Gilbert Laporte · Stefan Nickel
Francisco Saldanha da Gama *Editors*

Location Science

Springer

Location Science

Gilbert Laporte • Stefan Nickel •
Francisco Saldanha da Gama
Editors

Location Science

Springer

Editors
Gilbert Laporte
HEC Montréal
Montréal
Canada

Stefan Nickel
Institute of Operations Research (IOR)
Karlsruhe Institute of Technology
Karlsruhe
Germany

Francisco Saldanha da Gama
DEIO-FCUL
University of Lisbon
Lisbon
Portugal

ISBN 978-3-319-13110-8 ISBN 978-3-319-13111-5 (eBook)
DOI 10.1007/978-3-319-13111-5
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2015932096

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The idea of editing this book emerged during the fourth meeting on Combinatorial Optimization, Routing and Location, held in Benicassim, Spain, in May 2012 (CORAL 2012) and was formalized during the 12th International Symposium on Locational Decisions (ISOLDE XII), held in Nagoya and Kyoto, Japan, in July the same year.

Our goal was to edit a comprehensive and structured book gathering the essential knowledge on modern Location Science, as opposed to a collection of exhaustive surveys or a pedagogical textbook with worked examples and exercises. Rather, this is a book on "what you should know" about various aspects of Location Science. It provides the basic knowledge and structures of the field. It can be used either in standard academic programs or in specialized courses.

The book contains an introduction to modern Location Science and 23 chapters grouped under three main headings: basic concepts (five chapters), advanced concepts (12 chapters), and applications (six chapters).

We have identified some of the best reputed specialists in the field to write the different chapters of the book. Each chapter was reviewed several times by at least one of the editors. The process was completed within two years. Today we are glad to present to the location community a high quality book which we hope to update on a regular basis.

We thank all the authors who accepted our challenge to be involved in this book. The quality of their work together with their dedication and enthusiasm contributed to making this project a success.

Finally, thanks are also due to Mr Christian Rauscher and to the Springer staff for their help and encouragement throughout this project.

Montréal, Canada — Gilbert Laporte
Karlsruhe, Germany — Stefan Nickel
Lisbon, Portugal — Francisco Saldanha da Gama

Contents

Chapter 1
Introduction to Location Science

Gilbert Laporte, Stefan Nickel, and Francisco Saldanha da Gama

Abstract This chapter introduces modern Location Science. It traces the roots of the area and describes the path leading to the full establishment of this research field. It identifies several disciplines having strong links with Location Science and offers examples of areas in which the knowledge accumulated in the field of location has been applied with great success. It describes the purpose and structure of this volume. Finally, it provides suggestions on how to make use of the contents presented in this book, namely for organizing general or specialized location courses targeting different audiences.

Keywords Application areas • Foundations • Location courses • Location science, Related disciplines

1.1 Introduction

In the past decades, Location Science has become a very active research area, attracting the attention of many researchers and practitioners. Facility location problems lie at the core of this discipline. These consist of determining the "best" location for one or several facilities or equipments in order to serve a set of demand points. The meaning of "best" depends on the nature of the problem under study, namely in terms of the constraints and of the optimality criteria considered.

G. Laporte (✉)
HEC Montréal, Montréal, QC, Canada
e-mail: gilbert.laporte@cirrelt.ca

S. Nickel
Institute for Operations Research, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Fraunhofer Institute for Industrial Mathematics (ITWM), Kaiserslautern, Germany
e-mail: stefan.nickel@kit.edu

F. Saldanha da Gama
Centro de Investigação Operacional/Departamento de Estatística e Investigação Operacional, Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal
e-mail: fsgama@fc.ul.pt

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_1

Location Science is a rich and fruitful field, gathering a large variety of problems. The research conducted in this area has led to the creation of a considerable amount of knowledge, both in terms of theoretical properties and modeling frameworks, together with solution techniques. This knowledge has evolved over time, pushed by the need to solve practical location problems, by technical and theoretical challenges, and often by problems arising in various disciplines. In fact, the interaction with other disciplines such as economics, geography, regional science and logistics, just to mention a few, has always been a driving force behind the development of Location Science. Nowadays, the potential of this field of study in the context of many real-world systems is widely recognized. This book emerges from the need to gather in a single volume the basic knowledge on Location Science as well as from the importance of somehow structuring the field and showing how it interacts with other disciplines.

In this introductory chapter we start by tracing the roots of what is now known as Location Science. This is done is Sects. 1.2 and 1.3. In Sect. 1.4 we present the purpose and structure of this book. Finally, in Sect. 1.5 we provide some suggestions on how to make the best use of the book.

1.2 The Roots

In order to trace the roots of modern Location Science, one must go back to an old geometric problem which is simple to state: What is the point in the Euclidean plane minimizing the sum of its distances to three given points (Fig. 1.1)? This problem is widely credited to the French mathematician Pierre de Fermat (1601–1665)[1] although its origin is a matter of debate (see Wesolowsky 1993).

Since the seventeenth century, different solutions have been proposed for Fermat's problem. There is evidence that the first one is due to the Italian scientist Evangelista Torricelli (1608–1647). The geometric approach proposed by Torricelli is depicted in Fig. 1.2 and can be described as follows: By joining the three given points with line segments, a triangle is obtained. Equilateral triangles can now

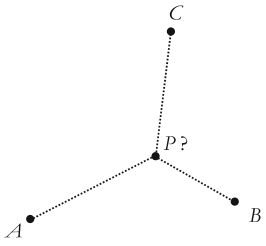


Fig. 1.1 Fermat's problem

[1]The problem is presented in his famous essay on maxima and minima.

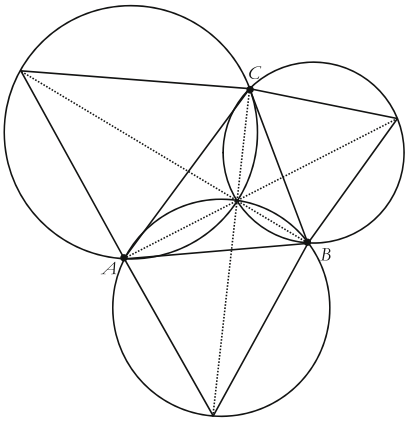
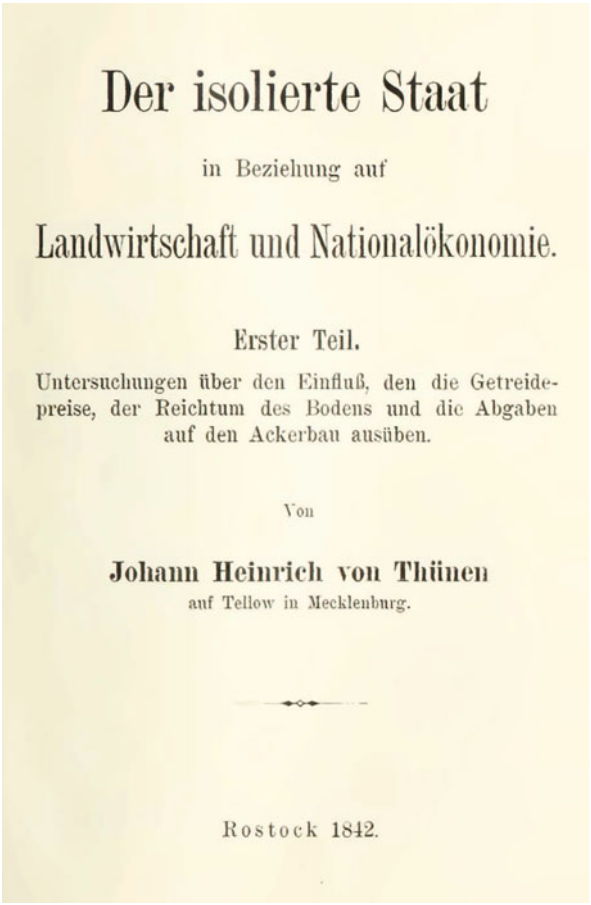


Fig. 1.2 Torricelli's geometric construction for the Fermat problem

be constructed on the sides of this triangle, their vertices pointing outwards. A circumscribing circle can then be drawn around each of these three triangles. The circles will intersect at a single point called the Torricelli point or, as some authors call it, the Fermat-Torricelli point. If all the angles in the initial triangle are at most equal to 120°, this point is the optimal solution to the problem; otherwise, the Torricelli point falls outside the initial triangle. In this case, the optimal solution is the initial point located at the apex of the angle greater than 120° (Heinen 1834).

It is interesting to note that nowadays this problem still attracts the attention of the scientific community (see, for instance, Nam 2013).

The first documented attempt to position location analysis within an economic context is due to Johann Heinrich von Thünen (1783–1850), an educated landowner in northern Germany. Von Thünen wished to understand the rural developments around an urban center. The results of his analysis were presented in 1826 in a treatise entitled *Die isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*, which was edited as a book in 1842 and translated into English in 1966 (von Thünen 1842). Figure 1.3 depicts the cover of the 1842 edition. Von Thünnen (1842) considered an isolated and homogeneous area with an urban center and aimed to discover laws which then governed agricultural prices translating them into land usage patterns. He also considered several types of agricultural activities (e.g., grain farming and livestock) grouped according to their relative economic yield per unit area, their perishability, and the difficulty in delivering the products to the (central) market. His findings led him to postulate that three factors should have a crucial impact on the spacial distribution of the activities: (1) the more perishable a product is, the closer to the market it will be grown; (2) the higher the economic productivity of a product per land area, the closer to the market it will be grown; (3) higher transportation difficulty leads to locating an activity closer to the market. One should therefore expect that the different agriculture activities will evolve in concentric rings around the urban center (Fig. 1.4).



Der isolierte Staat

in Beziehung auf

Landwirtschaft und Nationalökonomie.

Erster Teil.

Untersuchungen über den Einfluß, den die Getreidepreise, der Reichtum des Bodens und die Abgaben auf den Ackerbau ausüben.

Von

Johann Heinrich von Thünen

auf Tellow in Mecklenburg.

Rostock 1842.

Fig. 1.3 "Die Isolierte Staat" by Johann Heinrich von Thünen, Rostock, 1852 (*Source*: University of Toronto—Robarts Library, https://www.archive.org/details/derisoliertestaa00thuoft)

There still exists an intensive debate on the theory of von Thünen (Block and DuPuis 2001). Despite its merit, von Thünen's model is only descriptive, i.e., it is aimed at predicting the behavior of the system. In fact, at the time, models were mostly used to answer to questions such as "why do we do it?". Von Thünen's work can be viewed as fundamental in urban economics and location theory. Nowadays, it is still relevant in areas such as geography, agricultural economics and sociology (Block and DuPuis 2001). These authors emphasize that the centrality theory of von Thünen is still relevant for some dairy products such as milk. Other researchers have pursued von Thünen's centrality idea. The results are reviewed by Fischer (2011).

The first normative location models aimed at determining "what we should do", were proposed by Carl Friedrich Launhardt (1832–1918) and Alfred Weber (1868–1958). Launhardt (1900) introduced the problem of tracing an optimal rail route connecting three points. Interestingly, the author casted this problem within an industrial context. The problem was revisited by Pinto (1977) who stated it as follows: Consider the three nodes depicted in Fig. 1.5. Suppose that w_A tons of

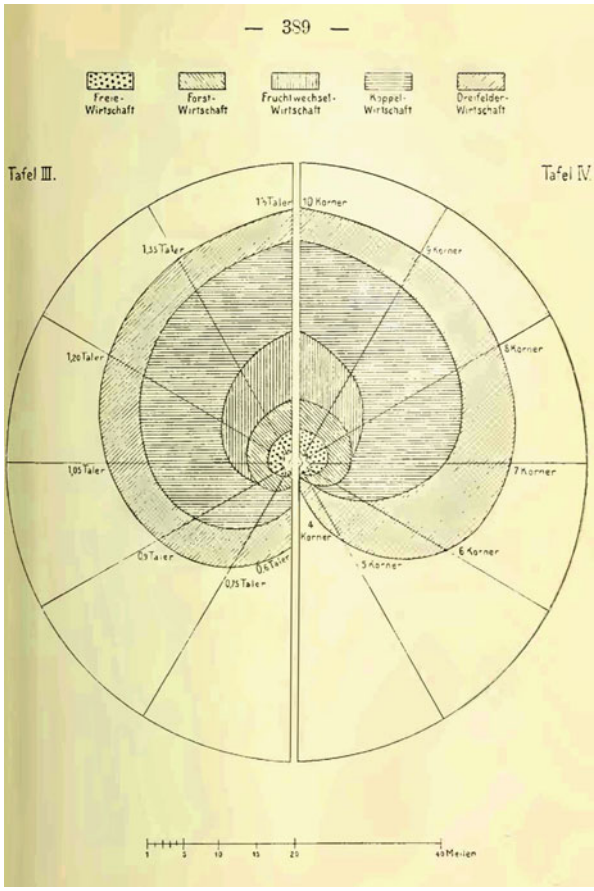


Fig. 1.4 Von Thünen's rings. From "Die Isolierte Staat" by Johann Heinrich von Thünen", Rostock 1842, p. 389 (*Source*: University of Toronto—Robarts Library, https://www.archive.org/details/derisoliertestaa00thuoft)

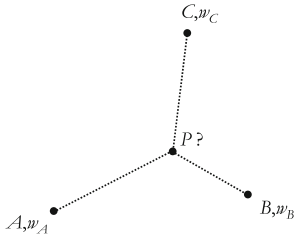


Fig. 1.5 Location problem proposed by Launhardt (1900) within an industrial context

iron ore (collected at A) have to be combined with w_B tons of coal (collected at B) to produce w_C tons of pig-iron to be dispatched to C. The problem calls for an industrial facility to be located somewhere between A, B and C. If d_A, d_B, d_C denote the Euclidean distances between the industrial location (to be determined) and nodes A, B, and C, respectively, then the goal is to determine the location of the industrial plant that will minimize the total weighted transportation cost given by $w_A d_A + w_B d_B + w_C d_C$.

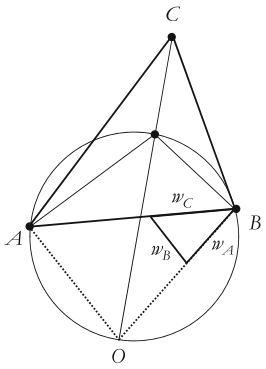


Fig. 1.6 Launhardt's geometric solution

This problem introduced by Launhardt is exactly what we now call the three-node Weber problem. However, as pointed out by Pinto (1977), the problem was introduced about 10 years before Weber (1909). Indeed, Launhardt (1900) proposed a simple geometric solution scheme for the problem. The solution is obtained as follows (see Fig. 1.6): Consider the triangle ABC defined by the original nodes (the locational triangle) and select one node, say C. Consider another triangle whose sides are proportional to the weights w_A, w_B and w_C.[2] Draw a triangle AOB similar (in the geometric sense) to the weight triangle but such that the edge proportional w_C has the same length as edge $\overline{AB}$, which is the one opposite to C in the locational triangle. The new triangle AOB is depicted in Fig. 1.6.[3] We can now circumscribe nodes A, B and O, by just touching each point. Finally, a straight line can be drawn connecting O and C. The intersection between the circle and this line yields the optimal location for the industrial facility.

This same problem was treated by Weber (1909) or, to be more accurate, by the mathematician Georg Pick (1859–1942), who is the author of the appendix in which the mathematical considerations of Weber's book are presented. The problem was solved in a different way but this resulted in the same solution. As put by Lösch (1944), the solution to this problem was discovered by Carl Friedrich Launhardt and rediscovered "one generation later" by Alfred Weber. Nevertheless, Weber (1909), presented a deeper analysis of the problem. He first noted that if the geometric construction leads to a point outside the original triangle, then the optimal solution lies on the boundary of the original triangle. Second, he observed that the pole method, which Launhardt (1900) believed should work for polygons

[2]This triangle is referred to by Weber (1909) as the *weight triangle*.

[3]Node O was called by Launhardt the *pole* of the locational triangle.

with more than three sides, does not necessarily yield the optimal solution when more than three nodes are involved. A practical algorithm for solving the problem with an arbitrary number of nodes was proposed by Weiszfeld (1937).[4] The iterative procedure proposed in this work was recently revisited in depth by Plastria (2011).

A synthesis of the first steps towards inserting location theory into an economic context is due to Lösch (1944). The importance of this work stems from the fact that, for the first time, location theory and the theory of market areas were connected. This work constitutes the first explicit recognition of the strong link that is often observed between these two areas.

1.3 Towards a New Science

The 1960s set the foundations of Location Science as new scientific area. We first witnessed the natural extension of the Weber problem to the multi-facility case. This was done, among others, by Miehle (1958) and Cooper (1963). In particular, the latter work introduced the planar p-median problem for which each demand node must be served by one out of p new facilities to be located. This became a fundamental problem in Location Science, which still attracts the attention of the scientific community (see the recent papers by Brimberg and Drezner 2013, Brimberg et al. 2014, and Drezner et al. 2014).

The seminal papers by Hakimi (1964, 1965) opened new important research directions. Hakimi (1964) introduced the concept of absolute median of a graph: a single facility is to be located anywhere in a network so as to minimize the sum of the distances of the nodes of the network to the facility. The author proved that there always exists an optimal solution for which the absolute median is a vertex of the graph. It is also in this paper that the concept of absolute center was first introduced: a single facility has to be located (anywhere in the network) in order to minimize the maximum distance between the facility and all the vertices. This work was extended to the multi-facility case by Hakimi (1965): now, p facilities are to be located. The vertex-optimality property is still valid for the resulting p-median problem. This property is of major importance because it means that many network location problems can be cast into a discrete setting which, in turn, leads to the possibility of using integer programming and combinatorial optimization techniques for tackling these problems.

[4]The author is now known to be Andrew Vázsonyi (1916–2003).

It is interesting to note that an important step toward the development of discrete facility location problems was taken in the same year when Balinski (1965) proposed the first mixed-integer linear programming (MILP) formulation for a discrete problem which also became classical in Location Science: the uncapacitated facility location problem (UFLP). Some inequalities proposed in this work were later used by ReVelle and Swain (1970) who formulated the first MILP model for the discrete p-median problem. One year later, Toregas et al. (1971) introduced the first integer programming formulation for a covering-location problem.

By the early 1970s, the foundations were laid for what would soon become a very active research field. The recent book by Eiselt and Marianov (2011) describes the works that can be considered to constitute the basis of Location Science.

In the past 40 years, significant advances have been made in several areas of Location Science, which is attested by several review papers, such as those by Brandeau and Chiu (1989), ReVelle and Laporte (1996), Avella et al. (1998), Hale and Moberg (2003), ReVelle and Eiselt (2005), ReVelle et al. (2008), and Smith et al. (2009).

Initially, the major concern of the researchers had to do with theoretical developments and properties of the problems and their solutions. Much work was developed on continuous and network location problems as well as on fundamental discrete facility location problems. Further links were created with other areas. For instance, the developments in continuous location problems led to the important connection between location analysis and computational geometry. This link remains quite strong to this day. In fact, one of the most relevant structures in computational geometry, the Voronoi diagram [after Georgy Feodosevich Voronoy (1868–1908)], is of major importance in the resolution of many continuous location problems (see, for instance, the review by Okabe and Suzuki 1997).

Nowadays, location problems can still be categorized according to the location space (continuous, network or discrete), but also according to their context, namely the objectives, constraints or type of facilities involved. Eiselt and Marianov (2011) highlight the three major forms of facility location problems according to the type of objective function: minsum, covering and minmax. For some time, it was also popular to distinguish between public, semi-public and private facility location.

Location Science is highly interconnected with other disciplines and has application in many areas. The theoretical foundations of this area lie in mathematics, economics, geography and computer science. The developments we have observed touch each of these areas.

More recently, stimulated by real-world problems, many areas have emerged where facility location has been applied with great success. Among these, we can point out logistics (see, for instance, Melo et al. 2006, for a problem in the context of logistics network design), telecommunications (see, for instance, Gollowitzer and Ljubić 2011, for a telecommunications network design problem), routing (e.g., in the truck and trailer routing problem introduced by Chao 2002, the location of the

trailer-parking places is one of the relevant decisions to make), and transportation (see, e.g., Nickel et al. 2001, for a location problem in the context of public transportation systems). The application of location theory in these areas partially explains why discrete facility location problems have progressively acquired a major relevance when compared with the early developments in Location Science.

Nowadays, Location Science is a very active and well-established research area with its own identity and research community. In addition to the fundamental problems, we observe different research branches being intensively investigated such has multi-criteria facility location, multi-period facility location, facility location under uncertainty, location-routing and competitive location, just to mention a few.

1.4 Purpose and Structure of This Book

As highlighted above, many location problems have applications in other disciplines. Researchers working in these disciplines often encounter location decisions as part of broader problems. From the point of view of researchers coming from the location community, the recent decades have shown that several very successful applications of the knowledge gathered in Location Science require a deep understanding of these disciplines.

In this book, readers will find a full coverage of basic aspects, fundamental problems and properties defining the field of Location Science, as well as advanced models and concepts that are crucial to the solution of many real-life complex problems. The book also presents applications of location problems to several fields. It is intended for researchers working on theory and applications involving location problems and models. It is also suitable as a textbook for graduate courses in facility location. This book is neither a typical textbook with worked examples and exercises, nor a collection of extensive surveys. It is more a book on "what you should know" about various aspects of Location Science; it provides the basic knowledge and structures the field. It is divided into three parts: basic concepts, advanced concepts and applications.

I. Basic concepts
   This part is devoted to the fundamental problems in Location Science, which include:

   - Chapter 2: p-median problems;
   - Chapter 3: Fixed-charge facility location problems;
   - Chapter 4: p-center problems;
   - Chapter 5: Covering location problems;
   - Chapter 6: Anti-covering location problems.

   The goal of this part is to provide the reader with the basic background of location theory. The problems described in Part I serve as a basis for much of the content of Parts II and III.

