this type of information systematically and provide benchmark reports to the industry. Furthermore, governments should provide a collaborative environment in the port industry, fostering the conformation of port-logistics communities as a public and private partnership and governance scheme.

As further research we propose to consider different case studies of logistics chains and determine the total landed costs incurred for cargo with more detail, under different circuits (with the participation or not of bonded warehouses, for instance), and perform a comparative analysis for different ports in the region. This requires the consideration of a different scope, operations, management, and equipment of the ports in order to provide a comparative analysis that may result

useful for port managers to understand what the rest of the ports are currently doing and their relative performance.

Acknowledgements We thank the undergraduate students Francisco Aldunate and Arnaldo Papaprieto for their work and dedication in their thesis that was the basis of this chapter. We also thank Luis M. Ascencio, the Technical Coordinator of the Program Network of Digital and Collaborative Network of Ports in Latin America and the Caribbean, for his valuable contributions in the validation of the results presented herein.

References

Avelar-Sosa L, García-Alcaraz JL, Cedillo-Campos MG, Adarme-Jaimes W (2014) Effects of regional infrastructure and offered services in the supply chains performance: case ciudad juarez. Dyna 81(186):208–217

Dou Z, Li H (2015) Optimization of the border port logistics and the key-factors recognition based-on HLA/SysML. J Coast Res 73(sp1):104–107

Elbert R, Pontow H, Benlian A (2017) The role of inter-organizational information systems in maritime transport chains. Electron Mark 27(2):157–173

Feng M, Mangan J, Lalwani C (2012) Comparing port performance: Western European versus Eastern Asian ports. Int J Phys Distrib Logist Manage 42(5):490–512

Heilig L, Lalla-Ruiz E, Voß S (2017) Digital transformation in maritime ports: analysis and a game theoretic framework. Econ Res Electron Netw 18:1–28

Ismail NW, Mahyideen JM (2015) The impact of infrastructure on trade and economic growth in selected economies in Asia. Working Paper Series 553, Asian Development Bank Institute, Manila

Jing Z, Jia-Wei X (2010) Study of Ningbo Zhoushan port logistics competitiveness based on factor analysis and cluster analysis. In: 2010 third international conference on information and computing (ICIC), vol 3. IEEE, New York, pp 123–126

Márquez-Ramos L, Martínez Zarzoso I, García EP, Wilmsmeier G (2007a) Determinantes de los costes de transporte marítimos. el caso de las exportaciones españolas. Informacion comercial espanola- Monthly Edition 834:79

Márquez-Ramos L, Martínez-Zarzoso I, Valenciaport F, Pérez-García E, Wilmsmeier G (2007b) Maritime knowledge network. Transporte marítimo: Costes de transporte y conectividad en el comercio exterior español. In: González-Laxe, Sánchez, Lecciones de Economía Marítima, Spain, Netbiblo, pp 105–144

Martínez Zarzoso I, Wilmsmeier G (2010) International transport costs and the margins of intra-Latin American maritime trade. Aussenwirtschaft 65(1):49

Martinez Zarzoso I, Wilmsmeier G et al (2011) Trade responses to freight rates: the case of intra latiamerican maritime trade. Eur Transp (Trasporti Europei) 2011(48):24–46

Nguyen H-O, Chin A, Tongzon J, Bandara M (2016) Analysis of strategic pricing in the port sector: the network approach. Marit Econ Logist 18(3):264–281

Octavio DN, Sánchez Ricardo J (2006) Indicadores de productividad para la industria portuaria. In: Aplicación de América Latina y el Caribe. CEPAL. Serie Recursos Naturales e Infraestructura, vol 112

Orlic Protega A, Rogic K, Vrdoljak J (2011) Logistic approaches to port management system. In: Annals of DAAAM & proceedings

Pérez-Salas G, González-Ramírez RG, Cedillo-Campos MG (2015) A framework to evaluate over-costs in natural resources logistics chains. Dyna 82(191):85–92

Sánchez RJ, Hoffmann J, Micco A, Pizzolitto GV, Sgut M, Wilmsmeier G (2003) Port efficiency and international trade: port efficiency as a determinant of maritime transport costs. Marit Econ Logist 5(2):199–218

Song D-W, Panayides PM (2008) Global supply chain and port/terminal: integration and competitiveness. Marit Policy Manage 35(1):73–87

Suárez-Alemán A, Serebrisky T, Ponce De León O (2018) Port competition in Latin America and the Caribbean: the role of concessions and competition policy. Marit Policy Manag. 45(5):665–683

Sujeta L, Navickas V (2014) The impact of port-logistics systems on a country's competitiveness (case of small countries). Econ Manage 19(1):44–53

UNCTAD (2017) Review of maritime transport, United Nations Conference on Trade and Development, UNCTAD/RMT/2017

Carla Vairetti She is a Professor and Researcher at the Universidad de Los Andes in Chile. She holds a bachelor's degree in Computer Science Engineering from the Universidad Nacional de La Plata in Argentina, a Master in Information Systems in Chile, and one PhD in Information and Telecommunication Technology in Trento—Italy and a second PhD in Engineering Sciences in Santiago—Chile.

She decided to start her professional career in Computer Engineering, since she was very interested in the software engineering tools that may have application in the industry. After completing her bachelor's degree, she decided to enroll immediately in a doctorate's program, a period in which she also worked as a teaching assistant and consulting projects in Chile. While she was taking her first doctorate she decided to take the Master examination obtaining the degree with a research related to web services composition.

She decided to pursue a scientific career in the area of business process and data mining research. During her doctoral studies, she had the opportunity to do a second PhD at the University of Trento in Italy under the supervision of Professor Fabio Casati. After completing her doctoral studies in Italy and Chile, she has worked as a professor and researcher at the University of Los Andes, Chile, which is her current assignment.

Her main research areas are: Business Process Management (BPM): Modeling and Process Simulation, Development of Technology Platforms and Big: Data Mining, Machine Learning; and Business Analytics. She has been working in several applied research projects with ports in Chile and is the author of scientific articles in journals.

For two years, she has participated in the organization of a women's event in Chile in order to attract students who learn about women leaders and gender issues in computing and technology. The synergy achieved among the participants generates actions and helps the different causes that work in isolation to join together to achieve greater goals to work for gender equality.

Rosa G. González-Ramírez She is a Professor and Researcher at the Universidad de Los Andes in Chile. She holds a bachelor's degree in Industrial Engineering from the Technologic Institute of Morelia, a Master in Industrial Engineering from Arizona State University and a Master in Quality Systems and Productivity, and a PhD in Engineering Sciences from Monterrey Tech in Mexico.

She decided to start her professional career in Industrial Engineering, as she was very interested in quantitative methodologies founded in mathematics and sciences, but with an application to the industries. After completing her Bachelor degree, she decided to immediately enroll into a Master program, period in which she also worked as teaching assistant and consulting projects at Monterrey Tech. She decided to pursue a scientific career in the area of operations research and logistics, and during her doctoral studies, she had the opportunity to do an academic exchange year in Arizona State University, obtaining a Master degree. After completing her doctoral studies in Mexico, she obtained a postdoctoral position in Chile. She has been working as professor and researcher, six years at the Catholic University of Valparaiso and 3 years at Universidad de Los Andes Chile, which is her current adscription.

Her research areas are logistics and transport of cargo, maritime shipping and port operations, and supply chain management and optimization. She has been working in several applied research projects with ports in Chile and has authored various scientific articles in journals such as Annals of Operations Research, International Journal of Production Research, Journal of Cleaner Production, International Journal of Shipping and Transport Logistics, and Transport Research Part E. She is a member of the Program "Latin American and Caribbean Network of Digital and Collaborative Ports" and is currently serving as a coordinator of the Scientific and Innovation Committee and one of the Organizers of the Regional Meeting of Port Logistics Communities that is being organized every year and gathers participants from the industry, government, and academy.

Luisa Fernanda Spaggiari She is a consultant and researcher for Empresa Multimodal SAS in Colombia, and assistant professor and researcher at Politécnico Grancolombiano in Colombia. She holds a bachelor's degree in Industrial Engineering from the Universidad Nacional de Colombia, a Graduate Certificate in Supply Chain Management at MIT's Center for Transportation and Logistics, and is currently coursing a Msc. in Industrial Engineering.

Her interest in the area of engineering and sciences started since her Bachelor degree, which has motivated her to work in the research areas of logistics and transport of cargo, maritime shipping and port operations, port competitiveness, and benchmarking.

She has been working in several research and consulting projects with ports, logistics, and transportation companies, as well as government agencies and regulators in Colombia.

She has participated in the consulting team that conformed the Port Logistics Community of the Port of Buenaventura, project that evolved under the scope of the Program "Latin American and Caribbean Network of Digital and Collaborative Ports" that is promoted by SELA and CAF. Furthermore, she has been an active collaborator of the program and participated in the Second Regional Meeting that was held in Cartagena, Colombia, in 2017.

Alejandra Gómez Padilla She is an Industrial Engineer from the Instituto Tecnológico y de Estudios Superiores de Occidente (Mexico), holds a Master's degree in Applied Sciences from the École Polytechnique de Montréal (Canada), and a PhD in industrial engineering from the Institut National Polytechnique de Grenoble (France).

She was first interested in Industrial Engineering because of the diversity of optimization methodologies and techniques that may be applied to improve the process in a wide variety of organizations. After obtaining her bachelor's degree, she worked as product engineer and process engineer for a company in electronics. This working experience interested her even more on optimization models to support decision making in logistics and supply chain management, so she decided to pursue on this fields a master's degree and a PhD immediately after.

Coping up with her research interests, she joined University of Guadalajara. As professor and researcher at Universidad de Guadalajara, she has lead research projects on optimization, supply chain management, and logistics, which are her research areas. She also leads an academic group that researches on systems analysis and optimization, achieving that the research background of the group positively influences the undergraduate programs of the university. She has authored several scientific articles in journals, book chapters, and participated in important international conferences.

Chapter 21
Using Simulation to Improve Container Depot Operations

Jimena Pascual and Alice E. Smith

Contents

21.1 Introduction 487
21.2 Literature Review on Simulation Projects 489
21.3 The Project 490
21.3.1 Overview 490
21.3.2 Related Technical Literature 493
21.3.3 Goals and Constraints 494
21.3.4 Project Team and Sponsorship 494
21.3.5 Project Team Interaction Modes 496
21.3.6 Data and Modeling 496
21.3.7 Verification and Validation of the Simulation Model 499
21.3.8 Using the Simulation Model for Policy Analysis 500
21.3.9 Results and Discussion 501
21.4 Retrospective Examination 504
21.4.1 Technical Lessons Learned 504
21.4.2 Project Participant Personal Growth 506
21.4.3 Tangible Outcomes 508
References 510

21.1 Introduction

This chapter will discuss a discrete-event simulation project in terms of both technical development and outcomes, and the effects on the project members involved in terms of lessons learned and professional growth. A holistic dissection

J. Pascual
School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso (PUCV), Valparaíso, Chile
e-mail: jimena.pascual@pucv.cl

A. E. Smith (✉)
Department of Industrial and Systems Engineering, Auburn University, Auburn, AL, USA
e-mail: smithae@auburn.edu

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_21

of these fundamental aspects of a typical simulation project is rarely seen though almost always experienced. The chapter will also focus on the international aspects of the project—a project team that spans two continents, two languages, and funding from both countries at different times. These global aspects are often encountered in contemporary simulation projects due to the international nature of organizations. This chapter intends to expressly link the technical project happenings and achievements with the qualitative and personal effects. Parts of this project have previously been presented in Hidalgo et al. (2016, 2017), and Pascual et al. (2016).

We do this through a case study recently experienced by the authors. This is the development of a practical simulation model to enable analysis and optimization of operations at an Empty Container Depot (ECD) located in Chile. This project is a suitable case study because the simulation model is relatively straightforward and is intended for decision-making support by an industrial user. It is especially applicable to the aims of the chapter because the project team includes faculty members at three universities in two countries, graduate (master's and Ph.D. level) and undergraduate students at two universities in two countries, and an industrial partner company with various involved employees including upper management, supervisory management, technical staff, and workers. Moreover, the project involved two funding agencies in different countries at different stages of the project and the project has spanned more than 4 years already.

The overall steps of this project (and those of most simulation projects) are:

1. Specify project goals and hypothesized outcomes.
2. Decide expected technical approach.
3. Gather, analyze, and interpret data.
4. Understand the system and its processes in detail.
5. Choose the level of complexity and accuracy in the simulation model.
6. Build, verify, and validate an initial model.
7. Consider abandonment or enhancement of the initial model.
8. Decide when the model building is complete.
9. Design experiments to fit the objectives of the project.
10. Conduct the experiments including identifying a proper warm-up time and a proper duration of the simulation (in time steps).
11. Analyze the results both statistically and also qualitatively.
12. With the outcomes of step 11 in mind, modify the simulation model if necessary. Then, design and conduct more experiments if warranted.
13. Choose how to portray results both tabular and graphical. Consider animations and videos.
14. Share the results and recommendations with the project members including the sponsoring/participating organization and other stakeholders.
15. Consider the responses and questions of the stakeholders from the organization, modify the simulation and/or conduct additional experiments, as needed.
16. Document the simulation model, the results, and the recommendations for all involved.
17. Disseminate publicly through published papers, presentations, online code, etc. as desired.

In our project, the principal investigators had plenty of time to invest in the long-term study of the issues, but the students did not as they needed to briskly complete their degrees. Therefore, we have used an "increasing complexity" approach for the four cohorts of students working on the project as detailed below. With the first cohort we worked on understanding the problem and the data available. We then developed a simplistic simulation model. In the second stage, the cohort's activities centered on gathering data from different sources, understanding operational decisions and procedures, building and validating a detailed simulation model of the system, and proposing an experimental design to analyze alternative policies and configurations through simulation. The third cohort worked on improving the input model, building an efficient logic model, and establishing a more complex design of experiments. The fourth cohort is currently working on redefining the scope of the system to include dependencies to other systems such as the seaport, reporting in greater detail the systemic analysis of the problem, expanding the design of experiments to include additional factors and levels, and considering the layout design of the ECD. We also want to test our approach on another ECD located in a different city.

21.2 Literature Review on Simulation Projects

A great many papers have been published on simulation models and their applications, to say nothing of the body of literature on the theoretic and empirical research to advance the science of simulation. However, there is a dearth of literature on project planning and project outcomes in a holistic sense for discrete-event simulation. Here, we consider the most related literature.

The simulation modeling process begins by understanding the character and scope of the problem to be solved. Pidd (2007) describes some systematic approaches used in simulation practice to diagnose and structure problems such that the right issues are addressed. A critical examination of a problem includes asking what, why, when, how, where, and who. What are people concerned with? Why are these issues important? Where/when is the problem happening? To whom? When will a solution be needed? How should the problem be solved?

Harmon (2014) presents formal tools to evaluate the nature of business process problems and help managers—and the modeling team—identify causes and effects of the problem at different hierarchy levels within the system. To describe this hierarchy, one needs to identify processes and sub-processes and their relationships and emergent properties, as well as the context and environment in which the general system operates. This systems thinking exercise, that includes consideration of points of view of different stakeholders invested in the problem, is relevant to defining the borders of the system to be modeled, the level of complexity to be represented, and the objectives of the simulation project. To figure out the nature of the problem some tools include gap model, cause-problem-consequence map, high-level process diagram, cause-effect (Ishikawa) diagram, and process scoping

diagram (Harmon 2014). Sturrock (2017) also elaborates on the importance of first determining who the stakeholders are and what their measure of success is, and describing the shared understanding of the system before starting the modeling process.

To develop a shared mental or conceptual representation of the system—including their subsystems, their elements, the relevant interactions, and the objectives they wish to evaluate—the modeling team must distinguish between the real system and the system representation. Robinson (2008a, b) presents a framework for developing conceptual models in simulation, discusses the requirements of such models and how to document them, and addresses their importance to simulation practice. A conceptual model involves the process of simplifying reality by abstracting elements of the real system to include in the proposed simulation model or virtual system. A conceptual model always exists (implicit in the mind of the modeler) and it is made explicit through diagrams, coding, and documentation. The conceptual model evolves over time from the project start and may continue to be analyzed and modified after the simulation results have been presented. The main components of a conceptual model are, according to Robinson (2017): the objectives (the purpose for the model and project), inputs (experimental factors to be studied), outputs (system performance measures to be evaluated), content (model logic scope and detail expressed by its elements and interrelations), assumptions about the system, and simplifications of the model. Case examples for conceptual modeling and simulation may be found in Furian et al. (2015), Chwif et al. (2013), and Gunal and Pidd (2007).

Sturrock (2017) describes characteristics of a well-managed simulation project. A functional specification is key to clarifying the project and its deliverables, including aspects such as the level of detail to be considered, the data requirements and collection, any assumptions on the logic of the system, and a specification of responsibilities. Some flexibility is needed as "simulation is often a process of discovery" (Sturrock 2017); however, considerations of scope and level of detail must be pondered against meeting the decision deadline with valuable insight from simulation output. One important insight from practice is the need for following an iterative process of model building, verifying, and validating before any experimentation begins.

21.3 The Project

21.3.1 Overview

In this section of the chapter, we present an overview of the current handling operations at the ECD that provides services for different shipping lines operating with the port of Valparaíso, Chile. The port of Valparaiso is the second largest port in Chile in terms of containerized cargo per year (measured in 20′ Equivalent

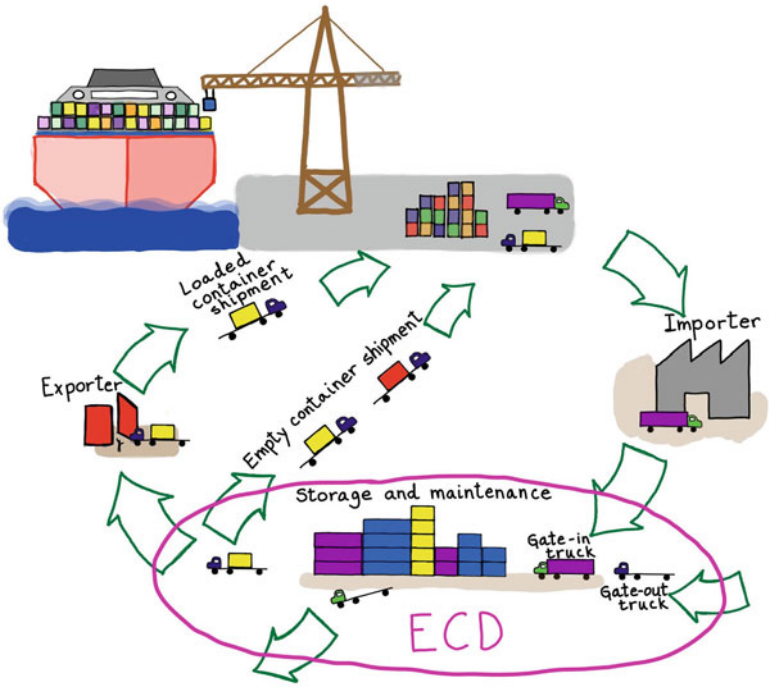


Fig. 21.1 Schematic of port operations and their relations to empty container depots

Units, TEUs), and is ranked 16th in the Latin America and Caribbean region by TEUs transferred according to the Economic Commission for Latin America and the Caribbean ECLAC (2017). The ECD is located in the inter-port area of Placilla (in the Valparaíso outskirts). A schematic of how the ECD relates to the operations at the port is shown in Fig. 21.1. Upon arrival of a vessel to the port, its containers are unloaded and stacked (either at the port yard or at a nearby bonded warehouse) to undergo several processes, such as customs clearance and other handling activities. Trucks then collect the containers to bring them to their destination (importer). The consignee or importer unloads the contents of the container and later the empty container is taken to an ECD (gate-in truck) for storage under a shipping line contract. When a shipping line requires a container for an export operation, a gate-out truck picks up the desired container at the ECD and transports it to the exporter (or shipper) to be filled. The loaded container is then trucked to the port where it later will be transferred to a vessel. Shipping lines also require moving containers for inventory repositioning operations around the world, and regularly request a large number of empty containers to be trucked from the ECD directly to the port for shipping (we call this a massive run).

ECDs are becoming more popular because of the congested nature of many ports and high value of the land in the immediate area around ports. Placilla is typical for ECDs in that the land is much more affordable and plentiful than near the port.

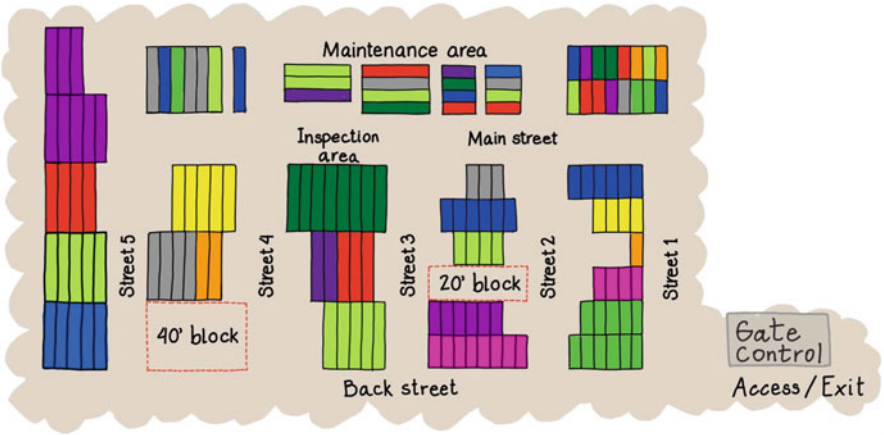


Fig. 21.2 Layout of the Empty Container Depot in Placilla, Chile (not to scale)

However, it also offers good proximity to major roads and is about a 20- to 30-min drive to the port.

The Placilla depot was selected because it is typical of ECDs and the company provided full information access and a true partnership with the project. The depot is divided into two primary areas: reefer (refrigerated containers) and dry (non-refrigerated containers). We considered only the latter area because it represents most of the activity. However, our approach and much of our analysis, is applicable to reefer-container operations as well. Figure 21.2 presents an overview of the ECD, where we can observe the dry area, as well as the gate access where the trucks enter and exit to pick up (gate-out processes) and deliver (gate-in process) containers. This facility stores roughly 2000 containers at any given time when it is medium busy (reefer-container operations are quite seasonal in Chile).

As indicated in the figure, there is an inspection zone located at the main street where all arriving containers are inspected. For the inspection process, containers are divided by size (40′ or 20′) and segregated into Operational or Damaged. Containers are further classified into three conditions: near perfect, acceptable, or marginally usable. There is a maintenance area where damaged containers may be repaired to become operational. The wide internal streets allow for reachstacker cranes (or toplifter cranes) to operate and handle either type of container (20′ or 40′), except for street No. 1. Street No. 1 is one way and used by the gate-in trucks to access the dry container inspection zone. The back street is used as the access lane for the gate-out trucks when retrieving a container from a particular block, and as an exit lane for all trucks. Different sections are organized into container blocks, each assigned to a particular customer according to size and class. Each block has a capacity of eight containers deep (row) and up to seven containers high (tier). A BAROTI (bay, row, and tier) notation gives the bay, row, and tier specification for each container so it may be located.

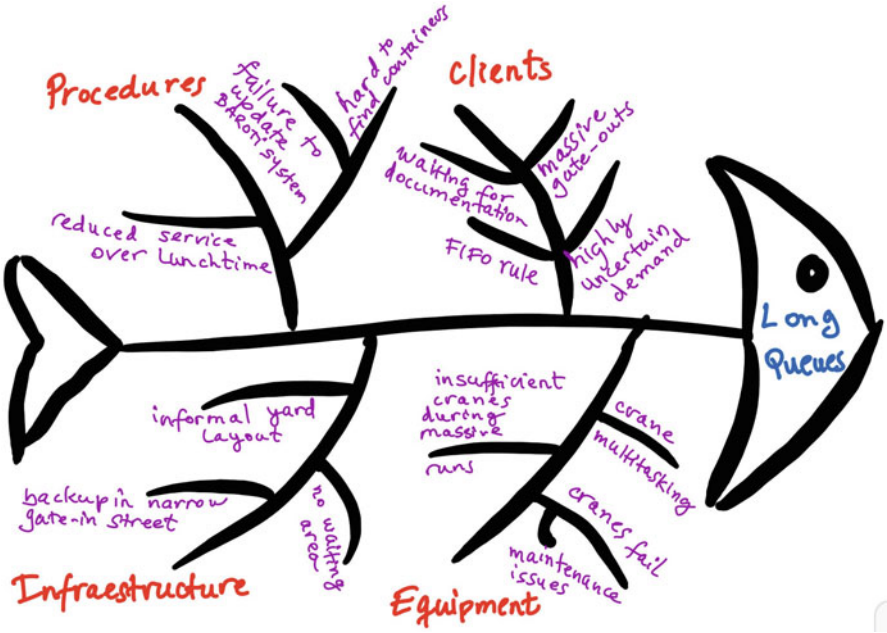


Fig. 21.3 Ishikawa (cause and effect) diagram of the problem

The problem detected by the ECD management, based on complaints by truck drivers and customers, was that the wait time for service at gate control was excessive, and that the access queue was often so long that spanned several streets adjacent to the depot. An Ishikawa cause-effect diagram was created to better understand this problem (Fig. 21.3).

21.3.2 Related Technical Literature

Carlo et al. (2014) provide an overview of port storage yard operations. They discuss the following main decision problems that occur in the storage yard operations: (1) yard design, (2) storage space assignment for containers, (3) dispatching and routing of material handling equipment to serve container storage and retrieval processes, and (4) optimizing the remarshalling of containers.

At the tactical and operational levels, container stacking policies and the storage space allocation problem (including the pre-marshaling problem and the blocks-relocation problem) have been addressed previously such as Kim and Kim (1999), Kang et al. (2006), Lee and Hsu (2007), Park et al. (2011), and Chen and Lu (2012). At the strategic level, yard design problems have been also addressed in the literature but not as often as container stacking strategies and policies. Yard design is an influential factor in the productivity of container handling operations (Kim et

al. 2008), thus requiring strategic decisions in yard layout and outline, and in the number and placement of aisles (surrounding container blocks). Some authors have focused on the analysis of layout design and strategic decisions related to equipment acquisition at port terminals including Kim et al. (2008), Wiese et al. (2011), Ranau (2011), Lee and Kim (2013), Kemme (2012), Wiese et al. (2013), and Taner et al. (2014). However, there is very little published work specifically directed at ECDs.

21.3.3 Goals and Constraints

The aim of the project is to identify recommendations for improved stacking policies and, potentially, the redesign of its yard layout. It is important to mention that stacking operations at this ECD are strongly influenced by the marketing strategy of the depot and the contracts in place with shipping lines. These contracts include a non-fee storage period, which creates a demand for a FIFO (First In, First Out) policy for the dispatching of empty containers.

The aspects we considered are (1) the assignment of block space for each customer of the depot (both shipping lines and leasing companies), (2) the traffic flow of trucks, and (3) stacking and retrieving policies. In the future, we plan to also consider (4) the layout design of the depot in terms of the dimensions of the stacking areas. To achieve all of this, we chose to develop a discrete-event simulation model (coded in Simio, a commercial software—see Smith et al. (2014)) with an interfaced database to evaluate different policies and configurations. The performance measures considered are (a) expected truck waiting times, (b) yard crane utilization, and (c) truck turnaround times. A major aspect is that different customers need their containers to be stacked in assigned and exclusive blocks, and that their contractual terms differ.

21.3.4 Project Team and Sponsorship

The project team was led by three faculty members (all women engineers)—the two authors and Dr. Rosa Gonzalez, who is an author of another chapter in this volume. Dr. Gonzalez and Dr. Pascual were colleagues at Pontifical Catholic University of Valparaiso, PUCV (Chile) while Dr. Smith is a faculty member at Auburn University (USA). Dr. Gonzalez left early in the project to work at another university located in Santiago, Chile. She has still been involved in this ECD project but at a lesser magnitude. At Auburn University, a student worked on the project for about a year before abandoning his Ph.D. studies after getting his master's degree. The position at Auburn was not vacant long, though, and another Ph.D. student was assigned to the project. At PUCV, two master's level students worked on the project for 2 years and then both graduated and became employed in industry. Taking their places were an undergraduate research assistant and a master's student. This master's student has

recently graduated so a new cohort of PUCV students is underway. This turnover is not unusual in a multi-year academic-based project. However, the discontinuity along with the different skill sets and levels of expertise of the students caused project slowdowns and even restarts.

In the partner organization, personnel changes were also frequent, both at the technical level and at the managerial level. Our business partner, a conglomerate of companies, provides diverse services in the port logistics system, including shipping agency services, air representation, terminal operations, empty container storage and maintenance, container leasing and sales, etc. Our initial contact at the conglomerate level was a PUCV alumnus who had detected the problem while working on his undergraduate thesis with Dr. Pascual and has subsequently been hired full time by the company. He introduced us to the ECD's general manager and arranged all of the initial visits to the depot. When he moved on to a different post within the conglomerate, we maintained contact with the depot general manager and we interacted often with the yard manager and logistics analyst, especially when building the conceptual model and all throughout the data gathering process. As the management at the conglomerate level changed, we have presented aspects of the project to them on several occasions, and we have now established a working partnership to develop other projects with our academic team.

The project was initially funded by the Comisión Nacional de Investigación Científica y Tecnológica, CONICYT, in Chile under a program that pays international travel and local subsistence for foreign researchers to come for short-term visits to Chile to conduct research. This program is named Programa de Cooperación Internacional, Concurso Nacional de Atracción de Capital Humano Avanzado del Extranjero, abbreviated MEC. It is a competitive program requiring a proposal. The MEC grant was awarded to Dr. Pascual and provided funding for Dr. Smith to visit for 2 consecutive years, staying for 1 month each year. PUCV provided internal funding for the students who worked on the project. At Auburn University, the students were funded as teaching assistants. Drs. Gonzalez, Pascual, and Smith were not compensated for their time spent on the project.

After the 2 years of MEC funding, Drs. Pascual and Smith were successful in securing funding from the Fulbright Scholar Program. This was done through a Senior Fulbright Fellowship for Dr. Smith to spend a semester at PUCV teaching a graduate class and working on this research project. The Fulbright provided for the international travel and gave a monthly stipend to Dr. Smith while PUCV provided a monthly housing allowance. After the Fulbright completion, the funding for the project became more ad hoc with each side looking opportunistically for ways to provide travel and student support.

During this time, PUCV was able to sponsor the Ph.D. student from Auburn University to reside at PUCV for about 6 weeks with some modest match from Auburn University. This residency, while furthering the technical and development aspects of the project, also provided a life changing experience to the student involved as will be discussed later in the chapter. Another residency, this time by a PUCV student to be resident at Auburn University, is planned for 2019.

21.3.5 Project Team Interaction Modes

While we outlined the residency time periods, the team was able to spend together in Valparaiso above, for the most part, the team was located many, many thousands of miles apart. There was also a language issue as Dr. Smith and her students spoke virtually no Spanish. Fortunately, the Chilean team members all possessed English skills ranging from excellent to good. Dr. Smith has a basic reading capability in Spanish allowing her to understand the documents and papers produced in that language.

The team's primary interaction modes were by video conferencing over the Internet. These were scheduled with each team member generally congregating in one office to video conference with the other country. While connections were not always reliable, this method worked surprisingly well. And, it incurred no cost.

Because of their location, PUCV had almost all of the interaction with the partner ECD company while the US participants visited as they could. Even on those visits because of the language barrier, the US team members were not fully participating.

Each faculty leader met with her student(s) as desired, usually at least weekly, to discuss project progress and identify issues and potential road blocks. As most of the students integrated this project with their degrees (three PUCV students and their master theses and the Auburn University Ph.D. student and his dissertation), the students have had an intrinsic motivation to move forward on the project.

Dr. Pascual interacted regularly with the management of the partner company through meetings at the partner's location. The depot general manager was readily available and enthusiastic about this project, but as often happens, it is hard to obtain detailed information about operations with just a few meetings. We met with him or other operations personnel once or twice a month for the first semester and one of the master's students was accepted as a summer intern at the ECD for a month. This allowed us to construct a rich model of the process and was helpful in defining the data collection process that followed. During the subsequent year, we mostly interacted through our students who weekly requested system state data from the logistics analyst. There were also occasional visits to the depot. These formal visits also included the US part of the team and allowed the complete team to present progress updates and to discuss operational policies to be included in the experimental design. The masters' students gained confidence throughout this period to present findings and ask questions of company representatives as well as to interact in English with Dr. Smith and her students and to present preliminary results at international conferences.

21.3.6 Data and Modeling

To represent the operations at the depot, data was collected from three sources. First, we considered a study already performed by the ECD to gather arrival process

data. Second, we collected our own field data on different process times, and, third, the depot provided access to data in their ERP database. We made detailed stochastic models of the following functions: (a) gate-in, (b) gate-out, (c) massive runs (when a customer requests a large number of containers to be shipped empty for repositioning), and (d) remarshaling of containers (reorganization of containers in a block to facilitate retrieval or to increase capacity).

Using this information, we developed the simulation model. A conceptual model summary is presented in Table 21.1.

The four toplifter yard cranes operate on a FIFO request order. Housekeeping operations (where containers are remarshalled to improve retrieval) are performed during idle times. The servicing of trucks arriving to the depot is FIFO. Each block of the yard is assigned to a single customer and containers are segregated according to the condition class. Two classes can be mixed in the same block if there is high demand at the yard. Damaged containers that have been authorized for maintenance or repair service are moved at 8:00 AM to the maintenance area, and at 7:00 PM containers that were repaired are taken back to the yard and stacked at appropriate locations.

We first integrated the simulation model with an SQL database. The two main databases used correspond to those used by the ECD. One contains the current state of the system (the precise location of all containers) and the other contains a history of all gate-in and gate-out operations. The simulation interacts with the database to update the state of the system and to perform calculations regarding container movements to retrieve a container or to remarshal and reorganize a block.

As described in the Model Scope section in Table 21.1, trucks and containers are represented as entities and they relate to each other through combiners and separators (for load and unload processes). Servers are used for gate-in, gate-out, and inspection operations. The depot layout is represented by paths, nodes, and detached queues and these elements are organized in a table. Cranes are represented as vehicles. The logic of depot operations is built within process modules, and these define the sequence or route of trucks and all crane requests for activities such as block remarshalling or maintenance.

The simulation is initialized with information regarding containers in the depot; this data is obtained from the real system at a particular date. The model interacts with the external database that performs all the data processing. During the simulation run, queries and updates of this database are performed in process modules. There are two external databases that may be used during the simulation run. The primary database registers the current position and all other attributes of containers stored in the depot. Several queries and procedures are implemented for this database. These procedures calculate the time it takes a crane to retrieve a container from a block or to remarshal a block; the procedures also determine positions for all incoming containers according to their classifications, and establish the final position of containers that have been moved or remarshalled. The other database contains registries of the history of all container or truck movements, and they are particularly useful for validation purposes as they mimic a database used by the real system.