II. Advanced concepts
This part covers models and concepts that aim at broadening and extending the basic knowledge presented in Part I, thus providing the reader with important tools to better understand and solve real-world location problems. The chapters in this part are the following:

- Chapter 7: Location of dimensional facilities in a continuous space;
- Chapter 8: Facility location under uncertainty;
- Chapter 9: Location problems with multiple criteria;
- Chapter 10: Ordered median location problems;
- Chapter 11: Multi-period facility location;
- Chapter 12: Hub location problems;
- Chapter 13: The quadratic assignment problem;
- Chapter 14: Competitive location;
- Chapter 15: Location-routing and location-arc routing;
- Chapter 16: Location and logistics;
- Chapter 17: Stochastic location models with congestion;
- Chapter 18: Aggregation in location.

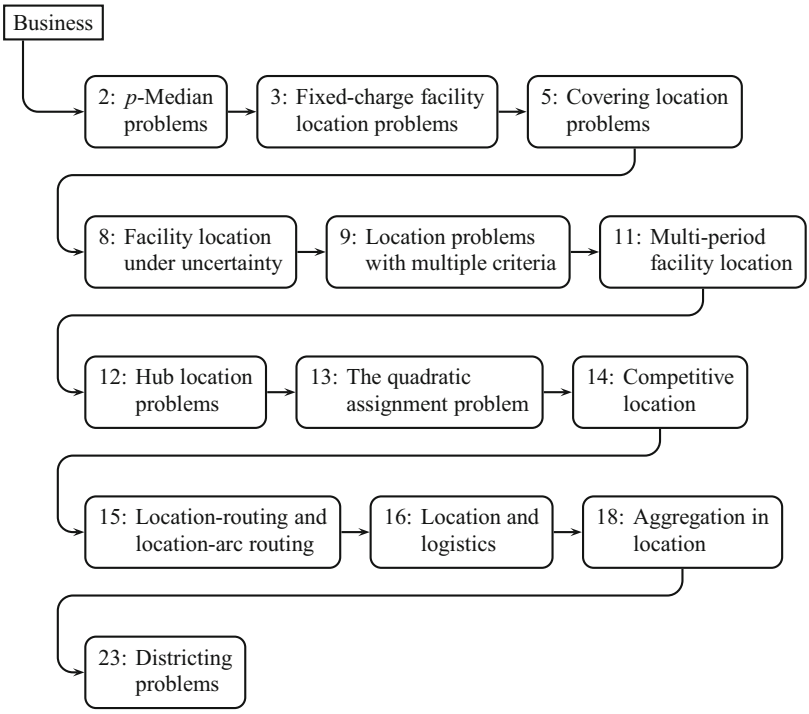
III. Applications
The links between Location Science and other areas are the focus of the third part. By presenting a wide range of applications, it is possible not only to understand the role of facility location in such areas, but also to show how to handle realistic location problems. These applications include:

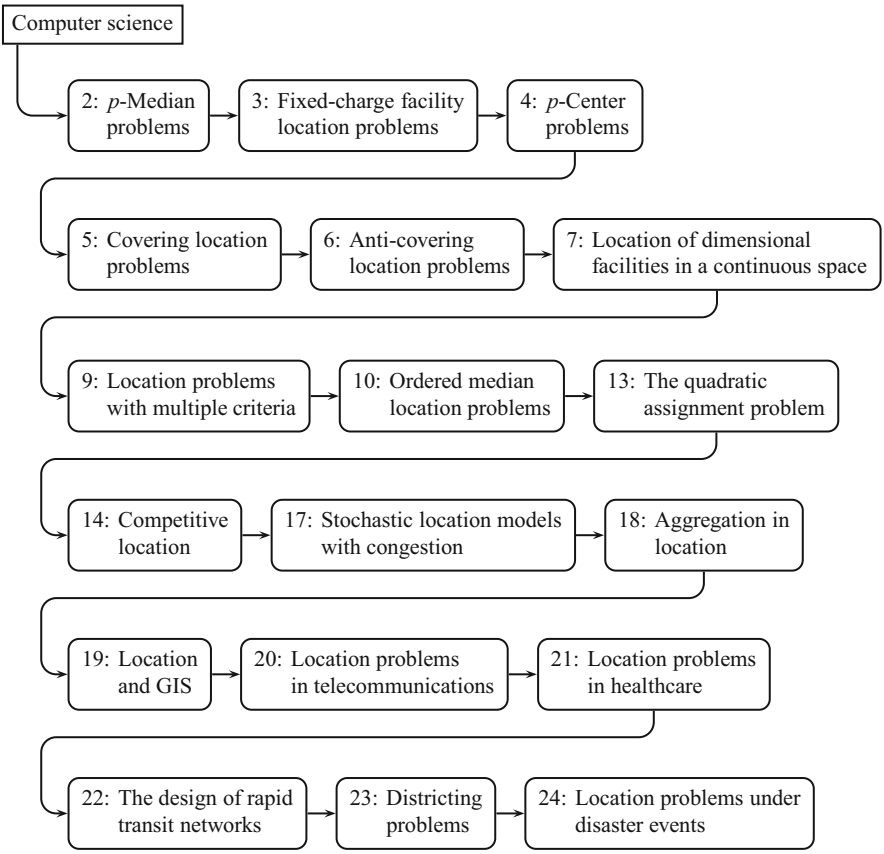
- Chapter 19: Location and GIS;
- Chapter 20: Location problems in telecommunications;
- Chapter 21: Location problems in healthcare;
- Chapter 22: The design of rapid transit networks;
- Chapter 23: Districting problems;
- Chapter 24: Location problems under disaster events.

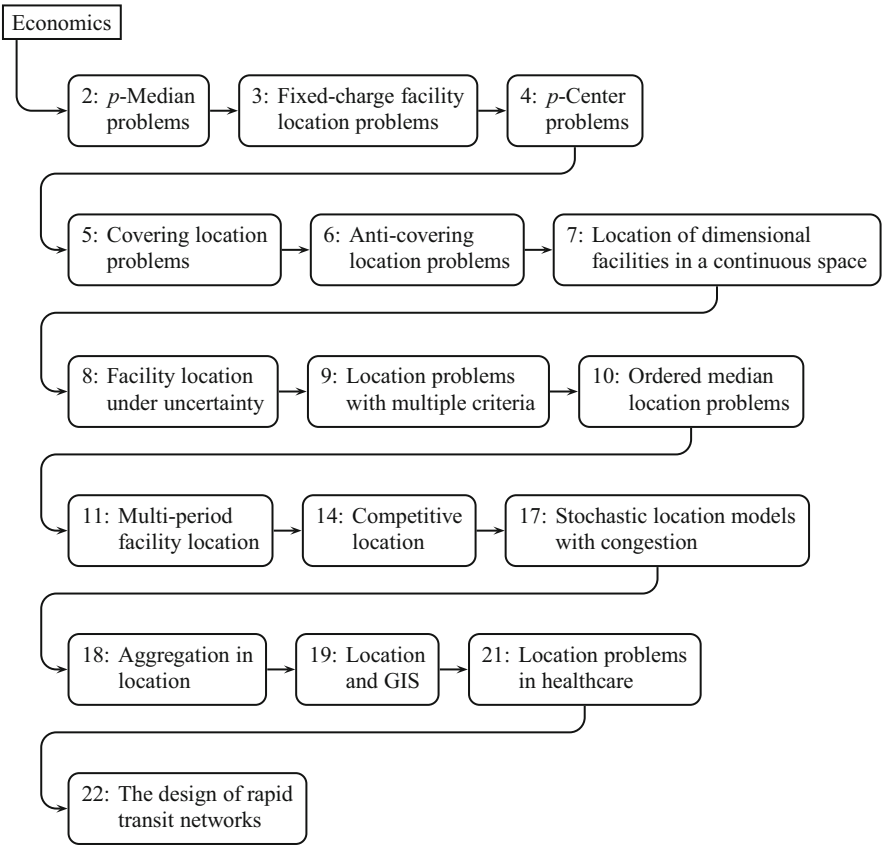
1.5 How to Use This Book

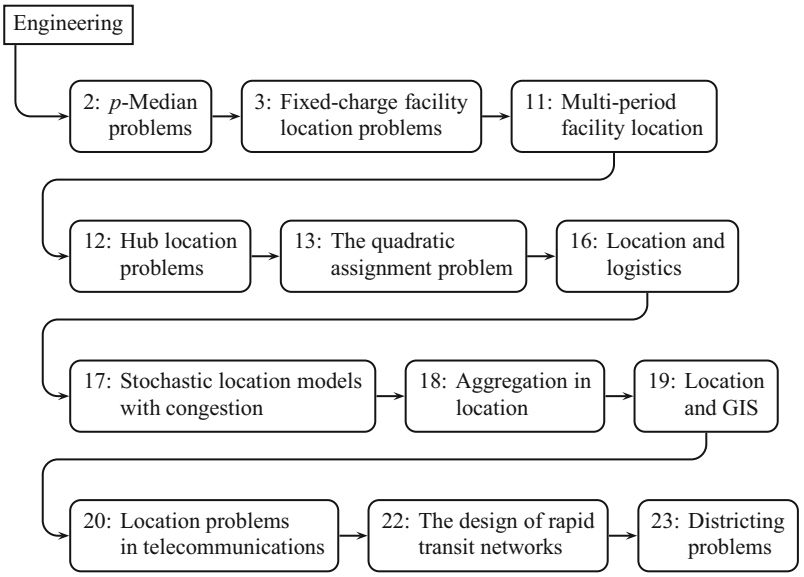
Over the past decades, problems, models, properties, and techniques from Location Science have been increasingly taught to students enrolled in different programs. We have identified six types of post-graduate curricula having a strong location content: business, computer science, economics, engineering, geography and mathematics.

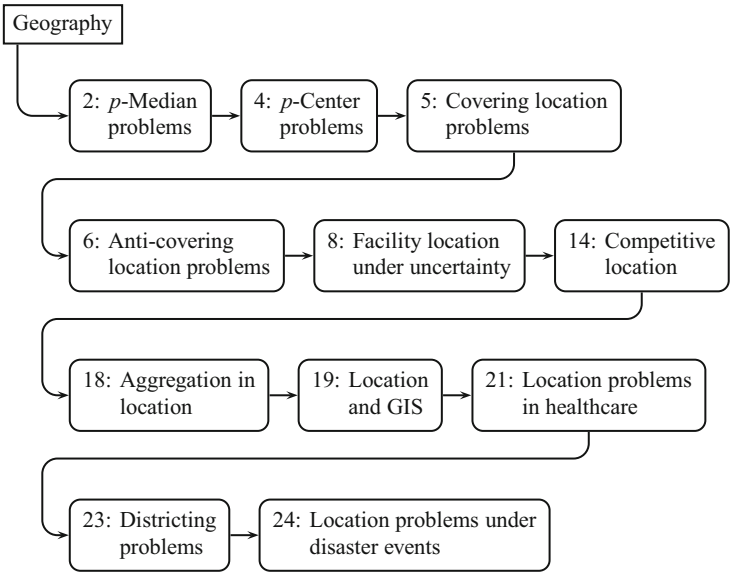
Depending on the audience, different contents emerge as the most appropriate. This book can be used with the purpose of organizing courses tuned for specialized targets by selecting specific combinations of chapters. Below, we offer some suggestions.

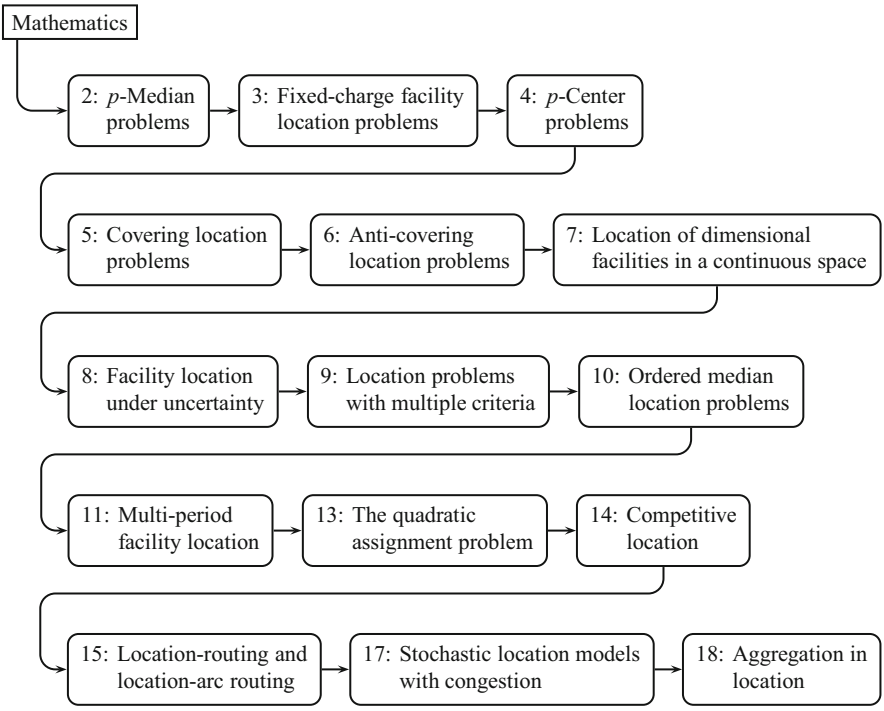



Computer science
2: p-Median problems
3: Fixed-charge facility location problems
4: p-Center problems
5: Covering location problems
6: Anti-covering location problems
7: Location of dimensional facilities in a continuous space
9: Location problems with multiple criteria
10: Ordered median location problems
13: The quadratic assignment problem
14: Competitive location
17: Stochastic location models with congestion
18: Aggregation in location
19: Location and GIS
20: Location problems in telecommunications
21: Location problems in healthcare
22: The design of rapid transit networks
23: Districting problems
24: Location problems under disaster events

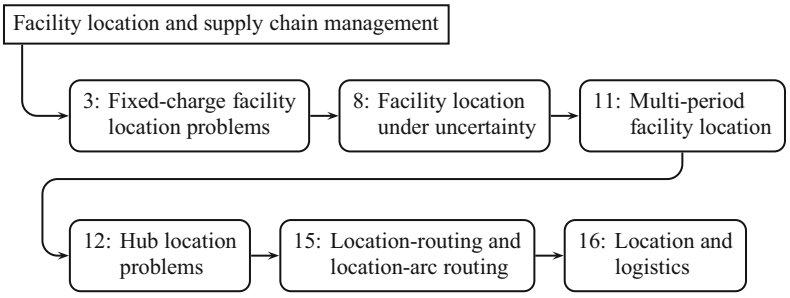

Economics
2: p-Median problems
3: Fixed-charge facility location problems
4: p-Center problems
5: Covering location problems
6: Anti-covering location problems
7: Location of dimensional facilities in a continuous space
8: Facility location under uncertainty
9: Location problems with multiple criteria
10: Ordered median location problems
11: Multi-period facility location
14: Competitive location
17: Stochastic location models with congestion
18: Aggregation in location
19: Location and GIS
21: Location problems in healthcare
22: The design of rapid transit networks

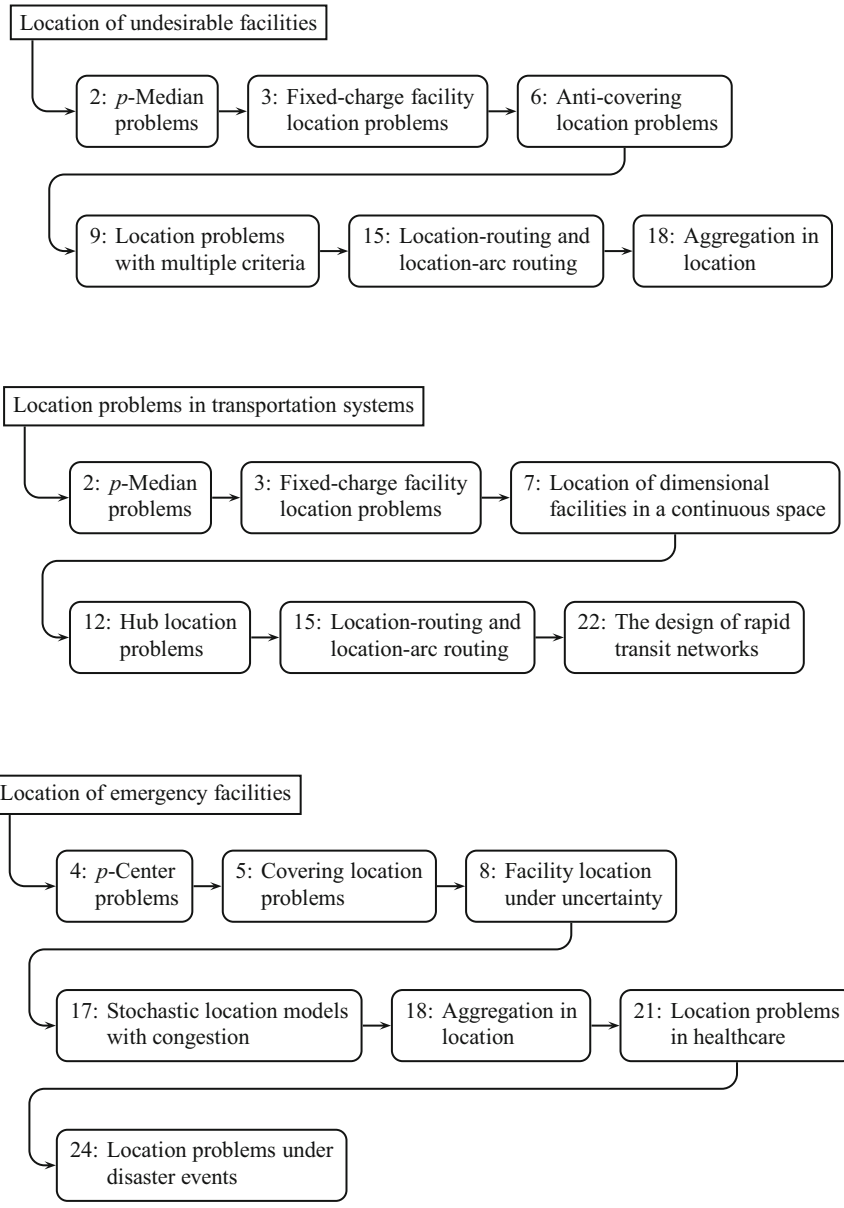

Engineering
2: p-Median problems
3: Fixed-charge facility location problems
11: Multi-period facility location
12: Hub location problems
13: The quadratic assignment problem
16: Location and logistics
17: Stochastic location models with congestion
18: Aggregation in location
19: Location and GIS
20: Location problems in telecommunications
22: The design of rapid transit networks
23: Districting problems


Geography
2: p-Median problems
4: p-Center problems
5: Covering location problems
6: Anti-covering location problems
8: Facility location under uncertainty
14: Competitive location
18: Aggregation in location
19: Location and GIS
21: Location problems in healthcare
23: Districting problems
24: Location problems under disaster events



This book can also be used to build specialized courses in specific areas. Below, we provide examples in four areas: facility location and supply chain management, location of undesirable facilities, location of emergency facilities, and location in transportation systems.





When used for teaching, this book should be complemented with examples and exercises; when used for research, it should be complemented with specialized readings. We found the following comprehensive references particularly relevant: Mirchandani and Francis (1990), Drezner (1995), Drezner and Hamacher (2002), Nickel and Puerto (2005), Eiselt and Marianov (2011), and Daskin (2013).