Table 21.1 Conceptual modeling elements for the ECD simulation

Project objectives • Organizational aim: to improve truck service times • Modeling objectives: to establish operational policies for container retrieval and remarshalling that reduce average truck service times to less than 30 min
General project parameters • Time-horizon: long term, 3 years and beyond (to include layout evaluation) • Model flexibility: should allow the implementation of new stacking/retrieval, remarshalling, and yard assignment policies • Runs: many experiments to be run • Animation: 3D • Model users: research team
Model results (outputs) • Outputs related to objectives: Mean truck turnaround time (cycle time) • Outputs for understanding causes of results or for validation: Crane utilization Percentage of container dwell times that are less than 10 days Average gate control queue length at different times of the day Number of container movements Average/percentiles buffer queue length
Experimental factors • Container remarshalling policy: 4 levels • Container retrieval policy: 4 levels
Model scope • Objects included in the model. – Entities: containers and trucks, flow through the depot and engage in service activities, their arrival pattern varies over the day – Servers: gate control (in and out) and inspection, with service time distributions – Vehicles: crane (4 units available) to move all containers, constant speed, they block streets, if not dedicated they serve requests in FIFO order and when more than one is available the closest is assigned – Activities: gate-in process, regular gate-out process, massive gate-out process, inspection process, container retrieval process, container drop-off/stacking process, container remarshalling process, container maintenance relocation process. All these activities impact objectives – Queues: gate control, gate-out control, inspection, trucks at blocks, containers at blocks, their discipline is FIFO, infinite capacity – Database: state table with the current location of each container in the system • Exclusions: – Detailed maintenance activities, they have little impact on objectives, only crane movements from and to the maintenance zone are necessary – Leasing/sales activities, their interaction with the system is not significant – Crane maintenance/repair, no breakdowns • Model simplifications: – Different crane stacking/unstacking times are estimated with a service time model that considers the current position/destination of containers to be moved (database calculation) – Containers that do not fit in appropriate blocks are temporarily stored in a buffer queue (representing streets/alleys)

Considering the historical database provided by the ECD under study, for each input variable we performed a goodness-of-fit test to determine an appropriate probability distribution to be used in the simulation experiments. Note that for the gate-out process we consider both the entry and exit service times of the trucks, while in the case of the gate-in process we only consider the time required at the gate.

21.3.7 Verification and Validation of the Simulation Model

Verification and validation are key elements of any simulation project (Law 2015). To increase model confidence, different stages of a simulation project require validation. Early in the project it needs to be established whether the conceptual model is a good representation of the real system, and for this purpose the research team held several meetings with company experts at different management levels. In a long project such as this one, the conceptual model has to be periodically revised as some of the system's logic, potentially including operational policies, changes according to the company's needs. The input model, describing the random variables and their distributions, must also be validated. This was done through goodness-of-fit tests but also by presenting the collected data and samples from the fitted models to the practitioner experts (yard manager, crane operators, and inspectors). The computational output must also be validated against historical performance data available for the system under study. As a result, input variables, assumptions, and the model's logic might require revisions. It might happen that the modelers do not have access to historical data related to all of the performance measures of interest. However, system data often exists for some measure that can be included in the computational model for validation purposes. In our project, the company's database had dwell time information (the time a container spends at the yard) and also the number of containers in the yard for the relevant period of analysis. The company had studied truck arrivals so a sample of truck wait times was available for gate-in operations during peak season. A third source of information that helped validate our model was the personnel themselves—the yard manager and the general manager provided expert opinions to validate the length of queue outcomes of the simulation model and they were shown an animation of the depot under their current policies to validate process flow. All of the quantitative information from the ECD was compared against their equivalent simulated output values.

Figure 21.4 shows the visualization that is standard with Simio for our simulation model of the ECD. Both visualization and animation can be useful in assessing the operation of the simulation model. They can also be valuable when explaining to the stakeholders about the structure and usefulness of stochastic simulation. This was the case when we shared the animation with the management of the ECD.

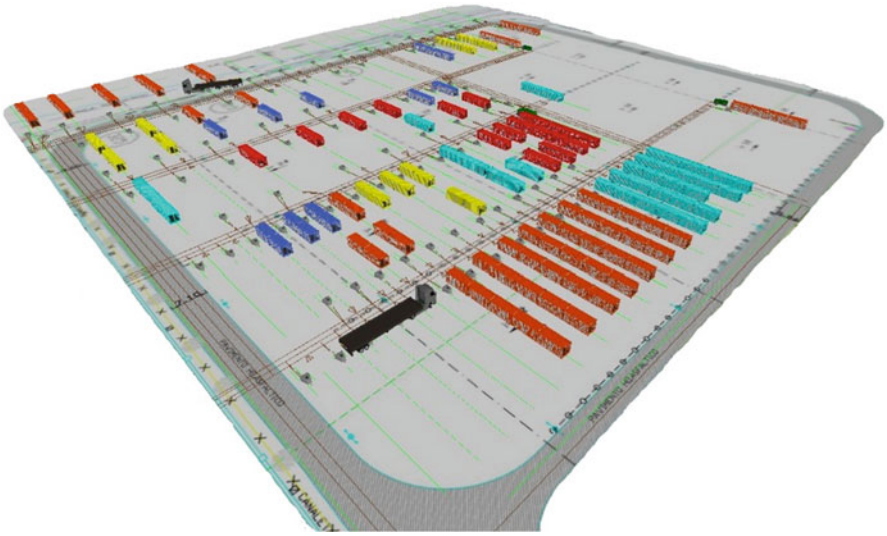


Fig. 21.4 Model visualization from Simio

21.3.8 Using the Simulation Model for Policy Analysis

An experimental design was specified to provide analysis of various policies in the depot and their effects on the outcome variables of interest. The design included the following input parameters:

Retrieval policy—we considered four versions: Easiest to retrieve (ET) and three variants of FIFO. The FIFO variants are: FIFO-strict (SF); FIFO-relaxed with a 5-day window (RF5) and FIFO-relaxed with a 10-day window (RF10). The relaxation means that FIFO was considered for any container within that specific time window. For example, if container X has been in the yard 4 days longer than container Y, but container Y is easier (i.e., quicker) to retrieve, the relaxed FIFO would choose to retrieve container Y. A strict FIFO would mandate that container X be retrieved before container Y (e.g., container labeled 9 in Fig. 21.5 (a) is retrieved first than container labeled 7 if strict FIFO is used, requiring unwanted movements of containers on top).

Remarshalling policy—we considered four versions: (1) remarshalling continuously (DR), (2) remarshalling once a day (1R), (3) remarshalling twice a day for shorter durations (2R), and (4) remarshalling twice a day for longer durations (3R). Either one or two cranes were assigned to perform the remarshalling operations.

The above design gives a full factorial of 16 combinations which were tested in the simulation model. The number of replications varied according to the complexity of the scenario tested but we maintained the same precision of mean estimate throughout. The confidence intervals are 95% and most scenarios required around 100 replications, while low variability scenarios (i.e., for the "easiest to pick"

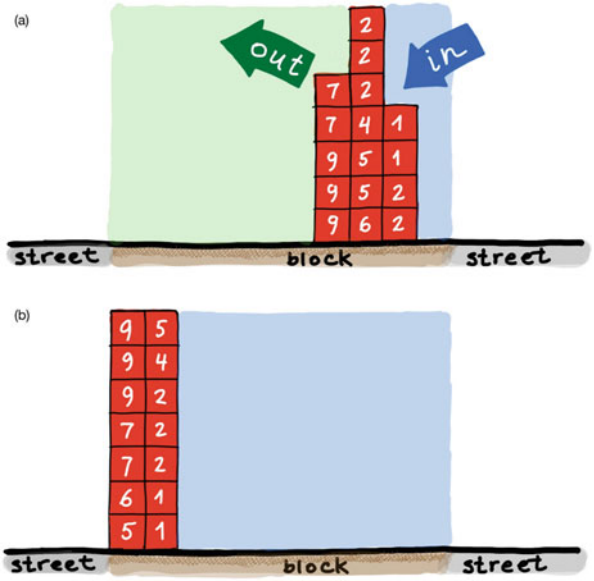


Fig. 21.5 The effects of remarshalling on a typical block of the depot. The top diagram (**a**) shows before remarshalling while the bottom diagram (**b**) shows after remarshalling

retrieval policy) required only 20 replications. Each replication consists of 365 days and considers 180 days of warm-up.

21.3.9 Results and Discussion

Using ANOVA, we found that the retrieval policy, the remarshalling policy, and the interaction between the two are highly statistically significant. This is true for all outcome variables considered. These results are not surprising. First, the outcome variables are somewhat correlated (e.g., a long queue will result in a longer time in system). Second, the impacts of which containers to retrieve and how much remarshalling is done are obviously important. To clarify the effects of remarshalling consider Fig. 21.5. The top picture (a) shows a typical block where containers have been retrieved from one side (left) but loaded from the other side (right). Numbers on the containers refer to the number of days they have been in the depot. This is not an efficient use of yard space. The lower picture shows the same situation after remarshalling. Now, the yard space used is compact and the containers are ready for retrieval by the cranes from the left-hand side. In this example, remarshalling also allowed the block to be organized such that the containers that have been in the depot the longest are the easiest to retrieve (which is particularly important under a strict FIFO retrieval policy).

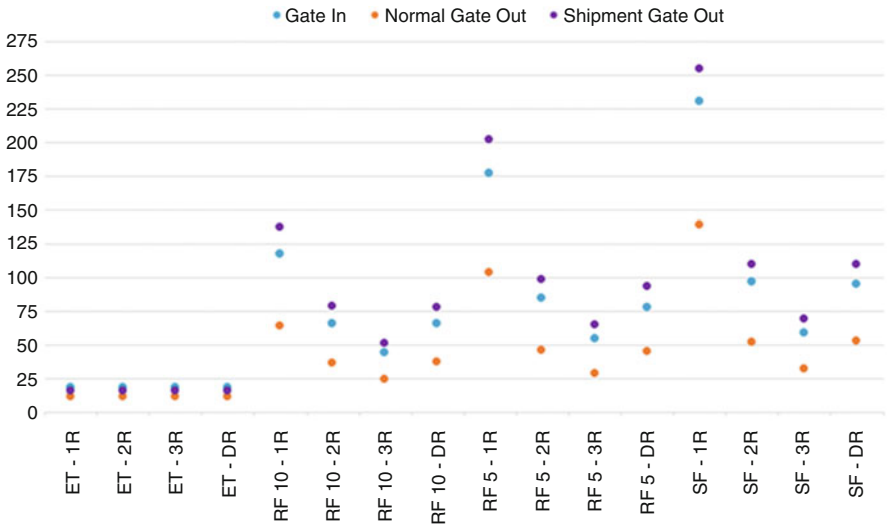


Fig. 21.6 Mean truck turnaround time (minutes) for the 16 policy combinations considered

What is more interesting are some of the detailed effects which are shown in the figures below. Figure 21.6 shows the effects of the various policy changes on the response variable, mean truck turnaround time (time in the system), for the operations of gate-in and gate-out. For each retrieval policy—from left to right these are easiest to retrieve (ET), relaxed FIFO within 5 (RF5) or 10 (RF105) days and strict FIFO (SF)—and for each remarshalling policy—once a day (1R), twice a day shorter (2R), twice a day longer (3R) and all day (DR)—are on the x axis while the truck turnaround time in minutes is on the y axis. The figure presents the resulting time in the system for three conditions: Gate-in, Normal Gate-out, and Shipment Gate-out. Gate-in corresponds to the time in the system for the trucks that are dropping off a container at the depot. It considers the time that starts when the truck enters the queue for inspection of the container and ends when that truck exits the depot. The Normal Gate-out measures the time in the system for those trucks that arrive at the depot to pick up a particular container that will be delivered to a shipper (exporter). It considers the time beginning when the truck enters the queue at the gate until the time the truck exits the depot. Shipment Gate-out corresponds to the time in the system of trucks that are performing a batch transport service of containers to the port (massive runs). Shipment Gate-out and Normal Gate-out differ in that trucks have different arrival patterns and the depot organizes the operations in different queues and servers.

As observed in Fig. 21.6, if the contractual arrangements with the shipping companies are ignored (ET case) the depot can operate most efficiently. This is ideal from the depot's perspective but ignores the costs imposed on the shipping companies by storage fees. At the other end of the spectrum, the strict FIFO which is most favorable to the shipping companies results in longer processing times,

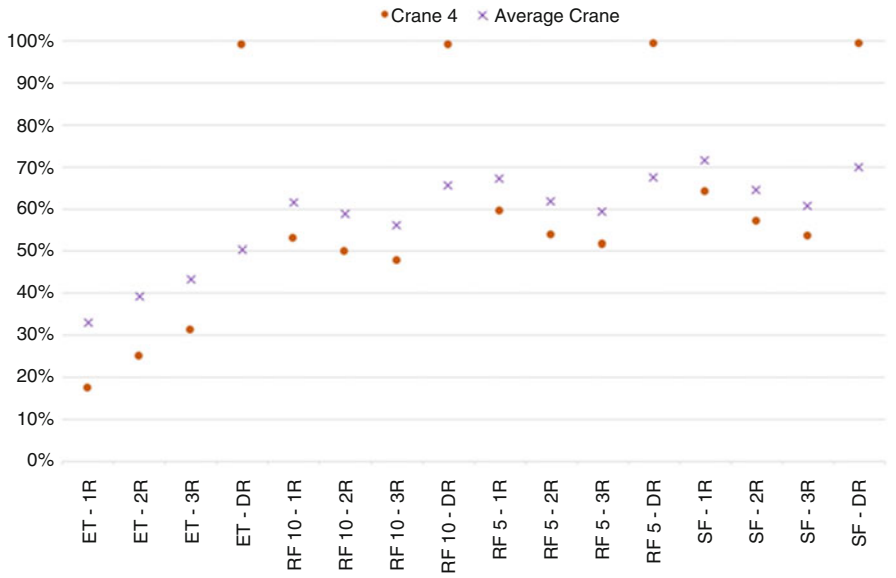


Fig. 21.7 Average crane utilization (%) (crane 4 does remarshalling tasks)

though remarshalling helps mitigate much of this. Relaxing the FIFO rule has improvements especially in the case of minimum remarshalling (1R). Turning to remarshalling, for the speediest service times it is best to remarshal twice a day for longer durations. In fact, for the option of remarshalling all day (DR), even during peak times one crane (crane 4) is dedicated to remarshalling so is unavailable to service pick-up or delivery trucks. This causes the average times (Gate-in and both types of Gate-outs) to increase.

Figure 21.7 has the same *x* axis but the *y* axis portrays the average utilization of four cranes and the utilization of the main crane used in remarshalling activities. Strict FIFO requires more active cranes, while retrieving the easiest to get container uses the cranes for the least proportion of time. What is interesting is that the remarshalling policies impact crane utilization for the three versions of FIFO. More remarshalling results in lower crane utilization except for crane 4 which is dedicated to all day remarshalling (notice the 100% utilization). This may seem counterintuitive but the extra time spent rearranging the containers more than compensates when it comes to retrieving specific containers.

Finally, we consider the number of days a container spends at the depot. Figure 21.8 shows this, again with the same *x* axis of the different policy combinations. The *y* axis shows the percentage of two aspects. With the first three series, it shows the dwell time of containers that have left the depot grouped in three statistics (%OUT): those that spend less than 8, 10, and 12 days at the depot. With the second three series, it shows the composition over time (in terms of how long have containers been at the depot) of the containers that remain at the depot (%DEPOT). This

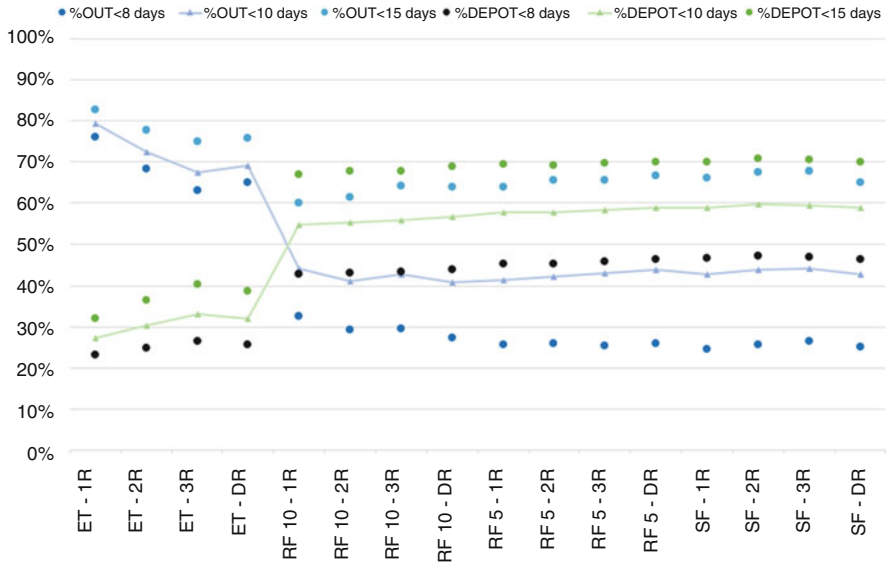


Fig. 21.8 Percentage of containers that spend less than a certain number of days at the depot under the 16 policy combinations

comparison visualizes the effect of different retrieval policies. Containers, for the most part, spend less time in the depot if they are retrieved by the easiest policy (ET). However, under this policy some containers spend a very long time at the depot. This is because containers that are easily accessible are retrieved quickly after arriving at the depot while those that are less accessible are only retrieved when needed (i.e., when fewer containers are in the depot). For the FIFO retrieval rules, there is a better container turnover ratio as fewer than 30% of the containers at the depot at any given time have been there more than 15 days (compared to the approximately 60% in the ET case). Remarshalling does not have a striking effect on this assessment variable.

21.4 Retrospective Examination

21.4.1 Technical Lessons Learned

The experience of working on a simulation project collaboratively, with participants that vary over time, has emphasized the importance of a systems thinking approach to the modeling effort. Communicating the problem and the overall project vision to different groups of people over time, including students, faculty, industrial partners, and other stakeholders, requires clarity on the objectives, and a conceptual

representation of the system and its borders, elements, and interactions. As in a relay race, carefully built conceptual models and model documentation reports are essential to the continuation of the project and allow the next team in the modeling "race" to capitalize on the experience gained when building earlier versions of the model. A smooth coordination over time gives the project the ability to adapt to changes in an agile manner. The challenge in this project has been to be able to communicate the models in both Spanish, for the benefit of our industrial partner, and English, for the benefit of the US team.

Communication to all stakeholders is also greatly improved when animations or dynamic visualizations of the system and its relevant elements are available. Support for this long-term project became so much stronger when middle and upper level management experienced the animation of the system, even when the model was still at a rough draft level. Gaining the industrial partner's trust is really relevant; however, the team must never be distracted by the animation alone and must focus on validation of the results and a solid design of experiments to learn insights on the effects of different system configurations. Students in this project have learned to challenge their pre-conceptions on expected results and to be rigorous and thorough about understanding the events or causes that trigger unforeseeable results. This knowledge is difficult to inherit to the next modeling team in a long-term project thus, although faculty provides the needed continuity for the project, time, and resources are often needed to train new team members to help them improve or acquire new skill sets for the job.

Another element of planning a long-term project is closely tied with deciding which aspect of the problem should be tackled in which stage. At first, our team had focused on the importance of considering the layout of the ECD as a design variable to be included in the study. However, as data was collected and a deeper understanding of the system emerged, the study of operating policies took precedence. The level of detail of system elements and the simplification of some interactions within the ECD and between the ECD and its macro-system (the import/export logistics system) had different considerations that helped decide the stages of the project. For example, when the amount of data for input modeling was large, as when fitting distributions for process times, students were tempted to build very detailed probabilistic models that were not matched with the quality of information available for other aspects of the system—e.g., how cranes move, turn, seize/release containers, etc. Revising these scale/level-of-detail issues was possible when beginning a new stage of the project. Also, leaving details out in one stage—such as the interaction between the size of a massive run and the space availability in the vessel that is planned to arrive at the port—allow their inclusion in subsequent stages of the project with a model that is verified and validated and ready to grow in complexity. The value of a progressive building of the model is an important lesson of this project.

Finally, working collaboratively with a diverse group of engineers has challenges and rewards. Each member contributes to the process with their unique background, time, observation skills, and motivations. Leaders emerge in turn for different aspects of the process. Learning about these characteristics is paramount

to project success, especially when the tasks are assigned in a way consistent to the skill set. Some of us are better writers, are more inquisitive, more patience, better communicators, have more time to gather data, like to make pretty graphs and figures, have previously studied the literature, have the industrial and port connections, etc. Several technical modeling lessons came from knowing in advance the modeling talents of the team. For example, an expert in computer systems architecture was decisive in integrating the simulation tool of choice (Simio) and a database management system (SQL Server). This combination, although increasing the simulation length due to the time it takes to read and write to an external file, eased the calculation of several process times that were related to the position of a container in a particular block in the ECD. The integration of engineering methods and tools to solve complex problems is one of the challenges that engineering teams increasingly face.

21.4.2 Project Participant Personal Growth

Because of the nature of this project—a long project with multiple countries and an industry partner, there were many opportunities for experiences outside of technical learning and development. Below are some observations and outcomes concerning the participants, beginning with the students involved.

The student from Auburn University is a Ph.D. candidate whose dissertation topic is this project. He is originally from Turkey but had the opportunity for a research residency at PUCV to work on this project with the Chilean team. He gives some thoughts concerning the impact of this:

> My visit to Chile was a great opportunity for me to get experience working with an international research group l. It was little scary for me to go another country which I had not known much about nor the language spoken there. However, I am grateful to my advisor, Dr. Alice Smith, for encouraging me to work there in person and to PUCV for hosting me so warmly and providing lodging for me.
>
> During my stay in Valparaiso, Chile, I used the opportunity as a chance not only to work and meet with different people, but also to learn different culture and language. Through the contacts I made in Chile I had a chance to visit Ecuador and Colombia. I made presentations about logistics and our ECD project at the Universidad Técnica del Norte (UTDN) and the Pontificia Universidad Católica del Ecuador-Ibarra, and the Chambers of Commerce in Tulcan, Ecuador, and Pasto, Colombia. Since I have got some experience and eliminated my concern about going to another country where I cannot speak their language, I was willing to try a different culture and get some experience.
>
> During my travels, seeing many different people and places has influenced my personal view about the reasons that why people behave as they do and this has helped me to overcome some difficulties in my life. I have been inspired by every single new person I have met during my travels.
>
> To sum up, I was glad to travel to Chile as a beginning of my adventure path in my academic and personal life, and continued to travel in South America with Ecuador and Colombia. I will look forward to joining any international project collaboration or meeting opportunity

> in my academic life. I truly believe that international collaborations will help people to improve personal and academic skills in any period of life. People learn so many things each day they travel even without realizing it.
>
> I think the quote by Mary Ritter Beard "Certainly, travel is more than the seeing of sights; it is a change that goes on, deep and permanent, in the ideas of living." is a good summary of my experiences from my South American visits.

The two PUCV students from the second cohort of the project completed their master's theses with this project and are now working. The first student reflects thus:

> This project was a challenging opportunity both at the personal level and for the research team. The size of the system to be modeled added to the complexity of the project and we had to learn how to manage it successfully. I left the project having learned a lot. For my personal development I learned how to set goals for myself and for the team, and to collaborate with others to achieve them. It was important to learn about different team member strengths and skillsets in order to divide and coordinate the work. My professional development was enhanced both because I had to immerse myself in the import/export, port logistics industry, and because I had to advance my knowledge in simulation and database integration.
>
> Now I work for a consulting firm providing simulation solutions to mining companies. My background was perfect for the job, not just because of my knowledge of simulation modeling and analysis but because of my experience presenting the results of our work at an international conference.

And the second student;

> The ECD simulation project meant a great professional contribution to me as I learned how to manage, model and analyze a complex problem. It took a long time for us to research the system and its operations, and to consider all the variables and parameters involved that were necessary to build a flexible model where to evaluate different policies and operating constraints. To properly analyze results I had to expand my knowledge in statistical analysis. I also had to improve my database management skills to extract relevant information from the company's ERP system for input modeling and validation purposes. This introduction to data analytics has motivated me to seek specialization in a new graduate program abroad.
>
> At a personal level, this project was a great opportunity for cultural exchange with team members of other nationalities (U.S. and Turkish). We had to learn how to work as a virtual team and to collaborate in the project with applied knowledge, which is something I had not learned during my Chilean education. This experience has been an asset for my professional work because industrial jobs expect you to be efficient, proactive and goal oriented. This cultural exchange also helped me practice social skills related to communication, empathy, patience and the planning of our team's joint effort. My current job has benefited from these project management skills because I have had to negotiate the implementation of regulations, policies and corporate management models with my company's affiliates that have different people and work cultures.

The industry partner is represented by the participant below, who was also a PUCV student during the initial cohort of the project.

> The experience of working as an industrial partner to this project helped me better understand the importance of bonded warehouses in the logistics chain of maritime transportation and understand how our level of service impacts other participants in this chain.

> During the implementation of the project I was able to see the results of the proposed policies and how these positively impacted the ECD's key performance indicators and its relationship with truck drivers and the transportation companies they work for, and with clients.
>
> Having had the opportunity to set aside time to analyze the problems of the company, thanks to the company's willingness to find alternatives that would differentiate them through an improved quality of service, I have been able to reflect on the importance of engineering for the support of business growth and sustainability.

For the two principal faculty members, even though we are senior faculty members, this project has brought profound changes. First, is the opportunity to work together and get to know each other very well, as collaborators first and now as close friends. We have shared cultures, families, travels, and more throughout these several years. Exposing our students to international teams and travel has gratified our professional accomplishments. The project has enhanced our professional stature through grants (CONICYT Chile and Fulbright), through publications and presentations, and through expanding our network of research collaborators both in academia and in industry/government. The other faculty member involved, though at a junior rank, benefited also from the international team and presenting this work at international conferences.

21.4.3 Tangible Outcomes

We continue meeting at the corporate level with the company and we are also invited for meetings with other actors in the port system. One of our first international poster presentations caught the eye of a young consultant on port technology at another country that wished to consider our insights in the design of a new depot. The master's student from the first cohort was hired by one of the port operators where his experience with the depot has been invaluable and where he sees other applications for simulation modeling. One of PUCV master's students in the second cohort, after working for a year in industry, leveraged his technical presentation at an international conference to facilitate a successful scholarship application to graduate school in Australia. His newly found confidence in English, thanks to this project, also impacted his interest in studying at an English speaking country. The other master's student from the second cohort was hired at an international consulting firm to work in the simulation department thanks to his simulation building skills developed during this project. The Auburn Ph.D. student presented his dissertation research proposal based on this project. A research residency at Auburn for one of the current (cohort 4) PUCV students is planned to commence soon. This will be his first experience on foreign soil.

Below are two photos of the project team (Figs. 21.9 and 21.10).

In summary, this project has had an abundance of outcomes, some expected and some surprising, both from a technical project perspective and from an educational perspective. Simulation projects are particularly interesting because a model of a

Fig. 21.9 The two principal faculty members (center) with their respective students at the ECD

Fig. 21.10 The entire second phase team of students and professors at the PUCV in Valparaiso, Chile

complex system is developed and used for decision-making. As such, there are many choices to be made regarding the level of fidelity, the type, and amount of data needed, assumptions to make, amount of analysis, validation with the actual system, and presentation of results and recommendations. Given that such projects tend to be long term with many team members, there are opportunities for personal and professional growth particularly when multiple countries are involved. Such was the case here. We are also proud to be a women led team engaging with a highly male dominated field—both the container depot domain and the simulation modeling domain.

References

Carlo HJ, Vis IFA, Roodbergen KJ (2014) Storage yard operations in container terminals: literature overview, trends, and research directions. Eur J Oper Res 235(2):412–430

Chen L, Lu Z (2012) The storage location assignment problem for outbound containers in a maritime terminal. Int J Prod Econ 135(1):73–80

Chwif L, Banks J, de Moura Filho JP, Santini B (2013) A framework for specifying a discrete-event simulation conceptual model. J Simul 7:50–60

ECLAC (2017). Ports ranking. The top 20 in Latin America and the Caribbean in 2017. https://www.cepal.org/en/infographics/ports-ranking-top-20-latin-america-and-caribbean-2017

Furian N, O'Sullivan M, Walker C, Vossner S, Neubacher D (2015) A conceptual modeling framework for discrete event simulation using hierarchical control structures. Simul Model Pract Theory 56:82–96

Gunal MM, Pidd M (2007) Interconnected DES models of emergency, outpatient, and inpatient departments of a hospital. In: Henderson SG, Biller B, Hsieh M-H, Shortle J, Tew JD, Barton RR (eds) Proceedings of the 2007 winter simulation conference, pp 1461–1466

Harmon P (2014) Business process change, 3rd edn. Elsevier, New York

Hidalgo F, Aranda D, Pascual J, Karakaya E, Smith AE, Gonzalez-Ramirez RG (2016) Empty container stacking operations: case study of an empty container depot in Valparaiso Chile. In: Proceedings of the international conference on production research, Valparaíso, Chile, October 2016

Hidalgo F, Aranda A, Pascual J, Smith AE, Gonzalez-Ramirez RG (2017) Empty container stacking operations: case study of an empty container depot in Valparaiso Chile. In: Chan WKV, D'Ambrogio A, Zacharewicz G, Mustafee N, Wainer G, Page E (eds) Proceedings of the 2017 winter simulation conference, pp 3114–3125

Kang J, Ryu KR, Kim KH (2006) Deriving stacking strategies for export containers with uncertain weight information. J Intell Manuf 17:399–410

Kemme N (2012) Effects of storage block layout and automated gantry crane systems on the performance of seaport container terminals. OR Spectr 34(3):563–591

Kim KH, Kim KY (1999) Routing straddle carriers for the loading operation of containers using a beam search algorithm. Comput Ind Eng 36(1):109–136

Kim KH, Park YM, Jin MJ (2008) An optimal layout of container yards. OR Spectr 30(4):675–695

Law A (2015) Simulation modeling and analysis, 5th edn. McGraw-Hill, New York

Lee Y, Hsu NY (2007) An optimization model for the container pre-marshalling problem. Comput Oper Res 34:3295–3313

Lee BK, Kim KH (2013) Optimizing the yard layout in container terminals. OR Spectr 35:363–398

Park T, Choe R, Kim YH, Ryu KR (2011) Dynamic adjustment of container stacking policy in an automated container terminal. Int J Prod Econ 133:385–392

Pascual J, Aranda D, Hidalgo F, Smith AE, Karakaya E, Gonzalez-Ramirez RG (2016) Empty container stacking operations: case study of an empty container depot in Valparaiso Chile. In: Proceedings of the 2016 winter simulation conference, Washington DC

Pidd M (2007) Making sure you tackle the right problem: linking hard and soft methods in simulation practice. In: Henderson SG, Biller B, Hsieh M-H, Shortle J, Tew JD, Barton RR (eds) Proceedings of the 2007 winter simulation conference, pp 195–204

Ranau M (2011) Planning approach for dimensioning of automated traffic areas at seaport container terminals. In: Bose JW (ed) Handbook of terminal planning (Chapter 10). Operations research/computer science interfaces series 49. Springer, New York, pp 179–193

Robinson S (2008a) Conceptual modelling for simulation part I: definition and requirements. J Oper Res Soc 59:278–290

Robinson S (2008b) Conceptual modelling for simulation part II: a framework for conceptual modeling. J Oper Res Soc 59:291–304

Robinson S (2017) A tutorial on simulation conceptual modeling. In: Chan WKV, D'Ambrogio A, Zacharewicz G, Mustafee N, Wainer G, Page E (eds) Proceedings of the 2017 winter simulation conference, pp 565–579

Smith JS, Sturrock DT, Kelton WD (2014) Simio and simulation: modeling, analysis, applications, 4th edn. Simio LLC, Sewickley

Sturrock DT (2017) Avoid failures! Tested success tips for simulation excellence. In: Chan WKV, D'Ambrogio A, Zacharewicz G, Mustafee N, Wainer G, Page E (eds) Proceedings of the 2017 Winter simulation conference, pp 588–596

Taner ME, Kulak O, Koyuncuoglu MU (2014) Layout analysis affecting strategic decisions in artificial container terminals. Comput Ind Eng 75:1–12

Wiese J, Suhl L, Kliewer N (2011) Planning container terminal layouts considering equipment types and storage block design. In: Handbook of terminal planning. Operations research, computer science interfaces series 49, pp 219–245

Wiese J, Suhl L, Kliewer N (2013) An analytical model for designing yard layouts of a straddle carrier based container terminal. Flex Serv Manuf J 25:466–502

Jimena Pascual is an Associate Professor of the School of Industrial Engineering at the Pontificia Universidad Católica de Valparaíso (PUCV), Chile. She holds Ph.D. and M.S.I.E. degrees in industrial engineering from Purdue University, and an industrial engineering degree from PUCV. Her research interests include the mathematical modeling and simulation of logistics and health systems, as well as research in engineering education applying active learning methodologies and systems thinking. Her recent collaborative projects include modeling logistics processes related to port operations, optimizing coverage and path planning of unmanned vehicle search-and-rescue missions, studying emergency room technology adoption and operations, and improving engineering education with a focus on innovation and entrepreneurship. Starting in 2019, she will lead her university's participation in the international project "Building the future of Latin America: engaging women into STEM (W-STEM)" funded by the European Union.

"I knew early in my life that I wanted to teach. I would create practice math exams to play with my friends at our made-up school, and during summer breaks I would teach the ants and bees at my grandmother's country estate. My parents were both math majors and teachers, and my dad was a statistics professor at the university, so I had a lot of inspiration at home and many cool books with puzzles and stories of all sorts. I was fascinated

with enigmas, riddles and brainteasers, but I was lazy about doing my home chores so I created several contraptions to ease my life. When I thought of following my parents' academic footsteps, they suggested that my practical use of creativity and my yearning for efficiency might make me a good engineer. I had never heard the word before and was clueless about this profession, but a tour of the university convinced me that I would have fun solving problems there. I later went on to graduate school to become a professor. Now I enjoy solving complex problems for which I believe a system's perspective is crucial when modeling with the mathematical toolset we acquire in our engineering training. And as for teaching, I believe it makes me as happy today as it did in my imaginary classroom as a child."

Alice E. Smith is the Joe W. Forehand/Accenture Distinguished Professor of the Industrial and Systems Engineering Department at Auburn University, where she served as Department Chair from 1999 to 2011. She also has a joint appointment with the Department of Computer Science and Software Engineering. Previously, she was on the faculty of the Department of Industrial Engineering at the University of Pittsburgh from 1991 to 1999, which she joined after industrial experience with Southwestern Bell Corporation. Dr. Smith has degrees from Rice University, Saint Louis University, and Missouri University of Science and Technology.

Dr. Smith's research focus is analysis, modeling, and optimization of complex systems with emphasis on computation inspired by natural systems. She holds one U.S. patent and several international patents and has authored more than 200 publications which have garnered over 3300 citations and an H Index of 25 (ISI Web of Science) and over 10,000 citations and an H Index of 44 (Google Scholar). Several of her papers are among the most highly cited in their respective journals including the most cited paper of *Reliability Engineering & System Safety* and the second most cited paper of *IEEE Transactions on Reliability*. She won the E. L. Grant Best Paper Awards in 1999 and 2006, and the William A. J. Golomski Best Paper Award in 2002. Dr. Smith is the Editor-in-Chief of *INFORMS Journal on Computing* and she is also an Area Editor of *Computers & Operations Research*.

Dr. Smith has been a principal investigator on over $8 million of sponsored research with funding by NASA, U.S. Department of Defense, Missile Defense Agency, National Security Agency, NIST, U.S. Department of Transportation, Lockheed Martin, Adtranz (now Bombardier Transportation), the Ben Franklin Technology Center of Western Pennsylvania and U.S. National Science Foundation, from which she has been awarded 17 distinct grants including a CAREER grant in 1995 and an ADVANCE Leadership grant in 2001. Her industrial partners on sponsored research projects have included DaimlerChrysler Electronics, Eljer Plumbingware, Extrude Hone, Ford Motor and Crucible Compaction Metals. In 2013, she was a Fulbright Senior Scholar at Bilkent University in Ankara,

Turkey; in 2016, a Fulbright Specialist at EAFIT in Medellin, Colombia, and in 2017 a Senior Fulbright Fellow at Pontifical Catholic University of Valparaíso, Chile.

For accomplishments in research, education, and service, she was named the Joe W. Forehand/Accenture Distinguished Professor in 2015. Previously, she was the H. Allen and Martha Reed Professor. In 2017, she was awarded the inaugural Auburn University 100 Women Strong Leadership in Diversity Faculty Award. Dr. Smith was awarded the Wellington Award in 2016, the IIE Albert G. Holzman Distinguished Educator Award in 2012 and the INFORMS WORMS Award for the Advancement of Women in OR/MS in 2009. Dr. Smith was named the Philpott-WestPoint Stevens Professor in 2001, received the Senior Research Award of the College of Engineering at Auburn University in 2001, and the University of Pittsburgh School of Engineering Board of Visitors Faculty Award for Research and Scholarly Activity in 1996.