References

Avella P, Benati S, Cánovas-Martinez L, Dalby K, Di Girolamo D, Dimitrijevic B, Giannikos I, Guttman N, Hultberg TH, Fliege J, Muñoz-Márquez M, Ndiaye MM, Nickel S, Peeters P, Pérez-Brito D, Policastro S, Saldanha da Gama F, Zidda P (1998) Some personal views on the current state and the future of locational analysis. Eur J Oper Res 104:269–287

Balinski ML (1965) Integer programming: methods, uses, computation. Manag Sci 12:253–313

Block D, DuPuis EM (2001) Making the country work for the city. Am J Econ Sociol 60:79–98

Brandeau ML, Chiu SS (1989) An overview of representative problems in location research. Manag Sci 35:645–674

Brimberg J, Drezner Z (2013) A new heuristic for solving the p-median problem in the plane. Comput Oper Res 40:427–437

Brimberg J, Drezner Z, Mladenović N, Salhi S (2014) A new local search for continuous location problems. Eur J Oper Res 232:256–265

Chao I-M (2002) A tabu search method for the truck and trailer routing problem. Comput Oper Res 29:33–51

Cooper L (1963) Location-allocation problems. Oper Res 11:331–343

Daskin MS (2013) Network and discrete location: models, algorithms and applications, 2nd edn. Wiley, Hoboken

Drezner Z (ed) (1995) Facility location: a survey of applications and methods. Springer, New York

Drezner Z, Hamacher H (eds) (2002) Facility location: applications and theory. Springer, Berlin/Heidelberg

Drezner Z, Brimberg J, Mladenović N, Salhi S (2014) New heuristic algorithms for solving the planar p-median problem. Comput Oper Res. doi: 10.1016/j.cor.2014.05.010

Eiselt HA, Marianov V (eds) (2011) Foundations of location analysis. Springer, New York

Fischer K (2011) Central places: the theories of von Thünen, Christaller, and Lösch. In: Eiselt HA, Marianov V (eds) Foundations of location analysis. Springer, New York, pp 471–505

Gollowitzer S, Ljubić I (2011) MIP models for connected facility location: a theoretical and computational study. Comput Oper Res 38:435–449

Hakimi SL (1964) Optimal location of switching centers and the absolute centers and medians of a graph. Oper Res 12:450–459

Hakimi SL (1965) Optimal distribution of switching centers in a communication network and some related graph theoretic problems. Oper Res 13:462–475

Hale TS, Moberg CR (2003) Location science research: a review. Ann Oper Res 123:21–35

Heinen F (1834) Über Systeme von Kräften, deren Intensitäten sich wie die n. Potenzen der Entfernungen gegebener Punkte von einem Central-Punkte verhalten, in Beziehung auf Punkte, für welche die Summe der n. Entfernungs-Potenzen ein Maximum oder Minimum ist. Bädeker, Essen

Launhardt C-F (1900) The principles of location: the theory of the trace. Part I: the commercial trace (trans: Bewley A, 1900). Lawrence Asylum Press, Madras

Lösch A (1944) The economics of location (trans: Woglom WH, 1954). Yale University Press, New Haven

Melo MT, Nickel S, Saldanha da Gama F (2006) Dynamic multi-commodity capacitated facility location: a mathematical modeling framework for strategic supply chain planning. Comput Oper Res 33:181–208

Mielhe W (1958) Link-length minimization in networks. Oper Res 6:232–243

Mirchandani PB, Francis RL (eds) (1990) Discrete location theory. Wiley, New York

Nam NM (2013) The Fermat-Torricelli problem in the light of convex analysis. ArXiv e-prints. Provided by the SAO/NASA astrophysics data system. http://adsabs.harvard.edu/abs/2013arXiv1302.5244M

Nickel S, Puerto J (2005) Location theory: a unified approach. Springer, Berlin/Heidelberg

Nickel S, Schöbel A, Sonneborn T (2001) Hub location problems in urban traffic networks. In: Niittymaki J, Pursula M (eds) Mathematical methods and optimization in transportation systems. Kluwer Academic Publishers, Dordrecht, pp 1–12

Okabe A, Suzuki A (1997) Locational optimization problems solved through Voronoi diagrams. Eur J Oper Res 98:445–456

Pinto JV (1977) Launhardt and location theory: rediscovery of a neglected book. J Reg Sci 17:17–29

Plastria F (2011) The Weiszfeld algorithm: proof, amendments, and extensions. In: Eiselt HA, Marianov V (eds) Foundations of location analysis. Springer, New York, pp 357–389

ReVelle CS, Eiselt HA (2005) Location analysis: a synthesis and survey. Eur J Oper Res 165:1–19

ReVelle CS, Laporte G (1996) The plant location problem: new models and research prospects. Oper Res 44:864–874

ReVelle CS, Swain RW (1970) Central facilities location. Geogr Anal 2:30–42

ReVelle CS, Eiselt HA, Daskin MS (2008) A bibliography for some fundamental problem categories in discrete location science. Eur J Oper Res 184:817–848

Smith HK, Laporte G, Harper PR (2009) Locational analysis: highlights of growth to maturity. J Oper Res Soc 60:S140–S148

Toregas C, Swain R, ReVelle CS, Bergman L (1971) The location of emergency service facilities. Oper Res 19:1363–1373

von Thünen JH (1842) The isolated state (trans: Wartenberg CM, 1966). Pergamon Press, Oxford

Weber A (1909) Theory of the location of industries (trans: Friedrich CJ, 1929). University of Chicago Press, Chicago

Weiszfeld EV (1937) Sur le point pour lequel la somme des distances de n points donnés est minimum. Tohoku Math J 43:335–386

Wesolowsky GO (1993) The Weber problem: history and perspectives. Locat Sci 1:5–23

Part I
Basic Concepts

Chapter 2
The *p*-Median Problem

Mark S. Daskin and Kayse Lee Maass

Abstract The *p*-median problem is central to much of discrete location modeling and theory. While the *p*-median problem is $\mathcal{NP}$-hard on a general graph, it can be solved in polynomial time on a tree. A linear time algorithm for the 1-median problem on a tree is described. We also present a classical formulation of the problem. Basic construction and improvement algorithms are outlined. Results from the literature using various metaheuristics including tabu search, heuristic concentration, genetic algorithms, and simulated annealing are summarized. A Lagrangian relaxation approach is presented and used for computational results on 40 classical test instances as well as a 500-node instance derived from the most populous counties in the contiguous United States. We conclude with a discussion of multi-objective extensions of the *p*-median problem.

Keywords Algorithm • Center • Covering • Lagrangian relaxation • Median • Multi-objective

2.1 Introduction

The *p*-median problem is that of locating *p* facilities to minimize the demand weighted average distance between demand nodes and the nearest of the selected facilities. The problem dates back to the seminal work of Hakimi (1964, 1965). The *p*-median problem is one of several classical location problems which also include the capacitated and uncapacitated facility location problems (Chap. 3), the *p*-center problem (Chap. 4), covering problems (Chap. 5) and anti-covering problems (Chap. 6). The *p*-median problem lies at the heart of many practical location problems, and, as shown below (Sect. 2.7), some of the other classical location problems can readily be formulated as *p*-median problems, leading to multicriteria location problems as outlined in Chap. 9.

Our objective is not to review every paper and every result related to this seminal problem. Rather, we summarize key results, algorithms and important extensions.

M.S. Daskin (✉) • K.L. Maass
IOE Department, University of Michigan, 1205 Beal Avenue, Ann Arbor, MI 48109, USA
e-mail: msdaskin@umich.edu; leekayse@umich.edu

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_2

We refer the reader to ReVelle et al. (2008) for a fairly recent annotated bibliography of the p-median and related models.

The remainder of this chapter is organized as follows. Section 2.2 outlines several key properties of the problem. Section 2.3 discusses optimal solution algorithms for the problem on a tree. Section 2.4 formulates the p-median problem as an optimization problem. Section 2.5 outlines algorithms for the problem on a general network. Section 2.6 presents selected computational results. Section 2.7 outlines two key multi-objective extensions of the p-median problem. Finally, conclusions are briefly presented in Sect. 2.8.

2.2 Model Properties

There are three key properties of the p-median problem that are important to know. First, Kariv and Hakimi (1979) showed that the p-median problem is $\mathcal{NP}$-hard on a general graph. This is the bad news. The good news, as outlined below, is that there are many effective algorithms and approaches to solving the p-median problem.

Second, Hakimi (1965) showed that at least one optimal solution to the p-median problem consists of locating only on the nodes. To see that this is true, consider a solution that entails locating a facility somewhere on an edge between nodes A and B. Let D_A be the total demand served by this facility that enters the edge via node A, and let D_B be the total demand served by the facility that enters via node B. Clearly, if $D_A > D_B$ we can move the facility to node A and reduce the objective function. This contradicts the assumed optimality of the facility at an intermediate location on the edge. Similar arguments hold if $D_B > D_A$ in which case we move the facility to node B. If $D_A = D_B$ we can move the facility to either node without adversely impacting the objective function value. Note that moving the facility to one of the nodes may result in the reassignment of demands to or from the facility if doing so will reduce the objective function. Such reassignments will only improve the objective function. Also note that moving the facility to one of the nodes may also result in some demands that were served by the facility, and that entered via the other node, to now enter the facility directly without traversing the edge between A and B. This would occur if traveling directly to the facility is shorter than traveling via the edge between A and B. Finally, we note that the nodal optimality property holds if the distance between a demand node and a candidate facility site is replaced by any concave function of the distance.

Finally, the demand weighted total cost or distance (or the demand weighted average cost or distance) decreases with the addition of each subsequent facility. This is clearly true since, if there exists an optimal solution to the problem with p facilities, then adding a $p + 1^{st}$ facility at any of the candidate nodes that does not have a facility will decrease the demand-weighted total cost or distance and therefore will also decrease the objective function. Locating the $p + 1$ facilities optimally is clearly as good or better than first locating p facilities optimally and adding a subsequent facility to that solution.

Table 2.1 Median results for top 100 counties in US

p	Demand weighted average distance	Change	Sites
1	969.45		St. Louis, MO
2	450.65	518.80	San Bernardino, CA; Allegheny, PA
3	320.15	130.50	Los Angeles, CA; Shelby, TN; Hudson, NJ
4	257.23	62.92	Los Angeles, CA; Tarrant, TX; New York, NY; Jefferson, KY
5	190.22	67.01	Los Angeles, CA; Cook, IL; Dallas, TX; New York, NY; Orange, FL

We would also expect that the marginal improvement in the demand weighted total (or average) cost or distance would decrease monotonically as we add facilities. This is frequently the case, but not always. As an example of a situation in which this is not so, consider the p-median problem with the 100 largest counties in the contiguous United States based on the 2010 census. While these counties represent only 3.2 % of the 3,109 counties in the contiguous United States, they account for 42.2 % of the total population. Using great circle distances and population as a proxy for demand, we obtain the results shown in Table 2.1. The demand weighted average distance decreases with the number of facilities as shown in the second column. However, the *change* in the demand weighted average distance increases from about 63 miles to 67 miles as we increase from four to five facilities.

2.3 The p-Median Problem on a Tree

While the p-median problem is $\mathcal{NP}$-hard on a general graph, the problem can be solved in polynomial time on a tree. We illustrate this with a linear time algorithm for finding the 1-median on a tree, which was proposed by Goldman (1971). This algorithm also helps explain why the problem is called the "median" problem. If any node of the tree has half or more of the total demand of all nodes on the tree, then it is clearly optimal to locate at that node. Moving away from that node will move the facility further from half or more of the demand and closer to less than half of the demand, thereby increasing the objective function value.

To outline this algorithm, we define the following sets:

$$\mathcal{I} = \{1, \ldots, i, \ldots, m\} \text{ the set of candidate locations}$$

$$\mathcal{J} = \{1, \ldots, j, \ldots, n\} \text{ the set of demand nodes.}$$

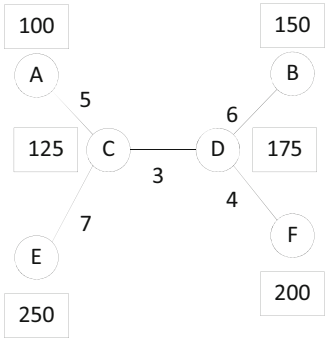


Fig. 2.1 Example tree

In addition, we define the following additional inputs:

d_j demand of customer j
c_{ij} unit cost of satisfying customer j from facility i.

Now suppose that no node has half or more of the total demand. We call any node that is connected to only one other node in the tree, a tip node. We let $d_j^{'}$ be the modified demand at node $j \in \mathcal{J}$. We also define $D_{total} = \sum_{j \in \mathcal{J}} d_j$. The algorithm is as follows.

Step 1: Let $d'_j = d_j$ for all nodes $j \in \mathcal{J}$.

Step 2: Select any tip node. Call the tip node, node A and the node to which it is connected node B. Remove node A and edge (A, B). Add the modified demand at node A to the modified demand at node B. If the new modified demand at node B equals or exceeds $D_{total}/2$, stop; node B is the 1-median of the tree. Otherwise repeat step 2.

This is clearly an $O(n)$ algorithm since Step 2 can be performed in constant time and each node is examined at most once in Step 2. The complexity of Step 1 is also clearly $O(n)$.

We can illustrate this algorithm with the tree shown in Fig. 2.1. The demand associated with each node is shown in a box beside the node and the edge distances are shown beside the edges. Nodes A, B, E and F are tip nodes. The total demand in the tree is $D_{total} = 1{,}000$. Clearly, no node has half or more of the total demand. We select node E as the first tip node to eliminate (since it has the largest demand of any tip node). We remove node E and link (C, E) from the tree and add 250 (the demand at node E) to the demand at node C. The modified demand at node C is now 375, which does not exceed half of the total demand. Next we can process node F, removing it as well as arc (D, F) and adding its demand to that of node D, resulting in a modified demand at node D of 375. Next we process node B, removing it as well as arc (B, D) and adding its demand to that of node D, resulting in a modified demand at node D of 525, which exceeds half of the total demand in the tree. Node D is therefore the 1-median of the tree.

Note that in computing the location of the 1-median we do not need to use the distances. In fact, node D would be the 1-median of the tree for any arc distances for the tree. To compute the objective function value, we clearly do need the distances. The objective function value for the 1-median located at node D in Fig. 2.1 is 5,375.

Kariv and Hakimi (1979) present an $O(n^2p^2)$ algorithm for the p-median problem on a tree. Tamir (1996) improved the computation time and presented an $O(pn^2)$ algorithm for the problem of locating p facilities on a tree.

2.4 Model Formulation

In this section, we formulate the p-median problem. In addition to the notation defined above, we define the following additional input:

p the number of facilities to locate.

Finally, we define the following decision variables:

$$y_i = \begin{cases} 1 \text{ if a facility is located at candidate site } i \\ 0 \qquad\qquad\qquad \text{otherwise} \end{cases}$$

x_{ij} the fraction of the demand of customer j that is supplied from facility i.

With this notation, we can formulate the p-median problem as follows:

$$\text{minimize} \quad \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} d_j c_{ij} x_{ij} \tag{2.1}$$

$$\text{subject to} \quad \sum_{i \in \mathcal{I}} x_{ij} = 1 \quad \forall j \in \mathcal{J} \tag{2.2}$$

$$\sum_{i \in \mathcal{I}} y_i = p \tag{2.3}$$

$$x_{ij} - y_i \leq 0 \quad \forall i \in \mathcal{I}; j \in \mathcal{J} \tag{2.4}$$

$$y_i \in \{0, 1\} \quad \forall i \in \mathcal{I} \tag{2.5}$$

$$x_{ij} \geq 0 \quad \forall i \in \mathcal{I}; j \in \mathcal{J}. \tag{2.6}$$

The objective function (2.1) minimizes the demand-weighted total cost. Constraints (2.2) mean that all of the demand at demand site j must be satisfied. Constraints (2.3) require exactly p facilities to be located. Constraints (2.4) state that demand nodes can only be assigned to open facilities. Constraints (2.5) stipulate that the location variables must be integer and binary. Finally, constraints (2.6) state that the assignment variables must be non-negative.

Note that we do not require the assignment variables to be binary variables. If the unit cost from a demand node to the nearest open facility is strictly less than the unit cost between that node and any other open facility, then the corresponding assignment variables for that demand node will naturally be binary. That is, all of the demand at that node will be assigned to the nearest open facility. If the unit costs between a demand node and two or more open facilities are the same, and the unit costs are less than the unit costs between the demand node and any other open facility, the assignment variables may indicate that the demand is to be split between the set of nearest facilities. We can always round all but one of these assignment variables down to 0 and round the last one up to 1 if we require all-or-nothing demand assignments or single sourcing.

2.5 Solution Heuristics for the *p*-Median Model on a General Network

In this section, we outline a number of heuristic algorithms for solving the p-median problem on a general network. We conclude the section by structuring a Lagrangian relaxation algorithm (Fisher 1981, 1985).

2.5.1 Basic Construction and Improvement Algorithms

The simplest algorithm is the myopic or greedy adding algorithm. In this algorithm, all candidate facility sites are examined and the one whose addition to the current solution reduces the demand-weighted total distance the most is added to the incumbent solution. The process continues until the solution includes p facilities. The following is pseudocode for the myopic algorithm. In this and all subsequent pseudocodes, we define $z(\mathcal{J}, X) = \sum_{j \in \mathcal{J}} d_j \min_{m \in X} \{c_{mj}\}$, where X is the current set of candidate facility sites. Note that the function depends on both the set of demand nodes to be considered and the candidate locations to be used.

Myopic Algorithm Pseudocode

1. **Set** $X \leftarrow \varnothing$. /* X is the set of locations to be used
2. **Find** $i^* = \text{argmin}_{i \in \mathcal{I}} \{z(\mathcal{J}, X \cup \{i\})\}$.
3. **Set** $X \leftarrow X \cup \{i^*\}$.
4. **If** $|X| < p$, go to Step 2; else stop.

Step 1 initializes the set of locations to the empty set. Step 2 finds the best node to add to the emerging solution. Step 3 adds that site to the solution. Step 4 asks if less than p facilities have been added to the emerging solution. If so, the algorithm continues with Step 2; if not, the algorithm stops.

The myopic algorithm can readily paint itself into a corner. There is no guarantee of optimality for the myopic algorithm. As illustrated below in the computational results, the algorithm can perform quite poorly. That said, it is clear that it is optimal if we are locating only a single facility.

Exploiting the optimality of the myopic algorithm for the 1-median problem, Maranzana (1964) proposed a neighborhood improvement algorithm. Starting with any feasible solution to the p-median problem, the algorithm assigns each demand node to its nearest facility. Ties are broken arbitrarily. The set of nodes assigned to a facility constitutes the neighborhood of that facility. Within each neighborhood, the algorithm examines each candidate node and selects the one that minimizes the demand-weighted total distance among all nodes in the neighborhood. In other words, within each neighborhood, the algorithm solves a 1-median problem. If no facility locations have changed, the algorithm stops; otherwise, if any facility locations have changed as a result of solving the 1-median problem, the algorithm re-assigns all demand nodes to the nearest open facility. If no assignments have changed, the algorithm stops; otherwise, the algorithm continues by solving the 1-median problem in each neighborhood. This process of determining neighborhoods and solving 1-median problems within each neighborhood continues until no further improvement is possible. The pseudocode below outlines the neighborhood search algorithm.

Neighborhood Search Algorithm Pseudocode

1. **Input:** X /* X is a set of p facility locations
2. **Set:** $N_i \leftarrow \phi,\ \forall i \in \mathcal{I}$ /* N_i is the set of demand nodes for which
   /* candidate site i is the closest open facility
3. **For** $j \in \mathcal{J}$ **do**
4. **Set** $i^* \leftarrow \text{argmin}_{i \in \mathcal{I}} \{c_{ij}\}$
5. **Set** $N_{i^*} \leftarrow N_{i^*} \cup \{j\}$
6. **End For**
7. **Set** $X^{new} \leftarrow \phi$ /* X^{new} is the set of new facility locations
8. **For** $i \in \mathcal{I}$ **do**
9. **If** $|N_i| > 0$ **then**
10. **Find** $k^* = \text{argmin}_{k \in N_i} z(N_i, \{k\})$
11. **Set** $X^{new} \leftarrow X^{new} \cup \{k^*\}$
12. **End If**
13. **End For**
14. **If** $X \neq X^{new}$ **then set** $X \leftarrow X^{new}$ **and go to Step 2; else stop**

Step 1 initializes the solution with any set of p facilities. Steps 2 through 6 initialize and then set the neighborhoods. Step 7 initializes a new candidate set of facility locations. Steps 8 through 13 find the new candidate locations. In particular, in Step 10, the algorithm finds the 1-median within each neighborhood and adds that vertex to the emerging new solution in Step 11. The algorithm, as written, assumes that the sets of demand locations and candidate sites are the same. While the neighborhood search algorithm finds the optimal location within each

neighborhood, there is no guarantee that it will find the global optimum for the problem.

The exchange algorithm, proposed by Teitz and Bart (1968), is another heuristic improvement algorithm that tends to do better than the neighborhood search algorithm. The algorithm attempts to improve the current solution by removing a node that is in the solution and replacing it with a node that is not in the solution. If an exchange of this sort can be found and improves the solution (i.e., reduces the demand-weighted total distance), it is implemented. The algorithm terminates when there is no such exchange that improves the solution. The pseudocode for one variant of the exchange algorithm is shown below.

Exchange Algorithm Pseudocode

1. **Input:** X /* X is a set of p facility locations
2. **For** $i \in X$ **do**
3. **For** $k \in \mathcal{I} \backslash X$
4. **If** $z(\mathcal{J}, X) > z(\mathcal{J}, X \cup \{k\} \setminus \{i\})$ **then**
5. **Set** $X \leftarrow X \cup \{k\} \setminus \{i\}$ **and stop**
6. **End If**
7. **End For**
8. **End For**

Step 1 initializes the solution with any set of p facilities. In Step 2 we loop over the sites in the current solution. In Step 3 we loop over candidate sites that are not in the solution. In Step 4, we ask if removing one site from the current solution and replacing it with a site not in the current solution will improve the objective function. If so, we make that substitution and the algorithm stops.

There are numerous ways of implementing an exchange algorithm. The algorithm might implement the *first* exchange that improves the solution, as shown in the pseudocode above. Alternatively, the algorithm might find the first node in the solution whose removal will result in an improvement to the solution and then find the *best* node to insert into the solution in place of the removed facility. Finally, the algorithm can find the best improving pair of nodes over all possible nodes to be removed and inserted into the solution.

If either of the first two approaches are adopted—that is, if the exchange algorithm does not find the best overall exchange possible—there are alternate ways in which the algorithm can proceed. One option is to continue the search with the next indexed node that is not in the solution, attempting to replace the node that was just inserted into the solution with another node. Another option is to proceed to the next node in the solution and attempt to find exchanges based on that node. A third option is to reinitiate the search from the first node in the solution. The various options for selecting an exchange to implement, as well as the different ways in which the algorithm can proceed once an improving exchange has been identified, result in numerous possible implementations of the exchange algorithm. Most of the literature does not identify which implementation was employed.

2.5.2 Metaheuristics for the p-Median Problem

The myopic algorithm is a construction algorithm. The neighborhood and exchange algorithms are improvement algorithms. A large variety of metaheuristic algorithms have been devised to find solutions to the *p*-median problem. Mladenović et al. (2007) provide a relatively recent review of these techniques. Below we highlight a few of the classic papers and approaches in this field.

Chiyoshi and Galvão (2000) present a statistical analysis of a simulated annealing algorithm (Kirkpatrick 1984) for the *p*-median model. They employed the 40-instance dataset proposed by Beasley (1990). The dataset includes instances ranging from 100 to 900 demand locations. They found that in 100 runs of a simulated annealing algorithm for each instance, the best solution found was the optimal solution in 26 of the 40 instances. The maximum deviation from optimality for the best of the 100 runs for the 40 instances was 1.62 %. Al-khedhairi (2008) also employed simulated annealing for the Beasley dataset and found the optimal solution in 33 of the cases. However, the maximum deviation was over 18 % for the seven instances for which the simulated annealing algorithm failed to find the optimal solution. Murray and Church (1996) also discuss the application of simulated annealing to the *p*-median problem as well as to the maximal covering problem.