Dr. Smith is a fellow of the Institute of Industrial and Systems Engineers (IISE), a fellow of the Institute of Electrical and Electronics Engineers (IEEE), a senior member of the Society of Women Engineers, a member of Tau Beta Pi and the Institute for Operations Research and Management Science (INFORMS), and a Registered Professional Engineer in Alabama and Pennsylvania. She is an IEEE Distinguished Lecturer in Computational Intelligence.

"As a child, I avidly read 'Cheaper by the Dozen' and its lesser known sequel, 'Belles on Their Toes.' The latter book was a signed copy by Dr. Lillian Gilbreth who was a professor at Purdue University when my mother was a mathematics major there in the 1950s. I was fascinated by the 'one, best way' to do everything. I formally entered STEM as an undergraduate civil engineering student at Rice University in the 1970s. During this time, programs to recruit more women into engineering were being implemented and I benefited from this visibility of engineering as a profession. I am the first engineer in my family so had no immediate role models. My affinity for mathematics and all things analytical took me on a pathway that led to systems engineering emphasizing computation. These research areas were and still are woefully underrepresented with women and one of the rewards of my professional career has been to mentor and assist women colleagues and women students. I have graduated 16 doctoral students of which five are women and all five are thriving academics. I am proud of them and even more proud of my three children, all of whom are engineers including my daughter who is an executive in the healthcare field. I firmly believe that engineering and related fields are great career choices for anyone."

Part VI
Production

Chapter 22
Sustainability and Life Cycle Product Design

Deborah Thurston and Sara Behdad

Contents

22.1 Introduction 518
22.2 Mathematical Modeling Approach to Sustainable Design 519
22.3 The Need to Consider Conflicting Objectives 521
22.3.1 Broadening the Set of Conflicting Objectives 522
22.3.2 Constrained Multiattribute Utility Optimization 523
22.3.3 Statistical Manufacturing Process Control to Identify Pollution Prevention Opportunities 523
22.4 The Need to Analyze the Entire Product Life Span 525
22.4.1 Maintenance, Repair, Replacement 525
22.4.2 Selective Disassembly, Value-Mining, and Sharing Disassembly Operations for Multiple Products 526
22.5 The Need to Use New Data Collection Tools 527
22.6 The Need to Investigate the Complex Role of Consumer Behavior 531
22.6.1 Consumer's Product Storing Behavior 532
22.6.2 Consumer's Product Usage Behavior 533
22.6.3 Consumer's Repair Behavior 534
22.6.4 Consumer's Product Return Behavior 535
22.7 Summary and Closure 537
References 537

This chapter addresses problems that arise during product design for sustainability and the life cycle. A description of the problem itself is provided from an industrial engineering viewpoint. The first section describes the problem elements, including

D. Thurston
Industrial and Enterprise Systems Engineering, University of Illinois, Urbana-Champaign, Champaign, IL, USA
e-mail: thurston@illinois.edu

S. Behdad (✉)
Industrial and Systems Engineering, Mechanical and Aerospace Engineering, University at Buffalo, SUNY, Buffalo, NY, USA
e-mail: sarabehd@buffalo.edu

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_22

the need to expand the set of conflicting objectives under consideration, the need to consider the entire product life cycle, the need to employ new data acquisition tools, and the need to investigate the complex role of consumer behavior before, during, and after the point of purchase. Subsequent sections summarize work the authors have done towards solving these problems. A general mathematical programming framework is first presented. Then, the chapter highlights several instances of the benefits of bringing the logic and mathematical rigor of industrial engineering methods to these problems. The authors' previous contributions to sustainable design are presented and include defining the concept of the product life cycle from a decision-based design point of view, developing different types of decision-making techniques for engineering design (both subjective and objective), normative decision analytic methods (e.g., multiattribute utility, constrained optimization), methods for environmentally conscious design to cover new environmental objectives (e.g., connection of design with the end-of-use phase), and immersive computing technologies to address challenges with information-intensive design procedures. The final section presents methods to consider heterogeneous consumer behavior during product selection, use, and disposal.

22.1 Introduction

Design for sustainability is more complicated that simply "doing the right thing." First, one must define which thing(s) should even be considered. Then, information about the current and possible future states of those things must be gathered. That information must then be analyzed, or processed in some way. Finally, decisions must be made on the basis of that analysis and action taken.

Industrial engineering provides the toolbox needed to tackle such problems. This toolbox includes a broad set of mathematical models that can be used to predict, control, and generally make things better. This chapter presents an overview of research conducted by the authors that employs industrial engineering tools towards the goal of making sustainable design as efficient, profitable, and sustainable as possible.

The engineering design process plays a significant role in both causing and solving sustainability problems. In the past, the traditional design process started with a set of technical performance specifications posed in terms of hard constraints. Then, the designer would create a configuration to satisfy these constraints. This step was often considered to be an art, rather than a science.

This configuration would then be evaluated, most often on the basis of cost to manufacture. What followed was an iterative configure/evaluate/reconfigure to improve/evaluate loop. Externalities such as environmental impacts were not considered, and the time frame of analysis typically encompassed only the manufacturing process, from the cost of materials entering the plant to the manufacturing cost of the finished product leaving the plant.

More recently, the engineering design process has employed mathematical modeling approaches to make the design process itself more efficient. For example,

the axiomatic design and the design structure matrix approach (Suh 1998; Tokunaga and Fujimura 2016) illuminate dependence relationships among physical design parameters and performance metrics, enabling designers to see how they might reconfigure the product to improve performance in one area without degrading performance in another. Also, normative decision analysis has brought mathematical rigor to the design evaluation process by formally modeling the decision-maker's willingness to make tradeoffs (Thurston 1991), the effect that uncertainty has on overall design utility, and integrating normative decision-making into constrained optimization (Thurston et al. 1994).

Design for sustainability poses new challenges that are not trivial. The time frame for analysis has been expanded quite significantly from just the manufacturing process to now include impacts occurring prior to manufacturing (such as those resulting from raw material extraction and processing), as well as during the consumer-use and end-of-life stages. The list of performance metrics has also expanded, from primarily cost (and perhaps quality) to now include environmental impacts before, during, and after the manufacturing process, including air, water, solid waste, and others. The degree, type, and range of uncertainty associated with estimating these impacts is significantly greater than that associated with estimating manufacturing costs.

Design for sustainability needs everything that industrial engineering tools can provide. In this chapter, we provide a summary and overview of work we have done to date towards these problems. The next section describes our general approach, which is decompositional in nature; breaking a difficult, unsolvable problem into smaller, solvable pieces and then reassembling those pieces into a whole.

What follows is a description of the mathematical modeling methods we have employed, beginning with those for addressing conflicting objectives. The following section presents methods for addressing the entire life cycle, focusing on disassembly. Then, a method for using new data collection tools is described, followed by more recent work in investigating the role of consumer behavior in design for sustainability.

22.2 Mathematical Modeling Approach to Sustainable Design

There are several key areas of focus for this chapter, as highlighted in Fig. 22.1. These are (1) the need to consider a broader set of conflicting objectives, (2) the need to assess and analyze sustainability-related outcomes across the entire product life span, (3) the need to recognize information-intensive nature of the sustainable design, and employ new data collection tools, and finally (4) the need to consider the complex and understudied role of consumer behavior.

Design for sustainability is difficult for the reasons described above. Work presented in this chapter employs mathematical models created to help predict,

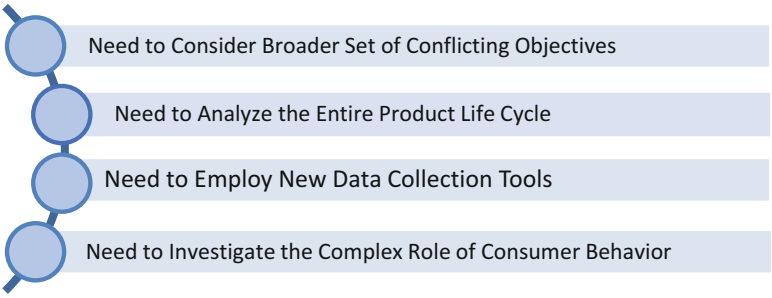


Fig. 22.1 Difficulties in sustainable engineering design

control, and improve the outcomes of design decisions. Then, results can then be analyzed to provide insight as to how to improve further. We center our discussion around the ideal of a bi-level optimization formulation to simultaneously consider the objectives of both manufacturers and consumers. Each section of this chapter describes work done in one area of the problem, beginning with identifying which objectives to include (or not) from the manufacturer's perspective, how to best satisfy different customer market segments, employing a designed experiment to reveal hidden causes of pollution, and using simulation methods to improve disassembly and reassembly operations.

The focus of optimization models used in the "design for sustainability" context is on identifying an optimal mix of strategies that result in minimum costs and environmental impacts with maximum customer satisfaction. The advantages of bi-level optimization models are that both the manufacturer's and customers' interests are considered. To incorporate the role of uncertainty, we can extend a simple bi-level optimization model to a stochastic optimization framework.

Ideas from probability theory and bi-level optimization can be employed to capture, quantify, and apply imperfect information about consumers' preferences in purchasing, using, and discarding products. The set of optimization models that designers build establishes a framework for evaluating intervention strategies (e.g., design for disassembly, design for longer lasting products) towards both environmental impact prevention and economic gain.

The results of simulation-based data generation techniques can also be integrated with optimization models. The integration of data generation tools within the optimization framework can provide designers with accurate estimates of the uncertainty associated with both consumers' decisions, expected profitability, and environmental impacts.

In our stochastic bi-level optimization models, consumers are the decision-makers in the lower level portion of the model, who make product purchase and usage decisions in response to design decisions made by the manufacturers in the upper level. One distinct feature of these models is their ability to find equilibria in the market system by employing a prospective approach. Manufacturers can make

decisions about the types of design features that should influence both purchasing and consumer-use behavior.

The standard stochastic bi-level optimization is in the following form:

$$\min_{x \in X} f(x) + \mathbb{E}\left[g\left(y\left(x, \omega\right)\right)\right] \tag{22.1}$$

$$\text{s.t.} \quad y(x, \omega) = \arg\min_{y \in Y(x)} h(x, y, \omega) \tag{22.2}$$

where x is the upper-level decision vector describing the manufacturers design interventions, ω is a random vector affecting probabilistic choices made by consumers, and $y(x, \omega)$ is the lower level decision vector of consumers given x and ω.

To solve the stochastic bi-level optimization problem, a computational approach can be developed that combines a single-level reformulation technique using Karush-Kuhn-Tucker (KKT) conditions for the lower level problems and a Monte Carlo simulation approach with sample average approximation (SAA) for the upper-level problem. The capability of bi-level optimization models has already been demonstrated in the literature and different solution algorithms have been developed. Examples include as reducing a bi-level model to a single-level model (Camacho-Vallejo et al. 2015), two-stage heuristic algorithms (Du and Peeta 2014), using meta-heuristic techniques such as progressive hedging (Yi et al. 2017), genetic algorithms (Mahmoodjanloo et al. 2016), simulated annealing algorithms (Starita and Scaparra 2018), and stochastic equilibrium constraints algorithms (Faturechi and Miller-Hooks 2014).

A look at the above-mentioned mathematical models reveals the importance of considering several points as highlighted in Fig. 22.1: (1) the importance of considering multiple objectives, (2) the need to consider sustainability effects of the entire product life span, (3) the need for collection and utilization of accurate and timely data for mathematical models, and (4) the role of consumer behavior as one of the main stakeholders. In the next several sections, we describe several studies conducted by the authors towards meeting these needs.

22.3 The Need to Consider Conflicting Objectives

This section presents the progression of our work in developing these models, beginning with simply broadening the set of objectives considered to evaluate a set of discrete options and formulating a multiattribute evaluation function, then: (1) using that function as the basis of the objective function in a constrained optimization formulation to (2) identify the best possible option subject to unavoidable cause and effect relations between decisions and attribute outcomes, and (3) finally using statistical manufacturing process control to discover opportunities for pollution prevention that can then be woven into the optimization formulation.

In our work, we have most often employed a multiattribute utility function as shown below in Eq. 22.3, where K and k_i are scaling constants and y_{ij} refers to the level of attribute i for decision j. The constraint functions (not shown) define the correlation between design decisions and each resulting attribute.

Despite misconceptions to the contrary, this functional form can be very useful during the design process (Thurston 2001). When properly assessed, it can accurately reflect whether utility is linear or nonlinear with respect to performance in any one attribute, the effect of uncertainty, and whether the designers' willingness to trade off one attribute against another remains constant or varies over the feasible design region.

Unlike other multi-objective evaluation methods (such as the weighted average method), this approach requires a rigorous, systematic process for defining the set of attributes that are both relevant and negotiable, their range of negotiability, the decision-maker's willingness to make tradeoffs (Thurston 1991), and the effect of uncertainty on the utility of design alternatives. This process requires some expertise in decision analysis, but is no more arduous than other analytic tools routinely employed by design engineers. We have also found that going through this process itself focuses the designer's effort where the payoff is greatest. In addition, defining the attributes and their ranges of negotiability in such a way that the preferential and utility independence conditions required so the Eq. 22.3 is valid has been especially useful.

$$1 + KU\left(y_1, y_2, \ldots, y_n\right) = \prod_{i=1}^{n}\left[Kk_iU_{ij}\left(y_{ij}\right) + 1\right] \tag{22.3}$$

22.3.1 Broadening the Set of Conflicting Objectives

Consideration of sustainability requires the designer to consider outcomes of the design decisions that were previously "outside the box" bounding their analytic frame. A formal method for design evaluation of multiple conflicting attributes was presented in (Thurston 1991). Material substitution in the automotive industry was investigated. Part of the motivation for this work was that collaborators (designer at several automotive companies) indicated that they wished to compare steel, aluminum, and polymer composite materials, and in particular wished to explicitly consider (and value) the sustainability of their material choices. A normative multiattribute utility function was formulated to help the designers make decisions that reflected their unique willingness to make tradeoffs. The assessment methodology employs a decompositional, "lottery" question and answer approach that measures the extent of decreasing marginal value as one attribute is improved, the subject's willingness to sacrifice performance in one attribute in order to gain in another, and the effect of uncertainty on their valuation of a design alternative. One interesting outcome of the analysis was that although the subjects all said that they

were concerned about the environment and wished to consider sustainable options, their responses to the lottery questions revealed that they, in fact, were not willing to make any tradeoffs to achieve sustainability, and as a result compliance with environmental regulations was then treated as a binary "must comply" constraint in the analysis.

22.3.2 Constrained Multiattribute Utility Optimization

In Mangun and Thurston (2002), the customer's willingness to make tradeoffs for sustainability in personal computers was explored. A constrained optimization problem was formulated, maximizing multiattribute utility (cost, reliability, and environmental impact), the structure of which is shown in Fig. 22.2. The binary decision variables reflected whether or not each of 88 components were to be new, reused, remanufactured, or recycled in a second product life cycle. It was determined that the customer was willing the make tradeoffs, and that, compared with all-new components, remanufacturing and recycling certain components for a second life cycle did increase customer utility. Further, the magnitude of this willingness to pay depended on which market segment a customer belonged to; technophile, utilitarian, or "green." The model structure enabled the identification of an optimal portfolio of products (a distinct product for each market segment) that presented each segment with the combination of remanufactured or recycled components that resulted in the set of tradeoffs that were best for them. Post-optimality analysis then revealed that if the manufacturer adopted a service-selling (rather than product-selling) approach, further efficiencies could be realized by controlling and fine-tuning the take-back period to each specific market segment. This was explored in (Thurston and De La Torre 2007), which determined that through a leasing program the specified the length of time of consumer-use, both profitability and consumer satisfaction could be improved.

22.3.3 Statistical Manufacturing Process Control to Identify Pollution Prevention Opportunities

In Carnahan and Thurston (1998), concurrent design for sustainability of both the product and the manufacturing process was considered. This work was motivated by a floor tile manufacturing plant that was seeking to expand production. At current production levels, they were in full compliance with air pollution control regulations although they sometimes observed unexplained spikes in emissions. If these spikes continued along with increased production levels, unacceptable levels of air pollution would be emitted.

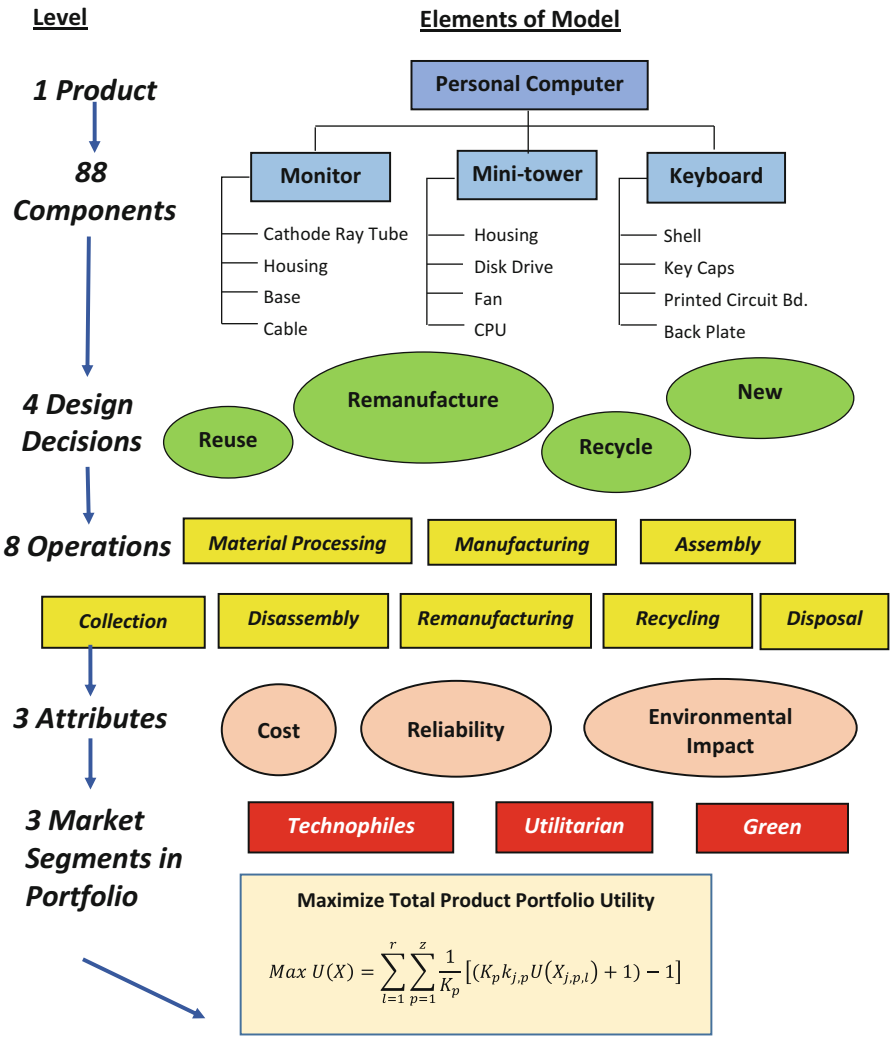


Fig. 22.2 Major elements of constrained optimization model for PC component reuse, remanufacturing, or recycling decisions

Again, the industrial engineering approach of mathematical modeling was employed to understand, predict, and control this situation. The tradeoffs here were between manufacturing cost, air pollution, and product quality. A constrained multiattribute utility function was employed, and the constraint functions reflected the cause and effect relationships between design decisions (13 raw material choices and 17 manufacturing process parameters) and resulting attributes of cost, pollution, and product quality. The variation in air pollution levels was seen as an opportunity to identify manufacturing process parameters that might be correlated with the

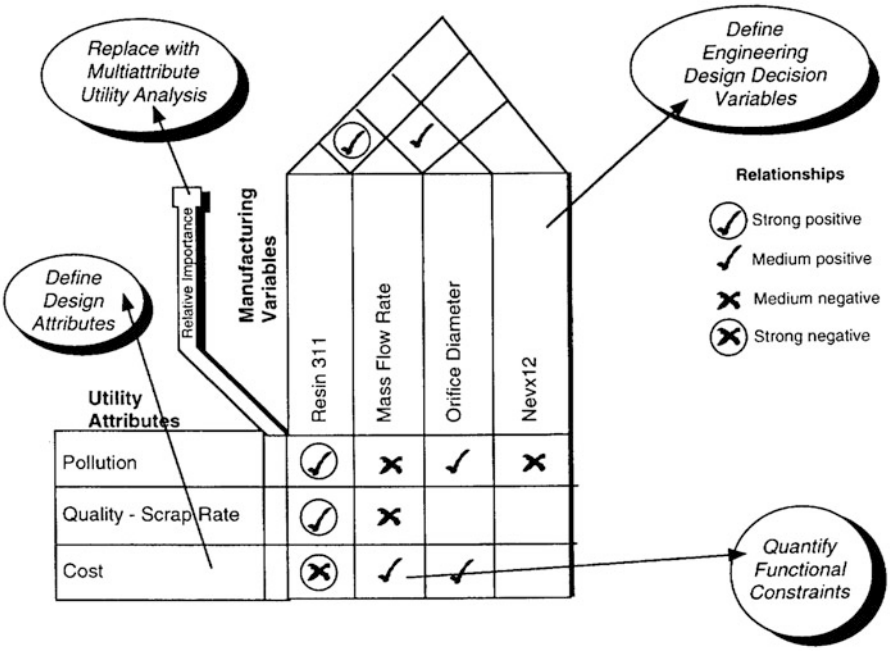


Fig. 22.3 House of quality, DOE and constrained optimization integrated to revealing raw material and manufacturing process parameters correlated to pollution, quality, and cost

variations, and perhaps be controllable. A statistical manufacturing process control experiment was conducted, which identified these correlated parameters, a selected subset of which are shown in Fig. 22.3. Information obtained from the experiment also helped to fine-tune the ranges over which the designer was truly able and willing to make tradeoffs. An ironic finding was that one of the raw materials that was correlated with higher air pollution levels had been a scrap material that the manufacturer was recycling on-site in order to reduce solid waste.

22.4 The Need to Analyze the Entire Product Life Span

22.4.1 Maintenance, Repair, Replacement

One possible strategy for reducing waste is to increase the product usage life span and promote repair and reuse practices among individual consumers. However, manufacturers currently do not view "design for longer lasting products" as a profitable strategy. In fact, cost-effective assembly processes are often irreversible (replacing screws with snap-fits, for example), and manufacturers' present strategies are sometimes focused on making products difficult and expensive to repair and reuse (Cooper 2012).

However, there are several economic reasons why manufacturers may be interested in design policies that facilitate repairs: (1) design for repairable products reduces the cost of after-sales services and warranty offers (Saccani et al. 2007; Fang and Hsu 2009), (2) repairability might be regarded as marketing and sales strategies (e.g., rated repairability attributes in online product reviews) (Gaiardelli et al. 2008; Stevels 2002), (3) new business models adopted by enterprises based on the concept of a sharing economy, the peer economy (e.g., renting, sharing, exchange) and service-based business models require durable products (Preston 2012), (4) design for repairable products is a strategy to secure the future supply of critical materials and rare earth elements (Peck et al. 2015), (5) repairability influences the future reusability of devices, the cost of remanufacturing, and the profitability of remarketing in the second-hand market (Cuthbert et al. 2016), and finally (6) the independent repair businesses and initiatives, e.g., the Digital Right to Repair Coalition, have been forming worldwide campaigns against the short-term profit-driven policies of manufacturers and urging them to produce more repairable electronics, to share the repair guides, and to supply the spare parts to the market (Rosner and Ames 2014). A product's repairability not only influences its first life cycle, but also it increases its future reusability and the opportunity to generate a new market for used devices.

Design for sustainability via facilitating cost-effective maintenance, repair or component replacement, can increase the length of product life cycles, decrease waste, and increase profitability. However, disassembly, repair, and reassembly processes incur costs and sometimes result in damage to one or more components. In Behdad and Thurston (2012), a graph-based integer linear programming and multiattribute utility analysis model is formulated to find the optimal sequence of disassembly operations. Conflicting attributes considered are disassembly time (and cost) under uncertainty, the probability of not incurring damage during disassembly, reassembly time (and cost), and the probability of not incurring damage during reassembly. A solar heating system example is presented. Analysis of the model and results revealed the auxiliary heater as a target for redesign to improve accessibility and/or longevity due to its need for frequent replacement.

22.4.2 Selective Disassembly, Value-Mining, and Sharing Disassembly Operations for Multiple Products

Returned end-of-life products arrive at the take-back facility with a great degree of variability and uncertainty in terms of quantity, age, and quality. The value of individual components is sometimes not worth the cost of full disassembly. Towards a solution to this problem, a stochastic chance constrained programming model converted to a mixed integer linear program for waste stream acquisition (as opposed to market-driven systems) is presented in (Behdad et al. 2012). The model treats returned product quantity as an uncertain parameter and determines

the optimal degree to which disassembly processes should be performed, as well as the optimal EOL option for each resulting subassembly. A stream of PCs received at a refurbishing company located in Chicago serves as an illustrative example.

In addition to uncertainty and variability in returned product quantity, age, and quality, incoming feedstock is often varies widely and includes several different product. This further hampers the profitability of product take-back operations. Two types of decisions are considered for multiple returned product streams in Behdad et al. (2010); how to efficiently perform selective disassembly operations, and how best to "mine" the value still embedded in components. An example using two cell phones illustrates the integration of a transition matrix with a mixed integer linear programming model for disassembly operations for multiple products. The solution simultaneously identifies the degree to which each product should be disassembled, and also the optimal end-of-life decision (disposal, reuse, recycle) for each component or subassembly. The example results indicated that sharing disassembly operations between different products can increase the cost-effectiveness of disassembly operations.

22.5 The Need to Use New Data Collection Tools

So far, we have discussed mathematical models for including new, conflicting objectives and for the considering sustainability issues throughout the entire life span of a product. This section discusses the role of data collection tools in gathering the information that these expanded models require. The focus of our discussion is on the role of Immersive Computing Technology (ICT) and virtual reality environments employed during the design process. ICT provides a relatively inexpensive technology (compared with multiple iterations of physical prototype testing) that allows designers to create virtual prototypes of design concepts, generate new ideas, test design concepts, evaluate them, and collect data. ICT has also has been used as a means of sharing information and experience among users. Research on ICT first was mainly focused on developing the technology, but more recently applications of the technology have also generated interest (Ong and Nee 2013), broadly ranging from engineering design through education, gaming, and other entertainments.

While the technology has advanced to the point that it offers many capabilities for data collection, product-user experiences, and visualization of artifacts, there is still much to be done towards improvements in the cognitive aspects of knowledge representation and processing. Advances in ICT need to be expanded to include mental models that designers employ, including how information provided by the ICT is perceived, processed, and acted upon.

Considering the example of design for disassembly, often the collection of the required information about the disassembly process is very challenging and time-consuming. Disassembly is an integral part of remanufacturing, repair and reuse activities, and disassembly times, cost and possible damage can vary from one

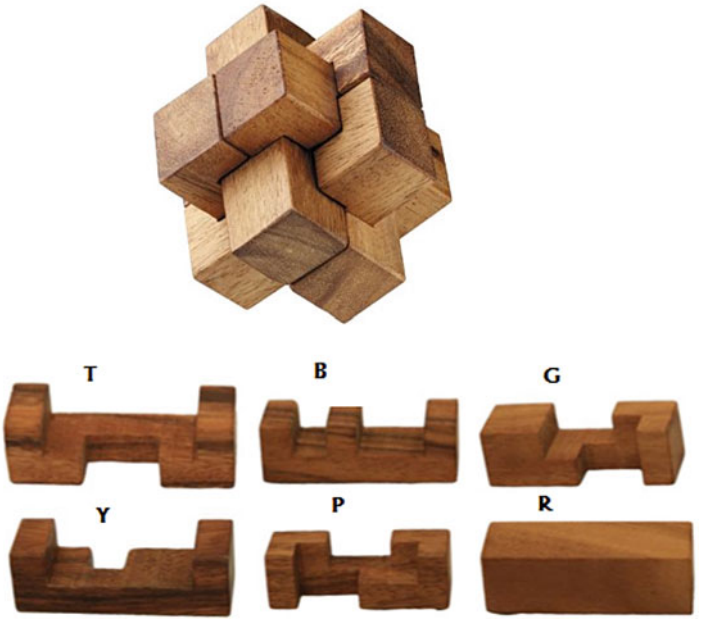


Fig. 22.4 A burr puzzle with six interlocking pieces

operation to another based on component and subassembly geometry, fastening methods, component condition, operator motions required, operator experience, and other factors. Designers most often do not have data on the probability distributions of disassembly times, costs, and possible damage to components during the disassembly process. Having access to such information during the early stages of product design can help designers create alternatives (e.g., the type of fasters, the number of joint parts) that facilitate particular disassembly sequences. To identify the best disassembly sequence (e.g., minimizing disassembly time and/or component damage), designers need to know the probability distribution of disassembly time and/or component damage during disassembly as a function of design and disassembly specifications.

Such data can be made available through the use of an ICT system, in which designers have the opportunity to create a virtual prototype of the product, disassemble the prototype, and collect the required data. In one study (Behdad et al. 2014a), we formulated a mixed integer nonlinear program equipped with data collected using ICT to determine an optimum disassembly sequence for a burr puzzle.

The six-component burr puzzle employed is shown in Fig. 22.4 has properties making it ideal for analyzing assembly and disassembly processes. Not all components can be moved at all times, and there are precedence relationships. Also, many movements are orthogonal to others. Each component is labelled according to its color in the ICT: (T)eal, (B)lue, (G)reen, (Y)ellow, (P)urple, and (R)ed.

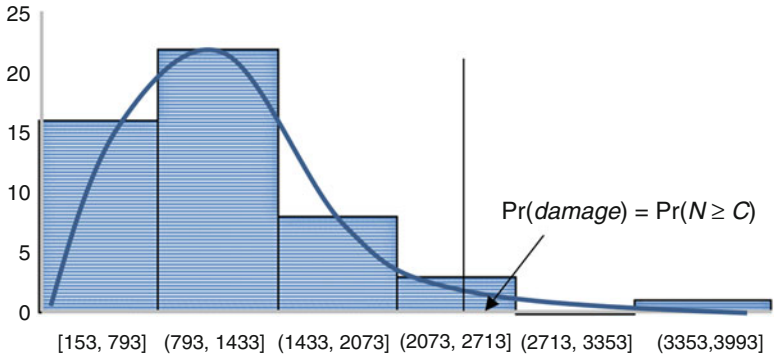


Fig. 22.5 The distribution of the number of collisions between parts

The first step was to determine all feasible disassembly sequences and present them in the form of a disassembly graph. The next step was to supplement the disassembly graph with some sort of cost structure (e.g., disassembly time, component damage for each disassembly operation). The third step was to obtain the data required by simulating the disassembly operations in the ICT environment, and the final stage was to use the data collected in a mathematical model in order to determine the optimum disassembly sequence.

We simulated feasible disassembly operations for the burr puzzle and conducted 30 trials of each feasible disassembly transition manually to collect 30 data points for each disassembly operation. The number of collisions between components was recorded, as a representative of component damage. This data enabled us to draw statistical distributions of the number of collisions between parts (N) for each disassembly operation, shown in Fig. 22.5.

The mathematical model developed in this study was a chance constrained program with the single objective of minimizing the number of collisions between different components during disassembly. The mean and standard deviation of the distribution of the number of collisions collected in the ICT was used to estimate the probability of damage. The distribution of the number of collisions was used to estimate the probability of damage directly. For example, if a threshold (C) is defined for the acceptable number of collisions between parts, using the distribution of the number of collisions and the defined threshold, the probability of components damage can be quantified as follows:

$$\Pr(\text{damage}) = \Pr(N \geq C)$$

Figure 22.6 shows the optimum sequence (1–2–3–5–7–8) derived from the mathematical model for the example of burr puzzle.

In addition to data collection, we also employed ICT capabilities to explore intuitive disassembly sequences and compare the results with the optimum disassembly

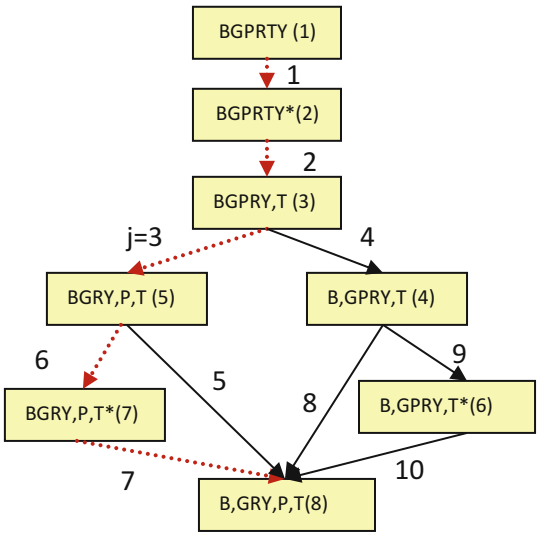


Fig. 22.6 The disassembly graph and the optimum sequence derived from the mathematical model (from Behdad et al. (2014a))

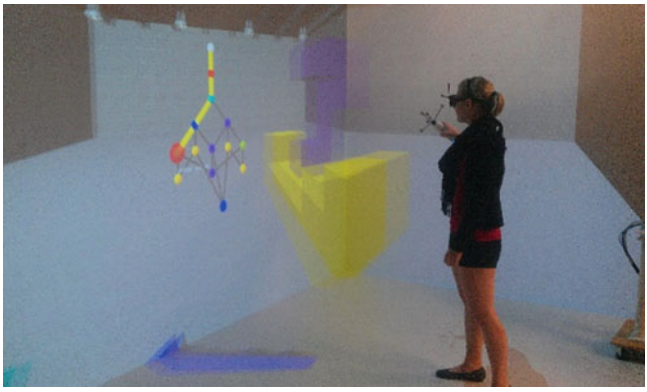


Fig. 22.7 An ICR experiment where a user disassembles a burr puzzle. The experiment has been simulated using the VR facility in Mechanical Engineering department at Iowa State University (courtesy of Dr. Judy Vance and Leif Berg)

sequence obtained above. The goal was to determine if there could be some synergy between the mathematical model and intuitive human expertise.

An experiment was conducted in the ICT environment as shown in Fig. 22.7 to determine what disassembly process a human might employ while "seeing" both the burr puzzle and its disassembly graph. The result of the experiment indicated that the user intuitively followed a disassembly path (1–2–3–5–8) that was different from the optimal path identified by the model.

For example, considering State 3 in the disassembly graph presented in Fig. 22.6. There are two options: remove Part B or remove Part P. In practice, the disassembly processes for these two parts are not the same. While Part P can be removed using one single horizontal manipulation, Part B requires manipulation across two axes,

however, this is not visible to the operator. If the optimum path is followed, then Part P is separated. From State 5, removing Part B appears to be a simple and intuitive one-piece removal operation; however, the optimal sequence suggests an intermediate operation in which Part P is reoriented to be visible to the operator and reveals the physical constraint for removing Part B to the operator. Thus, the fact that the intuitive path differs from the optimal path provides insights for redesign. Realigning the R component could provide a better operator perspective of B's interconnectedness. When the operator then removes B and is aware of the interconnectedness, greater care may be taken and damage minimized.

In a related study (Behdad et al. 2014b), we extended our data collection efforts in the ICT environment to collect not only the number of collisions as a representative of the probability of damage, but also the distance movement of parts to estimate the disassembly cost. We also developed a new multi-objective mathematical model using a multiattribute utility function as the objective function of a dynamic programming model. The probability of damage vs. disassembly time (or cost) are the two conflicting objectives. The faster is the pace of disassembly, the lower is the disassembly time (or cost) but the higher is the probability of damage and vice versa.

22.6 The Need to Investigate the Complex Role of Consumer Behavior

Consumer behavior plays a critical but understudied role in design for sustainability. Traditional views of the sustainable design, including many of those in the design for X domain, largely focus on improvement of end-of-life remanufacturing activities such as disassembly (Harjula et al. 1996) refurbishment (Nee 2015) and recycling (Gaustad et al. 2010), but fail to comprehensively consider the role of consumer behavior during product usage and end-of-use. The effectiveness of sustainable design policies depends on consumers' attitude and behavior, which are critical and are uncertain in nature.

Overall, there is a linkage between the product life cycle, design features, and consumer decisions to repair, reuse, or disposed of a product. Figure 22.8 illustrates the overall process of consumer decision-making during the product life span and the impact of consumers' behavior on usage and storage times.

It is important to develop analytical models that quantify the heterogeneity of consumer behavior and the deterioration of product-critical components, as well as integrate consumers' usage and disposal behaviors into design decisions.

Despite recent efforts to consider consumer marketing preferences, less attention has been paid to modeling the usage behavior of the consumers. Most of the previous studies have strived to address purchase or adoption behavior (Ma and Nakamori 2009; Miller et al. 2013; Davis et al. 2009; Haghnevis et al. 2016). Therefore, one research gap in the current engineering design literature is the lack of scientific methods for considering the heterogeneity and variability of

consumers' usage behavior during product design, or accurately quantifying the resulting environmental impacts. Future methods should make designers capable of calculating the total emissions of a system based on the aggregation of the micro-decisions made by individual consumers for individual products.

In this section, we discuss the importance of consumer behavior in reducing the sustainability-related issues during the entire product life cycle. The focus will be on four types of behavior based on several previous studies conducted by the authors: (1) consumer's product storing behavior at the end-of-use phase of the products, (2) usage behavior, (3) repair behavior and consumers decisions in extending product life cycle, and (4) return behavior and consumers decisions about timing, disposal, and waste removal channel for used products.

22.6.1 Consumer's Product Storing Behavior

Consumers often have a tendency to keep old products in storage before returning them back to waste removal chains. This behavior results to further technological obsolescence of used products where in many cases products that are finally returned are not resaleable in the second-hand market and must be shredded and sent to recycling and material recovery operations.

In one study (Sabbaghi et al. 2015a), data from about 10,000 used computers and laptops returned during 2011 and 2013 to an e-waste collection site located in Chicago, IL have been analyzed to determine the average storage times for used products and whether there are any connections between different design features and how long consumers kept their used electronics in storage. With the help of SMART software, we were able to retrieve data identifying the last time the operating system was used on each device, and since we also had information on the manufacturing date and return date, we were able to calculate the storage time (Storage time = Return date − Last time the OS was used), the upper bound of usage time (Usage time = Last time the OS was used − Manufacturing date), and the product age (Age = Return date − Manufacturing date). We also had information about the size of the hard disk drive, the brand, and the consumer type. We had information about two groups of consumers: individual households and commercial consumers (corporate) organizations who returned their used computers

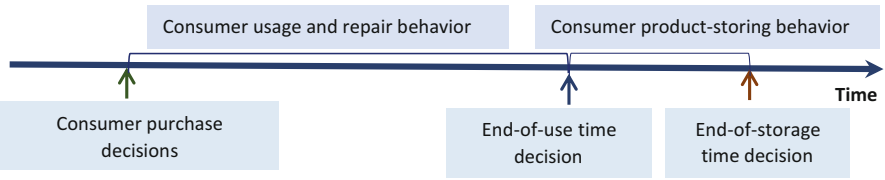


Fig. 22.8 The consumer decisions during the product life span (e.g., purchase, usage, repair, disposal)

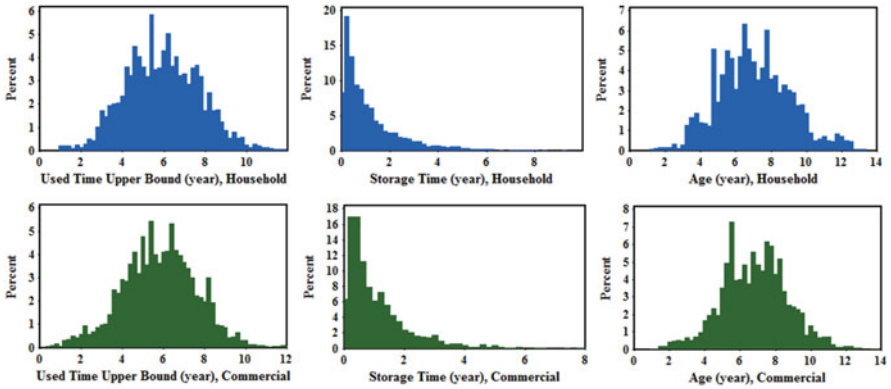


Fig. 22.9 The statistical distributions of the upper bound of usage time, storage time, and product age for two different groups of consumers (data from Sabbaghi et al. (2015a))

for recycling. Various statistical analyses were performed on the dataset and various insights have been derived.

It was found that the average product age and the average storage time was 6.9 years and 1.1 years, respectively. There was no statistically significant difference between different brands and sizes of the hard disk drives, but commercial users stored used electronics longer than households. Data security concerns and administrative processes might be the cause of this. Figure 22.9 shows the histograms of storage time, usage time, and product age for two groups of consumers.

22.6.2 Consumer's Product Usage Behavior

Another important consumer behavior is their usage behavior. The way that consumers use their products can influence sustainability outcomes such as energy consumption and future reusability. For example, the way that consumers charge and discharge their laptop batteries impacts the future reusability of used laptops, and this behavior varies significantly from customer to customer. This uncertainty must be considered. To shed light on the impact of consumer usage behavior, in one study (Sabbaghi et al. 2015b), about 500 same-brand laptop batteries used by students in a high school in Burbank, IL were studied over 3 years of usage to monitor the number of used cycles for each battery. The purpose of the study was to predict the future reusability of batteries based on age, battery type, number of used cycles, and consumer groups (e.g., students of the class of 2012, 2014, 2015, and 2016). For example, Fig. 22.10 illustrates the future reusability of batteries based on their age and the number of used cycles. Finally, we have determined what would be the most profitable end-of-use option (e.g., reuse, recycle, refurbish) for each battery based on their profile of usage.

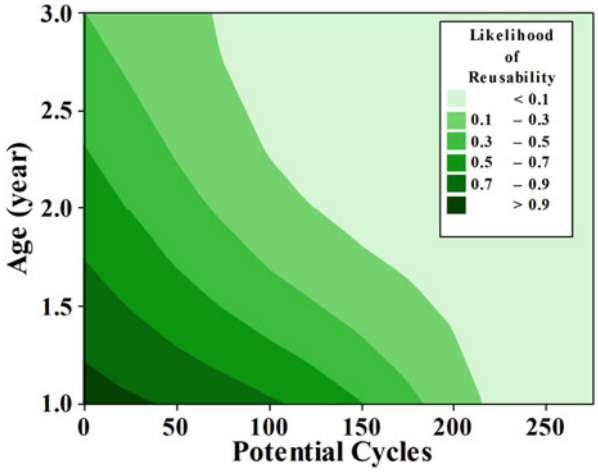


Fig. 22.10 The reusability of batteries based on their age and the number of potential used cycled (data from Sabbaghi et al. (2015b))

22.6.3 Consumer's Repair Behavior

The shortcomings of traditional design policies adopted by manufacturers are especially acute in the design for repairability domain (Cairns 2005). While different design concepts ranging from design for disassembly (Boothroyd and Alting 1992), reliability (Crowe and Feinberg 2001), reuse (Cowan and Lucena 1995), and recycling (Kriwet et al. 1995) have recently become better integrated with design efforts, the concept of "design for repair" has been overlooked. Manufactures in general do not consider "design for longer lasting products" a profitable strategy. In fact, manufacturers' policies are sometimes focused on making products difficult and expensive to repair and reuse with the aim of creating more market share for newly developed products (Cooper 2012).

In one study (Sabbaghi et al. 2016), an industry dataset was analyzed to extract consumer viewpoints towards product repairability, and the factors that make it difficult for consumers to repair products themselves under three main categories of product, economic, and consumer-related factors. The final objective was to test whether and how product repairability might influence consumers' future purchase choices and recommendations to family and friends. An online survey was conducted in collaboration with iFixit.com. iFixit has provided an initial dataset of around 11,500 respondents being surveyed from three different subject groups, including individual consumers (about 8000 respondents), employees of repair shops, and employers of repair businesses. This comprehensive iFixit survey includes a total 27 questions.

Among all 27 questions included in the iFixit survey, two directly inquire about the importance of product repairability, its associated cost, and the effect

Table 22.1 The relation between CLL and PRL given the prior repair experiences (data gotten from Sabbaghi et al. (2016))

		CLL: Consumer Loyalty Level		
		Low 379 (5%)	Medium 4039 (48%)	High 3985 (47%)
PRL: Purchase recommendation level	Low1086 (13%)	163(2%)	613(7.3%)	310(3.7%)
	Medium4874 (58%)	179(2.1%)	2688(32%)	2007(23.9%)
	High2418 (29%)	37(0.4%)	738(8.8%)	1668(19.8%)

on future purchase decisions. Measured in an ordinal scale, two questions were asked of iFixit survey respondents: "If you successfully repaired a product, are you more likely to buy new products from the same company in the future?" (CLL: Consumer Loyalty Level: low, medium, high), "Have your experiences fixing your own products impacted the purchasing recommendations you give to your friends? (PRL: Product Recommendation Level)". Assuming there is a significant correlation between these two questions, a bivariate ordered probit model was employed to estimate the probability that an observation (a consumer) with specific characteristics (repair experience) will come under one of the ordered categories (low, medium, and high loyalty and recommendation levels). Table 22.1 shows the relation between consumer loyalty and recommendation level for consumers with prior repair experiences.

Access to repair information, positive attitudes towards repairing electronics, product type, availability of spare parts, and unsuccessful repair experience (time-consuming repair and broken parts during repair) have been identified as important factors that influence future product choice and recommendation.

22.6.4 Consumer's Product Return Behavior

Another important consumer behavior is their decision about the type of disposal channel to employ, including storage, reselling, the trash bin, a formal collection site, trade-in programs, recycling centers, etc. It is important to estimate which portion of used products are returned back to each of these channels. Thus, in one study (Mashhadi et al. 2016), we developed a simulation framework to model consumer's behavior in returning used products in which consumers have four options: storage, resell, recycle, or discard. An agent-based simulation (ABS) framework integrated with a Discrete Choice Analysis (DCA) method was developed to predict consumer's disposal decisions. Consumers socio-demographic information and product design features were included in the model as possible predictive factors.

To identify the role of product design features on the consumer participation, an agent-based simulation model was created. The diverse set of decision-makers in

the take-back systems (consumers, products, OEMs) were represented as "agents." There is a mathematical model behind each agent. All agents have their own set of features, objectives, behavioral patterns, and decision-making rules.

The main objective of the simulation was to evaluate design alternatives in terms of consumers' participation in different take-back channels. Each individual consumer has been modeled as an agent. Consumer decisions on the selection (if any) of take-back systems (trade-in programs, store, trash bin) has been modeled using discrete choice analysis techniques and has been connected with product design strategies controllable by original manufacturers. DCA is commonly employed to study the individual decision-making process. The underlying assumption behind DCA is that individuals seek to maximize their utility considering two sets of attributes, their socio-demographic characteristics, and the features of alternatives available to them (Wassenaar and Chen 2003).

Through the DCA techniques, the choice probability for each decision (e.g., trade-in programs, trash bin, store) was determined based on product design attributes, take-back program features, and consumers' socio-demographic information. The capabilities of agent-based simulation also helped us consider the impact of other dynamic factors such as peer pressure, and consumer awareness of the decision made by each individual consumer. The interaction between agents has been modeled employing the capabilities of agent-based simulation. Figure 22.11 illustrates one example of the simulation results including the number of products stored, returned, sold, and thrown away over time with consideration of interactions between consumers.

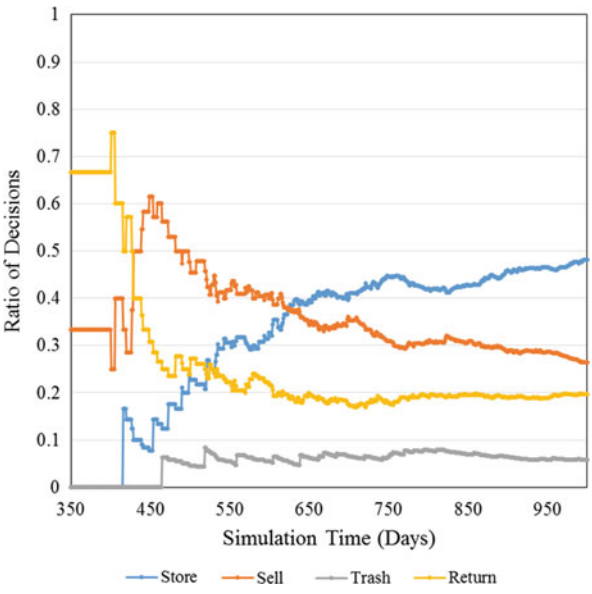


Fig. 22.11 The ratio of products stored, returned, sold, and thrown away over time with consideration of interactions between consumers (data from Mashhadi et al. 2016)

Building such simulation tools paves the way for achieving a more comprehensive understanding of consumer behavior and its impact on the used products' return paths. The final outcome of the study was to estimate the number of products stored, returned, sold, and disposed of over time.

22.7 Summary and Closure

This chapter has summarized a body of work that employs industrial engineering approaches to the problem of design for sustainability. These approaches began with normative multiattribute utility analysis to broaden the set of objectives under consideration and evaluate design alternatives, proceeded through constrained optimization to identify the best alternative, statistical process control to create new alternatives and immersive computing technology to gather data required for the optimization model, and conclude with studies of consumer behavior before, during, and after the use phase of the product life cycle.

This mathematical modeling approach requires designers to employ a decompositional strategy, which not only aids in solving previously intractable problems, but also facilitates gaining better insight to the problem as one analyzes the results and the structure of the model in detail. Lessons learned include that facts that although designers might be genuinely concerned about sustainability, their actual willingness to pay for it might be limited, that selling a service (through leasing) rather than selling a product can improve both profitability and customer satisfaction, that statistical manufacturing process control can reveal unexpected opportunities for pollution prevention, that virtual reality design environments can be used to quickly and efficiently gather some of the large amount of data needed to build comprehensive models, and that the customer also plays a significant role.

Acknowledgement This material is based upon work supported by the National Science Foundation—USA under grants DMI-9528627, DMI-9908406, DMI-0726934, CMMI-1100177, CMMI-1068926, CMMI-1435908, CMMI-1727190, and CBET-1705621. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Behdad S, Thurston D (2012) Disassembly and reassembly sequence planning tradeoffs under uncertainty for product maintenance. ASME J Mech Des 134(4):41011

Behdad S, Kwak M, Kim H, Thurston D (2010) Simultaneous selective disassembly and end-of-life decision making for multiple products that share disassembly operations. ASME J Mech Des 132(4):41002

Behdad S, Williams AS, Thurston D (2012) End-of-life decision making with uncertain product return quantity. ASME J Mech Des 134(10):100902

Behdad S, Berg LP, Thurston D, Vance J (2014a) Leveraging virtual reality experiences with mixed-integer nonlinear programming visualization of disassembly sequence planning under uncertainty. ASME J Mech Des 136(4):41005

Behdad S, Berg L, Vance J, Thurston D (2014b) Immersive computing technology to investigate tradeoffs under uncertainty in disassembly sequence planning. ASME J Mech Des 136(7):71001

Boothroyd G, Alting L (1992) Design for assembly and disassembly. CIRP Ann Manuf Technol 41(2):625–636

Cairns CN (2005) E-waste and the consumer: improving options to reduce, reuse and recycle. Electronics and the environment, 2005. In: Proceedings of the 2005 IEEE international symposium on, pp 237–242

Camacho-Vallejo J-F, González-Rodríguez E, Almaguer F-J, González-Ramírez RG (2015) A bi-level optimization model for aid distribution after the occurrence of a disaster. J Clean Prod 105:134–145

Carnahan JV, Thurston DL (1998) Trade-off modeling for product and manufacturing process design for the environment. J Ind Ecol 2(1):79–92

Cooper T (2012) Longer lasting products: alternatives to the throwaway society. Gower Publishing, Farnham

Cowan DD, Lucena CJP (1995) Abstract data views: an Interface specification concept to enhance design for reuse. IEEE Trans Softw Eng 21(3):229–243

Crowe D, Feinberg A (2001) Design for reliability. CRC Press, Boca Raton

Cuthbert R, Giannikas V, McFarlane D, Srinivasan R (2016) Repair services for domestic appliances. Springer, New York, pp 31–39

Davis C, Nikolić I, Dijkema GPJ (2009) Integration of life cycle assessment into agent-based modeling. J Ind Ecol 13(2):306–325

Du L, Peeta S (2014) A stochastic optimization model to reduce expected post-disaster response time through pre-disaster investment decisions. Netw Spat Econ 14(2):271–295

Fang C-C, Hsu C-C (2009) A study of making optimal marketing and warranty decisions for repairable products. In: 2009 IEEE international conference on industrial engineering and engineering management. IEEE, Piscatawa, pp 905–909

Faturechi R, Miller-Hooks E (2014) Travel time resilience of roadway networks under disaster. Transp Res B Methodol 70:47–64

Gaiardelli P, Cavalieri S, Saccani N (2008) Exploring the relationship between after-sales service strategies and design for X methodologies. Int J Prod Lifecycle Manag 3(4):261–278

Gaustad G, Olivetti E, Kirchain R (2010) Design for recycling. J Ind Ecol 14(2):286–308

Haghnevis M, Askin RG, Armbruster D (2016) An agent-based modeling optimization approach for understanding behavior of engineered complex adaptive systems. Socio Econ Plan Sci 56:67

Harjula T, Rapoza B, Knight WA, Boothroyd G (1996) Design for disassembly and the environment. CIRP Ann Manuf Technol 45(1):109–114

Kriwet A, Zussman E, Seliger G (1995) Systematic integration of design-for-recycling into product design. Int J Prod Econ 38(1):15–22

Ma T, Nakamori Y (2009) Modeling technological change in energy systems—from optimization to agent-based modeling. Energy 34(7):873–879

Mahmoodjanloo M, Parvasi SP, Ramezanian R (2016) A tri-level covering fortification model for facility protection against disturbance in R-interdiction median problem. Comput Ind Eng 102:219–232

Mangun D, Thurston DL (2002) Incorporating component reuse, remanufacture, and recycle into product portfolio design. IEEE Trans Eng Manag 49(4):479–490

Mashhadi AR, Esmaeilian B, Behdad S (2016) Simulation modeling of consumers' participation in product take-back systems. ASME J Mech Des 138(5):51403

Miller SA, Moysey S, Sharp B, Alfaro J (2013) A stochastic approach to model dynamic systems in life cycle assessment. J Ind Ecol 17(3):352–362

Nee AYC (ed) (2015) Handbook of manufacturing engineering and technology. Springer, London

Ong SK, Nee AYC (2013) Virtual and augmented reality applications in manufacturing. Springer Science & Business Media, New York

Peck D, Kandachar P, Tempelman E (2015) Critical materials from a product design perspective. Mater Des (1980–2015) 65:147–159

Preston F (2012) A global redesign? Shaping the circular economy. Energy Environ Res Gov 2:1–20
Rosner DK, Ames M (2014) Designing for repair? In: Proceedings of the 17th ACM conference on computer supported cooperative work & social computing—CSCW'14. ACM Press, New York, pp 319–331
Sabbaghi M, Esmaeilian B, Raihanian Mashhadi A, Behdad S, Cade W (2015a) An investigation of used electronics return flows: a data-driven approach to capture and predict consumers storage and utilization behavior. Waste Manag (New York, N.Y.) 36:305–315
Sabbaghi M, Esmaeilian B, Mashhadi AR, Cade W, Behdad S (2015b) Reusability assessment of lithium-ion laptop batteries based on consumers actual usage behavior. ASME J Mech Des 137(12):124501
Sabbaghi M, Esmaeilian B, Cade W, Wiens K, Behdad S (2016) Business outcomes of product repairability: a survey-based study of consumer repair experiences. Resour Conserv Recycl 109:114–122
Saccani N, Johansson P, Perona M (2007) Configuring the after-sales service supply chain: a multiple case study. Int J Prod Econ 110(1):52–69
Starita S, Scaparra MP (2018) Passenger railway network protection: a model with variable post-disruption demand service. J Oper Res Soc 69(4):603–618
Stevels A (2002) Integration of ecodesign into business. In: Hundal MS (ed) Mechanical life cycle handbook. Good environmental design and manufacturing. Marcel Deker, New York, pp 583–604
Suh NP (1998) Axiomatic design theory for systems. Res Eng Des 10(4):189–209
Thurston DL (1991) A formal method for subjective design evaluation with multiple attributes. Res Eng Des 3(2):105–122
Thurston DL (2001) Real and misconceived limitations to decision based design with utility analysis. ASME J Mech Des 123(2):176
Thurston DL, De La Torre JP (2007) Leasing and extended producer responsibility for personal computer component reuse. Int J Environ Pollut 29(1–3):104–126
Thurston DL, Carnahan JV, Liu T (1994) Optimization of design utility. ASME J Mech Des 116(3):801–808
Tokunaga T, Fujimura S (2016) A unified theory of design structure matrix and axiomatic design for product architecture. J Mod Project Manag 3(3):114–122
Wassenaar HJ, Chen W (2003) An approach to decision-based design with discrete choice analysis for demand modeling. ASME J Mech Des 125(3):490
Yi W, Nozick L, Davidson R, Blanton B, Colle B (2017) Optimization of the issuance of evacuation orders under evolving Hurricane conditions. Transp Res B Methodol 95:285–304

Deborah Thurston first learned about engineering as a possible career choice at a seminar for high school senior girls that she signed up for mainly to get out of a day of school. There, she heard from a practicing woman engineer about how fun and satisfying it was to *really* understand something that you wanted to understand, but at first found difficult. She had already been accepted to college with a different major, but switched to engineering. She is now a Gutsgell Professor of Industrial and Enterprise Systems Engineering at the University of Illinois. She serves as design node leader in the Institute for Reducing EMbodied-Energy And Decreasing Emissions (REMADE), Director of the Decision Systems Laboratory, and Co-Director of the Hoeft Technology and Management Program. She earned the M.S. and Ph.D. from MIT. At Illinois, Professor Thurston was instrumental in developing a new interdisciplinary Ph.D. program in Systems and Entrepreneurial Engineering, and also

managed the transfer of the Industrial Engineering program during a college reorganization. She has also served as CESUN Chair. As one of the first researchers in engineering design theory and methodology, she brought mathematical rigor to complex design decision problems by formalizing methods for making rational tradeoffs under uncertainty. Her research has been funded by NSF, EPA, and a number of industries. Professor Thurston received the NSF Presidential Young Investigator Award, two Xerox Awards for research excellence, four awards for undergraduate advising, and five best paper awards. She has served as area editor for the *ASME: Journal of Mechanical Design* and *The Engineering Economist*. She is a registered professional engineer and ASME Fellow.

Sara Behdad is Assistant Professor of Mechanical and Aerospace Engineering, and Industrial and Systems Engineering at the University at Buffalo (UB). She received her Ph.D. in Industrial and Systems Engineering from the University of Illinois at Urbana-Champaign in 2013 under the supervision of Dr. Deborah Thurston. She is the founding director of Green Engineering Technologies for the Community of Tomorrow (GETCOT) research lab at UB. Her recent research focuses on data-driven life cycle engineering, design for remanufacturing, design for additive manufacturing, and modeling the impacts of design policies on complex socio-economic systems. Her work has been covered in media in outlets such as PBS, Daily Mail, The Chicago Tribune, and Motherboard. She is the recipient of the 2017 International Life Cycle Academy Award for her contribution to sustainable consumption field. She is also an active member of the ASME International Design Engineering Technical Conferences (IDETC 2018).

Chapter 23
Dynamic Price and Lead Time Quotation Strategies to Match Demand and Supply in Make-to-Order Manufacturing Environments

Esma S. Gel, Pinar Keskinocak, and Tuba Yilmaz

Contents

23.1 Introduction 541
23.2 Overview of Literature on Dynamic Quotation 543
23.2.1 Quoting Lead Times 543
23.2.2 Quoting Prices 544
23.2.3 Joint Price and Lead Time Quotation 546
23.2.4 Inventory Decisions with Lead Time Quotation 551
23.2.5 Lead Time Quotation and Product Substitution 552
23.2.6 Dynamic Quotation and Cancellations 554
23.3 Conclusions 554
References 555

23.1 Introduction

There are a large number of studies in the operations management literature that point to the potential positive impact of dynamic quotation strategies towards matching supply and demand. These strategies target make-to-order environments where a quotation-order mechanism between the buyer and seller is possible, i.e., a buyer requests a quote for an order, the seller provides a quote for the transaction, and the buyer accepts or rejects the quote. Dynamic quotation allows the seller to adjust the terms of this transaction, so that the offer terms, generally price

E. S. Gel (✉)
Arizona State University, Tempe, AZ, USA
e-mail: esma.gel@asu.edu

P. Keskinocak
Georgia Institute of Technology, Atlanta, GA, USA
e-mail: pinar.keskinocak@isye.gatech.edu

T. Yilmaz
Ozyegin University, Istanbul, Turkey
e-mail: tuba.yilmaz@ozyegin.edu.tr

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_23

and lead-time, depend on the current system state. Changing market conditions and information technology advancements enable more flexibility for dynamic quotation, and hence, in this chapter we summarize the research on such strategies and the factors that influence effectiveness of dynamic quotation.

The two main components of a seller's quote are *price* for the product that the buyer is requesting, and the *lead time* that the seller is committing to deliver it by, subject to well-established delivery delay penalty terms. Depending on the context, the seller may be keeping some level of inventory or operating the system in a make-to-order or hybrid mode, which may bring up the question of how inventory levels should be set, given the quotation-order mechanism. If the seller produces a set of substitutable products, she may present additional quote(s) to the buyer for the other product(s) that the buyer may potentially be interested in.

Price and lead time quotation decisions are interdependent. A seller who quotes a low price for an order should be sufficiently confident that she can deliver the order "on-time," so that delivery delay penalties, which may render the order "unprofitable" for the seller, can be avoided. Early papers in the area have generally ignored this dependency, due to the fact that price and lead time decisions have been traditionally made by separate departments of companies (e.g., marketing and manufacturing groups) without coordination or information sharing. Therefore, academic papers following this tradition studied decisions from one of these departments' perspectives, looking at either (dynamic) price or (dynamic) lead time quotation decision. More recent papers have relaxed this assumption and explored the dependency and possible synergy between price and lead time decisions, also paying attention to the interaction and information flow between potentially separate entities making these two decisions.

It is important to note that the quotation-order framework that we describe here places the authority to accept or reject the transaction on the buyer. This implies that models that analyze this framework require a careful representation of the buyer's behavior in response to the quote, or after the acceptance of the quote, while the order is being processed/produced. Buyer's behavior in response to the quote (i.e., price and/or lead time presented in the quote) is generally modeled by an *acceptance probability function*, which indicates the probability that the buyer will accept a quote for a given price, p and/or lead time, l, where the possible values of the price and lead time quotes (i.e., minimum and maximum values) are defined. Numerous studies have shown that the benefits of dynamic quotation depend on the assumed form and parameters of the acceptance probability function that represents the buyer population's behavior, indicating that sellers who want to use dynamic quotation mechanisms need to develop relatively accurate characterization of buyer behavior.

In addition to buyers who are served using dynamically changing prices and/or lead times (e.g., in a spot market mode), the seller may be serving other types of buyers; for example, seller may have more stable buyers whose terms of engagement are determined through contractual agreements. It is important to study the effectiveness of dynamic quotation in settings with multiple types of buyers with differentiated or tiered access to the services offered by the seller. The use

of different channels (e.g., contracts and spot markets) allows the seller to take advantage of the stability of one channel while being able to take advantage of the ability to quote prices and lead times dynamically in the other channel, based on system status.

Dynamic quotation strategies are, in general, harder to implement and maintain than using "static" or constant price and/or lead times. For example, in many contexts, it is not clear how buyers may react to being quoted changing prices for the same product; it is harder for buyers to estimate costs of procurement, for example, since price and lead times may be higher (or lower) depending on the congestion level in the seller's system. Buyers may have increased expectations on delivery, quality, etc., depending on the prices quoted and paid for the order. Buyers' behavior in response to quotes may change over time, making it harder to develop accurate characterizations of acceptance probability functions. Competition for the buyers' orders may further complicate the price and lead time decisions of the seller.

Due to these and other reasons, the decision to implement dynamic quotation strategies need to be justified through a careful study of the performance improvement that dynamic quotation strategies can possibly offer. Performance may be measured by the long-run average profit (i.e., revenue minus delivery delay penalties) or costs, and improvement opportunity can be calculated by comparing the performance of (optimal) dynamic quotation strategies to that of the best "static" or constant price and/or lead time policy. An exploration of the synergy between dynamic quotation of prices and lead times is useful to understand if it is possible to only change one (price or lead time) dynamically, while keeping the other constant. Previous work in the area has shown that accurate characterization of buyers' preferences is very important in this respect; settings where buyers are more sensitive to the changes in price can be improved by quoting lead times dynamically, whereas settings where buyers are more sensitive to changes in lead time can exploit the use of dynamically changing prices.

The next section gives a general overview of the literature in this domain by providing a sample (rather than comprehensive) set of papers and the questions they seek to address. Fundamental results from some of these papers and our previous work are highlighted with the purpose of giving the reader an understanding of the important factors that make dynamic quotation strategies desirable. We conclude with potential on future research directions.

23.2 Overview of Literature on Dynamic Quotation

23.2.1 Quoting Lead Times

There is a large volume of work on setting lead times, or equivalently, *due dates* in production environments where, generally, it is assumed that all orders arriving to the system must be served. Cheng and Gupta (1989), Koulamas (1994) present

surveys of the scheduling literature that considers due dates. The broader area of *due date management* has addressed the problem of setting due dates along with other important questions such as lot sizes, inventory, and order sequencing. For detailed surveys of research on due date management, see Keskinocak and Tayur (2004), Kaminsky and Hochbaum (2004).

Papers that employ admission control strategies in combination with due date decisions allow the possibility of not serving incoming customer orders; examples include Duenyas and Hopp (1995), Duenyas (1995), Keskinocak et al. (2001), Charnsirisakskul et al. (2004), Charnsirisakskul et al. (2006), Ata (2006), Kapuscinski and Tayur (2007), Ata and Olsen (2009), Savasaneril et al. (2010), and Zheng et al. (2014). Among these studies, Keskinocak et al. (2001) and Charnsirisakskul et al. (2004, 2006) assume deterministic processing times, whereas Ata (2006) and Ata and Olsen (2009) employ heavy-traffic approximations to overcome the complexity of Markov Decision Process (MDP) formulations.

Modeling a make-to-order system as an $M/M/1$ queue and formulating the dynamic lead time quotation problem as an MDP, Duenyas and Hopp (1995) and Duenyas (1995) show that the lead time quotes should increase as a function of the number of orders in the system. Kapuscinski and Tayur (2007) use a finite-horizon discrete-time model of setting lead times in a make-to-order setting with demands from two different classes of customers. Savasaneril et al. (2010) extend the dynamic lead time quotation problem to a hybrid make-to-order/make-to-stock environment, i.e., a basestock inventory system characterized by lead time sensitive demand. They show that the lead time quotes are lower in a make-to-order system compared to a basestock system.

More recent work in this vein has continued deepening the characterizations of buyer behavior and response to quoted lead times, multiple buyer types, other relevant decisions such as scheduling and inventory levels (Kaman et al. 2013; Altendorfer and Minner 2015; Kahvecioglu and Balcioglu 2016; Nakade and Niwa 2017; Savasaneril and Sayin 2017).

23.2.2 Quoting Prices

There is even a larger volume of work on dynamic pricing in various different contexts; earlier work in this domain, and in particular, dynamic pricing decisions as related to inventory management practices, is summarized in Elmaghraby and Keskinocak (2003). Since then, dynamic pricing literature has continued to grow in many directions, in line with the growing interest in dynamically changing prices of products and services and the various capabilities that has become a reality through availability of supporting information technologies. Chen and Chen (2015b) provide a more recent review of the literature in this domain, with emphasis on problems with multiple products, competition, and contexts with limited demand information.

Several recent studies on dynamic pricing in make-to-order systems consider the option of accepting or rejecting price sensitive customers (Yoon and Lewis

2004; Son 2008; Aktaran-Kalayci and Ayhan 2009; Cil et al. 2009; Yildirim and Hasenbein 2010; Cil et al. 2011; Afeche and Ata 2013). A large number of these studies use admission control type MDP models, and present structural properties of optimal policies, such as the monotonicity of price quotes.

Yoon and Lewis (2004) study the pricing problem in non-stationary queues considering both total discounted reward and average reward objectives for an infinite horizon formulation. Similar to the findings on lead time quotation, they show that the optimal price and admission probability of an order is increasing and decreasing, respectively, in the number of orders in the system. Aktaran-Kalayci and Ayhan (2009) extend the monotonicity results to $M/M/s/K$ queues.

Cil et al. (2009) present a general framework of queuing admission control methods and structural properties. Cil et al. (2011) investigate an optimal sequencing and dynamic pricing problem for a two-class queueing system, and show that static pricing policies may have significant suboptimality. Afeche and Ata (2013) consider a dynamic pricing problem with impatient and patient customer classes in a setting where the proportion of these customers is not known. The authors show that the optimal Bayesian pricing policy has a nested-threshold structure and is sensitive to the maximum queue length.

Recent developments in dynamic pricing literature, in particular with multiple products, competition and limited demand information is reviewed in Chen and Chen (2015a). Other recent papers in this area include Liu et al. (2012), Lu et al. (2013), Chen et al. (2014), and Ardjmand et al. (2016).

Dynamic Pricing and Learning

The relationship between demand and price is assumed to be known to the seller in dynamic pricing problems, and this relationship defines the optimal prices according to the system state and the remaining time in the selling horizon. In practice, customer behavior in response to price changes is not available. As the seller quotes prices and observes customer behavior according to changing prices, seller can learn about the price and demand relationship and quote optimal prices according to this relationship.

den Boer and Zwart (2015) study dynamic pricing problem with finite inventory and use maximum likelihood estimation for demand parameter estimations. They show that if the prices seller quotes do not deviate much from optimal prices, then the relationship between demand and price is learned very fast, i.e., endogenous learning property. Cheung et al. (2017) argue that there are business constraints which limit extensive experimentation in practice. They assume that demand function is unknown but belongs to a finite set and seller can make limited number of changes in price in a given amount of time. The seller's trade-off is between adjusting prices to learn demand and optimize prices to maximize revenue. den Boer (2015) reviews dynamic pricing and learning literature with perspectives from different areas.

23.2.3 Joint Price and Lead Time Quotation

While less common than their counterparts that only consider dynamic quotation of lead times or prices, studies on joint lead and time price quotation have been considered in various contexts. In the following paragraphs, we outline the literature in this domain.

Both Price or Lead Time Fixed

Palaka et al. (1998) and Easton and Moodie (1999) consider make-to-order environments where demand is a function of price and lead time, and optimize expected profits by setting a fixed price and lead time. Fixed price and lead time quotation under decentralized supply chains is considered by Liu et al. (2007), Pekgun et al. (2008), Hua et al. (2010), Zhu (2015a) and Xiao et al. (2010), who discuss settings where price and lead times are determined by separate entities within the company. Xiao and Qi (2016) consider the impact of delivery reliability in addition to price and the announced delivery time in a two-stage supply chain, and focus on coordination of decisions of the manufacturer.