Alp et al. (2003) propose an effective genetic algorithm (Goldberg 1989; Haupt and Haupt 1998; Holland 1975; Michalewicz 1994; Mitchell 1998) for the *p*-median problem. For the 40-instance Beasley dataset, they ran their algorithm 10 times for each instance. They found the optimal solution at least once in 28 of the 40 cases. In six of the cases, the genetic algorithm always identified the optimal solution. In the 12 cases in which the genetic algorithm failed to find the optimal solution, the best of the ten runs resulted in objective functions that deviated from the optimal value by 0.02–0.4 %.

Rolland et al. (1996) applied tabu search (Glover 1990; Glover and Laguna 1997) to the *p*-median problem. They tested their algorithm using randomly generated datasets ranging in size from 13 to 500 nodes. For instances with 100 nodes or fewer, the results were compared to two-exchange heuristics as well as to the optimal solution found using an integer programming algorithm. For the larger instances, optimal solutions were not obtained and the three heuristics were compared with each other. In all cases, the tabu search algorithm outperformed the other two heuristics. For the smaller instances (100 nodes or fewer) the tabu search algorithm averaged 0.5 % from optimality with a maximum deviation of 6 %. Tabu search found the optimal solution in 66 % of the smaller test cases. For the 12 larger test cases, tabu search found the best solution in all but one case.

If an improvement (e.g., the neighborhood search or exchange algorithm outlined above) is started with many different randomly generated solutions, the p facilities that are selected are often similar across the various solutions. In other words, some sites are selected in many of the runs and many other candidate sites are never selected. Using this observation, Rosing and ReVelle (1997) developed a heuristic

concentration algorithm for the p-median problem. The idea is to generate a number of good solutions based on randomized starting solutions. A subset of the nodes that are selected in the various runs is then used to reduce the number of location variables in formulation (2.1)–(2.6) above. In other words, the concentration set, or the set of candidate sites, is reduced from $\mathcal{J}$ to a smaller set consisting of a subset of the nodes selected as facilities in the various randomized runs.

Heuristic concentration is based on eliminating some of the location variables. Church (2008) proposed the BEAMR approach which attempts to eliminate some of the assignment variables. BEAMR attempts to utilize only the h_j closest assignment variables for each demand node. To ensure feasibility, the model also includes a variable for each demand node allowing the assignment to a dummy facility further than the h_j closest candidate facilities. This assignment does not need to satisfy constraints (2.4). The resulting model provides a lower bound on the objective function value for the p-median problem. An upper bound can be found by simply assigning every demand node to the nearest of the selected facility sites. If the bounds are not close enough, then some of the h_j values can be increased, particularly for those nodes for which assignment to one of the nearest h_j candidate sites was not possible. The algorithm typically results in provably optimal solutions using a fraction of the constraints and variables of the original formulation (2.1)–(2.6).

Rosing et al. (1998) compared heuristic concentration to tabu search in problems with 100 and 200 demand nodes and candidate sites. Heuristic concentration found the optimal (or best known) solution in 17 of the 21 test cases, while tabu search found the optimal (or best known) solution in only two cases.

Mladenović and Hansen (1997) introduced a variable neighborhood search algorithm. Hansen and Mladenović (1997) applied this algorithm to the p-median problem. They found that variable neighborhood search outperformed both a greedy interchange algorithm and two different tabu search-based algorithms.

Hansen and Mladenović (2001) reviewed the basics of variable neighborhood search algorithms and compared a variety of metaheuristic algorithms, including variable neighborhood search for the 12 largest of the 40 Beasley instances. They found that variable neighborhood search and heuristic concentration outperformed tabu search and a greedy interchange algorithm. Variable neighborhood search was slightly better than heuristic concentration.

2.5.3 A Lagrangian Heuristic for the p-Median Problem

In this subsection, we outline a Lagrangian relaxation algorithm to the p-median problem. The advantage of Lagrangian relaxation over any heuristic approach is twofold. First, at every iteration of the Lagrangian procedure we obtain lower and upper bounds on the objective function value. Second, the Lagrangian procedure can readily be embedded in a branch-and-bound algorithm to obtain provably optimal solutions.

We relax constraint (2.2) to obtain the following Lagrangian problem:

$$\begin{aligned} \text{Max}_{\lambda}\text{Min}_{x,y}\mathcal{L} &= \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} d_j c_{ij} x_{ij} + \sum_{j\in\mathcal{J}} \lambda_j \left(1 - \sum_{i\in\mathcal{I}} x_{ij}\right) \\ &= \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} \left(d_j c_{ij} - \lambda_j\right) x_{ij} + \sum_{j\in\mathcal{J}} \lambda_j \end{aligned} \tag{2.7}$$

subject to (2.3)–(2.6).

For fixed values of the Lagrange multipliers, λ_j, we compute the value of being able to add a facility at node $i \in \mathcal{I}$. This value is given by $V_i = \sum_{j\in\mathcal{J}} \min\left\{0, d_j c_{ij} - \lambda_j\right\}$. We then select the p sites with the p most negative V_i values, breaking ties arbitrarily. This determines the values of the location variables, y_i. The assignment variables are determined by setting $x_{ij} = 1$ if (i) $y_i = 1$ and (ii) $d_j c_{ij} - \lambda_j < 0$, and setting $x_{ij} = 0$ otherwise. The resulting values can be used to evaluate (2.7), providing a lower bound on the objective function value. To obtain an upper bound on the objective function value, we simply assign every demand node to the nearest candidate facility for which $y_i = 1$ and evaluate (2.1) using these assignment values.

Some of constraints (2.2) are likely to be violated by the solution to the Lagrangian problem as outlined above. In particular, some demand nodes may not be assigned to any facility and some may be assigned to multiple facilities. This occurs when the Lagrange multipliers are not at their optimal values. Subgradient optimization can be used to improve the Lagrange multipliers. Daskin (2013) provides a detailed explanation of the Lagrangian algorithm for the p-median problem.

The Lagrange multipliers coupled with the best lower and upper bounds can be used to force candidate sites in and out of the solution at any point in the Lagrangian procedure. Typically, it is most useful to do so when the bounds are very close to each other but still differ by a small amount. Let LB and UB be the best-known lower and upper bounds, respectively. Using the Lagrange multipliers associated with LB, sort the V_i values so that $V_{[i]}$ is the ith smallest value. In other words, $V_{[1]}$ is the most negative value and $V_{[p]}$ is the last value that resulted in selecting a candidate facility site in the Lagrangian solution. Additionally, $V_{[p+1]}$ is the next largest value.

Consider a candidate site $i \in \mathcal{I}$ that is in the best-known solution. Then, if $UB < LB - V_i + V_{[p+1]}$, site $i \in \mathcal{I}$ can be forced into the solution; in other words, we can set $y_i = 1$ in all subsequent Lagrangian iterations and in any branching below the node at which this check is done (e.g., the root node of a branch-and-bound algorithm). Similarly, if site $i \in \mathcal{I}$ is not part of the best-known solution and $UB < LB - V_{[p]} + V_i$, then site $i \in \mathcal{I}$ can be forced out of the solution; in other words, we can set $y_i = 0$ in all subsequent Lagrangian iterations and in any branching below the node at which this check is done (e.g., the root node).

2.6 Computational Results

In this section, we provide sample results for some of the algorithms outlined above. We begin with Table 2.2 which shows the results of using a Lagrangian relaxation algorithm embedded in branch-and-bound for the Beasley dataset. The instances were all solved using an expanded version of the SITATION software (Daskin 2013) on a Macintosh computer running OS X version 10.8.5 with a 2.7 GHz Intel Core i7 processor and 16 GB of 1,600 MHz DDR3 memory using Parallels 7.0.15107. The average solution time was under 45 s. The longest solution time—for PMED36—was under 13 min. Seventeen of the 40 instances were solved at the root node and all but three of the instances required less than 40 branch-and-bound nodes. The average solution time is 44.9 s and the average number of branch-and-bound nodes needed is 21.5.

The second part of the table illustrates the impact of using the variable forcing rules outlined at the end of Sect. 2.5 at the end of the Lagrangian algorithm at the root node of the branch-and-bound tree. The rules are quite effective at eliminating candidate nodes; on average nearly 85 % of the candidate sites that could not be in the solution were excluded at the root node using these rules. (The number of candidate sites that could not be in the solution was equal to the total number of candidate sites minus the number of facilities). Overall, 81 % of the candidate sites were either forced in or out of the solution, on average.

Next we turn our attention to tests performed using the 500 most populous counties among the 3,109 counties in the contiguous United States. While these represent less than one sixth of the total counties, they encompass over 75 % of the population living in the contiguous United States. Great circle distances between the county centroids were employed. We used SITATION to solve the p-median problem for this dataset with the number of facilities increasing from 1 to 25. The solution time for each of these 25 runs was under 5 s and only two instances required branch-and-bound to obtain provably optimal solutions. In each of these two instances, only three nodes in the branch-and-bound tree needed to be explored after the root node forcing rules were employed. Figure 2.2 plots the results for five, 10, 15, 20 and 25 medians. The model locates the first five cities near the major cities of New York, Los Angeles, Dallas, Chicago and Miami. Additional facilities are then added to better serve the rest of the counties. Figure 2.3 plots the demand-weighted average distance versus the number of medians. As expected, the average distance decreases with the number of medians. Also, the marginal improvement decreases with the number of medians in this case.

Table 2.2 Lagrangian relaxation results for Beasley datasets

Dataset	# Dem	# Med.	Objective	Iterations	B&B nodes	CPU time (s)
Pmed1	100	5	5,819	1,200	1	2.94
Pmed2	100	10	4,093	3,500	9	8.92
Pmed3	100	10	4,250	2,958	7	7.70
Pmed4	100	20	3,034	1,200	1	3.06
Pmed5	100	33	1,355	1,200	1	3.03
Pmed6	200	5	7,824	5,758	19	15.09
Pmed7	200	10	5,631	1,200	1	3.08
Pmed8	200	20	4,445	1,200	1	3.09
Pmed9	200	40	2,734	4,981	15	14.73
Pmed10	200	67	1,255	1,200	1	5.31
Pmed11	300	5	7,696	1,788	3	4.81
Pmed12	300	10	6,634	5,927	19	17.3
Pmed13	300	30	4,374	1,200	1	4.80
Pmed14	300	60	2,968	1,747	3	8.70
Pmed15	300	100	1,729	1,200	1	7.94
Pmed16	400	5	8,162	8,447	29	24.55
Pmed17	400	10	6,999	9,220	29	27.89
Pmed18	400	40	4,809	1,200	1	6.55
Pmed19	400	80	2,845	1,200	1	9.30
Pmed20	400	133	1,789	2,401	5	24.50
Pmed21	500	5	9,138	1,200	1	3.70
Pmed22	500	10	8,579	13,687	39	55.86
Pmed23	500	50	4,619	1,200	1	8.64
Pmed24	500	100	2,961	3,995	10	41.42
Pmed25	500	167	1,828	4,721	11	72.44
Pmed26	600	5	9,917	5,380	15	22.25
Pmed27	600	10	8,307	2,925	7	12.53
Pmed28	600	60	4,498	1,200	1	12.30
Pmed29	600	120	3,033	1,200	1	18.81
Pmed30	600	200	1,989	2,001	4	57.55
Pmed31	700	5	10,086	6,517	19	29
Pmed32	700	10	9,297	3,212	7	15.41
Pmed33	700	70	4,700	1,200	1	19.88
Pmed34	700	140	3,013	1,200	1	33.02
Pmed35	800	5	10,400	9,680	31	47.64
Pmed36	800	10	9,934	140,011	437	767.16
Pmed37	800	80	5,057	5,754	14	97.06
Pmed38	900	5	11,060	17,905	57	107.78
Pmed39	900	10	9,423	22,018	65	136.27
Pmed40	900	90	6,128	1,200	1	32.89

(continued)

Table 2.2 (continued)

Dataset	# Dem	# Med.	No. sites forced in	No. forced out	% in	% out	% forced
Pmed1	100	5	4	94	80	99	98
Pmed2	100	10	2	79	20	88	81
Pmed3	100	10	3	71	30	79	74
Pmed4	100	20	15	75	75	94	90
Pmed5	100	33	25	59	76	88	84
Pmed6	200	5	0	161	0	83	81
Pmed7	200	10	8	189	80	99	99
Pmed8	200	20	18	178	90	99	98
Pmed9	200	40	1	71	3	44	36
Pmed10	200	67	52	116	78	87	84
Pmed11	300	5	0	280	0	95	93
Pmed12	300	10	0	257	0	89	86
Pmed13	300	30	27	267	90	99	98
Pmed14	300	60	9	160	15	67	56
Pmed15	300	100	78	178	78	89	85
Pmed16	400	5	0	336	0	85	84
Pmed17	400	10	0	327	0	84	82
Pmed18	400	40	24	314	60	87	85
Pmed19	400	80	67	307	84	96	94
Pmed20	400	133	49	163	37	61	53
Pmed21	500	5	5	495	100	100	100
Pmed22	500	10	0	397	0	81	79
Pmed23	500	50	44	444	88	99	98
Pmed24	500	100	14	308	14	77	64
Pmed25	500	167	36	191	22	57	45
Pmed26	600	5	0	542	0	91	90
Pmed27	600	10	0	539	0	91	90
Pmed28	600	60	50	496	83	92	91
Pmed29	600	120	97	450	81	94	91
Pmed30	600	200	24	131	12	33	26
Pmed31	700	5	0	639	0	92	91
Pmed32	700	10	0	645	0	93	92
Pmed33	700	70	13	603	19	96	88
Pmed34	700	140	98	459	70	82	80
Pmed35	800	5	0	684	0	86	86
Pmed36	800	10	0	478	0	61	60
Pmed37	800	80	10	610	13	85	78
Pmed38	900	5	0	780	0	87	87
Pmed39	900	10	0	707	0	79	79
Pmed40	900	90	85	805	94	99	99

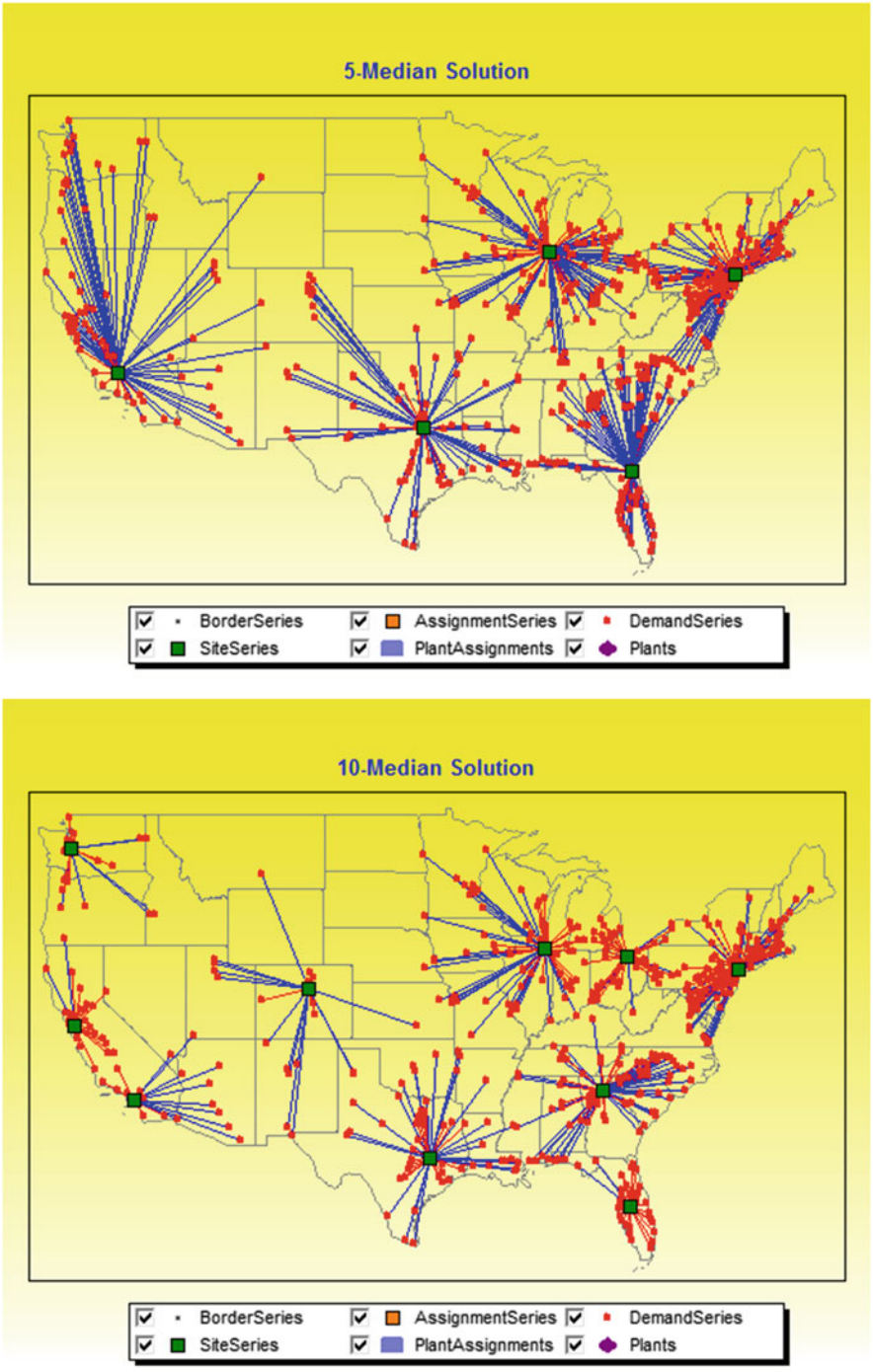


Fig. 2.2 (continued)

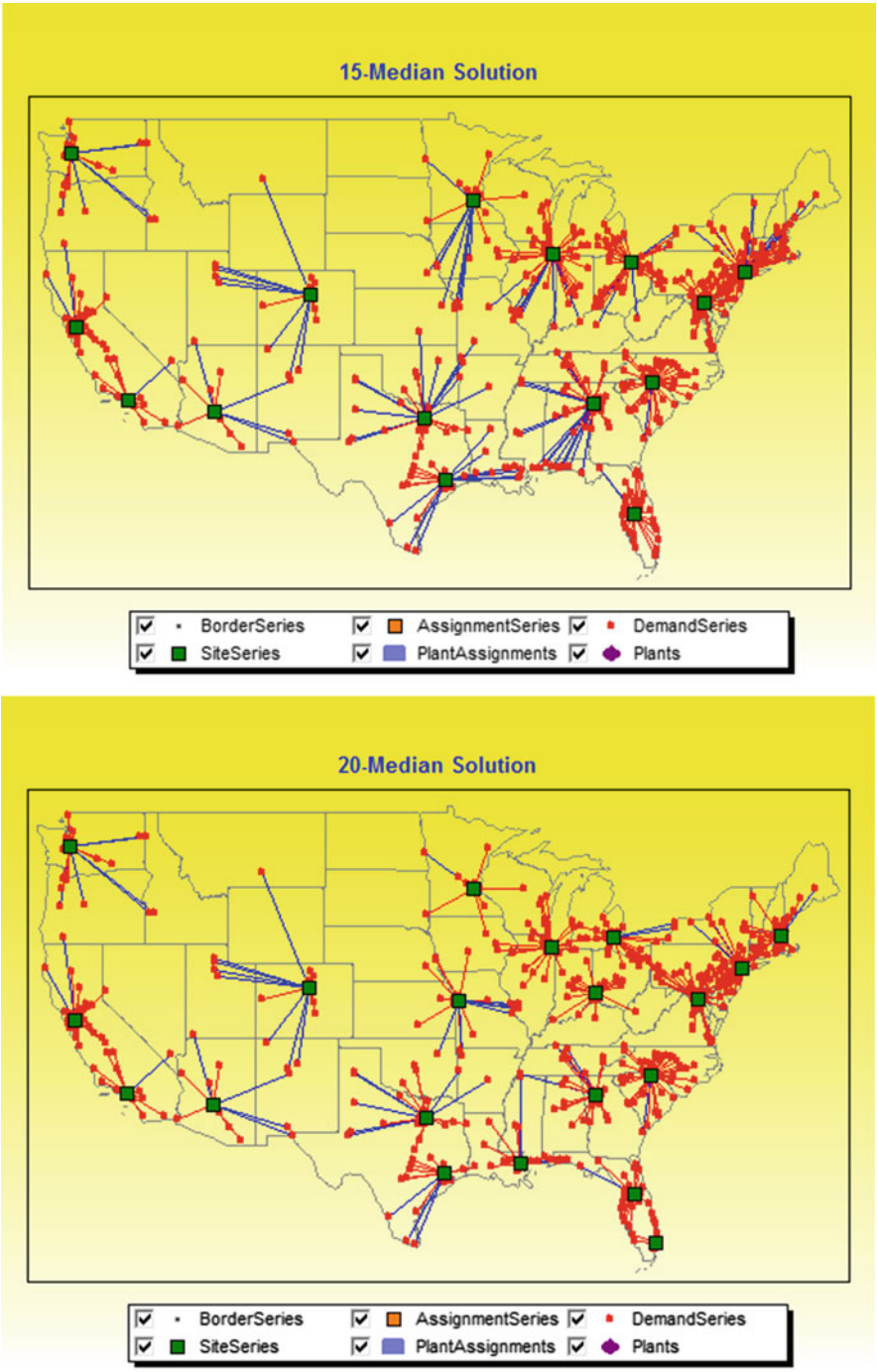


Fig. 2.2 (continued)