Either Price or Lead Time Fixed

Joint price and lead time decisions, where at least one of these decisions are constant over time, are considered by Palaka et al. (1998), Easton and Moodie (1999), ElHafsi (2000), Plambeck (2004), Ray and Jewkes (2004), Liu et al. (2007), Pekgun et al. (2008), Hua et al. (2010), Xiao et al. (2010), Chaharsooghi et al. (2011), Zhao et al. (2012), and Wu et al. (2012). Depending on the nuances of the problem that the authors consider, these studies present formulations with a variety of assumptions on customer response to changing prices and lead times, the departments making price and lead time decisions, types of customers and market segmentation, timing of price and lead time decisions, etc.

Price and Lead-Time Menus

Zhao et al. (2012) discuss the benefits of a uniform price and lead time policy in comparison to a differentiated policy where a menu of price/lead times are offered to customers. Akan et al. (2012), Ata and Olsen (2013), and Afeche and Pavlin (2016) assume customers are quoted price and lead-time menus. Akan et al. (2012) propose a cost-balancing policy for optimal dynamic choice of lead-time and price menus for customers with convex-concave delay costs. Ata and Olsen (2013) make heavy traffic assumptions and give an asymptotically optimal policy for customers with convex-concave delay costs. Afeche and Pavlin (2016) study the design of joint

price and lead-time menu and the corresponding scheduling policy problem within a queuing model, and allow lead-time dependent ranking of customer types.

Joint Dynamic Price and Lead Time Quotation

Setting prices and lead times jointly has been considered from the viewpoint of coordinating decisions made by different divisions in a firm, e.g., by marketing and manufacturing divisions, using various game theoretic (e.g., Stackelberg game) models (e.g., see Zhu 2015b; Ye et al. 2016), the following papers focus more on setting prices and lead times through the use of an MDP model, in a similar sense to admission control models, but rather than accepting or rejecting customers, dynamically changing prices and lead times considering the current system state as well as the response of customers to changing prices and lead times.

Celik and Maglaras (2008) study dynamic price and lead time quotation under heavy traffic assumptions, and discuss the effects of lead time flexibility, expediting orders subject to a high cost, and dynamic pricing. Feng et al. (2011) address dynamic price and lead time quotation in a make-to-order system using an $GI/M/1$ queuing model and an MDP formulation. They define an optimal policy structure including a threshold and reward maximizing lead time quote, and show that the latter is optimal under particular conditions.

Bekki and Gel (2005) study the quotation-order mechanism we described in Sect. 23.1; upon the arrival of a customer order, the following sequence of events happen: (1) the firm determines a price and/or lead time (i.e., due date) quote for the order and informs the customer, (2) the customer decides whether to accept the order or reject the order, and (3) if the customer accepts the price and lead time offer, processing of the order is performed in a first-come-first-serve sequence. If the item is available later than the due date quoted to the customer, a lateness cost is incurred per unit time. For any unassigned item that waits in inventory, a holding cost is incurred per unit time. The profit from a particular order is equal to the revenue received from the order (which is a fixed revenue, or the price quoted) minus the holding and late delivery penalties. If a customer is quoted lead time d and price p, then he places an order with probability $f(d, p)$, which is referred to as the acceptance probability function and is assumed to be decreasing in d and p. Furthermore, it is assumed that there exists a minimum price and lead time that the customer always (with probability 1) accepts, and maximum price and lead time that the customer always (with probability 1) rejects. Hence, quoting higher than the maximum acceptable price or lead time is equivalent to rejecting the order.

Bekki and Gel (2005) consider the problem in the context of a WIP-constrained make-to-order environment without capability to hold finished goods inventory. In particular, we model the production system as an $M/M/1$ server, as in So and Song (1998). The problem is formulated as an infinite horizon continuous-time MDP, with expected total discounted profit criterion. The state of the system is described by (n, i), where n denotes the number of jobs in the system, $n \in \{0, 1, \ldots, N\}$ and i denotes the type of the most recent event (order arrival or completion). At

each decision epoch, a price and lead time combination (a_{k1}, l_{k2}) is selected by the manufacturer, where $a_{k1} \in \mathscr{A}(n,i)$ and $l_{k2} \in \mathscr{L}(n,i)$ when the system is in state (n,i). The expected immediate reward for the state (n,i), when the action (a_{k1}, l_{k2}) is chosen, is equal to the expected revenue obtained from the order minus the expected holding and tardiness costs incurred for the order.

$$R(n,i,a_{k1},l_{k2}) = p(a_{k1},l_{k2})[a_{k1} - bE(CT_{n+1} - l_{k2})^{+} - hE(CT_{n+1})]\,, \quad (23.1)$$

where CT_{n+1} denotes the cycle time of the arriving order if it enters the system, b denotes the tardiness penalty per unit time of delay, and h denotes the unit holding cost per unit time. Note that $E(CT_{n+1} - l_{k2})^{+}$, $E(CT_{n+1})$ can be computed in closed form for the M/M/1 queue. The optimality equations for the (uniformized) dynamic price and lead time quotation problem are

$$\begin{aligned} V(n,i) = \max_{\substack{a_{k1}\in\mathscr{A}(n,i)\\ l_{k2}\in\mathscr{L}(n,i)}} \Big\{ R(n,i,a_{k1},l_{k2}) + \alpha' \Big[p(a_{k1},l_{k2}) \Big(\frac{\lambda}{\nu} V(n+i,1) \\ + \frac{\mu}{\nu} V(n+i-1,0) \Big) + \Big(1 - p(a_{k1},l_{k2})\Big) \Big(\frac{\lambda}{\nu} V(n,1) \\ + \frac{\mu}{\nu} V(n-1,0) \Big) \Big] \Big\}\,, \end{aligned} \quad (23.2)$$

where ν denotes the uniformization constant, $\nu = \mu + \lambda$. The value function is nonincreasing in the number of jobs in the system. Furthermore, it is possible to define a dominance relation between two price-lead time pairs, say, (a_1, l_1) and (a_2, l_2). Since the value function is nonincreasing in n, if $p(a_1, l_1) \geq p(a_2, l_2)$ and the immediate return for (a_2, l_2) is bigger than that for (a_1, l_1), i.e., $R(n, 1, a_2, l_2) \geq R(n, 1, a_1, l_1)$, then we can conclude that action (a_2, l_2) dominates action (a_1, l_1). Using this dominance relation, it is possible to determine cutoff levels for price and lead time pairs. We relabel actions which are price and lead time pairs. We rank possible pairs in their nondecreasing acceptance probability values. Let X_k denote the pair with the k^{th} highest acceptance probability. (For every congestion level, $0 \leq n \leq N$, we only need to consider price and/or lead time levels above the cutoff levels (Bekki and Gel 2005).)

Hence, we can show that for every congestion level, i.e., number of jobs already waiting to be processed, one only needs to consider price and/or lead time levels above a cutoff level, which is monotonically increasing. While cutoff levels can be calculated for each state in a short time and reduce the size of the action space dramatically, the use of the identified cutoff levels as a heuristic (rather than computing the optimal dynamic price and lead time policy) yields variable performance. While it performs almost optimally for certain cases (with low inventory buffers, and high number of price/lead time alternatives), at other times it performs worse than using fixed prices and lead times.

Given that the determination of optimal dynamic price and lead time policy can get computationally prohibitive, and simple heuristics such as the cutoff level

heuristic may not always work better than the practice of using fixed prices and lead times, it is important to know how much of a performance improvement (in terms of the expected total discounted profit) optimal dynamic price and/or lead time quotation offers. The numerical study given in Bekki and Gel (2005) indicates that there is strong synergy between price and lead time decisions, and the performance improvement can get as high as 12% for some instances. The numerical study further indicates that the use of dynamic price and/or lead time quotation policies does not deteriorate the long run average service level; on the contrary, by effectively controlling the congestion in the system, the firm is able to generally quote reasonable prices, which increases the long run average fraction of customers served by 1.0–6.0%. Hence, price and lead time quotation may even improve customer service and satisfaction in the long run.

More recently, Öner-Közen and Minner (2017) have considered a similar setting and analyzed the performance of a make-to-order manufacturing firm that dynamically quotes a price/leadtime pair and dispatches orders accepted by customers. Similar to our findings, the authors have shown that when tardiness of orders is penalized with a fixed cost and the customers differ significantly in their sensitivity to price and leadtime, the improvement offered by dynamic price and lead time quotation can be considerable. Other recent work in this area include (Albana et al. 2018; Altendorfer and Minner 2014).

Hybrid Systems with Multiple Customer Types/Channels

In many cases, it may be more appropriate to employ a hybrid strategy where some customers are served at fixed prices and lead times (due to, for example, contractual agreements) and while other customers are served using dynamically quoted prices and lead times. Such a hybrid strategy has the potential to take advantage of both channels through the "stability" that contracted customers may provide as well as the "opportunistic" serving of spot purchasers through dynamic prices and lead times at times of low system congestion.

Plambeck (2004) considers two customer classes with different price and lead time sensitivities. Price decisions are made at the beginning of the planning horizon, and lead times are quoted dynamically to the arriving customers.

In Hafızoğlu et al. (2016), we further explore joint price and lead time quotation using a similar method to the one discussed in Bekki and Gel (2005), in the more realistic context of a make-to-order or service environment with different types of customers, as noted above. In particular, the study considers the case when a company serves prioritized contract customers as well as spot purchasers that can be quoted changing prices and lead times over time, as congestion conditions dictate. In this case, it is important to hit the right capacity reservation trade-off by considering price and lead time decisions, which are dynamically made for spot purchasers and set at the beginning of planning horizon as contractual terms for contract customers. Hafizoğlu et al. (2016) expand the findings of previous studies (Easton and Moodie 1999; Ray and Jewkes 2004; Wu et al. 2012) on the impact of customer preferences

(i.e., customer's price and lead time sensitivities) by extending the results to more general demand functions and the flexibility to differentiate one type of customers (spot purchasers) from another one (contract customers) by changing prices and lead times dynamically.

Under relatively mild assumptions on the structure of the acceptance probability function, Hafızoğlu et al. (2016) show that price and lead time sensitivities of the buyer can be exploited for identifying effective dynamic quotation policies. More generally, there are three factors that pull the seller in different directions: (1) quoting short lead times and/or low prices to attract more customers to the system, depending on the customers' *price/lead time sensitivity*, (2) keeping congestion low by quoting higher prices and lead times to avoid paying high *delay penalties*, and (3) maintaining *profitability* by considering immediate and future impact of orders on the system.

In particular, Hafızoğlu et al. (2016) characterize the optimal dynamic price and lead time quotation policy as a function of these three parameters as follows. First, our results show that the optimal policy is strongly dependent on the ratio of the price/lead time sensitivity (defined by the ratio of the derivative of the acceptance probability function with respect to lead time to the derivative with respect to price) to the unit delay penalty incurred for one unit of delay for a spot purchaser order. We refer to this ratio as the "critical ratio," inspired by the similarity of the structure to the newsvendor policy. First, we show that we can use this critical ratio to identify situations in which joint dynamic price and lead time quotation is not necessary. For example, when spot purchasers are highly lead time sensitive (i.e., this critical ratio is greater than 1), firms should quote minimal lead times to attract customers, and focus only on dynamic pricing. On the other hand, when spot customers are highly price sensitive, however, firms should offer the minimal price acceptable to spot purchasers, and focus only on dynamic lead time quotation. The critical ratio of price/lead time sensitivity to unit delay penalty makes sense because a high (low) price/lead time sensitivity value indicates that a spot purchaser is more sensitive to lead time (price) changes. This then encourages the decision maker to quote shorter lead times to earn the spot purchaser's order. Conversely, a high value of delay penalty motivates higher lead time quotes to mitigate delay penalties. Hence, the critical ratio captures the trade-off between attracting spot purchasers and paying delay penalties. This useful result helps to identify cases in which the firm can avoid the burden of joint dynamic price/lead time quotation and only focus on one type of dynamic quotation strategy.

When spot purchasers are not highly price or lead time sensitive as indicated by the critical ratio defined above, our results show that the optimal solution often follows a newsvendor-like policy, where the critical ratio determines the optimal proportion of spot purchaser demand to be met, which in turn implies price and lead times to be quoted. The characterizations of an optimal policy also allow the reduction of decision space, which in turn allows significant computational efficiencies when calculating the optimal price and lead times to be quoted to spot purchasers.

Noting practical difficulties of joint dynamic quotation (such as adverse reactions from customers due to frequently changing prices, etc.), Hafizoğlu et al. (2016) present a comprehensive computational study to analyze the performance improvement that dynamic price and lead time quotation present over the use of static prices and lead times. The computational study reveals that in certain instances, using dynamic price and lead time quotation may make the difference between profitability versus being out of business, especially when the minimum willingness-to-pay among spot purchasers (i.e., a price level acceptable to all spot purchasers) is relatively small compared to the unit lateness penalty. The benefits of joint price and lead time quotation are most prominent when the profit margin is low, and when spot purchasers are lead time sensitive, and the acceptance probability function is strictly concave in lead time.

Finally, we explore the impact of prices and lead times used for the contract customers, and analyze the optimal mix of contract and spot customers that the firm should target to serve under different assumptions on the parameters. In particular, we identify cases where the probability that the customer order is met on time balances the trade-off between price and lead time sensitivity of the customer and the delay penalty. To determine optimal contracting terms given that the other customer class is served via optimal dynamic price and lead time quotation, we present an algorithm (via action space reduction) that is fast and produces near-optimal results. Using these tools, Hafizoğlu et al. (2016) provide an analysis of the benefits of achieving the optimal mix of contract customers and spot purchasers and present several managerial insights as to when firms should strive for the optimal mix versus settling for serving only one type of customer.

23.2.4 Inventory Decisions with Lead Time Quotation

Savasaneril et al. (2010) address lead time decisions jointly with inventory decisions. The study considers the problem in a setting with inventory, and shows that when the number of customers waiting in the system reaches a certain threshold, it is optimal to reject orders, since while this may result in loss of immediate revenue, it prevents future losses due to potential delays and lateness costs. Furthermore, assuming all customers are identical in terms of revenues and lateness penalties, keeping an inventory for a future customer would only increase the holding cost, and hence, the lead time quote should be zero whenever there are items in stock. The lead time quote increases with the number of customers in the system, and the increase eventually leads to a state at which the arriving customers are rejected. Other more recent works in this area include Kaman et al. (2013), Kahvecioglu and Balcioglu (2016), and Nakade and Niwa (2017).

23.2.5 Lead Time Quotation and Product Substitution

In contexts where the seller offers substitutable products or services to customers, it is interesting to consider how price and lead time quotation can leverage substitution flexibility to further improve system performance. The traditional wisdom from queueing theory makes us think that rather than committing production resources to two separate products, for example, the seller should pool production capacity and offer, if possible, a single type of product. There are many situations, however, companies would prefer to keep assigned production capacity to two or more different but possibly substitutable products targeting different niches of the market. It is interesting to consider the potential of price and lead time quotation-acceptance mechanisms to bring the performance of this kind of a setting as close as possible to that of the system with pooled production resources. In other words, is it possible to entice the buyer to put in an order for a substitutable product by offering favorable price and lead time quotes? Under what conditions are different combinations of quotation and substitution strategies beneficial?

van Ryzin and Mahajan (1999) and Smith and Agrawal (2000) study assortment and inventory decisions in a make-to-stock (MTS) environment assuming static substitution in which the buyer makes a choice without observing the current inventory levels. In dynamic substitution, the buyer observes the current inventory levels and makes a choice based on that observation. Gaur and Honhon (2006) consider both static and dynamic substitution for an assortment planning and inventory management problem. The study presents optimal policies for static substitution and offers effective heuristics for dynamic substitution. Mahajan and Ryzin (2001) propose a model under dynamic substitution for inventory decisions in retail assortments. Maddah and Bish (2007) jointly model the assortment, pricing and inventory decisions assuming static substitution. Talluri and Ryzin (2004) and Zhang and Cooper (2005) assume prices are fixed and determine which products to offer at each point in time. Ahiska and Kurtul (2014) consider one-way product substitution for remanufactured items when out of stock substituted by manufactured items in a periodic review inventory control problem. They conclude that the profitability depends on the remanufactured item price to manufacturing cost ratio. Kim and Bell (2015) consider joint pricing and inventory decisions with price-driven substitution and develop both deterministic and stochastic models. Bernstein et al. (2015) propose a model with multiple customer segments with different product preferences and decide which products to offer to arriving customers. They measure the potential revenue impact of assortment customization with different settings.

Dynamic pricing of substitutable products is considered in Bitran et al. (2006), Maglaras and Meissner (2006), Zhang and Cooper (2009), Dong et al. (2009), Akcay et al. (2010), Suh and Aydin (2011), Ceryan et al. (2013) and Chen and Chen (2017). Bitran et al. (2006) consider a family of substitutable perishable products with demand correlation and derive a price-sensitive demand function capturing the buyers' purchasing behavior. Maglaras and Meissner (2006) suggest

static and dynamic pricing heuristics for multi-product revenue management problems including pricing and capacity decisions. Zhang and Cooper (2009) study a dynamic pricing problem for multiple substitutable flights and propose value- and policy-approximation heuristics. Dong et al. (2009) and Akcay et al. (2010) study dynamic pricing of substitutable products using the multinomial logit (MNL) consumer choice model. Suh and Aydin (2011) consider dynamic pricing of two substitutable products assuming inventory levels are fixed. They show that optimal price difference between the two products and the optimal purchase probabilities can be useful to characterize the optimal policy. Ceryan et al. (2013) study joint price- and capacity-based substitution for a firm producing two products using capacitated, product-dedicated and flexible resources. They show that stable price differences can be maintained across items via flexible resources. Chen and Chen (2017) study dynamic pricing of two substitutable products considering substitution both across products and time periods, and propose a robust optimization model. Ceryan et al. (2018) consider price- and availability-based product substitution in a two-stage model. In first phase they decide price and replenishment levels for each product and in the second phase decide which customers to upgrade.

Yilmaz et al. (2014a) consider dynamic lead time quotation for two substitutable products in a make-to-order environment. There are two types of customers, and each of them prefers one of the substitutable products if both products are quoted the minimum lead-times. Arriving customers are quoted lead-times for both products and based on lead-times and choice probabilities of customers, they either place an order for one of the products or leave the system without placing an order. If a customer places an order, the order is processed in a first-come-first-served order; and a tardiness penalty is incurred if the order is completed after the quoted lead-time. The profit from a customer equals the revenue minus the tardiness penalty and the long-run average expected profit per unit time is maximized. The study characterizes an optimal policy and the performance improvement to be expected from this lead time quotation mechanism with respect to the profitability levels of products and the buyers' preferences for each product. Yilmaz et al. (2014a) build on the monotonicity of lead-times result in Duenyas and Hopp (1995) by making the observation that the lead-time quote of product i remains non-decreasing in number of orders of product i even when a substitutable product is added and either quoted fixed or dynamic, as in the case of single product. On the other hand, the lead-time quote of product i is not constant or monotonic in number of orders of product j in case of fixed and dynamic quoted substitutable product.

The study compares the benefits of dynamic lead time quotation and substitution by constructing scenarios with dynamic versus static lead time quotation with substitution and without substitution. The impacts of problem characteristics such as revenues, traffic intensity, arrival rates, lead-time sensitivities on the benefits obtained by dynamic quotation and substitution are analyzed. It is worthwhile to note that the performance improvement obtained by dynamic quotation and substitution in this case depends on the problem characteristics, similarly to the observations made by other studies. When revenues of products are close, the benefits of substitution decrease since there will not be additional gain from

directing the customer to the higher revenue product. When the customer mix gets unbalanced, substitution helps with offering the substitutable product to the customers with a longer queue instead of quoting them a high lead-time and rejecting, i.e., customers leave the system without placing an order.

23.2.6 Dynamic Quotation and Cancellations

Dynamic quotation of price and lead-time helps sellers to increase the resource utilization by adjusting the incoming customer orders according to system state. Order cancellations cause changes in the system state and should be taken into account when quoting prices and/or lead-times. If cancellations are ignored, the system state will be overestimated and hence resources will be underutilized and profits will be lower.

You (2003) discusses the newsboy problem with price-dependent demand and order cancellations and shows that the optimal price is nondecreasing with the amount of reserved units, and nonincreasing with the number of remaining decision periods. You and Wu (2007) assume two periods, namely advance sales and spot sales and determine the optimal advance sales prices, spot sales prices, order size, and replenishment frequency in which customer with reservations may cancel their orders before receiving them. They propose a solution algorithm for optimal decisions. Dye and Hsieh (2013) also assume two periods, and develop an inventory model for deteriorating items with price-dependent demand. They show that price in advance sales period is lower than spot price.

Rubino and Ata (2009) discuss a dynamic control problem for a make-to-order system with parallel servers, considering outsourcing and resource allocation decisions allowing order cancellations subject to a cancellation penalty. They make heavy traffic assumptions, and approximate the problem by a Brownian control problem. Ata and Peng (2018) study customer abandonments in queuing systems and characterize the equilibrium conditions by making heavy traffic assumptions.

Yilmaz et al. (2014b) considers a dynamic lead-time quotation problem for a make-to-order system with order cancellations. They assume order cancellations may occur either while order is in queue or process or after impact of cancellation on performance of dynamic quotation of lead times.

23.3 Conclusions

There is a large number of studies exploring the potential benefits of dynamic quotation strategies under which a seller provides price and/or lead time quotes to a potential buyer, who then either accepts or rejects the quote. The goal of dynamic quotation strategies is to encourage buyers to place an order at times of low system congestion and keep the system load in check by quoting high prices

and/or lead times at times of high system congestion. While prior studies consider a number of different contexts, a common finding is that the potential benefits of these strategies compared to the use of static price and/or lead times is highly dependent on how buyers react to dynamic changes in prices and/or lead times, as well as other important parameters such as the profitability of the product/service being offered.

An important direction to pursue further is to explore how buyers' preferences can be represented accurately in models that identify the right dynamic or static quotes to be used. While there are several different representations presented in the literature, not much attention has been given to how these models (e.g., acceptance probability function discussed above) will be parameterized in a particular context. How frequently should they be updated? How much accuracy is needed to obtain most of the benefits resulting from dynamic quotation strategies? What are some practical issues that one needs to consider for implementation?

Several extensions of the above-mentioned studies are also possible and worthwhile to make quotation strategies more effective and easier to implement, such as more realistic representations of substitutable products, contractual relationships and various flexibilities/privileges they offer to contracted customers, how buyers' behaviors change over time in response to previous experiences with the buyer. Computational advances to reduce the burden of identifying optimal dynamic quotation strategies are needed; efforts to develop approximate solution algorithms and to identify easy-to-implement, simple quotation policies with relatively few (or less frequent) price/lead time changes are also highly valuable for practicality and implementability of these approaches.

Acknowledgements This work has been supported in part by the National Science Foundation [Grant DMI-0621012].

References

Afeche P, Ata B (2013) Bayesian dynamic pricing in queueing systems with unknown delay cost characteristics. Manuf Serv Oper Manage 15(2):292–304. ISSN: 1523-4614

Afeche P, Pavlin JM (2016) Optimal price/lead-time menus for queues with customer choice: segmentation, pooling and strategic delay. Manage Sci 62(8):2412–2436

Ahiska SS, Kurtul E (2014) Modeling and analysis of a product substitution strategy for a stochastic manufacturing/remanufacturing system. Comput Ind Eng 72:1–11

Akan M, Ata B, Olsen T (2012) Congestion-based lead-time quotation for heterogenous customers with convex-concave delay costs: optimality of a cost-balancing policy based on convex hull functions. Oper Res 60(6):1505–1519. ISSN: 0030-364X

Akcay Y, Natarajan HP, Xu SH (2010) Joint dynamic pricing of multiple perishable products under consumer choice. Manage Sci 56:1345–1361

Aktaran-Kalayci T, Ayhan H (2009) Sensitivity of optimal prices to system parameters in a steady-state service facility. Eur J Oper Res 193(1):120–128. ISSN: 0377-2217

Albana AS, Frein Y, Hammami R (2018) Effect of a lead time-dependent cost on lead time quotation, pricing, and capacity decisions in a stochastic make-to-order system with endogenous demand. Int J Prod Econ 203:83–95

Altendorfer K, Minner S (2014) A comparison of make-to-stock and make-to-order in multi-product manufacturing systems with variable due dates. IIE Trans 46(3):197–212
Altendorfer K, Minner S (2015) Influence of order acceptance policies on optimal capacity investment with stochastic customer required lead times. Eur J Oper Res 243(2):555–565
Ardjmand E, Weckman GR, Young WA, Sanei Bajgiran O, Aminipour B (2016) A robust optimisation model for production planning and pricing under demand uncertainty. Int J Prod Res 54(13):3885–3905
Ata B (2006) Dynamic control of a multiclass queue with thin arrival streams. Oper Res 54(5):876–892. ISSN: 0030-364X
Ata B, Olsen TL (2009) Near-optimal dynamic lead-time quotation and scheduling under convex-concave customer delay costs. Oper Res 57(3):753–768 (2009). ISSN: 0030-364X
Ata B, Olsen TL (2013) Congestion-based leadtime quotation and pricing for revenue maximization with heterogeneous customers. Queueing Syst 73:35–78
Ata B, Peng X (2018) An equilibrium analysis of a multiclass queue with endogenous abandonments in heavy traffic. Oper Res 66(1):163–183
Bekki OB, Gel ES (2005) Dynamic quotation of price and lead time in make-to-order systems. Technical report, Department of Industrial Engineering, Arizona State University, Tempe, AZ
Bernstein F, Kok AG, Xie L (2015) Dynamic assortment customization with limited inventories. Manuf Serv Oper Manage 17(4):538–553
Bitran G, Caldentey R, Vial R (2006) Pricing policies for perishable products with demand substitution. Tech. rep., Stern School of Business
Celik S, Maglaras C (2008) Dynamic pricing and lead-time quotation for a multiclass make-to-order queue. Manage Sci 54(6):1132–1146. ISSN: 0025-1909
Ceryan O, Sahin O, Duenyas I (2013) Dynamic pricing of substitutable products in the presence of capacity flexibility. Manuf Serv Oper Manage 15(1):86–101
Ceryan O, Duenyas I, Sahin O (2018) Dynamic pricing and replenishment with customer upgrades. Prod Oper Manage. https://doi.org/10.1111/poms.12816
Chaharsooghi SK, Honarvar M, Modarres M, Kamalabadi IN (2011) Developing a two stage stochastic programming model of the price and lead-time decision problem in the multi-class make-to-order firm. Comput Ind Eng 61(4):1086–1097. ISSN: 0360-8352
Charnsirisakskul K, Griffin PM, Keskinocak P (2004) Order selection and scheduling with leadtime flexibility. IIE Trans 36(7):697–707. ISSN: 0740-817X
Charnsirisakskul K, Griffin PM, Keskinocak P (2006) Pricing and scheduling decisions with leadtime flexibility. Eur J Oper Res 171(1):153–169. ISSN: 0377-2217
Chen M, Chen ZL (2015a) Recent development in dynamic pricing research: multiple products, competition , and limited demand information. Prod Oper Manage 24(5):704–731
Chen M, Chen Z-L (2015b) Recent developments in dynamic pricing research: multiple products, competition, and limited demand information. Prod Oper Manage 24(5):704–731
Chen M, Chen ZL (2017) Robust dynamic pricing with two substitutable products. Manuf Serv Oper Manage: 1–20 (2017). https://doi.org/10.1287/msom.2017.0639
Chen X, Tai AH, Yang Y (2014) Optimal production and pricing policies in a combined make-to-order/make-to-stock system. Int J Prod Res 52(23):7027–7045
Cheng TCE, Gupta MC (1989) Survey of scheduling research involving due date determination decisions. Eur J Oper Res 38:156–166
Cheung WC, Simchi-Levi D, Wang H (2017) Technical note - dynamic pricing and demand learning with limited price experimentation. Oper Res 65(6):1722–1731
Cil EB, Ormeci EL, Karaesmen F (2009) Effects of system parameters on the optimal policy structure in a class of queueing control problems. Queueing Syst 61(4):273–304. ISSN: 0257-0130
Cil EB, Karaesmen F, Ormeci EL (2011) Dynamic pricing and scheduling in a multi-class single-server queueing system. Queueing Syst 67(4):305–331
den Boer AV (2015) Dynamic pricing and learning: Historical origins, current research, and new directions. Surv Oper Res Manage Sci 20(1):1–18

den Boer AV, Zwart B (2015) Dynamic pricing and learning with finite inventories. Oper Res 63(4):965–978
Dong L, Kouvelis P, Tian Z (2009) Dynamic pricing and inventory control of substitute products. Manuf Serv Oper Manage 11(2):317–339
Duenyas I (1995) Single facility due-date setting with multiple customer classes. Manage Sci 41(4):608–619. ISSN: 0025-1909
Duenyas I, Hopp WJ (1995) Quoting customer lead times. Manage Sci 41(1):43–57. ISSN: 0025-1909
Dye CY, Hsieh T-P (2013) Joint pricing and ordering policy for an advance booking system with order cancellations. Appl Math Modell 37:3645–3569
Easton FF, Moodie DR (1999) Pricing and lead time decisions for make-to-order firms with contingent orders. Eur J Oper Res 116(2):305–318. ISSN: 0377-2217
ElHafsi M (2000) An operational decision model for lead-time and price quotation in congested manufacturing systems. Eur J Oper Res 126(2):355–370. ISSN: 0377-2217
Elmaghraby W, Keskinocak P (2003) Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions. Manage Sci 49(10):1287–1309. ISSN: 0025-1909
Feng JJ, Liu LM, Liu XM (2011) An optimal policy for joint dynamic price and lead-time quotation. Oper Res 59(6):1523–1527. ISSN: 0030-364X
Gaur V, Honhon D (2006) Assortment planning and inventory decisions under a locational choice model. Manage Sci 52(10):1528–1543
Hafızoğlu AB, Gel ES, Keskinocak P (2016) Price and lead time quotation for contract and spot customers. Oper Res 64(2):406–415
Hua GW, Wang SY, Cheng TCE (2010) Price and lead time decisions in dual-channel supply chains. Eur J Oper Res 205(1):113–126. ISSN: 0377-2217
Kahvecioglu G, Balcioglu B (2016) Coping with production time variability via dynamic lead-time quotation. OR Spectrum 38(4):877–898
Kaman C, Savasaneril S, Serin Y (2013) Production and lead time quotation under imperfect shop floor information. Int J Prod Econ 144(2):422–431
Kaminsky P, Hochbaum D (2004) Due date quotation models and algorithms. In: Leung JY-T (ed) Handbook of scheduling: algorithms, models, and performance analysis edition. CRC Press, Boca Raton
Kapuscinski R, Tayur S (2007) Reliable due-date setting in a capacitated MTO system with two customer classes. Oper Res 55(1):56–74. ISSN: 0030-364X
Keskinocak P, Tayur S (2004) Due-date management policies. International series in operations research and management science edition. Kluwer Academic Publisher, Dordrecht
Keskinocak P, Ravi R, Tayur S (2001) Scheduling and reliable lead-time quotation for orders with availability intervals and lead-time sensitive revenues. Manage Sci 47(2):264–279. ISSN: 0025-1909
Kim SW, Bell PC (2015) A note on the optimal pricing and production decisions with price-driven substitution. Int Trans Oper Res 22:1097–1116
Koulamas C (1994) The total tardiness problem: review and extensions. Oper Res 42(6):1025–1041
Liu L, Parlar M, Zhu SX (2007) Pricing and lead time decisions in decentralized supply chains. Manage Sci 53(5):713–725. ISSN: 0025-1909
Liu Z, Lu L, Qi X (2012) Simultaneous and sequential price quotations for uncertain order inquiries with production scheduling cost. IIE Trans 44(10):820–833
Lu L, Liu Z, Qi X (2013) Coordinated price quotation and production scheduling for uncertain order inquiries. IIE Trans 45(12):1293–1308
Maddah B, Bish EK (2007) Joint pricing, assortment and inventory decisions for a retailer's product line. Oper Res 54:315–330
Maglaras C, Meissner J (2006) Dynamic pricing strategies for multi-product revenue management problems. Manuf Serv Oper Manage 8(2):136–148

Mahajan R, Van Ryzin G (2001) Stocking retail assortments under dynamic consumer substitution. Oper Res 49(3):334–351

Nakade K, Niwa H (2017) Optimization and customer utilities under dynamic lead time quotation in an type base stock system. Math Prob Eng 2017

Öner-Közen M, Minner S (2017) Dynamic pricing, leadtime quotation and due date based priority dispatching. Int J Prod Res 1–13

Palaka K, Erlebacher S, Kropp DH (1998) Lead-time setting, capacity utilization, and pricing decisions under lead-time dependent demand. IIE Trans 30(2):151–163. ISSN: 0740-817X

Pekgun P, Griffin PM, Keskinocak P (2008) Coordination of marketing and production for price and leadtime decisions. IIE Trans 40(1):12–30. ISSN: 0740-817X

Plambeck EL (2004) Optimal leadtime differentiation via diffusion approximations. Oper Res 52(2):213–228. ISSN: 0030-364X

Ray S, Jewkes EM (2014) Customer lead time management when both demand and price are lead time sensitive. Eur J Oper Res 153(3):769–781. ISSN: 0377-2217

Rubino M, Ata B (2009) Dynamic control of a make-to-order, parallel-server system with cancellations. Oper Res 57(1):94–108

Savasaneril S, Sayin E (2017) Dynamic lead time quotation under responsive inventory and multiple customer classes. OR Spectrum 39(1):95–135

Savasaneril S, Griffin PM, Keskinocak P (2010) Dynamic lead-time quotation for an M/M/1 base-stock inventory queue. Oper Res 58(2):383–395. ISSN: 0030-364X

Smith SA, Agrawal N (2000) Management of multi-item retail inventory systems with demand substitution. Oper Res 48(1):50–64

So KC, Song J-S (1998) Price, delivery time guarantees and capacity selection. Eur J Oper Res 111:28–49

Son J-D (2008) Optimal admission and pricing control problem with deterministic service times and sideline profit. Queueing Syst 60(1–2):71

Suh M, Aydin G (2011) Dynamic pricing of substitutable products with limited inventories under logit demand. IIE Trans 43(5):323–331

Talluri K, Van Ryzin G (2004) Revenue management under a general discrete choice model of consumer behavior. Manage Sci 50:15–33

van Ryzin G, Mahajan S (1999) On the relationship between inventory costs and variety benefits in retail assortments. Manage Sci 45(11):1496–1509

Wu ZP, Kazaz B, Webster S, Yang KK (2012) Ordering, pricing, and lead-time quotation under lead-time and demand uncertainty. Prod Oper Manage 21(3):576–589. ISSN: 1059-1478

Xiao T, Qi X (2016) A two-stage supply chain with demand sensitive to price, delivery time, and reliability of delivery. Ann Oper Res 241(1–2):475–496

Xiao TJ, Jin JA, Chen GH, Shi J, Xie MQ (2010) Ordering, wholesale pricing and lead-time decisions in a three-stage supply chain under demand uncertainty. Comput Ind Eng 59(4):840–852. ISSN: 0360-8352

Ye T, Sun H, Li Z (2016) Coordination of pricing and leadtime quotation under leadtime uncertainty. Comput Ind Eng 102:147–159

Yildirim U, Hasenbein JJ (2010) Admission control and pricing in a queue with batch arrivals. Oper Res Lett 38(5):427–431

Yilmaz T, Gel ES, Keskinocak P (2014a) Dynamic lead-time quotation for substitutable products. Technical report, Department of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA

Yilmaz T, Gel ES, Keskinocak P (2014b) Dynamic lead-time quotation considering customer order cancellations products. Technical report, Department of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA

Yoon S, Lewis ME (2004) Optimal pricing and admission control in a queueing system with periodically varying parameters. Queueing Syst 47(3):177–199. ISSN: 0257-0130

You P-S (2003) Dynamic pricing of inventory with cancellation demand. J Oper Res Soc 54(10):1093–1101

You PS, Wu MT (2007) Optimal ordering and pricing policy for an inventory system with order cancellations. OR Spectrum 29:661–679
Zhang D, Cooper WL (2005) Revenue management for parallel flights with customer-choice behavior. Oper Res 53(3):415–431
Zhang D, Cooper WL (2009) Pricing substitutable flights in airline revenue management. Eur J Oper Res 197:848–861
Zhao XY, Stecke KE, Prasad A (2012) Lead time and price quotation mode selection: uniform or differentiated? Prod Oper Manage 21(1):177–193. ISSN: 1059-1478
Zheng F, Zhang E, Xu Y, Hong W-C (2014) Competitive analysis for make-to-order scheduling with reliable lead time quotation. J Combin Optim 27:182–198
Zhu SX (2015a) Integration of capacity, pricing, and lead-time decisions in a decentralized supply chain. Int J Prod Econ 164:14–23
Zhu SX (2015b) Integration of capacity, pricing, and lead-time decisions in a decentralized supply chain. Int J Prod Econ 164:14–23

Dr. Esma S. Gel is an Associate Professor of Industrial Engineering in the School of Computing, Informatics and Decision Systems Engineering (CIDSE) at Arizona State University. Dr. Gel holds a B.S. degree in Industrial Engineering from Middle East Technical University, Turkey, and M.S. and Ph.D. degrees in Industrial Engineering from Northwestern University, obtained in 1995 and 1999, respectively. Her research focuses on the use of stochastic modeling and control techniques for the design, control, and management of operations in various settings, with emphasis on manufacturing and service systems, business and logistics processes, and health care systems. Dr. Gel's work has been funded by the National Science Foundation as well as several industrial partners such as Intel and Mayo Clinic. Her contributions were recognized with the Dr. Hamed K. Eldin Outstanding Young Industrial Engineer of the Year award in 2008 by the IIE. She is an active member of INFORMS and has served as a volunteer in many capacities such as a member of the INFORMS Board of Directors.