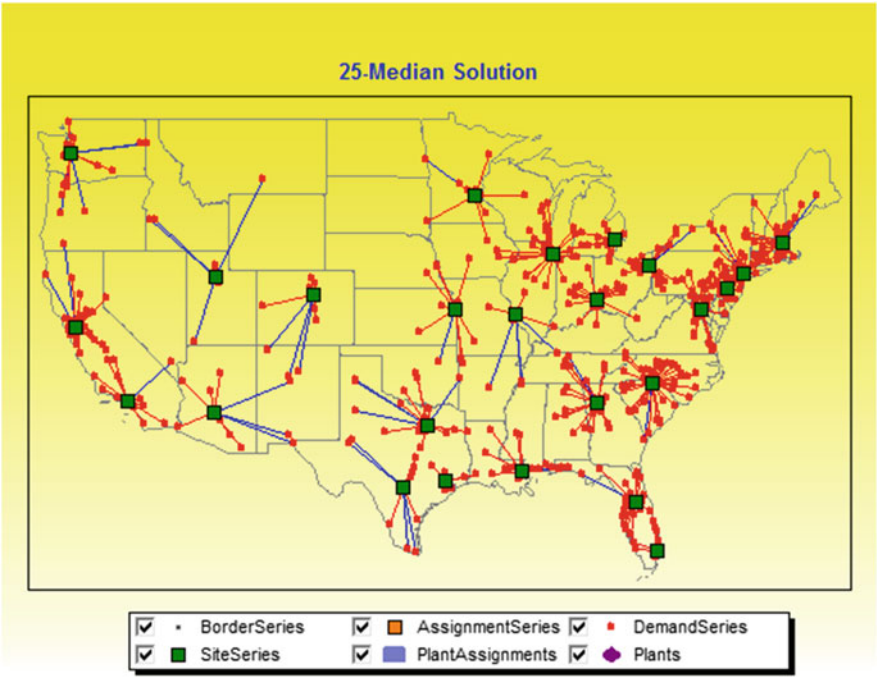


Fig. 2.2 Optimal locations for 5, 10, 15, 20 and 25 medians

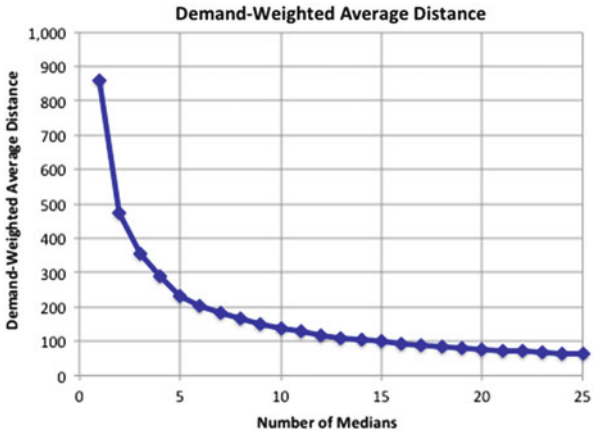


Fig. 2.3 Demand-weighted average distance versus number of medians

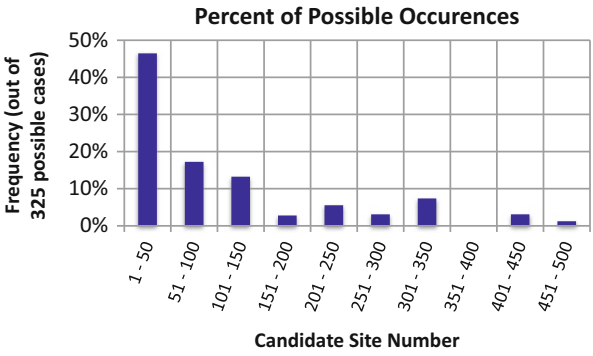


Fig. 2.4 Histogram of the frequency of county selection out of 325 possible cases

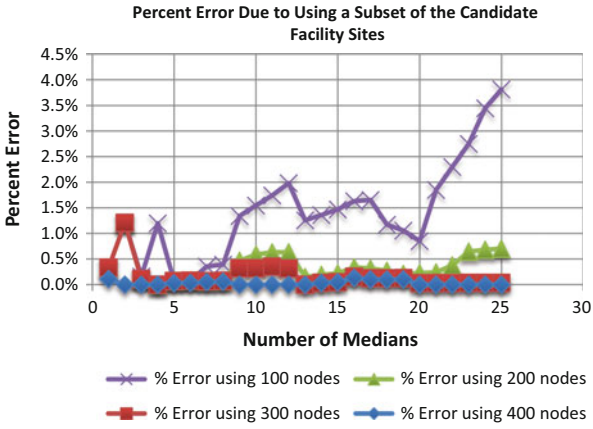


Fig. 2.5 Percent error due to limiting the candidate node set

With 1–25 medians being selected, there conceivably could be 325 unique nodes chosen as medians. This was not the case. Only 55 unique nodes were selected and these were biased toward the larger demand nodes. With the dataset sorted from the most populous to the least populous county, Fig. 2.4 plots the distribution of the number of times nodes in different groupings were selected. Nearly half of the counties selected were among the top 50 most populous counties. Over 75 % of the selected counties were in the top 150 most populous counties.

Figure 2.5 plots the percent error due to limiting the candidate solution set to the most populous 100, 200, 300 and 400 counties, compared to allowing all 500 counties to be in the solution. The errors are generally less than 1 % as long as at least 200 nodes are in the candidate set. Even when the candidate set is limited to only 100 nodes, the maximum error in the 25 runs was under 4 %, though the error seems to be growing with the number of medians in this case.

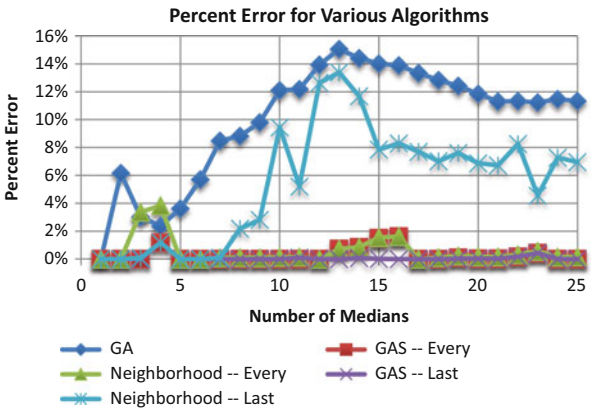


Fig. 2.6 Errors due to using various heuristic algorithms

We next consider the impact of using different construction and improvement algorithms to solve the problem. Five different algorithms were tested: the greedy adding or myopic algorithm (GA), the greedy adding algorithm with the node exchange algorithm applied after every median is added (GAS-Every), the greedy adding algorithm with the neighborhood algorithm applied after every median is added (Neighborhood-Every), the greedy adding algorithm with the exchange algorithm applied after all nodes have been added (GAS-Last), and the greedy adding algorithm using the neighborhood algorithm only after all nodes have been added to the solution (Neighborhood-Last).

Figure 2.6 plots the results. Both the greedy adding algorithm (GA) and the Neighborhood algorithm applied after all nodes have been added to the solution (Neighborhood-Last) result in large errors, often exceeding 10 %. The other three algorithms perform much better and result in errors that are under 4 % and often under 2 %.

Figure 2.7 plots the results of using a genetic algorithm similar to that proposed by Alp et al. (2003). The variant employs a standard crossover operator. To ensure feasibility of the solution generated by the crossover operator, we randomly drop nodes from any solution that has more than p facilities (always retaining facilities that are in both parent's solutions) and randomly add facilities from the parents when the operator results in fewer than p facilities being selected. The standard genetic algorithm can result in large errors and the errors seem to grow with the number of medians. However, if the final genetic algorithm solutions are subject to an exchange algorithm, the errors are under 1 % and average under 0.1 % for the 25 cases.

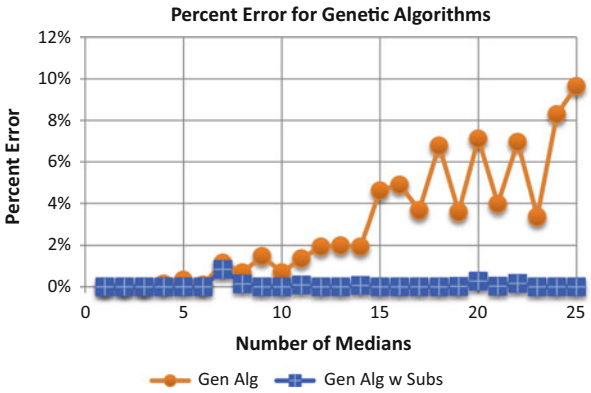


Fig. 2.7 Errors due to using a genetic algorithm

2.7 Multi-objective Extensions of the *p*-Median Model

The formulation above, (2.1)–(2.6), can be modified to obtain a formulation of the maximum covering problem (Church and ReVelle 1974). The maximum covering problem finds the location of p facilities to maximize the number of demand nodes that are covered within some coverage distance, d_c. In particular, we let d_{ij} be the distance between candidate site $i \in \mathcal{I}$ and demand node $j \in \mathcal{J}$. We then define

$$\hat{c}_{ij} = \begin{cases} 0 \ \textit{if} \ d_{ij} \leq d_c \\ 1 \ \textit{if} \ d_{ij} > d_c. \end{cases}$$

If we now solve (2.1)–(2.6) with c_{ij} replaced by $\hat{c}_{ij}$, we will be able to solve a maximum covering problem. In essence, we are minimizing the total number of uncovered demands, which is equivalent to maximizing the number of covered demands.

We can also find the tradeoff between the covering and average cost (or average distance) objective by minimizing a suitable linear combination of the two cost terms. In particular, we minimize a weighted sum $\tilde{c}_{ij} = \alpha c_{ij} + (1-\alpha)\,\hat{c}_{ij}$ of the original c_{ij} and the coverage term $\hat{c}_{ij}$, with $0 \leq \alpha \leq 1$. Clearly, if $\alpha = 1$, the model will simply minimize the demand weighted total distance or cost. Also, if $\alpha = 0$, the model will minimize the number of uncovered demands.

The choice of α is critical if we want to trace out the complete tradeoff curve. Many researchers and practitioners simply solve the problem for fixed values of α. For example, they might solve the problem for $\alpha = 0,\ 0.05,\ 0.1, \ldots, 1.0$. We do not recommend this approach because it is simultaneously likely to miss important points on the tradeoff curve and to result in obtaining many identical solutions.

Instead, one should solve the problem using $\alpha_1 = 1 - \epsilon_1$, where $\epsilon_1 > 0$ is a suitably small value so that we are guaranteed to get one of the (possibly) alternate optima for the p-median problem. Let Z^1 be the objective function value we obtain and let $D^1 = \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} d_j c_{ij} x_{ij}$ be the demand-weighted total distance corresponding to this solution and $U^1 = \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} d_j \hat{c}_{ij} x_{ij}$ be the total uncovered demand corresponding to this solution. Next, solve the problem with $\alpha_2 = \epsilon_2$, where $\epsilon_2 > 0$ is a suitably small value such that we are guaranteed to get one of the (possibly) alternate optima for the maximum covering problem. Let Z^2 be the corresponding objective function value and let D^2 and U^2 be the demand-weighted total distance and uncovered demand corresponding to this solution. We then solve $\alpha_3 D^1 + (1 - \alpha_3) U^1 = \alpha_3 D^2 + (1 - \alpha_3) U^2$ for α_3. This results in $\alpha_3 = \left(U^2 - U^1\right) / \left(D^1 - D^2 + U^2 - U^1\right)$. We then use this value of α to weight the two objectives. This will either result in a new solution being found with a demand-weighted total distance of D^3 and uncovered demand U^3, or the solution will return one of the two original solutions on the tradeoff curve. If a new solution is found, the procedure continues by exploring the region between solution 1 and solution 3 (i.e., using $\alpha_4 = (U^3 - U^1)/(D^1 - D^3 + U^3 - U^1)$) and then between solution 3 and solution 2. If no new solution is found, then no new solution can be identified between solutions 1 and 2. This process continues until all adjacent solutions have been explored in this manner. As a final note, we observe that this is the weighting method, which will fail to find the so-called duality gap solutions (Cohon 1978).

The tradeoff between the demand weighted total distance and the maximum distance—the p-center objective—can also be found using formulation (2.1)–(2.6) if we suitably modify the distance (or cost) matrix, assuming all distances are integer valued. (This is not an overly restrictive assumption since we can approximate any real distances by integer values. For example, if we need distances accurate to the nearest 0.01 mile (or about 50 ft) we just multiply all distances by 100 and round the resulting values.) We do so by initially solving the problem as formulated, letting c_{ij} be the distance between demand node $j \in \mathcal{J}$ and candidate location $i \in \mathcal{I}$. We record the maximum distance, D^0_{max}. We then modify the distance matrix so that $c^{new}_{ij} = \begin{cases} c_{ij} \text{ if } c_{ij} < D^0_{max} \\ M \text{ if } c_{ij} \geq D^0_{max} \end{cases}$, where M is a very large number. We then resolve formulation (2.1)–(2.6) replacing the original costs or distances c_{ij} by c^{new}_{ij}. If M is sufficiently large, the new solution will not entail assignments with distances greater than or equal to D^0_{max}. Let $D^1_{max} < D^0_{max}$ be the new maximum distance. The process continues in this manner until no feasible solution can be found, indicating that the final value of D_{max} that was obtained is the solution to the p-center problem. While

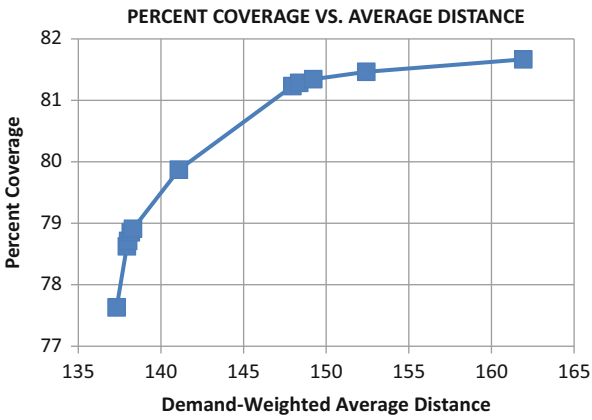


Fig. 2.8 Sample tradeoff between average distance and percent covered

this approach seems to also be a weighting approach since we are assigning a large weight to any distance greater than or equal to the most recently found maximum distance, it is really the constraint method (Cohon 1978) since we are precluding the assignment of demand nodes to facilities that are too far away. This approach will find all non-dominated solutions.

We close this section by illustrating these two multi-objective problems. Figure 2.8 plots the tradeoff between the average distance and the percent of the demand covered within 200 miles using ten facilities with demand represented by the 500 most populous counties of the contiguous United States. The maximum covering solution results in nearly an 18 % increase in the average distance from 137.32 to 161.93 miles, while increasing the percent covered by approximately 4 %. Obtaining the 12 solutions shown in the figure took under 10 min of solution time.

Figure 2.9 is a sample center-median tradeoff curve using the 250 most populous counties in the contiguous US. While this is under 10 % of the counties, it still encompasses over 61 % of the total population in the contiguous US. The algorithm above found 22 solutions (shown with squares and a solid line), only nine of which (shown with circles and a dashed line) could be found using a weighting method. The average distance ranges from about 125 miles to 152 miles, while the maximum distance ranges from a low of 349 miles to a high of 553 miles. Several good compromise solutions are clearly shown at the bend in the curve. Figure 2.10 is an example of one such compromise solution. Obtaining the 22 solutions shown in the figure took nearly 16 h of computing time.

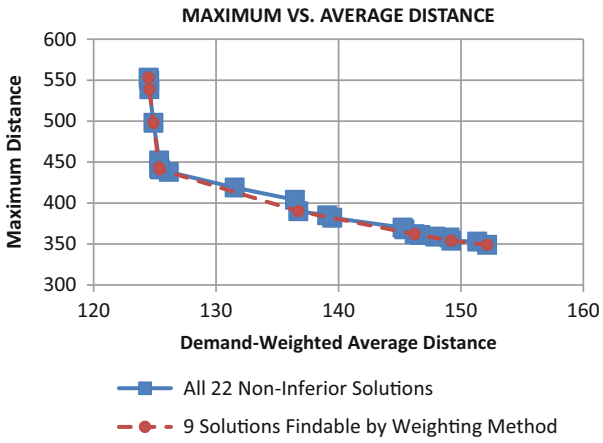


Fig. 2.9 Sample center-median tradeoff

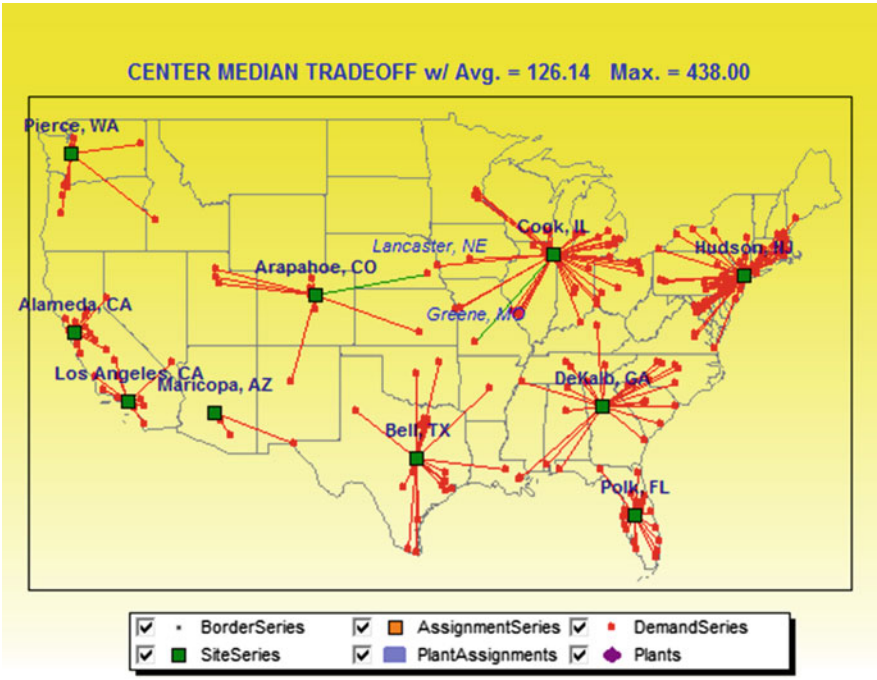


Fig. 2.10 Sample compromise center-median solution

2.8 Conclusions

The *p*-median problem is central to much of discrete location theory and modeling. This chapter has outlined several important model properties and has reviewed a classic formulation of the problem. While the problem is $\mathcal{NP}$-hard on a general graph, it can be solved in polynomial time on a tree. We summarized a linear-time algorithm for the 1-median on a tree and cited results for the general *p*-median problem on a tree. The chapter then presented classic construction and improvement algorithms for the *p*-median problem and pointed the reader to literature on a number of modern heuristic algorithms that have been employed in solving the problem on general graphs. Computational results were presented for both the classical Beasley datasets as well as a 500-node instance based on the most populous counties in the contiguous United States. A well-constructed Lagrangian algorithm embedded in a branch-and-bound algorithm can solve problem instances with up to 1,000 demand nodes and 1,000 candidate sites in reasonable time. (For $p = 1$ the myopic algorithm—which amounts to total enumeration in this case—will find provably optimal solutions.) Larger problem instances may require the use of heuristic algorithms such as tabu search or simulated annealing.

The chapter concluded with two multi-objective extensions of the *p*-median problem. The first examines the tradeoff between the *p*-median objective and the maximum covering objective, while the second explores the tradeoff between the *p*-median objective and the *p*-center objective. For small instances it is often possible to solve bi-objective problems using extensions of the Lagrangian algorithm outlined above. For larger instances, using a genetic algorithm is often advisable since the population of solutions in a genetic algorithm automatically gives an initial approximation of the non-dominated set of solutions.

References

Al-khedhairi A (2008) Simulated annealing metaheuristic for solving P-median problem. Int J Contemporary Math Sci 3:1357–1365

Alp O, Erkut E, Drezner Z (2003) An efficient genetic algorithm for the *p*-median problem. Ann Oper Res 122:21–42

Beasley JE (1990) OR-library: distributing test problems by electronic mail. J Oper Res Soc 41:1069–1072

Chiyoshi F, Galvão RD (2000) A statistical analysis of simulated annealing applied to the *p*-median problem. Ann Oper Res 96:61–74

Church RL (2008) BEAMR: an exact and approximate model for the *p*-median problem. Comput Oper Res 35:417–426

Church RL, ReVelle CS (1974) The maximal covering location problem. Pap Reg Sci Assoc 32:101–118

Cohon JL (1978) Multiobjective programming and planning. Academic, New York

Daskin MS (2013) Network and discrete location: models, algorithms and applications, 2nd edn. Wiley, New York

Fisher ML (1981) The Lagrangian relaxation method for solving integer programming problems. Manag Sci 27:1–18
Fisher ML (1985) An applications oriented guide to Lagrangian relaxation. Interfaces 15:10–21
Glover F (1990) Tabu search: a tutorial. Interfaces 20:74–94
Glover F, Laguna M (1997) Tabu search. Kluwer Academic Publishers, Boston
Goldberg DE (1989) Genetic algorithms in search, optimization and machine learning. Addison-Wesley, Reading
Goldman AJ (1971) Optimal center location in simple networks. Transp Sci 5:212–221
Hakimi SL (1964) Optimum location of switching centers and the absolute centers and medians of a graph. Oper Res 12:450–459
Hakimi SL (1965) Optimum distribution of switching centers in a communication network and some related graph theoretic problems. Oper Res 13:462–475
Hansen P, Mladenović N (1997) Variable neighborhood search for the P-median. Locat Sci 5: 207–226
Hansen P, Mladenović N (2001) Variable neighborhood search: principles and applications for the p-median. Eur J Oper Res 130:449–467
Haupt RL, Haupt SE (1998) Practical genetic algorithms. Wiley, New York
Holland J (1975) Adaption in natural and artificial systems. The University of Michigan Press, Ann Arbor
Kariv O, Hakimi SL (1979) An algorithmic approach to network location problems. II: the p-medians . SIAM J Appl Math 37:539–560
Kirkpatrick S (1984) Optimization by simulated annealing: quantitative studies. J Stat Phys 34:975–986
Maranzana FE (1964) On the location of supply points to minimize transport costs. Oper Res Q 15:261–270
Michalewicz Z (1994) Genetic algorithms + data structures = evolution programs, 2nd edn. Springer, Berlin
Mitchell M (1998) An introduction to genetic algorithms. MIT Press, Cambridge
Mladenović N, Hansen P (1997) Variable neighborhood search. Comput Oper Res 24:1097–1100
Mladenović N, Brimberg J, Hansen P, Moreno-Perez JA (2007) The p-median problem: a survey of metaheuristic approaches. Eur J Oper Res 179:927–939
Murray AT, Church RL (1996) Applying simulated annealing to location-planning problems. J Heuristics 2:31–53
ReVelle CS, Eiselt HA, Daskin MS (2008) A bibliography of some fundamental problem categories in discrete location science. Eur J Oper Res 184:817–848
Rolland E, Schilling DA, Current JR (1996) An efficient tabu search procedure for the p-median problem. Eur J Oper Res 96:329–342
Rosing KE, ReVelle CS (1997) Heuristic concentration: two stage solution construction. Eur J Oper Res 97:75–86
Rosing KE, ReVelle CS, Rolland E, Schilling DA, Current JR (1998) Heuristic concentration and tabu search: a head-to-head comparison. Eur J Oper Res 104:93–99
Tamir A (1996) An $O(pn^2)$ algorithm for the p-median and related problems on tree graphs. Oper Res Lett 19:59–64
Teitz MB, Bart P (1968) Heuristic methods for estimating generalized vertex median of a weighted graph. Oper Res 16:955–961

Chapter 3
Fixed-Charge Facility Location Problems

Elena Fernández and Mercedes Landete

Abstract Fixed-Charge Facility Location Problems are among core problems in Location Science. There is a finite set of users with demand of service and a finite set of potential locations for the facilities that will offer service to users. Two types of decisions must be made: Location decisions determine where to establish the facilities whereas allocation decisions dictate how to satisfy the users demand from the established facilities. Potential applications of various types arise in many different contexts. We provide an overview of the main elements that may intervene in the modeling and the solution process of Fixed-Charge Facility Location Problems, namely, modeling hypotheses and their implications, characteristics of formulations and their relation to other formulations, properties of the domains, and appropriate solution techniques.