Pinar Keskinocak is the William W. George Chair and Professor in the Stewart School of Industrial and Systems Engineering, and co-founder and Director of the Center for Health and Humanitarian Systems at Georgia Tech. She also serves as the College of Engineering ADVANCE Professor.

Dr. Keskinocak's research focuses on the applications of quantitative methods and analytics to have a positive impact in society, particularly in healthcare and humanitarian systems. Her recent work has addressed a broad range of topics such as infectious disease modeling, evaluating intervention strategies, and resource allocation; catch-up scheduling for vaccinations; decision-support for organ transplant; hospital operations management; and disaster preparedness and response. She has worked on projects with a variety of governmental and non-governmental organizations, and healthcare providers, including American Red Cross, CARE, Carter Center, CDC, Children's Healthcare of Atlanta, Emory Healthcare, Grady Hospital, Shepherd Center, and Task Force for Global Health.

She has served her professional community in various

roles, including Department Editor for Operations Research, INFORMS Secretary, INFORMS Vice President for Membership and Professional Recognition, co-founder and President of the INFORMS Public Sector Operations Research section, co-founder and President of the Junior Faculty Interaction Group Forum, and President of the Women in OR/MS Forum. In addition to research and educational activities, she also spends a significant amount of her time and efforts on promoting diversity and inclusion among faculty, students, and staff in higher education.

Tuba Yilmaz is an Assistant Professor of Operations Management in the Faculty of Business at Ozyegin University, Istanbul. Dr. Yilmaz received her B.S. in Industrial Engineering from Bogazici University in 2006. She received her Ph.D. from Industrial and Systems Engineering from Georgia Institute of Technology in 2013 and M.Sc. in Operations Research from the same university in 2011. She worked as an operations research consultant at a supply chain optimization software company, in various research projects in areas including inventory optimization, replenishment planning, transportation planning, and network design. She co-founded Optiyol Decision Analytics to develop SaaS solutions for supply chain optimization and revenue management.

Her research focuses on supply chain management, with an emphasis on resource allocation, pricing and revenue management, due date/lead-time decisions, and health-care applications.

Chapter 24
Oyster Mushroom Cultivation as an Economic and Nutritive Alternative for Rural Low-Income Women in Villapinzón (Colombia)

Natalia Vargas, Carmen Gutierrez, Silvia Restrepo, and Nubia Velasco

Contents

24.1 Introduction 561
24.2 Methodology 563
24.2.1 Study Site and the Villapinzón Women's Association (VWA) 564
24.2.2 Teaching the Cultivation Process 564
24.2.3 Business Plan: Industry and Sector Analyses 567
24.2.4 Perception of the Cultivation Process 568
24.3 Results 568
24.3.1 Cultivation Process 568
24.3.2 Business Plan 571
24.4 Discussion 577
24.4.1 How Did Rural Women Become Involved in the Cultivation Process? 578
24.4.2 Advantages of Cultivating Oyster Mushrooms 579
24.4.3 Future Perspectives: A Potential New Project Based on the Previous Case Experience 579
24.5 Concluding Remarks 580
Appendix 581
References 585

24.1 Introduction

Eradicating extreme poverty and hunger, *promoting gender equality, empowering women,* and *ensuring environmental sustainability* are among the eight Millennium Development Goals (MDGs 2018). A meaningful path out of poverty requires a

N. Vargas · S. Restrepo
Department of Biological Sciences, University of Los Andes, Bogotá, Colombia

C. Gutierrez
University of Los Andes, Bogotá, Colombia

N. Velasco (✉)
School of Management, University of Los Andes, Bogotá, Colombia
e-mail: nvelasco@uniandes.edu.co

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_24

strong economy that provides full and productive employment with good wages. In view of this, expanding women's opportunities in different sectors, and providing them with a stable source of income, can accelerate economic growth and turn into a vital path to poverty eradication. Sustainable growth must be environmentally sound and target the development of management practices aimed at biodiversity conservation, while, at the same time, meeting the necessary human production needs (Pilz and Molina 1996).

The project described in this study is the continuation of a mixed analysis that previously diagnosed poverty conditions in Villapinzón, Cundinamarca, performed by Bautista and Torres (2012). The authors identified priority areas for intervention and proposed alternatives for overcoming poverty in the municipality. Their results, according to a multidimensional poverty indicator, established that 38% of the population in Villapinzón is poor, and that housing (overcrowding), education (low level of education, educational lag, and illiteracy), and work (very high rate of informal employment) are priority areas for intervention (Bautista and Torres 2012).

The local government of Villapinzón and *Universidad de los Andes* have worked together to propose interventions for the priority areas identified. In December 2013, the Mayor of Villapinzón, the Villapinzón Women's Association (VWA), the School of Government, the School of Management, and the Department of Biological Sciences at *Universidad de los Andes*, all agreed to begin a pilot program for the production of *Pleurotus ostreatus* (commonly known as the oyster mushroom). The initiative focuses on the cultivation and marketing of these fungi as a stable and sustainable source of income for the inhabitants of Villapinzón, specifically for the women of the VWA. The process was participatory and implemented by the community and municipal government, once they defined their own priorities.

Until today, the main economic activities in Villapinzón are in the primary and secondary sectors, potato crop cultivation being one of the most important. The vast majority of farmers are smallholders, specifically, 48% own less than one hectare and 85% own less than three hectares (DANE 2001). The producers' economy is strongly undermined by the unstable and uncertain potato price cycle (FEDEPAPA 2013), and not all farmers are landowners; about 52% of the cultivated area is leased, leading to job instability. Lastly, according to DANE (2001), approximately 89% of the producers did not receive technical assistance in the previous year, leading to low levels of production and deficient marketing (DANE 2001).

The edible fungi industry has proven to be a viable alternative in many countries, as a means of providing incentives for biodiversity use, the diversification of products, and the improvement of income opportunities in marginal rural areas (Ortega-Martínez and Martínez-Peña 2008; Cai et al. 2011). According to the newsletter published by the "Corporación Internacional de Colombia" (2004), worldwide edible fungi consumption is approximately 3 million tons per year, consisting of 30 different fungal species. Latin American countries, such as Mexico, Guatemala, Chile, Peru, and Argentina are aware of the importance of using fungal resources (Boa 2004; Andrade et al. 2012), and they have traditionally introduced several species into their diets. According to FAO by 2004 the number of wild edible mushrooms reached 2166 species worldwide, proving their relevance as a food source (Boa 2004).

Along these lines, a strategy that explores crop diversification, establishing an alternative source of nutrition, and a stable source of income for farmers in the region, could contribute to the eradication of extreme poverty in Villapinzón. In this study, we used the *Pleurotus ostreatus* species (commonly known as the oyster mushroom) because (1) it is considered to be a very nutritive biological food source with a high content of amino acids, crude protein, vitamins, fiber, and unsaturated fatty acids (Cheung 2010; Michael et al. 2011; Rathee et al. 2012; Kalač 2013), (2) it has antioxidant and anti-inflammatory properties (Puttaraju et al. 2006; Rathee et al. 2012), and (3) it is an accessible, low-investment production option, which can be grown in a great variety of substrates.

Poverty conditions in towns like Villapinzón have become an additional obstacle for small producers to gain access to markets and subsidy programs. Partnerships and collaborations like the one between *Universidad de los Andes* and the VWA result in projects that can help overcome these barriers, meaning that they have a greater chance of succeeding. The Department of Biological Sciences at *Universidad de los Andes* provided the women with the required technical information and management practices to grow oyster mushrooms, ensuring a good end product. The School of Government monitored the group of women providing them with leadership workshops that targeted the development and management of emotional and self-awareness skills, as well as entrepreneurial tools to increase their chances of success. Finally, the School of Management helped designing the business plan and developed the production process and supply chain management. These partnerships set strategies in place to empower women with the appropriate knowledge, thus enabling them to sustainably use, and reproduce, the oyster mushroom cultivation. In this sense, the relevance of this project responds to the patterns of land ownership and the need for crop diversification and technical assistance in the municipality. More specifically, the initiative focused on generating a stable and sustainable source of income over time for the members of the Villapinzón Women's Association.

24.2 Methodology

In order to develop the project presented in this study, we proposed a methodological approach based on three main steps: (1) the description of the municipality and the VWA; (2) teaching all interested women how to produce the mushrooms, and (3) the business plan design. At the end of this process, a survey was applied to the women in order to identify the main difficulties that arose during the cultivation process, and the results were used to adjust the project as suggested.

24.2.1 Study Site and the Villapinzón Women's Association (VWA)

The municipality of Villapinzón has about 18,764 inhabitants, of which, according to the multi-dimensionally indicator, 67% reside in rural areas; 86% are classified as part of SISBEN[1] levels 1–3, and 38% are considered to be poor. Also, over 77% of the labor market in Villapinzón is informal, which is a higher rate than the 53% national average (DANE 2001) reflecting a high rate of job instability. Studies have shown a high rate of informal work and a higher rate of unemployment, mostly among women, leads to income instability (Joumard and Londoño 2013).

The Villapinzón Women's Association (VWA) has been working informally with the Mayor's office since 2012, and was only legally constituted in 2014. The purpose of the VWA is to provide employment alternatives to single-parent households, and its legal documents include details on 35 women who live mostly in the rural areas of the municipality with limited and unstable sources of income. The Association has worked for 3 consecutive years with the Mayor's office to determine possible employment alternatives and to train them with specific technical farming skills.

24.2.2 Teaching the Cultivation Process

The applied teaching techniques were very important for our methodological approach. We exercised a high degree of caution throughout the steps detailed and during the cultivation process, ensuring that the new technical language shared with the VWA was clear and understandable. Rural women participated during the cultivation teaching process, carried out in six rural homes. The cultivation process baseline was according to Guzmán et al. (1993), Albertó (2008), and Gaitán-Hernández et al. (2006), with modifications.

Grain Sterilization

Over the course of the first visits, we taught the women how to undertake the complete process under sterilized conditions. The methodology described was used in six pilots, in which we used special instructions—given in the appropriate language for the women's level of understanding—using the equipment and materials available in each house. For grain preparation, pearl barley, wheat, and rice were used to determine which of these was more efficient for mycelial growth. Previous to the sterilization of the grain, this must be washed with water, drained and served

[1]SISBEN is an instrument that obtains socioeconomic information on specific groups in the country. It is the main instrument for targeting social programs to poor and vulnerable groups.

in clean jars until filling 3/4 parts of a jar. For the sterilization process, we suggested that they use a pressure cooker. We constructed a handmade mesh with four screws, to be put at the bottom of the pot (Appendix, Fig. 24.5a, b), so the jars were not touching the bottom of the pot, and a maximum temperature was used to heat the pot for 1 h. This process was repeated for an additional hour, replacing the water used before. After 2 h the bottles were allowed to cool at room temperature.

Preparing the Spawn

The production process should be carried out in an aseptic area that is free of contaminants. On a prewashed table with two lighters surrounding the inoculation area (Fig. 24.1a), the women, using the tip of a scalpel or knife (previously briefly held in the fire), placed 1 × 1 cm square portions of mycelium growing on the *Petri* dish on the surface of each of the jars containing the wet and sterilized grain. Then, they closed the jars with foil and rubber bands and left them in the designed area for the incubation periods.

Use of the Available Resources: Sterilization of the Substrates

The women used the substrates available in their houses. During the pilots, they worked with 6 Kg of dry substrate, that was pre-hydrated by submerging it in water for 10–16 hrs and drained previous to the sterilization process. They added 300 g of calcium carbonate (5%—dry weight of substrate), and 15% (dry weight of substrate) of supplement (molasses and wheat bran) to the drained substrate, placing these in a 70 L pot (all packed in a clean potato bag) with water, covered and heated to 80 C for 1 hour to sterilize the substrate. The sterilized substrate and supplements must be well drained before use in the next step.

Bags Incubation, Fruiting Bodies Collection, and Packaging

On a disinfected table with two lighters surrounding the area of inoculation, the women placed a layer of sterilized substrate inside 31 × 45 cm transparent bags using the tip of a scalpel or knife. Then, they placed portions of the spawn on a layer of sterilized substrate, repeating this process until making each layer homogeneous (Fig. 24.1b, c). They inoculated the bags with 5% grain spawn (wet weight of substrate). They pressed each bag down slightly and used a piece of PVC pipe with a gauze, tied with a rubber band, to allow gas exchange in the top of the bag (Fig. 24.1d). The cultivation process consisted of two phases: (1) a darkness phase for bag incubation and (2) a light phase for the production of fruiting bodies. The

Fig. 24.1 (**a**) Clean table for spawn (grain + fungus mycelium) preparation. (**b** and **c**) Women of the AWV during the bags assemblage; (**d**) Bags with the substrate and spawn

VWA women designed a zone covered with black bags for the darkness phase area, and another area in the room for the light phase.

Dark phase, after preparing the bags The women injected 10 mL of boiled water into one of the holes in the gauze every day in order to keep the bags damp, and, as instructed, they recorded the days when the white mycelium started to spread and to cover the whole substrate.

Light phase Once the white mycelium had covered the whole substrate, the women opened 2 cm-holes randomly distributed over the bag surface. They observed the substrate on a daily basis, while keeping the bags damp as explained above, until they started to produce primordia and then mature fruiting bodies (Fig. 24.2a, b, respectively).

Fig. 24.2 Fruiting bodies production: (**a**) young fruiting bodies, (**b**) mature fruiting bodies, (**c–e**) different packaging types

Collection Once the fungi reached of a good size (pileus [cap]: 5–8 cm) and before becoming dry, the women harvested them by cutting with a clean knife or razor, and then storing them in a container in the freezer. The women measured the following parameters during the incubation and production phases: time for the mycelium to cover the grain and substrate, number of production days, and fresh weight (g) of fruiting bodies/bag (Albertó 2008; Sher et al. 2011). The statistical summary parameters of the grain coverage, substrate coverage and biological efficiency were obtained by using R (R Development Core Team, http://www.r-project.org/). The mushroom yield was calculated according to Albertó (2008) as a percentage: e.g., a 3 kg/bag of fresh mushrooms from 10 kg of wet substrate = 30% yield.

Packaging Once individual 250 g quantities were collected, the women packed the fungal product according to the clients' requirements.

Microbiological and Statistical Analyses

Experts at the Laboratory of Food Microbiology (LEMA) at *Universidad de los Andes* carried out routine analyses of the fresh and dehydrated fruiting bodies, and water from two localities in the municipality, as well as direct microscopic examinations of green appearance colonies growing in the cultivated bags in order to describe and classify the bacterial and fungal genera.

24.2.3 Business Plan: Industry and Sector Analyses

We proposed the development of a business plan in order to identify whether oyster mushroom production is a good opportunity for the VWA women. The different tools we used in the methodology allowed us to diagnose the appropriate conditions for a feasible business idea. According to Ehmke and Akridge (2005), a business plan has six components: business description, market analysis, competitor assessment, marketing plan, operating plan, and financial plan.

The business description includes a mission statement, the company objectives, the legal structure and the owner(s), information about the nature of the business, how the company will start, a general description of the products/services, and the

target market. The market analysis section describes the market characteristics and customer profiles. We carried out a competitor assessment to find the competitors' profiles and their strengths and weaknesses. The marketing plan describes the products and services, identifies their features and benefits, discusses the needs and problems they addressed, and describes and specifies the pricing. Finally, we present the logistic management plan, focusing on the distribution channels and client approaches.

The operating plan section includes two main components: ownership and management, and resources and production. The former describes the key people in the organization and the external advisors. The resources and production section presents the production process, equipment, and facilities.

Finally, the financial plan describes the current financial status of the project, and presents forecasts of future financial statements as well as the potential return on investment. This section shows how, or under which conditions, the proposal is feasible.

24.2.4 Perception of the Cultivation Process

In order to identify the women's perception at different parts of the cultivation process as well as their acceptance of a new product in their daily nutrition, we informally surveyed the six groups of VWA women during the visits.

24.3 Results

24.3.1 Cultivation Process

Rural women participated in the cultivation pilot programs carried out in six rural households, where small rooms adapted for bag incubation, and small greenhouses were built and adapted with the women's consent when necessary. During the first visits, in an attempt to promote sustainable production, we found available resources in the region that could be used as substrates: hay, rice husk, and sawdust.

Grain and Substrate Coverage

For grain assembly, the women were instructed not to fill the entire jar with grain, as this would not have left the necessary space for the fungus to breathe. The average time for *P. ostreatus* to cover the grains was 21.92 days ($SD = 3.47$), from 12 jars (two in each of the six houses). The substrate that showed the highest biological efficiency was rice husk supplemented with molasses (24%, 1.2 kg fresh

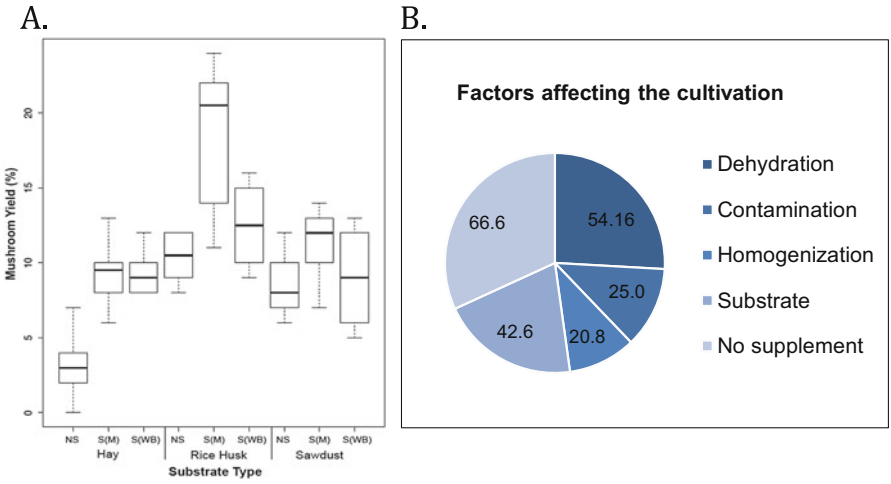


Fig. 24.3 Production process. (**a**) Percentage of fruiting bodies production by treatment. Dark horizontal lines represent the mean, with the box representing the 25th and 75th percentiles, and whiskers the minimum and maximum values. *NS* no supplement, *SM* supplemented with molasses, *SWB* supplemented with Wheat Bran. (**b**) Percentage of bags affected by common problems: dehydration (26 out of 48 bags); contamination (12 out of 48 bags); homogenization (10 out of 48 bags); substrate (23 bags showing less than 10% of production, out of 54 bags; no supplement added (12 bags showing less than 10% of production out of 18 bags)

mushroom/5 kg fresh substrate, during the first cycle of production) (Fig. 24.3a), and the mycelia proved to be highly efficient, covering the substrate in an average of 30.25 days ($SD = 2.78$) (48 bags, eight bags in each of the six houses).

Common Problems Associated to the Cultivation Process

The women came up with many ideas to solve problems such as contamination, dehydration (above 40 °C), and homogenization of the substrate (Fig. 24.3b). The most common problem was dehydration of the fruiting bodies once the bags were opened and the women attributed this to the warm days, which caused a slower production of fruiting bodies during the production period. To solve this, the women increased the frequency with which they sprayed the substrate. Some women also noticed green colonies inside some of the bags, indicating a source of contamination (Table 24.1). As such, we instructed the women to check the efficacy of the sterilization process in order to decrease contamination by other fungal species, identified as *Penicillium* and *Trichoderma* spp. by the Laboratory of Mycology and Plant Pathology (*Universidad de Los Andes*).

Microbiological analyses *The* results of the water and mushroom analyses showed the presence of *Listeria* sp. and a high fecal coliform loading. Hence, following *Médecins Sans Frontières* (MSF) recommendations, we suggested the women to

Table 24.1 Women's perception of oyster mushrooms cultivation

Category	Women's answers appear in *italics*
Which were the main difficulties you found associated to the first cultivation pilot?	*"The routine of the process is sometimes exhausting"* "*The control of mosquitoes is hard to handle*" "*Sometimes I forgot to boil the water and use the syringe"* *"It is not as simple as it is to cultivate potatoes, we have to be more careful about contamination"* Contamination was a common problem "*This green color in the bag is growing"* *"I take the green part from the bag with a cleaned knife"* *"All the bags filled with hay were contaminated"*
Could you teach the whole process to another person?	Every woman said that she was prepared to replicate the whole process by herself and to teach it to other farmers: "The *first time I did not understand very much, but you learn as you do it*" *"It is not a complex process, and rice husk is easier to handle"* *"I will teach the whole process since it is easy"* *"Yes, I will teach the process with what I have learned so far"* *"Yes, it is like raising chickens or plants, is not complicated"*
Which is your first perception of the oyster mushrooms harvesting process?	"I *thought that the culture was damaged since a brown color appeared in the cap*" *"In the morning, I found like an old white dust on the tables"*—We explained that this was the color of fungal spores, and that spores from different fungi are different colors, additionally we explained that if we were growing common champignons, the *dust* would be brown *"They are like little umbrellas, they are beauty"* *"They are like grey ears, I can't wait to eat them"* *"For me they are like cup-shaped trees"* One woman named the spawn: *"el cocido"*
Did anyone help you in the process or did you carry out the complete procedure?	*"I did everything by myself"* *"A friend helped me with the irrigation of the bags"* *"Sometimes my children helped to accommodate the bags"*
What do you think about the oyster mushroom product	"*They taste like chicken"*, "*it is delicious,*" "*they are harder to chew than common champignons*" "*My mom loves this . . .*" "*. . . we like this protein source because it is healthy,"* *"My godfather stopped by and said, I am interested in this fungus,"* *"on occasions the oyster mushrooms replace meat," "my daughter likes the taste, and it is good for her because she suffers from hypoglycemia."*
How many bags are you willing to prepare?	"*Maybe 20*" *"I will make 10 bags"* "*I will culture all the bags that fit on the shelves"*

treat each liter of water, used during incubation and spraying steps, with a drop of hypochlorite.

Women's Perception of the Cultivation Process and the Inclusion of a New Fungal Product

In every meeting with the group of women, we explained the nutritional properties of the oyster mushrooms as an alternative to supplement the absence of meat-based protein. Some common perceptions and comments about the cultivation process are summarized in Table 24.1. The general perception regarding the production process was positive and the women agreed that it was a promising option for their families' daily nutrition. Progressively, they began to consume the mushrooms as creamy soups, grilled with butter and garlic, with chicken breasts, and cooked in soups and spaghetti. The women perceived that the mushrooms taste like chicken, and that they are harder to chew than normal champignons.

24.3.2 Business Plan

This section outlines the components of the business plan: First, we present a general description of the company, followed by the market study. We then detailed the technical and organizational study as well as the economic analysis shown in section "Organizational Study." Finally, we propose an implementation plan in section "Implementation Plan."

Business Description

The VWA women chose the name *"Orellanas de la Villa"* for their company, which seeks to satisfy the demand of gourmet restaurants, organic markets in Bogotá, local customers, and ultimately improve the economic conditions of the Villapinzón population (Díaz and Mesa 2013).

We applied an in-depth SWOT analysis to the company in order to study its Strengths, Weaknesses, Opportunities and Threats (SWOT) (Fig. 24.4). The tool assessed internal and external aspects and identified the positive characteristics that could help develop the business, as well as the negative issues that could be harmful to the company. Figure 24.4 summarizes this analysis.

Orellanas de la Villa's mission is to capture and retain customers, providing them with a quality product that is 100% organic and contributes to a healthy diet. It should do this while, at the same time, maintaining high quality standards that are reflected in an outstanding business performance, adequately rewarding all those who invest ideas and work in the company. By 2019, *Orellanas de la Villa* will be one of the leading brands in the organic mushroom market, serving major markets

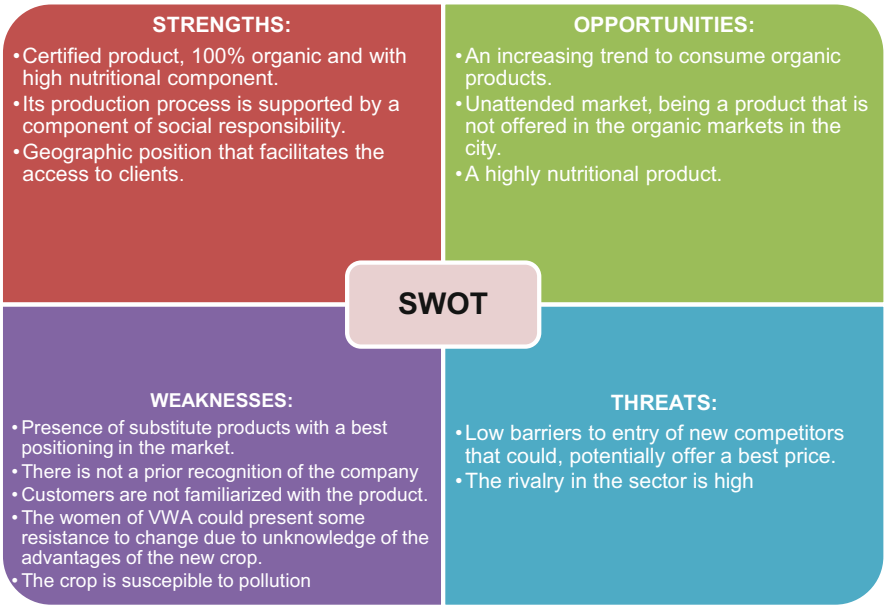


Fig. 24.4 SWOT matrix

and restaurants in Colombia. This will be represented by a gradual increase in sales, and a reduction of poverty in the municipality of Villapinzón, as mushroom sales become a stable source of income for the workers and stakeholders. The main strategy is to reduce the number of intermediaries by using the municipality's geographical positioning as an advantage to quickly access markets, and develop its own distribution channel from the producer directly to the client. The company's production system guarantees an organic product that will soon be certified.

Market Study

According to Macro Setas Colombia (2012), the oyster mushroom market in the country is growing at a rate of 0.0015 tons a year, meaning that the projected demand for the coming years is estimated at 728 tons a year. This market is satisfied by national production that currently has a capacity to harvest 11 tons a month (Perilla et al. 2005), or 132 tons per year, representing less than 20% of the market. Other companies could potentially cover the remaining 80% gap in the market. More specifically, in Bogotá alone, and only considering gourmet restaurants and organic markets—with more than 4000 restaurants located near 12 organic markets (Acodres 2006) and more than one million inhabitants in the high-income bracket—the demand is estimated at 18 tons a year. This potential market of 18 tons a year is seven tons more than the existing national production.

Once we identified the target market, we contacted the most important restaurants and markets in Bogotá (Appendix, Table 24.6), finding that *Orellanas de la Villa* has the opportunity to start producing and commercializing approximately 765 kg a month, representing 50% of the demand in Bogotá. The company's 10-year goal is to be supplying both national and international markets.

Nationally, the competitors' analysis shows that there are four potential competitors: (1) *ASOFUNGICOL*: a farmers' association in the department of Huila that produces and commercializes oyster mushrooms; (2) *CasOrellana*: a group of companies located in the Valle del Cauca region that produces oyster mushrooms on local and regional levels; (3) *Setas de Boyacá*: a producers' network that collects from more than 100 farming families in the Boyacá region; (4) *AMUSEF*: a women's association that produces a diversity of edible mushrooms in Usme, Bogotá.

In order to evaluate the competitive positioning of *Orellanas de la Villa,* we developed a curve value chart (Appendix, Fig. 24.6), which shows seven features associated with the production, commercialization, and distribution of the fungal product: price, organic certification, social responsibility, market access, distribution channels, product presentation, and technical support. We can see that, for these seven attributes, *Orellanas de la Villa* has a better competitive position than other producers. Given that the company's proposal is to (a) reduce intermediaries, (b) use geographical positioning to quickly access markets, and (c) take control of distribution activities, the price will be favorable for both farmers and clients.

The marketing strategy will be developed based on organic certification and the fact that the growers themselves harvest and commercialize the product. Also, given that oyster mushroom consumption is not common in Colombia, *Orellanas de la Villa* will add an oyster mushroom recipe on their labels for the product (Appendix, Fig. 24.7).

Technical Study: Production Process

We structured the production process into eight stages—from grain preparation to distribution—based on the cultivation process undertaken in six households. Table 24.2 specifies the raw materials, the resources required, and the processing time for each stage. The production process is currently being undertaken at the women's houses, and stages 1–4 are executed in their kitchens using the equipment with which they usually cook. Before each stage, they must sanitize all the equipment and clean the kitchen to guarantee the conditions required. During the dark and light phases, the bags must be irrigated at least twice a day. For these stages, each woman has adapted a space of approximately 10 m^2, where they can simultaneously assemble ten plastic bags.

Table 24.2 Oyster mushroom production process requirements

Stage	Raw material	Resource	Time
1. Grain preparation and sterilization (section "Grain Sterilization")	Wheat Grain (rice) Hot water	Pot Glass bottle Marmite	150 min
2. Inoculation (section "Preparing the Spawn")	Mycelium Grain preparation	Sterilized storage area Burner	24 days
3. Substrate sterilization (section "Use of the Available Resources: Sterilization of the Substrates")	Substrate (sawdust, rice husk, hay) Supplements (molasses, coffee grounds waste)	Marmite (70 L) Fabric sac	10.5 h
4. Bags setting (section "Bags Incubation, Fruiting Bodies Collection, and Packaging")	Spawn Sterilized substrate Plastic bags Tube Chiffon	Burner	15 min a bag
5. Dark phase (section "Bags Incubation, Fruiting Bodies Collection, and Packaging")	Sterilized water	Syringes Black plastic Greenhouse	35 min several times a day until . . .
6. Light phase (section "Bags Incubation, Fruiting Bodies Collection, and Packaging")	Sterilized water	Syringes Greenhouse	35 min several times a day until . . .
7. Collection (section "Bags Incubation, Fruiting Bodies Collection, and Packaging")	Collected oyster mushroom	Cutter Freezer	10 min a bag
8. Package (section "Bags Incubation, Fruiting Bodies Collection, and Packaging")	Bags, tray, labels	Balance	20 min

Raw Materials, Suppliers and Production Equipment

Universidad de los Andes and the local government provided technical support to identify who in the community could supply the most important raw materials required. *Universidad de los Andes* provided the fungal strain; the plastic bags to be used in steps 4–7 were sometimes recycled from the women's homes (i.e., from empty bags of sugar or rice); and the remaining materials were provided by local markets (see a summary of materials, suppliers, and prices in Appendix, Table 24.7). The equipment required to produce the oyster mushrooms is summarized in Appendix, Table 24.8, which also outlines additional information regarding costs.

For a fixed production level of 765 kg a month (see section Technical Study: Production Process), the following materials are required: 765 plastic bags, 1913 kg of substrate, 38 kg of supplement, 3.8 m of PVC piping, 6.1 m of chiffon, 38 kg of grain, and 3825 m^3 of water. If the production process were to be concentrated in one specific area, it would require 142 m^2. This distribution is presented in Appendix, Fig. 24.8.

Organizational Study

As part of the technical support, an organizational analysis was carried out to identify the relationships within the community, and to determine the appropriate organizational structure for the development of company activities.

Organizational structure *Orellanas de la Villa* is a company created with and for the VWA, with the technical support of *Universidad de los Andes*. Its purpose is to promote the production, marketing, and consumption of organic oyster mushrooms, using the available raw substrates in rural Villapinzón. At the same time, the company was designed to provide an economic alternative to the women in the community, initially guaranteeing a monthly income of a quarter of the Colombian minimum wage.

The proposed organizational structure for *Orellanas de la Villa* is based on the VWA's current established organization. High ranked positions are to be filled by external staff, as they require more specific knowledge in the management and operational processes. It is important to highlight that some of these management positions could be filled with people from the community or the association if they have the necessary expertise, thus giving them the opportunity to climb the organizational structure.

The salaries associated to each job are presented in Appendix, Table 24.9. The General Manager and Operations Manager are currently from the *Universidad de los Andes* team, but, in the near future, it should be the women who assume this responsibility. As the production system is not yet working, a financial advisor is not yet necessary. The VWA also has a legal advisor who works with the local government and usually supports the women's activities.

In order to legally establish *Orellanas de la Villa*, and according to national policies, a set of activities need to be carried out such as commercial registration, the payment of commercial fees, and adherence to the required regulations that guarantee adequate food handling and processing. For this process, the company must seek advice from a legal consultant or the Bogotá Chamber of Commerce (CCB).

Financial Study

To formally establish *Orellanas de la Villa,* an estimated initial investment of 10 million COP is required. The financial analysis was developed on the basis that: (1) according to the market study, *Orellanas de la Villa* will sell 765 kg a month, or 51% of the current market; (2) *Orellanas de la Villa's* target market will consist of 80% restaurants (the product will be provided in 1 kg bags) and 20% organic markets (the product will be provided in a 250-g tray); (3) Oyster mushroom consumption will present a growth of 13% (Macro Setas Colombia 2012); (4) the safety stock will be 1% of supplies required; (5) the clients will pay on delivery; (6) the annual

Table 24.3 Total cost production of 1 kg of oyster mushrooms

Production cost (1 kg)			Observation
Material	Quantity	Cost (US$)	
Plastic bag	1	0.20	Polypropylene bags
Substrate	2.5 kg	0.16	55 kg rice hulls US3.2$
Supplement	50 g	0.01	8 kg molasses US$ 2.4
PVC pipe	5 cm	0.02	2 m of PVC pipe of 6 cm of diameter—US$ 0.80
Chiffon	8 cm	0.04	Box of 4 m US$ 1.6
Rubber	1	0.03	Box of 100 units US$ 3.0
Grain	50 g	0.08	2 kg US$ 1.60
Water	5 m^3	0.40	1 L US$ 0.08
Total production cost		**0.94**	

increase in expenses will be of 3% CPI (Consumer Price Index); and (7) *Orellanas de la Villa* will pay suppliers when the raw materials are acquired.

Considering only the variable costs, the production cost of 1 kg of oyster mushroom is US$ 0.94, as shown in Table 24.3. The equipment required for the production process and investment in additional equipment are given in Appendix, Tables 24.8 and A.5, respectively, including the packaging costs, (Appendix, Table 24.11). We can then estimate that the total production cost of a 1-kg bag is US$ 1.07 and US$ 0.38 for a 250-g tray. This final investment considers all administrative requirements to establish the company and to obtain an organic certificate. Taking into consideration the production, packing, and equipment costs, the analysis estimates an approximate selling price of US$ 3.91 for a 1-kg bag of oyster mushroom, and US$ 1.96 for each tray. Given these prices, and assuming that 80% of the product is stored in 1-kg bags, the equilibrium point for the two types of packaged products will be to produce and sell 7294 1-kg bags and 5608 250-g trays a year, corresponding to a yearly production of 8696 kg.