Keywords Discrete location • Models and formulations • Solution Algorithms • Inequalities and facets

3.1 Introduction

Fixed-Charge Facility Location Problems (FLPs) are among core problems in Location Science. In FLPs there is a finite set of users with demand of service and a finite set of potential locations for the facilities that will offer service to users. Two types of decisions must be made. Location decisions determine where to establish the facilities whereas allocation decisions dictate how to satisfy the users demand from the established facilities. Each possible decision incurs fixed-charge costs for the facilities that are established and assignment costs for the allocation decisions. In FLPs the aim is to make optimal decisions with respect to the considered costs.

E. Fernández (✉)
Department of Statistics and Operations Research, Universitat Politècnica de Catalunya–Barcelona Tech, Barcelona, Spain
e-mail: e.fernandez@upc.edu

M. Landete
Department of Statistics, Mathematics and Computer Science, University Miguel Hernández, Elche, Spain
e-mail: landete@umh.es

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_3

Applications of FLPs arise in an wide variety of contexts. The book by Drezner and Hamacher (2002) surveys different applications of fixed-charge facility location in such diverse areas as the public sector, software for GIS or robotics. Furthermore, fixed-charge facility location also plays a critical role in many other areas like supply chain management, distributed systems, humanitarian relief, emergency systems, location-routing problems or freight transportation. Melo et al. (2009) survey facility location models in the context of supply chain management until 2009. Klose and Drexl (2005) summarize applications of FLPs within distributed system design. The paper by Balcik and Beamon (2008) is a recent sign of the interest of the combination of both humanitarian relief analysis and facility location models. Further examples of applications can be found in Owen and Daskin (1998), Daskin et al. (2002), Nagy and Salhi (2007) and Jiaa et al. (2007). In fact, the applicability of fixed-charge facility location models goes beyond the area of Location Analysis. Some fixed-charge facility location models are also valid within other fields like machine scheduling, cluster analysis or combinatorial auctions (Escudero et al. 2009; Klose and Drexl 2005; Singh 2008).

It has been traditionally assumed that in FLPs location decisions are strategic, whereas allocation decisions are tactical or operational. There are potential applications, however, in which location and allocation decisions are at the same hierarchy level in the decision making process. One example of application in which both decisions are strategic can be found in the design of a backbone network in telecommunications. An example of application in which both decisions are operational can be faced by some logistic companies which, at each time period, have to solve a FLP to determine the warehouses locations and the distribution pattern to be applied within the corresponding period.

Because FLPs are difficult optimization problems with many potential applications the study of their properties and efficient solution methods is of interest on its own. A further motivation for this study is that it sets the basis for the analysis of more complex models related to FLP extensions. In some cases, these extensions can, in turn, be modeled as some basic FLP. For example, some multi-period facility location problems (see Chap. 11) or some hub-arc location problems (see Chap. 12) can be can be reduced to the FLPs studied here (see, for instance Albareda-Sambola et al. 2009a; Contreras and Fernández 2013).

There are indeed a number of issues that define the characteristics of FLPs. These will be discussed in this chapter and include the possibility of satisfying the demand of each of the users from more than one facility, or capacity limits on the maximum demand that can be served from any selected facility, among others. Furthermore, several alternative formulations can be valid for a given FLP. Usually, none of these alternatives has a clear advantage over the others although, as it often happens with other discrete optimization problems, each of them is better suited for a certain solution technique. We aim to give the reader a broad overview of the main elements that may intervene in the solution process of FLPs, namely, modeling assumptions and their implications, characteristics of formulations and their relation to other formulations, properties of the domains, and appropriate solution techniques. However, in order to keep the length of the chapter within

a reasonable limit, it has been impossible to address all relevant variants and extensions of the problem. As a consequence, we have selected some topics which, in our opinion, cover most of the major issues related to fixed-charge facility location. Diversity among the selected topics has been a major guideline as well.

The material presented in this chapter is the result of the research carried out by many authors in this area over the last 60 years. Most of it has been published but occasionally we present and prove some unpublished results which are either adaptations of well-known results for other cases, or simple results that can be easily derived from the existing state of knowledge.

The remainder of this chapter is structured as follows. In Sect. 3.2 we introduce our notation and we provide an overview of the problems we study. Section 3.2 also discusses modeling issues leading to standard formulations or to alternative Set Partitioning formulations and properties of the domains. A sample of possible solution methods, namely Lagrangean relaxation and column generation is presented in Sect. 3.3. Some of the major difficulties of FLPs that will offer service to users derive from the assumption that individual facilities do not have enough capacity to satisfy the demand of all customers. Releasing this assumption yields a particular FLP known as the Uncapacitated Facility Location Problem (UFLP), which is studied in Sects. 3.4 and 3.5. The UFLP satisfies some specific properties that do not hold for general FLPs. These properties can be exploited for modeling purposes or for deriving specific solution techniques. In particular, Sect. 3.4.1 studies some properties derived from Linear Programming duality, whereas Sect. 3.4.2 presents a formulation for the UFLP based on its supermodular property and relates it with the so-called radius based formulations. Finally, Sect. 3.5 gives some polyhedral results related to the UFLP. The chapter closes in Sect. 3.6 with some comments.

3.2 Overview and Modeling Issues

In this chapter we will use indistinctively the term service center when referring to a facility, and customer or demand point when referring to a user. Let $I = \{1, \ldots, i, \ldots, m\}$ denote the index set for the potential locations for the facilities and $J = \{1, \ldots, j, \ldots, n\}$ the index set for the users. We will refer to potential locations by their indices, so we will say that a facility is open at location i, or simply that facility i is open, if the decision to establish a service center at the potential location i is made. We will also denote users by their indices and simply refer to user j. Associated with each $i \in I$, q_i denotes the maximum capacity of facility i, if it is opened. The service demand of user $j \in J$ is denoted by d_j. As mentioned, there are two types of costs. The decision to establish a facility at $i \in I$ incurs a fixed-charge (setup) cost f_i. For $i \in I$ and $j \in J$, c_{ij} is the cost for serving all the demand of customer j from facility i.

Classical formulations for FLPs use two sets of decision variables: one set for the selection of the facilities to open and another set for the allocation of users demand to open facilities. For the location decisions, associated with each $i \in I$ we define

$$y_i = \begin{cases} 1 \text{ if a facility is open at location } i \\ 0 \text{ otherwise.} \end{cases}$$

For the allocation decisions, associated with $i \in I$, $j \in J$ we define

$$x_{ij} = \begin{cases} 1 \text{ if the demand at user } j \text{ is served by facility } i \\ 0 \text{ otherwise.} \end{cases}$$

A standard integer programming formulation for the FLP is as follows:

$$\text{minimize } z = \sum_{i \in I} f_i y_i + \sum_{i \in I} \sum_{j \in J} c_{ij} x_{ij} \tag{3.1}$$

$$\text{subject to } \sum_{i \in I} x_{ij} = 1 \qquad j \in J \tag{3.2}$$

$$\sum_{j \in J} d_j x_{ij} \leq q_i y_i \qquad i \in I \tag{3.3}$$

$$y_i \in \{0, 1\} \qquad i \in I \tag{3.4}$$

$$x_{ij} \in \{0, 1\} \qquad i \in I,\ j \in J. \tag{3.5}$$

Constraints (3.2) guarantee that each customer is served from one facility, while constraints (3.3) play a double role: (1) they ensure that the capacity of facilities is not exceeded; and (2) they prevent users from being allocated to non-open facilities. Constraints (3.4) and (3.5) define the domains of the decision variables. In the above formulation inequalities (3.3) can be substituted by the two sets:

$$\sum_{j \in J} d_j x_{ij} \leq q_i \qquad i \in I \tag{3.6}$$

$$x_{ij} \leq y_i \qquad i \in I, j \in J. \tag{3.7}$$

Now the set of knapsack constraints (3.6) enforce that facility capacities are not violated, whereas inequalities (3.7) relate the two sets of decision variables. While constraints (3.3) are equivalent to (3.6) and (3.7) when the binary condition of the y variables (3.4) is enforced, the compact set of constraints (3.3) dominates (3.6) and (3.7) when the integrality of the location variables is relaxed to $0 \leq y_i \leq 1$, $i \in I$.

Formulation (3.1)–(3.5) is appropriate for models requiring that the total demand of each customer be served from the same facility. A number of situations exist where such a requirement is justified, the most obvious one being the case

where the demand of each customer represents a physical object that cannot be split. This case is known as the *single allocation* FLP (SFLP). Equations (3.1)–(3.5) is a valid formulation for the SFLP. Many FLP models, however, allow splitting the demand at users among several open facilities. Such models, which are referred to as *multiple allocation* FLPs (MFLPs), arise, for instance, when customers represent population areas and not all the individuals in a given area need to be served from the same service center. In MFLPs allocating customer j to facility i means that some positive fraction of d_j is served from facility i. Hence, for $i \in I$, $j \in J$ the allocation decision variables x_{ij} are defined as the fraction of demand of user j served by facility i, and the domain for the x variables is thus substituted by its continuous relaxation

$$0 \leq x_{ij} \leq 1, \qquad i \in I, j \in J. \tag{3.8}$$

With the above definition of the allocation decision variables, constraints (3.2) have a slightly more general interpretation than in the single allocation case. Since they impose that the sum of all the fractions served from the different facilities be one, they also guarantee that the total demand at each user is satisfied. Therefore, in order to obtain a valid formulation for the MFLP, in formulation (3.1)–(3.5) we "only" have to change the domain of the allocation variables x. It then follows that (3.1)–(3.4) together with (3.8) is a valid formulation for the MFLP.

The FLP is $\mathscr{NP}$-hard since a polynomial transformation can be used to reduce the node cover problem, which is known to be $\mathscr{NP}$-hard (Garey and Johnson 1979), into the FLP (see, for instance, Cornuéjols et al. 1990).

The reader may note that the "difficult" decision in FLPs is the selection of the facilities to open. This is readily seen in the multiple allocation case where, if the set of facilities to open is given, $S \subset I$, the best allocation of customers within S can easily be obtained by solving the following transportation problem:

$$TP(S) \quad \text{minimize } z = \sum_{i \in S} \sum_{j \in J} (c_{ij}/d_j) s_{ij} \tag{3.9}$$

$$\text{subject to} \quad \sum_{i \in S} s_{ij} \geq d_j \qquad j \in J \tag{3.10}$$

$$\sum_{j \in J} s_{ij} \leq q_i \qquad i \in S \tag{3.11}$$

$$s_{ij} \geq 0 \qquad i \in S,\ j \in J. \tag{3.12}$$

In formulation (3.9)–(3.12) above the continuous decision variable s_{ij} denotes the amount of demand of customer j which is served from facility i. Hence we have the relation, $x_{ij} = s_{ij}/d_j$.

In the single allocation case, finding an optimal allocation of customers to a given set of open facilities $S \subset I$ is still a difficult problem, namely a Generalized Assignment Problem, which is also $\mathscr{NP}$-hard (Fisher et al. 1986).

Now, a formulation for finding the best allocation of customers within the set of facilities S is given by:

$$GAP(S) \quad \text{minimize } z = \sum_{i \in S} \sum_{j \in J} c_{ij} x_{ij} \tag{3.13}$$

$$\text{subject to } \sum_{i \in S} x_{ij} = 1 \qquad j \in J \tag{3.14}$$

$$\sum_{j \in J} d_j x_{ij} \leq q_i \qquad i \in S \tag{3.15}$$

$$x_{ij} \in \{0, 1\} \qquad i \in S,\ j \in J. \tag{3.16}$$

So far we have presented FLPs as a minimization problems in which both types of decisions incur costs. However, the type of objective function depends on the decision maker. Minimization FLPs usually appear in the public sector when locating facilities for essential services: public hospitals or schools, dumps for garbage collection, etc. In the private sector, however, service to customers produces a profit to companies so that the objective of companies facing location decisions for their service centers is to maximize the net profit, defined as the difference between the revenue derived from the serviced customers and the cost for the location of the selected facilities. There is indeed an essential difference between these two models: while minimization FLPs impose that all customers be served (no demand point can be excluded from an essential service), in maximization FLPs not all users necessarily have to be served. The company may not have enough incentive for servicing all customers and only those generating a profit in an optimal location setting will be served. However, as we will next see, from a mathematical programming point of view the maximization and minimization versions of the FLP are equivalent.

Consider a maximization FLP where b_{ij} denotes the profit for servicing customer $j \in J$ from facility $i \in I$. As indicated in Cornuéjols et al. (1990), typically, b_{ij} is a function of the unit production costs at facility i (h_i), the unit transportation costs from facility i to customer j (t_{ij}), and the service price for customer j (s_j). That is, $b_{ij} = d_j(s_j - h_i - t_{ij})$. Then, the objective function for a maximization FLP is

$$\text{maximize } z = -\sum_{i \in I} f_i y_i + \sum_{i \in I} \sum_{j \in J} b_{ij} x_{ij}. \tag{3.17}$$

In principle, if not all customers have to be served, allocation constraints should be stated as inequalities, i.e. $\sum_{i \in I} x_{ij} \leq 1$, $j \in J$. However, such constraints are easily transformed into equalities by simply defining a fictitious potential facility 0, representing the facility to which all unserved demand is allocated. To this end, we assume a sufficiently large capacity for the fictitious facility, $q_0 = \sum_{j \in J} d_j$,

and set to zero, both the fixed-charge cost of the fictitious facility ($f_0 = 0$) and the allocation profits of all customers ($b_{0j} = 0$, $j \in J$). Thus, without loss of generality we can assume that in the maximization FLP allocation constraints must also be satisfied as equality.

Taking into account the expression of the coefficients b_{ij} and because of the equality allocation constraints, the second term in (3.17) can be rewritten as

$$\begin{aligned}\sum_{i\in I}\sum_{j\in J} b_{ij}x_{ij} &= \sum_{i\in I}\sum_{j\in J} d_j(s_j - h_i - t_{ij})x_{ij} \\ &= \sum_{i\in I}\sum_{j\in J} d_j s_j x_{ij} - \sum_{i\in I}\sum_{j\in J} d_j(h_i + t_{ij})x_{ij} \\ &= \sum_{j\in J} d_j s_j - \sum_{i\in I}\sum_{j\in J} c'_{ij}x_{ij}.\end{aligned}$$

Hence objective (3.17) reduces to

$$\sum_{j\in J} d_j s_j - \min\left[\sum_{i\in I} f_i y_i + \sum_{i\in I}\sum_{j\in J} c'_{ij}x_{ij}\right]. \tag{3.18}$$

Since the first term in (3.18) is a constant, the maximization FLP is equivalent to a minimization FLP.

3.2.1 Set Partitioning Formulation of FLPs

Below we present alternative formulations for FLPs which use decision variables to model the overall customers demand allocated to open facilities. Consider for the moment the single allocation case and note that feasible assignments to a given facility $i \in I$ are associated with subsets of customers $T \subset J$ such that $\sum_{j\in T} d_j \leq q_i$. We will use the notation K_i to denote the index set of feasible assignment subsets for facility $i \in I$, $T_k \subset J$ the index set of the customers served in feasible assignment $k \in K_i$, and p_{ki} for the fixed-charge cost of facility i plus the cost for assigning to i all the customers indexed in T_k, i.e. $p_{ki} = f_i + \sum_{j\in T_k} c_{ij}$. Also, for $i \in I$, $k \in K_i$, $j \in J$, let $a_{ijk} = 1$ if $j \in T_k$ and 0 otherwise. Consider now the following decision variables:

$$z_{ki} = \begin{cases} 1 \text{ if the subset of customers } T_k \text{ is assigned to facility } i \\ 0 \text{ otherwise.}\end{cases}$$

Then, a set partitioning formulation for the SFLP is

$$SPSFLP \quad \text{minimize} \quad \sum_{i \in I} \sum_{k \in K_i} p_{ki} z_{ki} \tag{3.19}$$

$$\text{subject to} \quad \sum_{i \in I} \sum_{k \in K_i} a_{ijk} z_{ki} = 1 \quad j \in J \tag{3.20}$$

$$\sum_{k \in K_i} z_{ki} = y_i \quad i \in I \tag{3.21}$$

$$y_i \in \{0, 1\} \quad i \in I \tag{3.22}$$

$$z_{ki} \in \{0, 1\} \quad i \in I, k \in K_i. \tag{3.23}$$

Constraints (3.20) ensure that each customer is assigned to exactly one facility. Constraints (3.21) guarantee that no assignment is selected for a non-open facility and also that one feasible assignment is selected for each open facility. Observe that, because of (3.20), constraints (3.21) can be written as $\leq$ inequalities and will still be satisfied as equalities. Constraints (3.22) and (3.23) define the domain of the decision variables. The above a formulation will be referred to as SPSFLP.

A set partitioning formulation for the multiple allocation case can be obtained from the above formulation by simple relaxing the integrality conditions on the z variables to $0 \leq z_{ki} \leq 1, i \in I, k \in K_i$. It is now necessary to use the $\leq$ expression for constraints (3.21), since optimal solutions may exist with some open facility only serving fractions of demand of the allocated customers. This formulation will be referred to as SPMFLP.

The large number of variables both in SPSFLP and in SPMFLP make these formulations suitable for column generation.

3.3 Solution Algorithms for Fixed-Charge Facility Location

In this section we overview solution methods for FLPs. Several heuristic and exact algorithms have been proposed for FLPs and an exhaustive survey on the related literature is outside the scope of this chapter. Branch-and-bound methods proposed in the early papers (Sá 1969; Davis and Ray 1969; Ellwein and Gray 1977; Akinc and Khumawala 1977; Nauss 1978; Neebe and Rao 1983) were followed by many algorithms based on Lagrangean relaxation (Geoffrion and McBride 1978; Christofides and Beasley 1983; Guignard and Kim 1983; Barceló and Casanovas 1984; Klincewicz and Luss 1986; Pirkul 1987; Beasley 1988; Guignard and Opaswongkarn 1990; Barceló et al. 1990, 1991; Cornuéjols et al. 1991; Beasley 1993; Sridharan 1993, 1995; Holmberg et al. 1999). Some of the first works on approximation algorithms are those of Shetty (1990), Shmoys et al. (1997) and Chudak and Shmoys (1999). Algorithms based on Benders and cross

decomposition have been respectively proposed in Wentges (1996) and Van Roy (1986), whereas branch-and-price has been applied by Díaz and Fernández (2002) and Klose and Görtz (2007). Some recent works are Barahona and Chudak (2005), Sankaran (2007), Sharma and Berry (2007), Ghiani et al. (2012), and Zhen et al. (2012). For an overview of heuristics for FLPs the interested reader is addressed to Jacobsen (1983), Filho and Galvão (1998), Delmaire et al. (1999a,b), Hindi and Pienkosz (1999), Cortinhal and Captivo (2003), Ahuja et al. (2004) and references therein.

The most obvious strategy for solving an FLP instance to optimality is to use a standard mixed integer programming (MIP) solver with formulation SFLP or MFLP, depending on the case. This approach may, however, fail on large instances, especially for the single source case. Some alternatives are presented below, which somehow exploit the structure of the problem and lead either to an exact algorithm or to methods that can be embedded within an exact algorithm. First we study Lagrangean relaxation, which has been used by a number of authors both for the single and multiple allocation cases. Then we address the pricing problem for the set partitioning formulation SPSFLP, which is one of the main ingredients of the branch-and-price algorithm of Díaz and Fernández (2002).

3.3.1 *Lagrangean Relaxation*

We next present a Lagrangean relaxation of model SFLP in which the assignment constraints (3.2) are relaxed. This relaxation has been used by a number of authors (see, for instance, Pirkul 1987; Barceló et al. 1990, 1991; Beasley 1993; Holmberg et al. 1999). The Lagrangean subproblem associated with a given set of multipliers $\pi \in \mathbf{R}^n$, is

$$L_{SFLP}(\pi) = \quad \text{minimize} \quad \sum_{i\in I}\left(f_i\, y_i + \sum_{j\in J} c_{ij}x_{ij}\right) + \sum_{j\in J} u_j\left(1 - \sum_{i\in I} x_{ij}\right) \tag{3.24}$$

$$\text{subject to} \quad \sum_{j\in J} d_j\, x_{ij} \le q_i\, y_i \qquad i \in I \tag{3.25}$$

$$x_{ij} \in \{0,1\} \qquad i \in I,\, j \in J \tag{3.26}$$

$$y_i \in \{0,1\} \qquad i \in I. \tag{3.27}$$

After rearranging its terms the objective function can be rewritten as

$$\sum_{j\in J} \pi_j + \min \quad \sum_{i\in I}\left(f_i\, y_i + \sum_{j\in J}\left(c_{ij} - \pi_j\right) x_{ij}\right).$$

A solution to $L_{SFLP}(\pi)$ can be obtained applying the following two steps:

(1) For each $i \in I$ solve the knapsack problem

$$KP(i): \quad \text{maximize} \sum_{j \in J} \left(c_{ij} - \pi_j\right) x_{ij} \tag{3.28}$$

$$\text{subject to} \sum_{j \in J} d_j x_{ij} \leq q_i \tag{3.29}$$

$$x_{ij} \in \{0,1\} \quad j \in J. \tag{3.30}$$

Let $J(i)$ denote the index set of variables at value 1 in an optimal solution to $KP(i)$ and $v(i) = \sum_{j \in J(i)} (c_{ij} - \pi_j)$ its associated optimal value.

(2) For each $i \in I$, with $f_i + v(i) < 0$ then $y_i = 1$, and $x_{ij} = 1$, for $j \in J(i)$.

The Lagrangean dual associated with $L_{SFLP}(\pi)$ is

$$D_{SFLP} \quad \max_{\pi \in \mathbf{R}^n} L_{SFLP}(\pi).$$

Proposition 3.1 *The optimal value of the Lagrangean dual D_{SFLP} coincides with the value of the linear programming (LP) relaxation of program SPSFLP.*

Proof Consider the following Lagrangean function resulting from relaxing constraints (3.20) in *SPSFLP* in a Lagrangean fashion:

$$L_{SPSFLP}(\pi) = \text{minimize} \quad \sum_{i \in I} \sum_{k \in K} p_{ki} z_{ki} + \sum_{j \in J} \pi_j \left(1 - \sum_{i \in I} \sum_{k \in K_i} a_{ijk} z_{ki}\right) \tag{3.31}$$

$$\text{subject to} \quad \sum_{k \in K_i} z_{ki} \leq y_i \qquad i \in I \tag{3.32}$$

$$z_{ki} \geq 0 \qquad i \in I,\ k \in K_i \tag{3.33}$$

$$y_i \in \{0,1\} \qquad i \in I. \tag{3.34}$$

The objective function (3.31) can be expressed as

$$\sum_{j \in J} \pi_j + \min \left[\sum_{i \in I} \sum_{k \in K_i} p_{ki} z_{ki} - \sum_{i \in I} \sum_{k \in K_i} \sum_{j \in J} \pi_j a_{ijk} z_{ki} \right] =$$

$$\sum_{j \in J} \pi_j + \min \left[\sum_{i \in I} \sum_{k \in K_i} (p_{ki} - \sum_{j \in T_k} \pi_j) z_{ki} \right].$$

Thus, for a given vector π, the solution to $L_{SPSFLP}(\pi)$ can be obtained as follows:

- For $i \in I$, do
  - Find $k(i) \in \arg\max_{k \in K_i}\{p_{ki} - \sum_{j \in T_k} \pi_j\}$.
  - If $p_{k(i)i} - \sum_{j \in T_{k(i)}} \pi_j < 0$ then $y_i = 1, z_{k(i)i} = 1, z_{ki} = 0,\ k \in K_i \setminus \{k(i)\}$.
    If $p_{k(i)i} - \sum_{j \in T_{k(i)}} \pi_j \geq 0$ then $y_i = 0, z_{ki} = 0, k \in K_i$.

Note that, for each feasible solution $(\hat{z}, \hat{y})$ to (3.32)–(3.34), for each $i \in I$ there exists a one-to-one correspondence between $(\hat{y}_i, (\hat{z}_{ki})_{k \in K_i})$, and a vector $(\hat{y}_i, (\hat{x}_{ij})_{j \in J})$, that satisfies constraints (3.25). In particular, $\hat{x}_{ij} = \sum_{k \in K_i} a_{ijk}\hat{z}_{ki}$ for all $i \in I$, $j \in J$. Note that the above solution is well defined since for $i \in I$ there is at most one $k \in K_i$ with $\hat{z}_{ki} = 1$. Furthermore, by definition of the z variables, for $i \in I$, $(\hat{x}_{ij})_{j \in J}$ represents a feasible assignment to facility i, i.e. $\sum_{j \in J} d_j \hat{x}_{ij} \leq q_i \hat{y}_i$. Finally, the objective function values of the two solutions coincide since for $i \in I$ fixed, $\sum_{k \in K_i} p_{ki}\hat{z}_{ki} = f_i \hat{y}_i + \sum_{j \in J} c_{ij}\hat{x}_{ij}$. Therefore, taking into account the above considerations, $L_{SPSFLP}(\pi)$ can be rewritten as

$$\sum_{i \in I} \pi_i + \text{minimize} \sum_{j \in J} \left(f_i y_i + \sum_{j \in J} c_{ij} x_{ij} \right) - \sum_{j \in J} \sum_{i \in I} \pi_j x_{ij} \tag{3.35}$$

$$\text{subject to} \sum_{j \in J} d_j x_{ij} \leq q_i y_i \quad i \in I$$

$$x_{ij} \in \{0, 1\} \quad i \in I, j \in J$$

$$y_i \in \{0, 1\} \quad i \in I,$$

which is indeed $L_{SFLP}(\pi)$. □

The reader will immediately conclude that a similar result holds for the MFLP.

Proposition 3.1 establishes that D_{SFLP} and the LP relaxation of SPSFLP are equally tight in terms of the lower bounds they produce (the same is true for D_{MFLP} and the LP relaxation of SPMFLP). Now, the question that arises naturally is how to compare both types of formulations from an algorithmic point of view.

As we have seen, the Lagrangean subproblem $L_{SFLP}(\pi)$ is rather easy to solve and subgradients are easy to compute at each point. For a given vector π, let $(y(\pi), x(\pi))$ denote an optimal solution to $L_{SFLP}(\pi)$. Then, a subgradient of $L_{SFLP}(\pi)$ is given by $\varphi = (\varphi_j)_{j \in J}$, where $\varphi_j = 1 - \sum_{i \in I} x_{ij}(\pi)$. Therefore, D_{SFLP} can be efficiently solved with subgradient optimization. However, when looking for an exact algorithm, the Lagrangean dual D_{MFLP} may not be very handy within an enumeration scheme. In contrast the LP relaxation of SPSFLP may be more demanding than D_{SFLP} from a computational point of view (the pricing subproblem must be solved repeatedly to generate all the needed columns), but it can be very

well integrated within a branch-and-price scheme. For this reason, the next section studies the pricing problem for generating columns for SPSFLP, which is the main component of an exact branch-and-price algorithm for the SFLP based on this formulation (Díaz and Fernández 2002).

3.3.2 The Pricing Problem for SPSFLP

Suppose we have solved the LP relaxation of the subproblem of SPSFLP associated with a subset of columns $\overline{K} = (\overline{K}_i)_{i \in I}$. Let π, and λ denote the optimal values of the dual variables associated with constraints (3.20) and (3.21), respectively. Then in order to know whether there exists a z variable of the overall formulation that, if added to the current set of columns, would improve the current LP solution, we must find the column of the coefficient matrix of *SPSFLP* with the smallest reduced cost. The reduced cost of variable $z_{ki}, i \in I, k \in K_i$, is given by $r_{ki} = p_{ki} - \sum_{j \in J} \pi_j a_{ijk} - \lambda_i$. Thus, in order to find the column that yields the smallest reduced cost we must solve the following pricing problem:

$$(PP) \qquad \min_{i \in I,\, k \in K_i} \quad r_{ki} = p_{ki} - \sum_{j \in J} \pi_j a_{ijk} - \lambda_i .$$

Since $p_{ki} = f_i + \sum_{j \in T_k} c_{ij}$, then $r_{ki} = f_i + \sum_{j \in J} \left(c_{ij} - \pi_j\right) a_{ijk} - \lambda_i$. Note also that feasible columns $\mathbf{a}_{ik}, k \in K_i, i \in I$, are characterized by the condition $\sum_{j \in J} d_j a_{ijk} \leq q_i$. Thus, the solution to *PP* can be obtained by solving a series of independent problems, one for each $i \in I$. Since, for a given $i \in I$, the value $f_i - \lambda_i$ is fixed, then the corresponding problem reduces to

$$\begin{aligned} PP_i \qquad & \text{minimize} && \sum_{j \in J} \left(c_{ij} - \pi_j\right) a_{ijk} \\ & \text{subject to} && \sum_{j \in J} d_j a_{ijk} \leq q_i \\ & && a_{ijk} \in \{0, 1\} \quad j \in J. \end{aligned}$$

3.4 The Uncapacitated Facility Location Problem

An important particular case of the FLP arises under the assumption that the capacity of any open facility is sufficient to satisfy the demand of all customers, i.e. $q_i \geq \sum_{j \in J} d_j, i \in I$, so that the capacity constraints (3.3) are not needed. This particular case is known as the *Uncapacitated Facility Location Problem* (UFLP) and has received a considerable amount of attention. Next we focus on the UFLP

and study some of its properties. The interested reader is addressed to Cornuéjols et al. (1990) for a deeper analysis and further details.

A first observation is that the UFLP basically involves one main decision: finding the set of facilities to open. Note that an optimal allocation of customers within a given set of open facilities, say S, is trivial, and consists of serving all the demand of each customer from a facility in S with minimum allocation cost, with ties broken arbitrarily. That is, for $j \in J$, let $i(j) \in \arg\min\{c_{ij} \mid i \in S\}$ be arbitrarily chosen, then $x_{i(j)j} = 1$, $x_{ij} = 0$, $i \in I \setminus i(j)$ is an optimal allocation of customers within the set of facilities S. Thus, a closed expression for the objective function value for a set of facilities $S \subseteq I$ is $z(S) = \sum_{i \in S} f_i + \sum_{j \in J} \min_{i \in S} c_{ij}$. The main implication of this observation is that the UFLP can be stated as the minimization of a known set function. Before addressing this issue, we study some properties and algorithmic alternatives, derived from a standard MIP formulation for the UFLP.

Indeed a MIP formulation for the UFLP can be obtained with the y and x decision variables of the previous sections. Now it is no longer necessary to impose the binary condition on the allocation variables, even if single allocation is imposed. The argument is simple: if some customer is allocated to more than one facility in an optimal solution, the allocation costs of that customer to all its allocated facilities must be equal (otherwise the solution would not be optimal). Thus the customer can be fully served from any arbitrarily selected open facility of minimum allocation cost. On the other hand, even if capacity constraints are no longer needed, it is still necessary to impose that no customer is assigned to a non-open facility. Thus, by replacing constraints (3.3) by (3.7) we obtain the following valid formulation for the UFLP:

$$
\begin{aligned}
UFLP \quad \text{minimize} \quad & \sum_{i \in I} f_i y_i + \sum_{i \in I} \sum_{j \in J} c_{ij} x_{ij} && && (3.36)\\
\text{subject to} \quad & \sum_{i \in I} x_{ij} = 1 && j \in J && (3.37)\\
& x_{ij} \leq y_i && i \in I, j \in J && (3.38)\\
& 0 \leq x_{ij} && i \in I,\ j \in J && (3.39)\\
& y_i \in \{0, 1\} && i \in I. && (3.40)
\end{aligned}
$$

A broad literature exists on the UFLP. From seminal papers (Kuehn and Hamburger 1963; Stollsteimer 1963; Manne 1964; Balinski 1966; Efroymson 1966; Spielberg 1969a,b; Khumawala 1972; Bilde and Krarup 1977; Cornuéjols et al. 1977; Guignard and Spielberg 1977; Nemhauser et al. 1978) and other early contributions (Guignard 1980; Cornuéjols and Thizy 1982; Guignard 1988; Beasley 1988; Körkel 1989; Beasley 1993; Aardal 1998), to more recent works (Goldengorin et al. 2004; Klose and Drexl 2005; Mladenović et al. 2006; Janacek and Buzna 2008; Beltran-Royo et al. 2012; Letchford and Miller 2012, 2014), virtually any type of solution algorithm has been proposed for it. As with the general facility location problem, an extensive literature review is outside the scope of this chapter. The

interested reader is referred to Krarup and Pruzan (1983), Cornuéjols et al. (1990), Labbé et al. (1995), ReVelle and Laporte (1996) or Verter (2011) for overviews of the main contributions.

3.4.1 Bounds for UFLP Derived from LP Duality

Consider the LP relaxation of UFLP, in which constraints (3.38) have been written as $y_i - x_{ij} \geq 0$, and the upper bound constraints on the y variables as $-y_i \geq -1$, $i \in I$. Let u, w and t denote the vectors of dual variables of appropriate dimensions associated with constraints (3.37), (3.38) and the upper bound constraints, respectively. Then, the dual of the LP relaxation of UFLP is

$$DUFLP \quad \text{maximize} \quad \sum_{j\in J} u_j - \sum_{i\in I} t_i \tag{3.41}$$

$$\text{subject to} \quad \sum_{j\in J} w_{ij} - t_i \leq f_i \quad i \in I \tag{3.42}$$

$$u_j - w_{ij} \leq c_{ij} \qquad i \in I, j \in J \tag{3.43}$$

$$w_{ij} \geq 0 \qquad i \in I,\ j \in J \tag{3.44}$$

$$t_i \geq 0 \qquad i \in I. \tag{3.45}$$

The optimal values for the t variables can be determined from the optimal w values as $t_i = \left(\sum_{j\in J} w_{ij} - f_i\right)^+$, $i \in I$, where $(a)^+ = \max\{0, a\}$. In turn, the optimal w values can be determined from the optimal u values as $w_{ij} = \left(u_j - c_{ij}\right)^+$, $i \in I$, $j \in J$. Therefore, DUFLP can be expressed in terms of only u variables as

$$DUFLP \quad \max D(u) = \sum_{j\in J} u_j - \sum_{i\in I} \left(\sum_{j\in J} \left(u_j - c_{ij}\right)^+ - f_i \right)^+ .$$

Furthermore, the following optimality conditions hold:

(a) There exists an optimal DUFLP solution where $u_j \geq \min_{i\in I} c_{ij}$ for all $j \in J$.
If $u_j < \min_{i\in I} c_{ij}$ for some $j \in J$, then we can increase the value of u_j without decreasing the objective function value.

(b) There exists an optimal DUFLP solution where $\sum_{j\in J} \left(u_j - c_{ij}\right)^+ - f_i \leq 0$ for all $i \in I$.
If $\sum_{j\in J} \left(u_j - c_{ij}\right)^+ - f_i > 0$ for some $i \in I$, we can decrease the value of some component u_j (with $u_j > c_{ij}$) without decreasing the objective function value.

Condition (b) means that the objective function value of an optimal dual solution reduces to $\sum_{j\in J} u_j$. In other words, an optimal dual solution exists with $t_i = 0$ for all $i \in I$. Hence, the complementarity slackness conditions for constraints (3.42) are

$$\left(f_i - \sum_{j\in J} \left(u_j - c_{ij}\right)^+ \right) y_i = 0 \qquad i \in I. \tag{3.46}$$

These conditions, which apply to any primal-dual optimal pair to the LP relaxation of UFLP, hold trivially for all $i \in I$ with $y_i = 0$. When $y_i > 0$, (3.46) holds provided that $\sum_{j\in J} \left(u_j - c_{ij}\right)^+ = f_i$. For the integer UFLP the complementarity slackness conditions (3.46) give the guidelines for primal-dual heuristics. Two alternative strategies may be applied: (1) the primal solution is obtained first and then a vector u is built to satisfy $\sum_{j\in J} \left(u_j - c_{ij}\right)^+ = f_i$ for all $i \in I$ with $y_i = 1$; or (2) the dual solution u is first obtained and then the primal solution sets $y_i = 1$ for all $i \in I$ with $\sum_{j\in J} \left(u_j - c_{ij}\right)^+ = f_i$. The first strategy can be applied starting from any set of open facilities S (which can be obtained, for instance, with a greedy heuristic). The associated dual solution $u(S)$ can be obtained by setting $u_j(S) = \min_{i\in S} c_{ij}$ for all $j \in J$ (note that this solution need not satisfy condition (b)). The DUFLP objective function value for $u_j(S)$ is

$$\begin{aligned} D(u(S)) = \sum_{j\in J} u_j(S) - \sum_{i\in I} \left(\sum_{j\in J} \left(u_j(S) - c_{ij}\right)^+ - f_i \right)^+ = \\ \sum_{j\in J} \min_{i'\in S} c_{i'j} - \sum_{i\in I} \left(\sum_{j\in J} \left(\min_{i'\in S} c_{i'j} - c_{ij} \right)^+ - f_i \right)^+ = \\ \sum_{j\in J} \min_{i'\in S} c_{i'j} - \sum_{i\notin S} \left(\sum_{j\in J} \left(\min_{i'\in S} c_{i'j} - c_{ij} \right)^+ - f_i \right)^+ . \end{aligned}$$

Since the value of the primal solution associated with S is $Z(S) = \sum_{i\in S} f_i + \sum_{j\in J} \min_{i\in S} c_{ij}$, the deviation between the primal/dual values of S and $u(S)$ is

$$Z(S) - D(u(S)) = \sum_{i\in S} f_i + \sum_{i\notin S} \left(\sum_{j\in J} \left(\min_{i'\in S} c_{i'j} - c_{ij} \right)^+ - f_i \right)^+ .$$

The above expression for the deviation suggests choosing S in order to satisfy $\sum_{j\in J} \left(\min_{i'\in S} c_{i'j} - c_{ij}\right)^+ - f_i \leq 0$ for all $i \notin S$, since in this case the above deviation reduces to $\sum_{i\in S} f_i$.

To illustrate the second strategy let u be a dual solution satisfying the optimality condition (b) above and define $I(u) = \{i \in I \mid \sum_{j\in J}(c_{ij}-u_j)^+ - f_i = 0\}$. Assume further that $u_j \geq \min_{i\in I(u)} c_{ij}$. Consider now a set of facilities $S(u) \subseteq I(u)$ satisfying $\max_{i\in I(u)} c_{ij} = \max_{i\in S(u)} c_{ij}$, for all $i \in I$ and let $s_j = \{i \in S(u) \mid c_{ij} < u_j\}$, $j \in J$. Then, $D(u) = Z(S(u))$ (see Proposition 3.2. in Cornuéjols et al. 1990). This means that under the above assumptions, $S(u)$ is an optimal UFLP solution.

Note that $D(u) = Z(S(u))$ means that the optimal UFLP value coincides with that of its LP relaxation. Thus, in general, one should not expect to find a solution u that together with $S(u)$ satisfies the conditions stated above. However the DUALOC heuristic (see Erlenkotter 1978; Bilde and Krarup 1977), which follows this spirit has proved to be extremely effective for finding optimal or near-optimal solutions for the UFLP. The basic idea is to start with $u = (u_j)_{j\in J} = (\min_{i\in I} c_{ij})_{j\in J}$, and then progressively attempt to increase each component u_j while satisfying condition (b). If u_j can be increased, then its next value is $\min\{c_{ij} \mid c_{ij} > u_j\}$, provided that this value satisfies (b). If not, u_j is increased to the maximum possible value. Indeed, the outcome of the above heuristic depends on the order in which the indices in $j \in J$ are considered. Necessary and sufficient conditions for the duality LP gap to be zero, which may lead to tighter bounds have been proposed in Mladenović et al. (2006). Heuristics in the same spirit have been proposed for other discrete facility location problems, like the one for the stochastic version of the FLP proposed in Louveaux and Peeters (1992).

3.4.2 *The UFLP as the Optimization of a Supermodular Set Function*

As mentioned, the UFLP can be stated as the minimization of a set function. In this section we see that an alternative formulation for the UFLP can be obtained by exploiting the supermodularity property of this set function, which has been observed by several authors, namely Spielberg (1969a), Frieze (1974), Babayev (1974), Fisher et al. (1978), and we relate such a formulation with a radius based formulation. We start by recalling some well-known results on supermodular set functions (see, e.g., Section III.3.1 in Nemhauser and Wolsey 1988) and introduce some additional notation.

Definition 3.1 Let N be a finite set, and Z a real-valued function on the subsets of N. The function Z is *supermodular* if $Z(S) + Z(T) \leq Z(S \cup T) + Z(S \cap T), \quad \forall S, T \subseteq N$.

For $i \in N$ let $\rho_i(S) = Z(S \cup \{i\}) - Z(S)$ be the *incremental value* of adding element i to the set S.

Lemma 3.1 *Each of the following statements is equivalent and defines a supermodular set function.*

(a) $Z(S) + Z(T) \leq Z(S \cup T) + Z(S \cap T), \quad \forall S, T \subseteq N.$
(b) $Z(S \cup \{i\}) - Z(S) \leq Z(T \cup \{i\}) - Z(T), \quad \forall S \subset T \subset N$ *and* $i \in N.$
(c) If, in addition, Z *is non-increasing, then* $Z(T) \geq Z(S) + \sum_{i \in T \setminus S} \rho_i(S)$, $\forall S, T \subset I.$

In the following we suppose that N is the set of potential facilities, i.e. $N = I$, and we consider as set function Z the cost function of UFLP solutions. That is $Z(S) = \sum_{i \in S} f_i + \sum_{j \in J} \min_{i \in I} c_{ij}$. To see that $Z(.)$ is supermodular we recall that a positive linear combination of supermodular functions is supermodular and we observe that $Z(S) = f(S) + c(S)$ with $f(S) = \sum_{i \in S} f_i$ and $c(S) = \sum_{j \in J} \min_{i \in I} c_{ij}$. Thus, it is enough to see that both $f(.)$ and $c(.)$ are supermodular. Because $f(S)$ is linear, it is clear that it is supermodular. We next see that $c(.)$ is also supermodular.