Finally, we carried out a financial simulation to assess the financial sustainability of *Orellanas de la Villa,* by assessing three scenarios, summarized in Table 24.5: (1) a pessimistic scenario that considers a 5% market growth rate, and in which there is no company expansion; (2) a neutral scenario, which considers a normal annual growth rate of 13%; and (3) an optimistic scenario, which considers a 16% annual market growth rate.

In these scenarios, and contemplating the Net Present Value (NPV) for a period of up to 10 years, the results shows that even in the most pessimistic scenario the company will generate profit (Table 24.4). According to the business plan and merchandising analyses, each woman has to cultivate 60 bags per month in order to earn a quarter of a Colombian minimum wage (Table 24.5).

Table 24.4 NPV for three scenarios analyzed

Scenario (NPV 10 years)	Value (COP)
Pessimistic: considers a market growth rate of 5%. There is no company expansion	$5.962.180
Neutral: considers the normal annual growth rate of 13%	$110.826.615
Optimistic: considers that the market growth rate is of 16% a year	$242.588.452

Table 24.5 Bags produced related to the Colombian salary

	Half a salary–344.000 (COP)[a]	Quarter of a salary–172.000 (COP)
Workers	9	9
Bags per woman	60	45
Total kg per month	765	405

[a]Salary in the year 2015, Banco de la Republica (2015)

Implementation Plan

We propose a plan of action to develop the company over a 10-year period. In the short term, the company will have a monthly income of US$ 119, considering the economic constraints that oyster mushrooms will be under at the beginning of the project. During this phase, the company will develop all microbiological tests of the fungal product, obtain the organic certification, develop an intensive marketing strategy, and increase its production. In the midterm, after the fifth year, the company will increase workers' salaries, guaranteeing a minimum wage for each technician, and it will be able to build four new greenhouses, allowing for an increase in production to 940 kg a month. In the long term, at the end of tenth year, the company will be well established in the national market, and will have developed a plan to export to the USA, Canada, the United Kingdom, and Germany.

24.4 Discussion

We are convinced that in order to meet the millennium development goals (MDG), rural communities must be aware of the importance of diversifying their products, and of expanding the possibilities to take greater advantage of nutritive biological sources. In this project, we have proposed a sustainable food-based approach for low-income rural women in Villapinzón (VP), a municipality in Colombia, by establishing the conditions for oyster mushroom production with local resources. To do this, we analyzed the viability of the product's commercialization in the proposed business plan, as well as in local stakeholder meetings in rural communities in VP. Despite mushrooms not being regularly consumed in the area, the new information we offered to rural women dealt with the technical aspects of cultivation, mushroom biology, and their health-promoting properties. These have been essential steps in establishing sustainable management of a daily dietary source and income.

24.4.1 How Did Rural Women Become Involved in the Cultivation Process?

Teaching the women how to cultivate the mushrooms was successfully implemented using the substrates available in the rural communities. This is due, in part, to the high adaptability of *P. ostreatus* to a wide variety of substrates, making it possible to take advantage of a sub-utilized substrate. The women aim to reach a biological efficiency of 25%, which will make supplying the product economically viable.

Low-income rural women from the VWA intend to replicate the process by teaching other women and by overcoming the difficulties that occur during the cultivation process. Although production is not labor-intensive, the women had to acquire new skills that are very different from those they are used to, for example those required for potato cultivation. These new skills include sterilization with pressure and heat, disinfecting surfaces, the use of tools that are always boiled in water, the use of gloves and surgical masks, the maintenance and replication of the fungal source (mycelium), and the recognition of every stage in the mushroom production, among others.

The learning of new alternatives to the common potato crop for product diversification involved providing information about oyster mushrooms' nutritional facts as well as therapeutic benefits, and post-harvest uses. On the other hand, information of the nutritional facts, therapeutic benefits and post-harvest uses of *Pleurotus ostreatus*, was strategic in getting the women interested in the mushroom cultiviation as a new alternative to the traditional potato crop. Particularly, *Pleurotus ostreatus* is a nutritive product having high content of protein, vitamins, polyunsaturated fatty acids, antioxidant and anti-inflammatory properties (Rathee et al. 2012). The post-harvesting activities are also beneficial in that they provide opportunities to promote the sustainability of small farming systems, given that residues can be used as a growing substrate and then returned to the land as fertilizer (Marshal and Nail 2009), compost, or animal feed (i.e., for cattle).

Threats to the cultivation process always demanded VWA action. The proactive initiative of women when facing problems such as contamination and dehydration helped to solve these issues; for example, they learned how to eliminate contaminants (commonly associated to species of mold such as *Trichoderma* spp.) when they appeared in the bags. A number of workshops that are currently being run with the women have highlighted the need for incubating rooms, in which the humidity and temperature can be controlled, allowing the mushrooms to withstand high dry-season temperatures and dehydration. This will lead to an optimal production of *P. ostreatus,* and ensure the demand for personal nutrition and business activities. The women recently visited the *Universidad de Los Andes* campus in what proved to be a very enriching experience, as they were able to compare their homemade process with a laboratory one.

24.4.2 Advantages of Cultivating Oyster Mushrooms

There are a few advantages of growing this fungal species over other fungal species. The cultivation process is easy when compared to the infrastructure required for other edible fungi such as traditional champignon mushrooms. Also, it does not require a large initial investment, and the oyster mushroom is a fairly complete nutritional source. Moreover, according to financial and marketing studies, there is a high level of unmet demand in the Colombian oyster mushroom market, and more specifically, demand was found to exist in the organic market. In Colombia, oyster mushrooms producers recorded a monthly production of around 11 tons, or 132 tons per year; however, in 2014, the same studies established a potential national market of 728 tons per year, showing an opportunity for growth (Mesa and Diaz 2014). Whereas continents such as Europe and Asia are traditionally known for their high levels of production and consumption, in Colombia, mushroom cultivation is still an emerging activity that is considered a good and viable alternative to traditional cultivation, given the sustainability of substrates and the null environmental impact. Small farmers are the main producers; however, the low quantities produced and the high demand of gourmet restaurants still make imports an important option when it comes to accessing the product. As such, oyster mushroom production is a highly viable business opportunity for the municipality of Villapinzón.

24.4.3 Future Perspectives: A Potential New Project Based on the Previous Case Experience

Several countries have attempted to promote the edible fungi industry as a means of providing incentives for forest conservation and improving the earning opportunities of people living in marginal rural areas (Ortega-Martínez and Martínez-Peña 2008; Cai et al. 2011). For many decades, interest in the commercial harvesting of wild edible fungi has increased considerably in many regions. For instance, in the USA, recreational and subsistence mushroom harvesters have found an abundance of mushrooms in nearby forests, which has led unemployed timber workers to sell mushrooms as an alternative source of income (Pilz et al. 2003).

Following this perspective, we argue that proposing strategies to characterize and recognize the species of saprotrophic edible fungi that can be cultivated will enhance local knowledge about the use of fungal diversity, and the understanding of its importance as part of the ecosystem. Considering the above, rural farmers have been flexible in incorporating new alternatives, such as the oyster mushroom, to compensate for the lack of nutritive food and to potentially improve their economic situations with the establishment of *Orellanas de la Villa* company.

Additionally, the Colombian State is in the process of building the Plan for Food Security and Nutrition, and, as such, the search for alternatives to contribute to the management of the community's complementary nutritional needs as a relevant issue for the country's development. This project will seek to further explore the potential of saprotrophic fungal species as a Non-Wood Forest Product (NWFP). As stated by FAO (2001): it is . . . "an interesting product to be used by human society, regarding its nutrition values."

24.5 Concluding Remarks

The interdisciplinary project that received conjunct contribution from three research areas, proved to be a strategic approach to promote the knowledge appropriation of producing a new product, and its subsequent development. The case study shown here was proposed as a promising option to help mitigate the effects of poverty, hunger, inequality, and consequently, as a basis for further studies related to the emergent research field in conservation and the sustainable use of fungi in Colombia. Moreover, the community involved in the project was very interested in learning about an unfamiliar agricultural activity, and so they were attentive, ready to solve cultivation issues, and creative in the way they integrated oyster mushroom consumption.

According to the municipality development plan (Development Plan 2012–2016), entitled "Villapinzón, the path to progress," the government is aiming to produce favorable employment conditions for companies to increase the productive sector (Development Plan 2012). Drawing on this precept, this study used a strategy that relies on the promotion of economic and nutritional improvement by turning a cultivation activity into a company, *Orellanas de la Villa*. The successful results obtained from the experience and the organic product itself are being shown to potential consumers, who have agreed to taste and to classify a high-quality product.

Acknowledgements This research was partially supported by Vice-Presidency of Research, *Universidad de los Andes*. We thank all the participants of the Asociación de Mujeres de Villapinzon, the UMATA of Villapinzon, and the Alcaldía de Villapinzón who gave us the opportunity to work with the women. We also thank the Laboratory of Food Microbiology (LEMA) at *Universidad de los Andes* that carried out the routine microbiological analyses.

Appendix

Table 24.6 Potential clients and situation of negotiation

Establishment	Activity	State
Restaurants		
Wok	The procurement manager is interested in acquiring Orellana from *Orellanas de la Villa*. She has been constantly in contact	It is necessary to produce a sample
Balzac (Harry Sasson)	The Main Chef is interested in acquiring oyster mushrooms from *Orellanas de la Villa*	It is necessary to produce a sample
Café Renault	They already know the product and the company	Continuing to contact them
Teriyaki	They already know the product and the company	Continuing to contact them
La Monferrina	The procurement manager is interested in buying the product	It is necessary to produce a sample and guarantee procurement
Organic markets		
Escarola	The client was visited	It is important to generate an agreement. It is important to define the time and volume to procure each week and define when production will start
Bioplaza	The client was visited	It is necessary to generate a proposal with the price and the lead-time defined
Vivir Bonito	The client was visited	It is necessary to generate a proposal with the price and the lead-time defined
Clorofila	The client was visited	They demanded a proposal and a sample
Supermarkets		
Jumbo	The client was contacted	They demanded a proposal following company criteria

Table 24.7 Raw materials and their suppliers

Raw material	Supplier	Price
Grain (rice)	Local market	2 US$/kg
Hot water	Local drinking water distribution system	0.1 US$/L
Mycelium	The Laboratory of Mycology and Phytopathology of Universidad de los Andes taught the women to produce the seed from the one used in previous assays	10 units—5 US$
Substrate (sawdust, rice-husk, hay)	The rural community	55kg—4 US$
Supplements (molasses, coffee grounds)	The rural community	8 kg—3 US$
Plastic bags	Local market Home recycling bags	0.25 US$/Unit
Chiffon	Local market	
Tube	Local market—it could be reused	2 m—1 US$
Sterilized water	Local drinking water distribution system	0.1 US$/l
Bags to pack	Local market	5000 units—4.5 US$
Tray to pack	Local market	500 units—12.4 US$
Labels to pack	Printers from Bogota	1000 units—150 US$

Table 24.8 Equipment required for the production process

Equipment	Quantity	Cost (US$)[a]	Total (US$)
Steel tables	2	100.0	200.0
Fridge	1	416.0	416.0
Cooker	2	32.0	64.0
Industrial hob	2	107.0	214.0
Plastic containers	20	4.0	80.0
Balance	2	16.0	32.0
Buckets	4	2.5	10.0
Packing machine	1	218.0	218.0
Burners	4	1.5	6.0
Thermometer	1	4.5	4.5
Thermos PS	5	8.5	42.5
Total investment in equipment			1287.0

[a]Costs have been estimated in Colombian pesos (COP), this table shows the prices in USD using a TRM of 1 USD = 2522 $COL

Table 24.9 Organizational team and salaries at *Orellanas de la Villa*

Position	Number of employees	Salary[a] (US$)	Observation
General manager	1	460.0	
Production manager	1	131.5	
Financial advisor	1	198.2	Advisors are hourly paid
Juridical advisor	1	198.2	Advisors are hourly paid
Technicians	8	920.5	*Orellanas de la Villa* starts its production with eight women that belong to the association

[a]Salaries include all benefits according to Colombian labor laws

Table 24.10 Other initial investment

Item	Cost (US$)
Computer	294.0
Software	63.5
Company establishment	307.0
Working capital	79.5
Initial inventory	630.0
Greenhouse adjustments	397.5
Organic certification	516.5
Organoleptic tests	274.0
Total	2562.0

Table 24.11 Packing cost

Packing material	Presentation	Cost (US$)
Polystyrene tray	500 units	10.0
Cling film	1500 m	18.5
Plastic Bag 1 kg	100 units	3.5
Label	1000	119.0

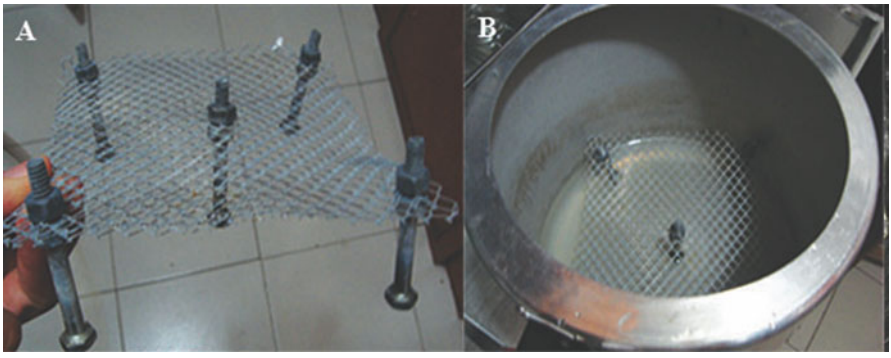


Fig. 24.5 Handmade mesh

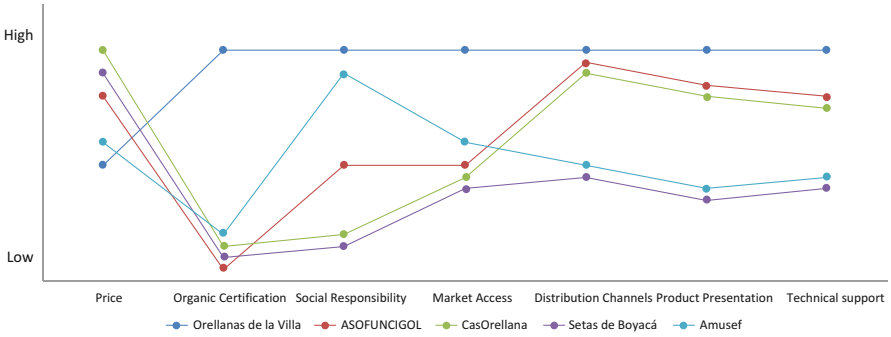


Fig. 24.6 Comparison between traditional mushrooms providers

Fig. 24.7 Product label

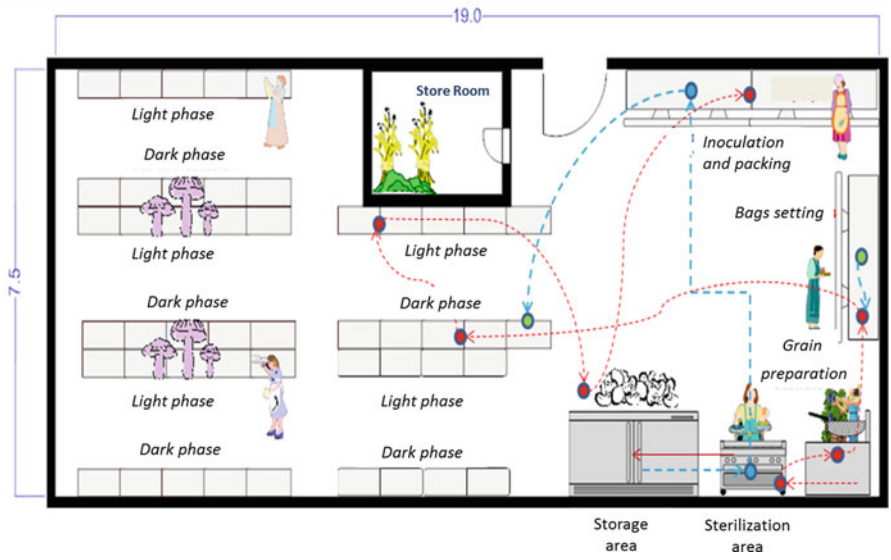


Fig. 24.8 Facility layout

References

Acodres (2006) Informe restaurantes. Dinámica del sector. Restaurantes estrato 5 y 6. http://www.catering.com.co/ediciones_catering/EDICION18/manteles.pdf. Accessed 10 Feb 2014

Albertó E (2008) Cultivo intensivo de los Hongos comestibles. Editorial Hemisferio Sur, Buenos Aires

Andrade R, Mata G, Sánchez J (2012) La producción iberoamericana de hongos comestibles en el contexto internacional. In: Sánchez J, Mata G (eds) *Hongos Comestibles y Medicinales en Iberoamerica*. El instituto de Ecología, Chiapas

Banco de la Republica (2015) Salario mínimo legal en Colombia. Serie histórica. http://www.banrep.gov.co/es/mercado-laboral/salarios. Accessed 15 May 2018

Bautista E, Torres MF (2012) Diagnostico mixto para la superación de pobreza en Villapinzon, Cundinamarca: Identificación de algunas áreas prioritarias de intervención. Universidad de los Andes, Bogota

Boa E (2004) Wild edible fungi: global A overview of their use and importance to people. In: *Non-wood forest products 17*. FAO, Rome. http://www.fao.org/docrep/007/y5489e/y5489e00.htm#TopOfPage. Accessed 4 May 2018

Cai M, Pettenella D, Vidale E (2011) Income generation from wild mushrooms in marginal rural areas. Forest Policy Econ 13:221–226

Cheung PC (2010) The nutritional and health benefits of mushrooms. Nutr Bull 35:292–299

Corporación Colombia Internacional (2004) Setas y Hongos Boletín 21. http://www.agronet.gov.co/www/docs_agronet/200511314480_perfil_producto_setas.pdf. Accessed 26 Jan 2014

DANE (2001). http://www.dane.gov.co/files/investigaciones/agropecuario/ena/censo_papa_villapinzon.pdf. Accessed 13 Sept 2013

Development Plan (2012). Plan de desarrollo del municipio. "Villapinzón, el camino del progreso". Consejo Municipal, Villapinzón

Díaz MA, Mesa L (2013) Orellanas de la Vida. Monografía de grado Ingeniería Industrial. Universidad de los Andes, Bogotá

Ehmke C, Akridge J (2005) The elements of a business plan: first steps for new entrepreneurs. Agricultural Innovation and Commercialization Center. Purdue Extension. EC-735. https://www.extension.purdue.edu/extmedia/ec/ec-735.pdf. Accessed 10 Feb 2015

FAO (2001) Resource assessment of non-wood forest products. http://www.fao.org/DOCREP/004/Y1457E/Y1457E00.HTM. Accessed 27 April 2013

FAO (2018) Reduce rural poverty. http://www.fao.org/about/what-we-do/en/. Accessed 3 May 2018

FEDEPAPA (2013) Papas nativas con valor agregado. http://www.fedepapa.com/wp-content/uploads/pdf/revistas/ed29.pdf. Accessed 5 May 2014

Gaitán-Hernández R, Salmones D, Pérez Merlo R, Mata G (2006) Manual práctico del cultivo de setas: aislamiento, siembra y producción, 2a. reimp. Instituto de Ecología, Xalapa

Guzmán G, Mata G, Salmones D, Soto-Velasco C, Guzmán-Dávalos L (1993) El cultivo de los Hongos comestibles. Instituto politécnico nacional, Xalapa

Joumard I, Londoño J (2013) Income inequality and poverty in Colombia—Part 1. The role of the labour market. OECD Economics Department Working Papers, No. 1036. OECD Publishing, Paris

Kalač P (2013) A review of chemical composition and nutritional value of wild-growing and cultivated mushrooms: chemical composition of edible mushrooms. J Sci Food Agric 93:209–218. https://doi.org/10.1002/jsfa.5960

Macro Setas Colombia (2012) Macro Setas Colombia. http://macrosetascolombia.com/paginas/mercado.html. Accessed 26 Jan 2014

Marshal E, Nail N (2009) Make money by growing mushrooms. FAO diversification booklets

Mesa L, Diaz MA (2014) Plan de negocio Orellanas de la Villa. Capstone project. Industrial Engineering Department, Universidad de los Andes

Michael HW, Bultosa G, Pant LM (2011) Nutritional contents of three edible oyster mushrooms grown on two substrates at Haramaya, Ethiopia, and sensory properties of boiled mushroom and mushroom sauce: nutrient of edible oyster mushrooms. Int J Food Sci Technol 46:732–738. https://doi.org/10.1111/j.1365-2621.2010.02543.x

Millennium Development Goals–MDGs (2018) The eight Milleinum Development Goal (MDGs). http://www.un.org/millenniumgoals/. Accessed 3 May 2018

Ortega-Martínez P, Martínez-Peña F (2008) A sampling method for estimating sporocarps production of wild edible mushrooms of social and economic interest. Investigación Agraria: Sistemas y Recursos Forestales 17:228–237

Perilla C, Palomino C, Orozco M (2005) Creación de la empresa comercializadora de orellanas. Fundacion Universitaria Ceipa, Neiva

Pilz D, Molina R (1996) Managing forest ecosystems to conserve fungus diversity and sustain wild mushroom harvests. General Technical Report PNW-GTR-371. U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station, Portland

Pilz D, Norvell L, Danell E, Molina R (2003) Ecology and management of commercially harvested Chanterelle mushrooms. General Technical Reports PNW-GTR-576. U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station, Portland

Puttaraju NG, Venkateshaiah SU, Dharmesh SM, Urs SMN, Somasundaram R (2006) Antioxidant activity of indigenous edible mushrooms. J Agric Food Chem 54:9764–9772

Rathee S, Rathee D, Rathee D, Kumar V, Rathee P (2012) Mushrooms as therapeutic agents. Rev Bras 22:459–474

Sher H, Al-Yemeni M, Khan K (2011) Cultivation of the oyster mushroom (Pleurotus ostreatus (Jacq) p. Kumm.) in two different agro-ecological zones of Pakistan. Afr J Biotechnol 10: 183–188

Natalia Vargas Estupiñán is a Microbiologist working as a postdoctoral researcher in the Laboratory of Mycology and Plant Pathology in the Biological Sciences Department at Universidad de Los Andes, during the last two years. She has a M.Sc. in Microbiology and developed her PhD in Biology with a scholarship supported by the Administrative Department of Science, Technology and Innovation—Colciencias. Her PhD project focused in the taxonomy and phylogeography of fungal species, invasive ecology of introduced fungi, social aspects and recommendations for protecting macrofungi and teaching techniques for cultivating of oyster mushrooms. The project developed with low income rural women in Villapinzón was part of her PhD, integrating other disciplines including Engineering and the Governance and Management School.

Dr. Silvia Restrepo is a Full professor in the Biological Sciences department, at Universidad de los Andes. Currently, she is the Vice president for Research and was the head of the Biological Sciences department and Dean of the School of Sciences. Dr. Restrepo is a leading expert in plant pathology and has developed research on diseases of cassava and solanaceous crops. Dr. Restrepo teaches plant pathology for undergraduate students and advanced molecular biology for graduate students. Dr. Restrepo has received several awards, like the Elizabeth Grose prize from the School of Sciences for her distinguished research, the Third World Academy of Sciences award to young scientists, the Prize from the Institut de Recherche pour le Développement for her doctoral thesis, and the Louis Malassis

International Scientific Prize for agriculture and food. She also received the Merit Ordre from the French Government.

Dr. Restrepo has worked a lot for modernizing the career of microbiology at Universidad de los Andes and designed a minor in bioinformatics and a graduate program in Computational Biology. She enjoys teaching young kids in schools the scientific thinking and proposed to start a research program in STEM as an interdisciplinary endeavor gathering the schools of engineering, sciences and architecture and design. Writing a chapter for this series has been a wonderful opportunity for her to work once again with her former doctoral advisee, Dr. Natalia Vargas and her colleague from the Management School, Nubia Velasco.

Nubia Velasco has been an assistant professor in the Management School at Universidad de los Andes for the last three years. Prior to this, she held the same position at the Engineering School for nine years. She is a Chemical Engineer, holds an M.Sc. in Industrial Engineering and earned her PhD in Applied Automation and Information from Université de Nantes. Professor Velasco leads the Supply Chain Management and Technology research group. Her research focused on the use of OR techniques to optimize the distribution activities in production and service organizations. Currently, she is working to develop optimization decision tools to improve the operation in health organizations. With the work presented in this series, she has had the opportunity to integrate different disciplines: microbiology, engineering, governance, and management in a project developed with the Villapinzon Women Association, together with her colleagues Silvia Restrepo, Natalia Vargas, and Carmen Gutierrez. This experience has motivated professor Velasco in continuing to support other women communities.

Chapter 25
Data-Driven Intelligent Predictive Maintenance of Industrial Assets

Olga Fink

Contents

25.1 Introduction to Predictive Maintenance 589
25.2 Five Levels of Condition-Based and Predictive Maintenance 591
25.3 From Feature Engineering to Feature Learning 593
25.4 Autoencoders for Fault Detection Based on Signal Reconstruction 596
25.5 Deep Learning in Predictive Maintenance 597
25.6 Health Indicator Learning Combined with Feature Learning for Predictive Maintenance Applications 599
25.7 Discussion and Outlook 602
References 604

25.1 Introduction to Predictive Maintenance

Falling costs and increased reliability of sensing devices, data transmission and data storage have fostered condition monitoring systems to become near ubiquitous for many complex engineered systems. Concurrently, internet of things (IoT) is enabling a real-time transmission of the information on system condition captured by numerous diverse condition monitoring devices, including images and video streams. This development provides a great opportunity to use these condition monitoring data intelligently within condition-based and predictive maintenance regimes, thereby improving the availability of the systems, reducing the maintenance costs, improving the operational performance and supporting the decision maker in selecting the optimal point of time and the optimal action for the maintenance intervention.

Condition-based and predictive maintenance rely on the availability of the information on the system condition, its evolution in time and the environmental

O. Fink (✉)
Chair of Intelligent Maintenance Systems, ETH Zürich, Zurich, Switzerland
e-mail: ofink@ethz.ch

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2_25

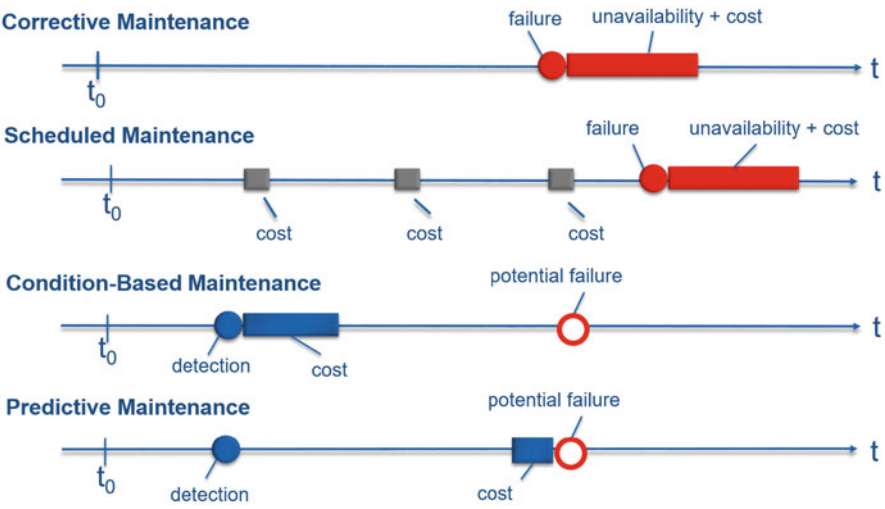


Fig. 25.1 Overview of the different maintenance strategies (It should be noted that installation costs are not considered in this presentation)

conditions and operating context, which can be acquired either in a continuous way or at discrete points in time, e.g. during inspections, or initiated by triggering events.

Once a faulty system condition has been detected, the intervention is typically performed either immediately or with a short delay in the case of condition-based maintenance. The information of the remaining useful lifetime is not available in such cases and therefore, a part of the potentially usable lifetime of a component is wasted. Also, due to the short-term planning of the required maintenance intervention, the selected point in time of the intervention may not be optimal from the operational point of view which may cause additional cost. For this reason, condition-based maintenance has higher maintenance costs compared to predictive maintenance, where the useful lifetime of a component can be fully exploited (Fig. 25.1). The main difference between condition-based and predictive maintenance is the ability to predict the remaining useful lifetime and thereby to act proactively by optimally using the lifetime of the component.

In the context of condition-based and predictive maintenance systems, also the terms prognostics and health management (PHM) are often used. However, PHM goes one step beyond predictive maintenance. PHM is defined as an engineering discipline that aims at providing users with an integrated view of the health state of an asset (Lee et al. 2014). PHM comprises fault detection, diagnostics and prognostics (Kadry 2013). Detection is the ability to detect the occurrence of a fault and the need for a maintenance intervention. The main tasks of diagnostics are fault isolation and identification of failure root causes. Prognostics is defined as the use of automatic methods to predict system performances and anticipate future faults or failures before their occurrence. This process of prognostics comprises detecting an incipient fault and predicting the remaining useful lifetime of a component. While

detection and diagnostics are part of the condition-based maintenance, prognostics can be considered as part of the predictive maintenance.

However, PHM does not end with the prediction of the remaining useful life. The system health management goes one step beyond the predictions of failure times and supports optimal maintenance and logistics decisions, based on outputs from diagnostics and prognostics, by considering the available resources, the operating context and the economic consequences of different faults. Health management is the process of taking timely and optimal maintenance actions taking into consideration available resources and operational demand (Lee et al. 2014).

Many different approaches have been introduced for condition-based and predictive maintenance applications that can be divided into three main categories:

- model-based, typically physics of failure models
- data-driven, based on condition monitoring data
- knowledge-based, relying on domain expert judgment.

There are also several hybrid approaches, combining two different categories. Due to the increased availability of condition monitoring data, the application of data-driven approaches, particularly, machine learning has been recently increasing.

25.2 Five Levels of Condition-Based and Predictive Maintenance

The very first step of any condition-based and predictive maintenance application and a basic prerequisite is a robust and cost-effective condition monitoring system that is able to capture the system condition and deviations from the expected behaviour (Fig. 25.2). This step integrates an effective sensor selection, selection of their location and measurement frequency and the implementation of the internet of things architecture (if required), the connectivity, aggregation of the measurements, etc. A commonly applied approach to decide about the selection of the suitable sensors is the fault symptom analysis that links symptoms, based on which the faults and the degradation can be detected, to the condition monitoring sensors. Depending on the considered system, the measurements of the system condition can be acquired from the control system and no additional sensors are required. This is particularly true for critical and complex systems, such as power plants. Particularly for systems that are already operating and need to be retrofitted with additional sensors, special requirements and conditions apply. Because, for example, in the case of safety critical systems if any change is taken on the system, a new safety homologation may be required which would prevent the operator from implementing additional changes to the operated system.

The second level in Fig. 25.2 comprises the actual PHM system implementation that contains fault detection, diagnostics and the prediction of the remaining useful lifetime. One of the challenges for the implementation of data-driven PHM systems

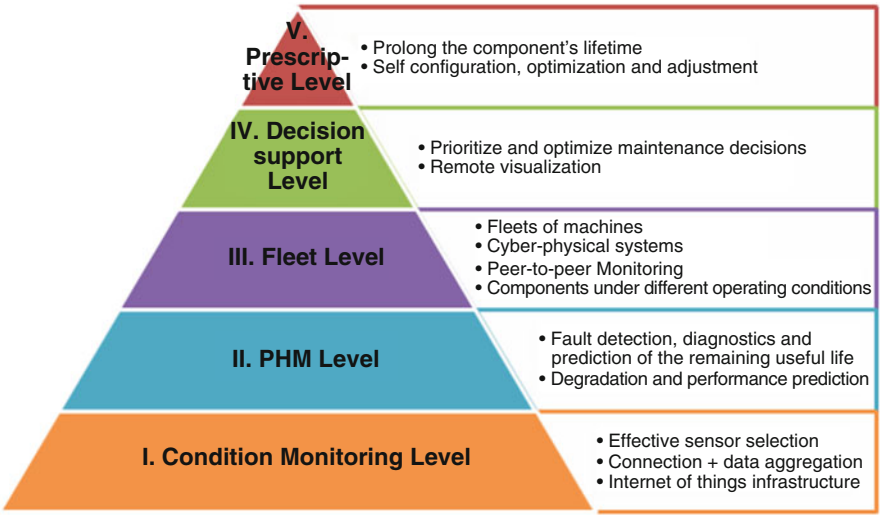


Fig. 25.2 Five levels of condition-based and predictive maintenance, demonstrating an increasing level of complexity from bottom to top, modified from Lee et al. (2014)

is the availability of representative datasets for training the developed models. The performance of these models highly depends on the training data and thereby also on the representativeness of the operating experience during the time period of data collection. This is particularly challenging if a significant change in operating conditions occurs or if the PHM system is developed for a system that is newly taken into operation. In this case, condition monitoring data for a single asset is either not available, is insufficient or is not sufficiently representative. In such cases, the number of false alarms may be significant since the previously unseen operating conditions can be now mistaken with faulty system conditions.

To overcome some of the limitations of the single asset perspective, the operating experience on the entire fleet is required. The main challenge of the third level is, therefore, to learn from the experience of single systems within a fleet. This is required because in critical systems, faults are seldom and, therefore, the information on the fault should be used within the entire fleet to shorten the required observation time period. If the systems have similar configurations and similar operating conditions, the transfer of knowledge between the single systems is possible. However, in case of diverse system configurations and diverse operating conditions, transferring fault patterns between different systems of a fleet is challenging and is a current research topic.

While one challenge is to transfer experience of the fault patterns between the single assets of a fleet, a second challenge is to ensemble a sufficiently representative training dataset for developing robust PHM algorithms from several assets of a fleet with sufficiently similar operating conditions (Michau et al. 2018). This challenge addresses the case described above, where the operating experience of a

single asset is not sufficiently representative for the expected operating regimes and can, therefore, not be used alone to train a robust PHM algorithm. The operating experience of a single asset needs to be enriched by the operating experience from other assets. The challenge here is to select representative operating points from other assets that are, on the one hand, sufficiently similar (integrating dissimilar operating conditions in the training dataset may make the algorithm confuse faulty system conditions from the considered asset with healthy system conditions from other assets) but not too similar since the operating experience needs to be enlarged in order to ensure an improved representativeness.

It is typically not sufficient just to provide the information on the system condition and detect, diagnose faults or predict the remaining useful life, a decision support is also required to integrate the decision in the production plan and take resource constraints into consideration. This is part of the fourth level of predictive maintenance systems. Even though some decision support systems have been proposed for condition-based and predictive maintenance systems (Yam et al. 2001; Thomas et al. 2009) they have been focusing on some specific boundary conditions and specific applications with defined inputs and are therefore not easily transferable to dissimilar applications. Additionally, many of the approaches focused only on one system and not on a fleet of systems.

The last level in the pyramid aims at extending predictive maintenance to prescriptive maintenance, representing the highest level of complexity but also the biggest potential impact on the system availability and operating costs. Contrary to the predictive maintenance, prescriptive maintenance does not stop with predicting the remaining useful lifetime but takes the information on the influencing parameters and the failure mechanisms and influences the remaining useful life proactively, by controlling the operating parameters. In these cases, also the term of reliability adaptive systems is used (Meyer and Sextro 2014).

Condition-based and predictive maintenance systems typically cover at least two of the levels (first and second level). Some type of (simplified) decision support is typically also often additionally integrated. The difference is typically made in the complexity of the decision support systems manifested by the number, simplification and the restrictiveness of the assumptions the decision support system is based on. Particularly the level of fleet PHM and the prescriptive level are currently subject of research.

25.3 From Feature Engineering to Feature Learning

The condition of complex systems is typically monitored by a large number of different types of sensors, such as temperature, pressure, vibration, images or even video streams of system conditions, resulting in very heterogeneous condition monitoring data. Additionally, the signals are affected by measurement and transmission noise. In many cases, the sensors are partly redundant, having several sensors measuring the same system parameter. Not all of the signals contain information

on the specific fault type since different fault types are affecting different signals. In most of the cases, raw condition monitoring data cannot be directly used to detect the faults since the information content in raw data is limited and cannot be directly processed by machine learning applications.

A traditional implementation of predictive maintenance applications typically requires a manual engineering of features. Feature engineering comprises (a) feature extraction, a step of transforming raw data into more informative representations, e.g., by statistical indicators or other signal processing approaches; followed by (b) dimensionality reduction, either with manual or automatic feature selection (filter, wrapper or embedded approaches) (Forman 2003). Feature selection depends on the past observations or knowledge of the possible fault types and their affect on the signals. Selecting too few or too many features may result in missed alarms, particularly for those fault types that have not been previously observed.

Due to the multiplicity of the possible fault types that can occur, feature engineering may face limitations to design a set of representative features that are able to depict the differences between all the possible fault types. Handcrafted features typically lack generalization ability and transferability to other systems and partly even to other fault types. They also have a limited scalability due to the expert-driven manual approach. Additionally, the performance of feature engineering is highly dependent on the experience and expertise of the domain experts performing the task. The quality of the extracted features highly influences the performance of machine learning approaches that are using the extracted features. As the number of monitored parameters increases so does the difficulty of feature engineering for diagnostics engineers and consequently the urge for automatic processes (Yan and Yu 2015).

To overcome several limitations of the traditional feature engineering approaches, feature learning or representation learning (Bengio et al. 2013) has been recently progressing and has been gaining importance in many application domains, such as speech recognition, visual object recognition, object detection and clinical predictive modeling (Bengio et al. 2013; Vincent et al. 2010).

Contrary to feature engineering, feature learning is an integrated learning process: the algorithms learn to automatically transform raw data into condensed and informative features that simplify the learning task and improve the performance of the algorithms. End-to-end learning architectures of deep learning algorithms enable to use raw data for complex learning tasks, with several steps of feature learning within the neural network architectures (Khan and Yairi 2018).

Generally, the requirements for feature learning applied for predictive maintenance based on high-dimensional condition monitoring signals include:

- the ability to learn the features without any prior information on the type and nature of condition monitoring signals,
- the possibility to use different types of condition monitoring signals as inputs, impacted by different degrees of noise;
- a good generalization ability for different types of faults;
- scalability of the feature learning approaches;

- robustness of the learned features;
- ability to learn the features without any knowledge on the possible faults.

There are two main types of feature learning: supervised and unsupervised. This is similar to the categories of machine learning applications. In supervised learning, the features are learned in an integrated way while learning to predict the provided labels, which can be either within classification or regression tasks. An inevitable requirement for supervised feature learning is the availability of a sufficient number of labeled data. In PHM applications, labels can be acquired from time-to-failure trajectories, maintenance and inspection reports or provided by domain experts. However particularly for critical systems, faults are scarce and often not representative, the operating conditions are diverse and the faults types numerous. Therefore, labels are not available or not sufficient for many applications. This is where unsupervised feature learning steps in. It can be directly applied to unlabeled input data. Autoencoders are some of the most popular unsupervised feature learning approaches.

Obviously, the learning task is simpler for the supervised feature learning applications since the learned representation can already be optimized for the learning task, for example achieving separability on a classification learning task already within the feature space. For classification tasks, higher layers of representation amplify aspects of the input that are important for discrimination and suppress irrelevant variations (Bengio et al. 2013). Unsupervised feature learning can only transform information that is already contained in the data making it easier to process it for the subsequent learning step. Also in these cases, different fault types can become separable in the feature space. However, the separability cannot be enforced since the features are not optimized for the separability objective as in the case of supervised learning.

Also in predictive maintenance applications, the use of feature learning has been progressing (Li et al. 2018; Zhang et al. 2018). While raw condition monitoring data contains many irrelevant variations, representation learning is able to amplify those characteristics of the input data that are responsible for changes in system conditions. For complex non-linear relationships in condition monitoring data, these characteristics can only be extracted in higher layers of representations and enable a discrimination between healthy and faulty conditions.

Remarkable improvements could be observed with end-to-end deep learning architectures on benchmark datasets within PHM applications (Sateesh Babu et al. 2016; Zhang et al. 2018). However, these are particularly supervised learning tasks with existing labels and features being optimized for the learning task. This indicates the potential not only for further developments in this field but also for progress on unsupervised learning.

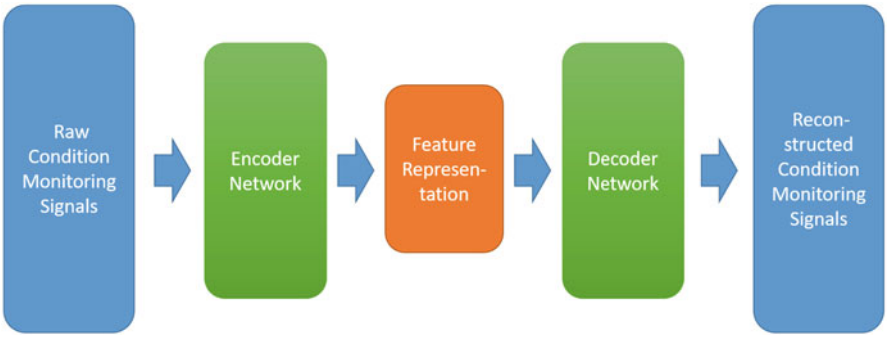


Fig. 25.3 Basic principle of an autoencoder

25.4 Autoencoders for Fault Detection Based on Signal Reconstruction

One of the data-driven approaches in the field of fault detection and diagnostics is signal reconstruction. The general idea behind the signal reconstruction is to define a model that is able to learn the normal system behaviour and to distinguish it afterwards from system states that are dissimilar to those observed under normal operating conditions. The implementation of signal reconstruction can be not only with physics-based but also increasingly with data-driven approaches. Particularly for the data-driven signal reconstruction approaches, it is important to identify representative system conditions and to use condition monitoring data from these conditions to train the algorithms to learn the underlying functional relationships.

Several approaches have been introduced for signal reconstruction, including principal component analysis (PCA), autoassociative kernel-regression (AAKR) and neural networks (Hu et al. 2017). Autoencoding neural networks are trained to reproduce their own input and contain two main parts: an encoding network and a decoding network (Fig. 25.3). The task of the encoding network is to transform the input into a dense representation of the input, while the task of the decoding network is to take this dense representation and to reconstruct the original input. Both, encoding and decoding networks can contain several layers and are typically designed symmetrically.

Given now the trained networks, which were designed to reconstruct signals under healthy operating conditions, based on the latent representation of the learned healthy conditions, the autoencoder will also reconstruct faulty signals to the expected normal system conditions, which it has been trained on, since this is the only learned latent representation. For faulty conditions, it will result in deviations between the observed and the reconstructed signals. The detection is typically performed at the level of resulting residuals between the model output and the measured condition monitoring signals. The complexity of a detection decision arises in cases of high-dimensional input signals since the decision is based on a large number of residuals with possibly different behaviour. In such cases, an

additional decision algorithm may be required combining the information contained in all the residuals including a possible prioritization of selected signals.

Besides using autoencoders as stand-alone algorithms, stacked autoencoders are a promising approach to exploit the potential of multi-layer feature learning. Stacked autoencoders are one of the deep learning approaches enabling an efficient feature learning. By learning to reproduce their own input, autoencoders compress the contained information to a lower-dimensional representation, also referred to as latent-space representation. By using this representation learned in the first autoencoder as the input to the next autoencoder, a higher level of compression can be achieved and more informative underlying relationships can be revealed. The number of used layers required to reveal the informative features depends on the complexity of the underlying relationships and the dimension of the input data.

Stacked autoencoders can be considered as unsupervised learning approaches because they do not require any additional labels (compared to supervised learning approaches).

25.5 Deep Learning in Predictive Maintenance

The application of machine learning approaches in predictive maintenance has been increasing due to the availability of a vast amount of condition monitoring data (Zhao et al. 2016). Some of the applications for fault detection for simple components have become state of the art with several industrial applications, for example the analysis of bearing vibration data for fault detection and diagnosis (Lee et al. 2014). In these cases, the relevant features and also the different fault types are well understood. However, for complex stochastic systems with multi-dimensional complex condition monitoring signals, the transfer to real industrial applications has been challenging. Some of the reasons for it are the limited scalability and transferability of the developed approaches.

Overcoming some of the challenges faced by traditional machine learning approaches, the end-to-end learning architectures of deep learning algorithms have been recently becoming one of the promising research directions in predictive maintenance (Khan and Yairi 2018; Zhao et al. 2019; Remadna et al. 2018).

Deep learning is a class of machine learning and comprises algorithms with multiple non-linear processing layers that are able to learn the underlying representations automatically from raw data. The underlying representations are typically learned by composing several layers of simple non-linear modules. At each layer, a higher level of abstraction is achieved enabling thereby an extraction of the features that are relevant for the discrimination between the classes, for example in the case of fault type classification tasks. Deep-learning methods are representation learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level.

Several deep learning approaches have been increasingly applied in different applications. One of the broad application domains of deep learning approaches is object detection in images and image segmentation (LeCun et al. 2015). Such approaches as convolutional neural networks (CNN) are particularly suitable and were specifically developed for image type 2D- or 3D-data. CNN are composed of several types of layers: convolutional layers, pooling and a fully connected layer. There are few research studies in which CNN have been applied to predictive maintenance applications (Chen et al. 2015; Li et al. 2018). In many cases, either (a) 1D-CNN have been applied, which do not fully exploit the potential of the CNN filter architecture that captures the local dependencies or (b) additional feature extraction was required before applying the CNN to the condition monitoring data. Some recent approaches proposed to transform the input into a data structure suitable for feature learning by CNN. These include grammar angular fields and 2D-Graphs (Krummenacher et al. 2018). Besides the widely used CNN approaches, several other deep learning approaches have been applied for feature learning, including deep belief networks (DBN) (Zhang et al. 2017), deep long short-term memory (LSTM) networks (Zhao et al. 2017), and different types of autoencoders (Chen and Li 2017), e.g., variational autoencoders (Yoon et al. 2017). The deeper the structure of the applied algorithms, the larger the required dataset to train all the parameters within the neural networks.

A further promising research direction in the field of deep learning are generative algorithms. One of the most widely known generative networks is the generative adversarial network (Goodfellow et al. 2014). Generative networks are a type of unsupervised neural networks that enable to generate random new outputs that are similar and follow the same characteristics as the inputs used to train the algorithm. Thereby, new previously unseen samples can be generated. This approach has been particularly used on images and the images generated by the neural networks look realistic to the observers (Goodfellow et al. 2014).

Another type of generative neural networks are the variational autoencoders (VAE) (Yoon et al. 2017). Contrary to just randomly generating samples, the VAE enable to use variations contained in the data and to control how the new samples are generated. Their basic principle is similar to the normal autoencoders. However, there is an additional constraint on the encoding network that forces the encoder to generate latent vectors that roughly follow a unit gaussian distribution (Fig. 25.4). This is enforced by integrating the Kullback–Leibler divergence (that is typically used to measure the distance between two distributions) into the loss function. This constraint enables to sample a latent vector from the gaussian distribution. The decoder can then generate a new sample based on the latent representation of the sample. While VAE are particularly beneficial for generative models, they also enforce a good feature representation, enabling also to use VAE for feature learning tasks.

One of the challenges faced by supervised deep learning applications for predictive maintenance is the dependence on the availability of labeled datasets. Particularly given the size of the required training datasets for deep architectures and the challenge of unbalanced data (with very few faults and long observation

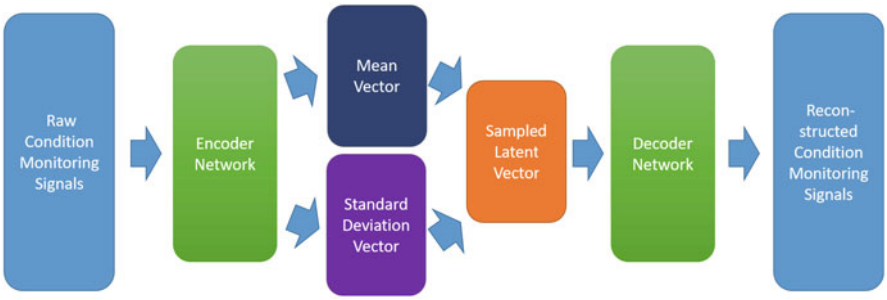


Fig. 25.4 Basic principle of a variational autoencoder

periods with healthy system conditions), deep learning architectures have had limited application fields.

One way to tackle this challenge is by obtaining additional labels for the condition monitoring data. The labels in predictive maintenance applications could be possibly obtained from time-to-failure data. However, failures are very seldom for critical systems and fault patterns cannot be easily generalized. Additionally, preventive maintenance is still performed on critical systems, only enabling to observe time-to-maintenance trajectories with limited information for fault predictions. The other possible source of labels could be obtained from a directly or indirectly measurable health indicator. However, the information on the exact system condition is either very expensive (requiring either expertise or very specific devices to estimate health indicator) or in many cases even impossible to acquire since the health condition is represented by a multi-dimensional vector and cannot be easily combined into a single health indicator. A further way of obtaining representative training datasets could be by simulating different operating conditions and using these datasets for initially training the algorithms and then incrementally updating the algorithms with real operating conditions. However, detailed simulations are very expensive and typically contain several assumptions thereby having large deviations to real processes.

25.6 Health Indicator Learning Combined with Feature Learning for Predictive Maintenance Applications

As discussed above, large representative labeled datasets are only very seldom available for predictive maintenance applications, particularly due to the specificity of the system design, system configuration and operating conditions, resulting in the fact that fault patterns are typically quite unique and faults are seldom. Particularly representative samples of fault types, continuously or discretely measured health indicators and time-to-failure trajectories are typically scarce. However, condition monitoring data from the healthy operating conditions is often abundantly available

(given the condition monitoring system is in place). The challenge here is to make as much use as possible of the unlabeled data, i.e. healthy system conditions for the predictive maintenance applications. Using healthy system conditions enables to use data-driven condition-based and predictive maintenance systems without the need to wait until a representative training dataset is collected. This makes data-driven predictive maintenance systems applicable to cases that were not suitable for their application before.

One approach to implementing fully unsupervised condition-based and predictive maintenance applications is by using in the first step an unsupervised feature learning approach and condensing the information contained in the raw data into a small number of informative features and in the second step monitoring deviations from the normal system conditions directly in the latent feature space. While this can be a good approach for some of the systems, the latent feature space has typically still several dimensions and interpreting the deviations in the single features can be difficult, even for domain experts. One possible way to make the features easier to monitor and interpret for the domain expert is to apply dimensionality reduction approaches. One possible and nonlinear approach for visualizing high-dimensional features' embedding in a lower dimensional space (particularly for two or three dimensions) is the t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten and Hinton 2008).

However, even in cases where fault detection can be applied directly in feature space, additional decision boundaries for fault detection and diagnostics need to be implemented, for example by first clustering the feature space and subsequently using the identified clusters to implement classification algorithms classifying each newly observed condition monitoring pattern as healthy or unhealthy.

A more intuitive approach for supporting the decision making based on the learned features is to design a health indicator and subsequently only monitor this health indicator, comparing it to defined thresholds which are then used to raise alarms (Michau et al. 2017).

The only additional piece of information it uses compared to the fully unsupervised approaches is the information that the system was operated under normal operating conditions during the time period of data collection. This information can be normally easily obtained for data collected in the past where it is known that no faults have occurred. This information is typically easily available and abundant.

One of the challenges valid for all the condition-based and predictive maintenance solutions also holds here: ensuring the representativeness of the training dataset (the composition of the training dataset needs to be representative for all the possible occurring healthy system states under different operating regimes). Otherwise, misclassification and false alarms can occur by confusing new operating conditions with faulty system states. If the information on the actual health status of the system is not available due to the lack of feedback from maintainers or lack of documentation, a step-wise approach can be chosen, in which small parts of the historically collected data are used to test against the rest if there are any major deviations or differences in the health indicators of the entire observation period.

The information on the healthy system condition can be used to formulate the decision problem as a one-class classifier (contrary to the binary classifiers where also a representative dataset for the second label needs to be provided to the learning algorithm): by giving all the condition monitoring signals the label "healthy".

The basic task of the one-class classifier is to learn the representation of the healthy system conditions. The interpretation of the learning task can be considered in the following way: the information contained in the features is condensed into a single piece of information—an indication that the features represent a healthy system condition. Once the algorithm has been trained to map the healthy system conditions to a single indicator, the trained algorithm can be used to compare all the newly measured system states in terms of their similarity to the healthy system states during training. This indicator goes further than just providing the information if the new observation is similar or dissimilar to the training data (binary information), it measures the distance to the healthy system condition. This enables to interpret this indicator as a health indicator. It can be used to monitor the evolution of the system condition over time and distinguish between different degrees of abnormalities and thereby different degrees of severity of the faults affecting the system. Ultimately, this indicator can be even used to not just detect the deviation from the healthy system conditions but also by trending it to predict the remaining useful life of the system (which would require learning the thresholds for the end of life of the specific system based on examples).

The proposed framework is based in the first step on feature learning with stacked autoencoders, providing informative and robust features to the second learning step—the one-class classifier that condenses the information into a single indicator, interpreted for the predictive maintenance applications as a health indicator. The approach proposed in Michau et al. (2017) is based on hierarchical extreme learning machines (HELM). However, the framework of feature learning and health indicator learning can be applied with any machine learning algorithm. HELM comprise stacked extreme learning machines (ELM) (Tang et al. 2016) (Fig. 25.5). The particular benefit of stacking ELM is the efficiency of their training process. ELM are feed forward neural networks with a single hidden layer and randomly generated weights between the input and the hidden layer. Thereby, only the weights between the hidden layer and the output need to be learned. This is contrary to the typical back-propagation learning regimes that typically require iterative learning processes.

As described in Sect. 25.4, autoencoders have also been used as stand-alone fault detection applications using the residuals between the observed and the reconstructed signals to detect changes. Since different fault types will affect different condition monitoring signals, different fault types will also result in different patterns of residuals. By comparing the combination of affected signals and their residuals, also different fault types can be distinguished a-posteriori. However, this will not provide a label for the different fault types. The fault types can be grouped and then presented to a domain expert, who are typically able to at least narrow down the possible fault causes to only a few possible hypotheses on the fault types. This can be subsequently confirmed during an inspection, testing only a

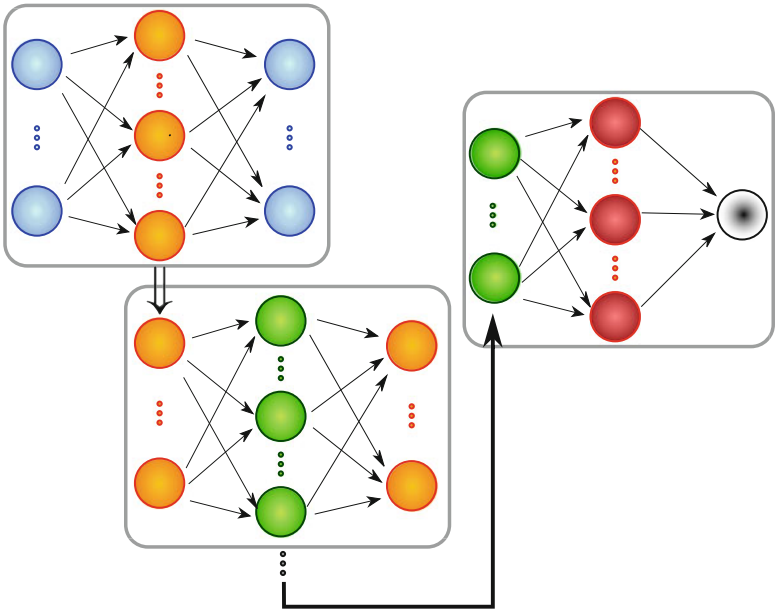


Fig. 25.5 Basic principle of an integrated feature learning and health indicator extraction approach (Michau et al. 2017)

limited number of hypotheses compared to a wide range of possible fault types that would need to be investigated otherwise.

In the approach proposed in Michau et al. (2017) based on HELM, fault detection and isolation are integrated in a single architecture: in the first step the fault is detected by the one-class classifier. Once the fault has been detected, the residuals from the reconstructed autoencoder output (within the feature learning part) are used to isolate the most affected signals and can assist the diagnostics engineer to evaluate the specific fault type and the possible root causes of the occurred faults (Fig. 25.6). All this is performed at no additional computational cost since it is an integrated approach.

25.7 Discussion and Outlook

The approaches discussed in this chapter overcome some of the challenges of practical predictive maintenance applications, including lack of labeled data, complexity and high-dimensionality of the condition monitoring signals, combination of fault detection and fault diagnosis.

However, the proposed approaches impose some requirements on the training datasets to ensure their applicability and robustness to real applications. One of the requirements is on the representativeness of the training dataset since data-driven approaches require to learn the functional relationships from representative

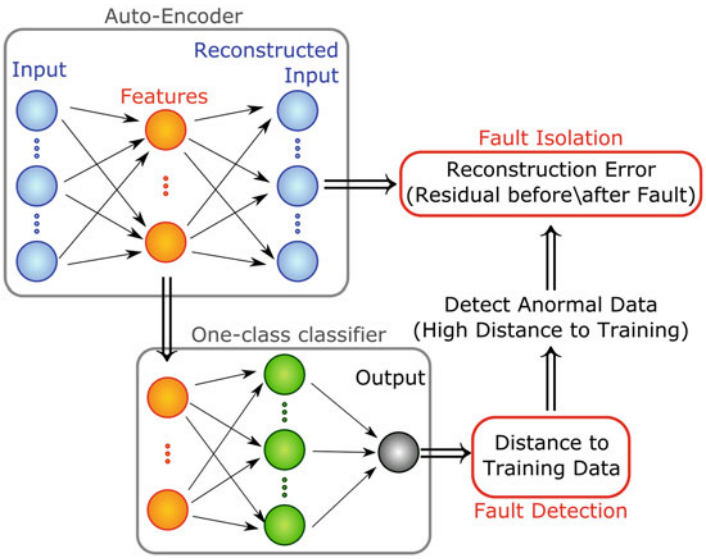


Fig. 25.6 Integrated fault detection and isolation (Michau et al. 2017)

samples. One of the limitations arising from this requirement is that predictive maintenance systems cannot be developed for systems taken newly into operation due to the lack of a representative training dataset. This is where particularly the operational experience of other systems within a fleet can be used to enrich the operational experience of the single system (Michau et al. 2018). Additionally, also hybrid approaches combining physical models and machine learning provide a promising research direction to ensure representativeness of the training datasets and the transferability of the operating experience and the experienced fault types within a fleet.

Given the scarcity of labeled data for condition-based and predictive maintenance applications, unsupervised and semi-supervised approaches are particularly promising and need to be further developed. There have been several promising developments in other domains which can be also transferred to predictive maintenance applications.

A further challenge is making the learned features more interpretable overcoming by that the black-box character of machine learning and in particular of deep learning.

The use of generative models that are currently considered by some of the experts as one of the most promising developments in neural networks has the capability to solve some of the challenges discussed in this chapter.

Reinforcement learning is a type of neural networks that learn how to achieve a complex objective without having any label but only by receiving feedbacks (or rewards) from the environment that the agents interact with. This research direction has not yet been applied in predictive maintenance applications. However, it has the potential to solve several types of problems.

Also, the last two levels of the pyramid presented in Sect. 25.2: decision support and prescriptive maintenance need to be further developed in order to ensure a wider acceptance of data-driven predictive maintenance applications by practitioners.

References

Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828

Chen Z, Li W (2017) Multisensor feature fusion for bearing fault diagnosis using sparse autoencoder and deep belief network. IEEE Trans Instrumen Meas 66(7):1693–1702

Chen Z, Li C, Sanchez R-V (2015) Gearbox fault identification and classification with convolutional neural networks. Shock Vib 2015:1–10

Forman G (2003) An extensive empirical study of feature selection metrics for text classification. J Mach Learn Res 3:1289–1305

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680

Hu Y, Palmé T, Fink O (2017) Fault detection based on signal reconstruction with auto-associative extreme learning machines. Eng Appl Artif Intell 57:105–117

Kadry S (2013) Diagnostics and prognostics of engineering systems: methods and techniques. Engineering Science Reference, p 433

Khan S, Yairi T (2018) A review on the application of deep learning in system health management. Mech Syst Signal Process 107:241–265

Krummenacher G, Ong CS, Koller S, Kobayashi S, Buhmann JM (2018) Wheel defect detection with machine learning. IEEE Trans Intell Transp Syst 19(4):1176–1187

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444

Lee J, Wu F, Zhao W, Ghaffari M, Liao L, Siegel D (2014) Prognostics and health management design for rotary machinery systems—reviews. Method Appl Mech Syst Signal Process 42(1):314–334

Li X, Ding Q, Sun J-Q (2018) Remaining useful life estimation in prognostics using deep convolution neural networks. Reliab Eng Syst Saf 172:1–11

Meyer T, Sextro W (2014) Closed-loop control system for the reliability of intelligent mechatronic systems. In: European conference of the prognostics and health management society, vol 5

Michau G, Palmé T, Fink O (2017) Deep feature learning network for fault detection and isolation. In: Conference of the PHM society

Michau G, Palmé T, Fink O (2018) Fleet PHM for critical systems: bi-level deep learning approach for fault detection. In: Proceedings of the European conference of the PHM society, vol 4. Utrecht, The Netherlands

Remadna I, Terrissa SL, Zemouri R, Ayad S (2018) An overview on the deep learning based prognostic. In: 2018 International conference on advanced systems and electric technologies (IC_ASET), IEEE, pp 196–200

Sateesh Babu G, Zhao P, Li X-L (2016) Deep convolutional neural network based regression approach for estimation of remaining useful life. Springer, Cham, pp 214–228

Tang J, Deng C, Huang G-B (2016) Extreme learning machine for multilayer perceptron. IEEE Trans Neural Netw Learn Syst 27(4):809–821

Thomas É, Levrat É, Iung B, Cocheteux P (2009) Opportune maintenance and predictive maintenance decision support. In: IFAC proceedings volumes, vol 42, no 4, pp 1603–1608

van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9(Nov):2579–2605

Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol P-A (2010) Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res 11:3371–3408

Yan W, Yu L (2015) On accurate and reliable anomaly detection for gas turbine combustors: a deep learning approach. In: Proceedings of the annual conference of the prognostics and health management society

Yam RCM, Tse P, Li L, Tu P (2001) Intelligent predictive decision support system for condition-based maintenance. Int J Adv Manuf Technol 17(5):383–391

Yoon AS, Lee T, Lim Y, Jung D, Kang P, Kim D, Park K, Choi Y (2017) Semi-supervised learning with deep generative models for asset failure prediction. arXiv: 1709.00845

Zhang C, Lim P, Qin AK, Tan KC (2017) Multiobjective deep belief networks ensemble for remaining useful life estimation in prognostics. IEEE Trans Neural Netw Learn Syst 28(10):2306–2318

Zhang J, Wang P, Yan R, Gao RX (2018) Deep learning for improved system remaining life prediction. Proc CIRP 72:1033–1038

Zhao R, Yan R, Chen Z, Mao K, Wang P, Gao RX (2016) Deep learning and its applications to machine health monitoring: a survey. arXiv: 1612.07640

Zhao R, Yan R, Wang J, Mao K (2017) Learning to monitor machine health with convolutional Bi-directional LSTM networks. Sensors 17(2):273 (Switzerland)

Zhao R, Yan R, Chen Z, Mao K, Wang P, Gao RX (2019) Deep learning and its applications to machine health monitoring. Mech Syst Signal Process 115:213–237

Dr. Olga Fink is SNSF (Swiss National Science Foundation) professor for intelligent maintenance systems at ETH Zürich. Before joining ETH faculty, she was heading the research group "Smart Maintenance" at the Zurich University of Applied Sciences (ZHAW). She is also research affiliate at the Department of Civil and Environmental Engineering at Massachusetts Institute of Technology (MIT), Cambridge.

Olga received her Ph.D. degree in civil engineering from ETH Zurich, and Diploma degree in industrial engineering from Hamburg University of Technology.

She has gained valuable industrial experience as reliability engineer for railway rolling stock and as reliability and maintenance expert for railway systems.

Olga's research focuses on data-driven condition-based and predictive maintenance of complex industrial systems; recently with a particular focus to deep learning applications in predictive maintenance.

Why These Studies?

I decided to study industrial engineering because it provided to me a perfect combination of analytics and engineering with a practical impact and direct applications. Later in my Ph.D. studies, I became fascinated by the enormous potential that artificial intelligence provides for solving complex problems. Therefore, I focused my research on artificial intelligence applications for maintenance systems. While my Ph.D. focused on data-driven predictive maintenance applied to railway applications, I extended my research direction later to other critical industrial assets. I enjoy doing innovative research and at the same time providing solutions to industrial applications.

Index

A
Adaptive search, 154

B
Big data, 27–42, 319
Branch and bound, 60, 113–152
Business model, 526
Business process modeling (BPM), 461

C
Calibration uncertainty, 347–365
Capacity expansion, 51–55, 59, 66
Care process, 273, 278, 279, 282, 283
Care transition, 211–226, 279, 283, 285, 289, 292
Career interest tests, 3
Clinical trials, 321–341
Compliance, 391–413, 466, 523
Constraint programming, 423, 436, 437
Consumer behavior, 391–413, 518, 519, 531–537
Container depot, 487–510
Contamination, 392, 394–399, 401–403, 409–413, 568–570, 578
Critical access hospital (CAHs), 212
Critical thinking, 185–193, 195
Cross-disciplinary, 185–193

D
Decision analytics, 560
Deep learning, 29, 34–36, 41, 303, 307–311, 314, 594–595, 597–599, 603
Dimension reduction, 29, 33–34, 40–41
Dynamic pricing, 544, 545, 547, 550, 552, 553
Dynamic programming, 55, 80, 331, 340, 531

E
Efficiency expert, 14
Electronic data interchange, 458
Emergency department utilization, 211, 212
Engineering design, 164, 182, 192, 426, 518, 520, 527, 531, 540
Engineering education, 22, 165, 185–193, 195, 274, 429, 511, 512
Engineering student needs questionnaire, 163–173
Environmental genome, 209
Equity, 160, 236, 257, 258, 266, 371–387
Ergonomics, 3, 17, 282

F
False negatives, 332–334, 340
Fault detection, 30, 31, 590–592, 596–597, 600, 620, 603
Fault diagnostics, 30, 35, 602
Feature learning, 35, 593–602
Food, 15, 16, 164, 236, 255–260, 382–383, 386, 391–413, 562, 575, 579
Food insecurity, 255–260
Forecasting, 29, 51, 53, 61–63, 115, 272, 381

G
Global optimization, 114, 124, 126, 144, 154
Graph embedding(s), 73, 75

© Springer Nature Switzerland AG 2020

A. E. Smith (ed.), *Women in Industrial and Systems Engineering*,
Women in Engineering and Science, https://doi.org/10.1007/978-3-030-11866-2

H
Healthcare life cycle analysis, 209
Healthcare supply chain, 274
Healthcare sustainability, 201, 204
Hospital readmission, 212, 221, 224
Human factors engineering, 275–293
Humanitarian, 234, 236, 271, 272, 371–387, 389, 390, 398, 415, 559
Humanitarian logistics, 234, 236, 271, 272, 373, 374, 381–382, 385, 389
Hybrid algorithms, 454

I
Industrial psychology, 5, 11, 13, 19
Inference, 29, 36, 38–41
International collaborations, 456, 507
Interventions, 200, 203, 213, 235, 242–244, 247, 251, 264, 276, 289, 347, 351, 358, 392, 396–397, 405, 407, 520, 562

K
Kitchen triangle, 3

L
Large-scale networks, 76, 91
Lead time quotation, 541–555
Local search, 93–107
Logistics cost analysis, 459, 462

M
Machine learning (ML), 28–36, 209, 307, 319, 483, 591, 594, 595, 597, 601, 603
Management theory, 9
Manufacturing, 27–36, 41, 42, 66, 207, 331, 458, 518, 523, 532, 541–555
Markov decision processes (MDPs), 234, 252, 253
Markovian disease models, 353
Matching demand and supply, 541–555
Medical decision making (MDM), 235, 236, 272, 347, 367
Medical imaging, 301–314
Mentoring, 385, 419, 421, 422
Metaheuristics, 93–107, 454
Mitigation, 391–413
Mixed-integer nonlinear programming, 525
Mixed-integer programming, 60, 436, 454
Multi-attribute decision making, 427, 430
Multilevel logistic regression, 222
Multimodal transportation, 417–431
Mushroom production, 567, 574, 577–579

N
Network representations, 73–75
Non-profit, 256, 257, 371–387, 389

O
Optimization, 34, 41, 52–54, 57–61, 65, 66, 72, 76, 83, 93–96, 107, 113–152, 234, 236, 254, 259, 265, 330, 340, 356, 357, 380, 381, 384–386, 423, 436, 442, 453, 454, 460, 488, 519–521, 537, 553
Optimization under uncertainty, 155
Ovarian cancer, 347–365

P
Parameter estimation, 115, 351, 545
Parameter tuning, 93–107
Patient journey, 275–293
Patient risk factors, 221
Patient safety, 201, 212, 247, 248, 251, 275–293, 299
Pharmaceutical industry, 325, 330–331
Pharmacy, 91, 235, 252–254, 274
Port logistics, 457–482, 495
Prediction, 28–32, 34, 36–38, 42, 53, 58, 62, 64, 87, 213, 222, 302, 591, 592
Predictive maintenance, 589–604
Project management, 195, 507

R
Re-enactment, 61, 64–66
Representation learning,71–90, 594, 595, 597
Resource allocation, 241, 321–340, 373, 376, 382, 554, 559, 560
Risk stratification, 211–226, 230, 248, 255, 272
Robust representations, 72, 89
Rolling horizon, 51
Rural women empowerment, 564, 568, 577–578, 586

S
Scenario generation, 56, 61–66
Scenario reduction, 56–58, 61–64

Scientific management, 6, 8, 9, 12, 14
Self-adaptive, 93–107
Sequential testing, 323–328, 332, 335
Simulation, 51, 54, 62, 64, 65, 113–116, 144, 234, 235, 238–244, 252–254, 331, 333, 340, 349, 487–510, 520, 535, 576
Statistical model, 95, 242, 317–319, 352
Stochastic mixed-integer programming (SMIPs), 59, 60
Stochastic optimization, 51, 56, 65, 66, 380, 381, 385, 520
Stochastic process model, 50, 52–55, 57, 61, 63–65
Student needs, 163–175, 178
Student performance, 234, 236, 260–265
Supply chain, 200–201, 252–258, 376, 377, 384–387, 391–413, 425, 426, 435, 458, 460, 546
Sustainable lifecycle engineering, 176
System dynamics, 332, 340, 345, 426
Systems perspective, 247, 345

T
Teamwork, 15, 185–193
Texture features, 302–307, 312, 313
Therbligs, 7, 8, 15
Thinking style, 186–190
Time and motion studies, 11, 12, 14, 19
Time windows (TW), 436, 452–454, 500
Trade facilitation, 572

U
Unit commitment, 51–52, 57–56, 65

V
Vehicle routing, 96, 97, 373, 379, 435–454

W
Women research impacts, 372, 382, 386
Women workers, 13, 14
Work system, 285, 286, 290