Proposition 3.2 $c(.)$ *is supermodular and non-increasing.*

Proof We will use the characterization of supermodular functions of Lemma 3.1b. For $S \subset T \subset I$, and $i \in I \setminus T$,

$$c(S \cup \{i\}) - c(S) = \sum_{j \in J} \left[\min_{i' \in S \cup \{i\}} c_{i'j} - \min_{i' \in S} c_{i'j} \right] = \sum_{j \in J} \min \left\{ 0, c_{ij} - \min_{i' \in S} c_{i'j} \right\} \leq$$

$$\sum_{j \in J} \min \left\{ 0, c_{ij} - \min_{i' \in T} c_{i'j} \right\} = \sum_{j \in J} \left[\min_{i' \in T \cup \{i\}} c_{i'j} - \min_{i' \in T} c_{i'j} \right] =$$

$$c(T \cup \{i\}) - c(T),$$

where the inequality follows since $\min_{i' \in S} c_{i'j} \geq \min_{i' \in T} c_{i'j}$ for all $j \in J$. Furthermore, c is non-increasing since

$$c(S \cup \{i\}) - c(S) = \sum_{j \in J} \left[\min_{i' \in S \cup \{i\}} c_{i'j} - \min_{i' \in S} c_{i'j} \right] \leq 0. \qquad \square$$

For the function $c(.)$ the incremental value of adding element i to the set S is $c(S \cup \{i\}) - c(S)$. Hence, statement (b) of Lemma 3.1 can be rewritten as

$$c(T) \geq c(S) + \sum_{i \in T \setminus S} [c(S \cup \{i\}) - c(S)] = c(S) + \sum_{i \in T \setminus S} [c(S \cup \{i\}) - c(S)], \forall S, T \subset I. \tag{3.47}$$

The UFLP formulation below exploits the supermodular property of $z(.)$ and $c(.)$ as well as the non-increasing property of $c(.)$. Consider the polyhedron

$$P_{SF} = \left\{ (\eta, x, y) \in \mathbb{R} \times \mathbb{B}^{|I|\times|J|} \times B^{|I|} : \eta \geq \sum_{i\in S} f_i y_i + c(S) + \sum_{i\notin S} \rho_i(S) y_i, \forall S \subseteq I \right\},$$

where η is a continuous variable and $\mathbb{B}^{|I|\times|J|}$ and $\mathbb{B}^{|J|}$ are the domains of the binary vectors associated with the location and allocation variables x and y, respectively.

Theorem 3.1 *Let $T \subset I$ and $(\eta, x^T, y^T) \in \mathbb{R} \times \mathbb{B}^{|I|\times|J|} \times \mathbb{B}^{|I|}$, with x and y the incidence vectors of the UFLP solution associated with subset T. Then, $(\eta, x^T, y^T) \in P_{SF}$ if and only if $\eta \geq Z(T)$.*

Proof If $(\eta, x^T, y^T) \in P_{SF}$ then

$$\eta \geq \sum_{i\in T} f_i y_i^T + c(T) + \sum_{i\notin T} \rho_i(T) y_i^T = \sum_{i\in T} f_i + c(T) = Z(T).$$

Suppose now that $\eta \geq Z(T)$. We have

$$\begin{aligned} f(T) &= \sum_{i\in T} f_i y_i^T = \sum_{i\in T\cap S} f_i y_i^T + \sum_{i\in T\setminus S} f_i y_i^T \\ &= \sum_{i\in S} f_i y_i^T + \sum_{i\in T\setminus S} f_i y_i^T, \qquad \text{for all } S \subseteq I. \end{aligned}$$

Since c is non-increasing supermodular, by (3.47), we also have

$$c(T) \geq c(S) + \sum_{i\in T\setminus S} \Big[c(S\cup\{i\}) - c(S) \Big] = c(S) + \sum_{i\notin S} \Big[c(S\cup\{i\}) - c(S) \Big] y_i^T, \quad \text{for all } S \subseteq I.$$

Thus, for all $S \subseteq I$

$$Z(T) = f(T) + c(T) \geq \sum_{i\in S} f_i y_i^T + \sum_{i\in T\setminus S} f_i y_i^T + c(S) + \sum_{i\notin S} \Big[c(S\cup\{i\}) - c(S) \Big] y_i^T.$$

Hence, $\eta \geq Z(T) \geq \sum_{i\in S} f_i y_i^T + c(S) + \sum_{i\notin S} \rho_i(S) y_i^T$, for all $S \subseteq I$.
Therefore, $(\eta, y^T, x^T) \in P_{SF}$ and the result follows. □

As a consequence of Theorem 3.1, the UFLP can be stated as the following MIP (see Nemhauser and Wolsey 1981):

$$\text{minimize} \quad \eta \tag{3.48}$$

$$\text{subject to} \quad \eta \geq \sum_{i\in I} f_i y_i + c(S) + \sum_{e\notin S} \rho_i(S) y_i \qquad \forall S \subseteq I^* \tag{3.49}$$

$$\eta \geq 0 \tag{3.50}$$

$$y_i \in \{0,1\} \qquad i \in I, \tag{3.51}$$

where $I^* = I \cup \{i^*\}$ and i^* is a fictitious facility such that (1) $c_{i^*k} > \max_{i \in I} c_{ij}$, for all $j \in J$; and (2) $\sum_{j \in J} c_{i^*j} > \max_{i \in I}(f_i + \sum_{j \in J} c_{ij})$. This assumption guarantees that at least one variable y_i is at value one in any optimal solution to the above formulation.

Taking into account the supermodularity of $c(.)$ we can obtain a tighter formulation by respectively substituting objective (3.48) and constraints (3.49) by

$$\text{minimize} \quad \sum_{i \in I} f_i y_i + \sum_{j \in J} \eta^j, \tag{3.52}$$

$$\text{and} \quad \eta^j \geq \min_{i \in S} c_{ij} + \sum_{i \notin S} \left[\min_{i' \in S \cup \{i\}} c_{i'j} - \min_{i' \in S} c_{i'j} \right] y_i, \quad \forall S \subseteq I^*, j \in J. \tag{3.53}$$

The following observation indicates that only a polynomial number of constraints (3.53) is required to obtain a valid formulation for the UFLP.

Remark 3.1 For $S \subset I$ and $j \in J$ given, the right-hand side of their associated constraint (3.53) does not change if the summation is taken over all $i \in I$, since $\min_{i' \in S \cup \{i\}} c_{i'j} - \min_{i' \in S} c_{i'j} = 0$, for $i \in S$. Moreover, for any $S \subset I$, the value of $\min_{i \in S} c_{ij}$ will be one of the values c_{ij}, with $i \in S$. That is, for any S its associated constraint (3.53) can be written as

$$\eta^j \geq c_{sj} + \sum_{i \in I} (c_{ij} - c_{sj})^- y_i, \qquad \text{for some } s \in S.$$

To apply the above remark and obtain a formulation with a polynomial number of constraints, for each $j \in J$, we order the elements of I in non-decreasing values of their coefficients c_{ij}, and we denote by i_r the rth index according to that ordering. That is, $c_{i_1 j} \leq c_{i_2 j} \leq \ldots \leq c_{i_m j} \leq c_{i_{m+1} j}$, where $c_{i_{m+1} j} = c_{i^* j}$ is the allocation cost of customer j to the fictitious facility i^*.

Theorem 3.2 *The UFLP can be formulated as*

$$(SUFLP) \quad v_S = \textit{minimize} \quad \sum_{i \in I} f_i y_i + \sum_{j \in J} \eta^j \tag{3.54}$$

$$\textit{subject to} \quad \eta^j \geq c_{i_r j} + \sum_{i \in I} (c_{ij} - c_{i_r j})^- y_i \quad r = 1, \ldots, m+1, \ j \in J \tag{3.55}$$

$$\eta^j \geq 0 \qquad j \in J \tag{3.56}$$

$$y_i \in \{0,1\} \qquad i \in I. \tag{3.57}$$

The proof which is based on Remark 3.1 is left to the reader. Formulation (3.54)–(3.57) involves $|m|$ binary variables y and $|J|$ continuous variables η. Its total number of constraints is $(m+1)|J|$.

The reader familiar with Benders type reformulations (Benders 1962) will immediately observe that, in fact, constraints (3.55) are nothing but Benders cuts. Thus formulation (3.54)–(3.57) admits an alternative interpretation as a Benders type reformulation for the UFLP. The interested reader is addressed to the inspiring chapter by Magnanti and Wong (1990) for an extensive description of the application of Benders reformulations to the UFLP.

We close this section by interpreting SUFLP as a radius-based formulation. Such formulations have been broadly used in recent years for different types of location and hub location problems, after the work by Elloumi et al. (2004). Their main characteristic is the use of decision variables to model the service cost for customers. Using the above notation, in which, for $j \in J$, $c_{i_r j}$ denotes the rth smallest allocation cost for customer j, we define a new set of binary decision variables z_{rj}, $r = 1, \ldots, m$, where $z_{rj} = 1$ if and only if the allocation cost of customer j is at least $c_{i_r j}$. With these decision variables, the allocation cost of customer j can be written as the telescopic sum $c_{i_1 j} + \sum_{r=2}^{m}(c_{i_r j} - c_{i_{r-1} j})z_{rj}$, so that an alternative UFLP formulation is

$$(RUFLP) \quad v_R = \text{minimize} \quad \sum_{i \in I} f_i y_i + \sum_{j \in J}\left(c_{i_1 j} + \sum_{r=2}^{m}(c_{i_r j} - c_{i_{r-1} j})z_{rj}\right) \tag{3.58}$$

$$\text{subject to} \quad z_{rj} + \sum_{\substack{i \in I \\ c_{ij} < c_{i_r j}}} y_i \geq 1 \qquad r = 1, \ldots, m+1, \; j \in J \tag{3.59}$$

$$z_{rj} \in \{0,1\} \qquad j \in J, r = 1, \ldots, m+1 \tag{3.60}$$

$$y_i \in \{0,1\} \qquad i \in I. \tag{3.61}$$

The equivalence between both formulations can be established by observing that feasible solutions to SUFLP define feasible solutions to RUFLP and vice versa. Indeed, if (η, y) is feasible for SUFLP we obtain a feasible RUFLP solution by setting, for each $j \in J$, $z_{rj} = 0$ for all r with $c_{i_r j} \geq \eta^j$, and zero otherwise. Constraints (3.55) guarantee that (z, y) satisfies constraints (3.59) and is feasible for RUFLP. Conversely, we can also check that a feasible SUFLP solution can be obtained from a feasible RUFLP solution by setting for, $j \in J$, $\eta^j = c_{i_{r^*} j}$ with $r^* = \arg\min\{c_{i_r j} : y_{i_r} = 1\}$.

3.5 Polyhedral Analysis of the UFLP

This section concentrates on the polyhedral analysis of the UFLP. We assume the reader is familiar with the basic polyhedral concepts (an exposition can be found, for instance in Nemhauser and Wolsey 1988). Although any UFLP formulation can be analyzed from a polyhedral perspective, we focus on the set packing formulation for the UFLP, because it is the one that has received more attention from a polyhedral point of view. An alternative analysis to the one we develop next, based on a set partitioning UFLP formulation, can be found in Guignard (1980).

As indicated in Sect. 3.2 facility location problems can also be modeled as maximization problems in which the expression of the objective function is (3.17). In the case of the UFLP such a formulation can be easily transformed into a set packing one by doing the change of variables $\bar{y}_i = 1 - y_i$, $i \in I$; i.e. $\bar{y}_i = 1$ if and only if facility i is not opened. The objective function can be rewritten in terms of the new variables as $-\sum_{i\in I} f_i + \sum_{i\in I} f_i \bar{y}_i + \sum_{i\in I}\sum_{j\in J} p_{ij}x_{ij}$, whose maximization reduces to maximizing the objective $\sum_{i\in I} f_i \bar{y}_i + \sum_{i\in I}\sum_{j\in J} p_{ij}x_{ij}$ within the appropriate domain. Hence, a set packing formulation for the UFLP is

$$(KUFLP) \quad \text{maximize } z = \sum_{i\in I} f_i \bar{y}_i + \sum_{i\in I}\sum_{j\in J} p_{ij}x_{ij} \tag{3.62}$$

$$\text{subject to } \sum_{i\in I} x_{ij} \le 1 \qquad j \in J \tag{3.63}$$

$$x_{ij} + \bar{y}_i \le 1 \qquad i \in I, \forall\, j \in J \tag{3.64}$$

$$x_{ij} \in \{0,1\} \qquad i \in I,\ \forall\, j \in J \tag{3.65}$$

$$y_i \in \{0,1\} \qquad i \in I. \tag{3.66}$$

Formulation KUFLP can be viewed as a set packing formulation and thus its set packing properties are inherited. For this we will consider the intersection graph, that we denote by $G(m,n)$, with a node for each variable of KUFLP and with an edge for each pair of variables sharing a constraint in KUFLP.

In the following P^{mn} and F^{mn} denote the convex hull of the feasible solutions of KUFLP and its LP relaxation, LKUFLP, respectively. For $m^* \le m$ and $n^* \le n$, we call $m^* \times n^*$ *adjacency matrix* S to any $m^* \times n^*$, 0-1 matrix with no zero row and no zero column. Given an adjacency matrix S and two ordered sets $I^S \subseteq I$ and $J^S \subseteq J$, we denote by $G^S = (V^S, E^S)$ the subgraph of $G(m,n)$ given by $V^S = \{x_{ij} : i \in I^S, j \in J^S, s_{ij} \neq 0\} \cup \{\bar{y}_i : i \in I^S\}$, $E^S = \{(x_{ij}, x_{kj}) : i,k \in I^S, i < k, j \in J^S, s_{ij} = s_{kj} = 1\} \cup \{(\bar{y}_i, x_{ij}) : i \in I^S, j \in J^S, s_{ij} = 1\}$. Finally, $\alpha(G)$ denotes the independence number of graph G, i.e., the maximal cardinality of a packing of nodes in G, and B denotes a cyclic matrix of type (k,t), i.e. its size is $k \times k$ and its rows are 0-1 vectors with t adjacent 1's, which move one position to the right in each row.

Some relevant contributions on the polyhedral analysis of KUFLP are (in chronological order): Cornuéjols et al. (1977), Guignard (1980), Cornuéjols and Thizy (1982), Cho et al. (1983a,b), Myung and Tcha (1996), Cánovas et al. (2000, 2001, 2002, 2003), Baiou and Barahona (2009a) and Chen et al. (2012). New trends in this area relate to the study of how to adapt the known polyhedral properties of the UFLP to problems generalizing it. Nice examples are the papers by Hamacher et al. (2004) and by Baiou and Barahona (2009b). In both cases the authors give results allowing to directly adapt any valid inequality of the UFLP to the Hub Location Problem and the Two-Level Facility Location Problem, respectively. Next we summarize the main results in this area.

First of all, P^{mn} is full-dimensional, i.e., $\dim(P^{mn}) = mn + p$. Thus, two different facets of P^{mn} always define two different sets of feasible solutions for KUFLP.

Cho et al. (1983a) have proven that for $m \leq 2$ or $n \leq 2$ the coefficients matrix of KUFLP is totally unimodular, so the polyhedral analysis is of little interest. They have also given a complete description of the facets of P^{mn} when $m = 3$ or $n = 3$. Recently, Baiou and Barahona (2009a) and Chen et al. (2012) have presented new conditions for F^{mn} to be integral, i.e., to have all its extreme points integral. Both papers define a particular type of odd cycles in the intersection graph of KUFLP without which the extreme points of the polyhedron F^{mn} are integral.

The remainder of this section is divided in three parts: extreme points of F^{mn}, valid inequalities and facets of P^{mn}, and lifting procedures.

3.5.1 Extreme Points

We are aware of two papers dealing with the characterization of the fractional extreme points. Cornuéjols et al. (1977) give a characterization for the extreme points of F^{mn}. Let $I_f = \{i \in I : 0 < \bar{y}_i < 1\}$, $J_0 = \{j \in J : x_{ij} \in \{0, 1 - \bar{y}_i\}$ for all i and x_{ij} non-integer for some $i\}$ and let U be the $|I_f| \times |J_0|$ matrix whose elements are

$$u_{ij} = \begin{cases} 1 \text{ if } x_{ij} > 0, \\ 0 \text{ if } x_{ij} = 0. \end{cases}$$

Theorem 3.3 (Cornuéjols et al. 1977) *The fractional feasible solution* (x, y) *of LKUFLP is an extreme point of* F^{mn} *if and only if*

(a) $1 - \bar{y}_i = \max_j \{x_{ij}\}$ *for all* $i \in I_f$,
(b) for each $j \in J$, *there is at most one* i *with* $0 < x_{ij} < 1 - \bar{y}_i$,
(c) the rank of U *equals* $|I_f|$.

Cánovas et al. (2001) have later provided a characterization for the extreme points of a more general polyhedron and proved that condition (a) of Theorem 3.3 follows from conditions (b) and (c). Cho et al. (1983a) make use of this characterization

to prove that a certain family of valid inequalities can cut fractional solutions of LKUFLP. The results of Cánovas et al. (2001) also characterize the extreme points of the polyhedra associated with the FLP formulation in Leung and Magnanti (1989) and of other related problems.

3.5.2 Valid Inequalities and Facets

Next we present several families of valid inequalities of P^{mn}. Further details and results can be found in Cho et al. (1983a) and Cánovas et al. (2002).

Cornuéjols et al. (1977) presented the first polyhedral study of the KUFLP. They proposed, without proof, the following family of valid inequalities of P^{mn}

$$\sum_{i \in I^C} b_{ij} x_{ij} + \sum_{i \in I^C} \bar{y}_i \leq 2k - \lceil k/t \rceil, \tag{3.67}$$

where k and t are integers such that $k = tp + 1$ for some integer p, B is a cyclic matrix of type (k, t) and $I^B \subseteq I$, $J^B \subseteq J$ are subsets of cardinality k. Later, Cornuéjols and Thizy (1982) proved that (3.67) is a facet.

Several well-known families of facets for the KUFLP with binary coefficients are discussed below:

Theorem 3.4 (Cho et al. 1983b) *Consider $I^S \subseteq I$ and $J^S \subseteq J$. Then, the inequality*

$$\sum_{i \in I^S} \sum_{j \in J^S} s_{ij} x_{ij} + \sum_{i \in I^S} \bar{y}_i \leq \alpha(G^S),$$

where $s_{ij} = 0$ or 1, is facet-defining for P^{mn} (and different from a clique facet) if and only if S is a $|I^S| \times |J^S|$, maximal mn-adjacency matrix.

A characterization of maximal *mn*-adjacency matrices can be found in Cho et al. (1983b). A special case of maximal *mn*-adjacency matrix gives rise to a concrete family of facet-defining inequalities of P^{mn}:

Theorem 3.5 (Cornuéjols and Thizy 1982) *Consider ℓ and t such that $2 \leq t < \ell \leq m$ and subsets $P \subseteq I$, $D \subseteq J$, such that $|D| = \binom{\ell}{t}$, $|P| = \ell$. Let $A^{\ell t}$ be the matrix whose columns are all vectors 0-1 with t ones and $\ell - t$ zeros. Then,*

$$\sum_{i \in I} \sum_{j \in J} a_{ij}^{\ell t} x_{ij} + \sum_{i \in I} \bar{y}_i \leq \binom{\ell}{t} + t - 1$$

is a facet-defining inequality of P^{mn}.

By exploiting the set packing structure of KUFLP, the odd holes in the intersection graph of KUFLP allow to define two new families of valid inequalities.

Theorem 3.6 (Cornuéjols and Thizy 1982) *The inequality*

$$\sum_{i=1}^{3} x_{ii} + \sum_{i=1}^{3} x_{(i+1) \bmod 3,i} + \sum_{i=1}^{3} \bar{y}_i \leq 4$$

is facet-defining for P^{33}.

Theorem 3.7 (Cornuéjols and Thizy 1982) *The inequality*

$$x_{13} + x_{41} + \sum_{i=1}^{5} x_{ii} + \sum_{i=1}^{5} x_{(i+1) \bmod 5,i} + \sum_{i=1}^{5} \bar{y}_i \leq 7$$

is facet-defining for P^{55}.

Families of facet defining inequalities for KUFLP with general integer coefficients are also known.

Theorem 3.8 (Cánovas et al. 2000) *Let* S *be an* $r \times c$ *adjacency matrix satisfying*

(i) $\forall i_1, i_2 \in I^S \; \exists j \in J^S$ *such that* $s_{i_1 j} s_{i_2 j} = 1$ *and*
(ii) $\forall (i, j) \in I^S \times J^S$ *with* $s_{ij} = 1 \; \exists \ell \in I^S$, $\ell \neq i$, *such that* $s_{\ell j} = 1$ *and* $s_{ih} s_{\ell h} = 0 \; \forall h \neq j$.

Then,

$$\sum_{i \in I^S} \sum_{j \in J^S} s_{ij} x_{ij} + \sum_{i \in I^S} \left(\sum_{j \in J^S} s_{ij} - 1 \right) \bar{y}_i \leq \sum_{i \in I^S} \sum_{j \in J^S} s_{ij} - |I^S| + 1$$

is a facet-defining inequality of P^{rc}.

Theorem 3.9 (Cánovas et al. 2002) *Let* S *be the* $k \times k$ *adjacency matrix,* $k \geq 3$, *given by*

$$S = \begin{pmatrix} 0 & \mathbf{1}_{1 \times (k-1)} \\ \mathbf{1}_{(k-1) \times 1} & I_{(k-1) \times (k-1)} \end{pmatrix}$$

Then,

$$\sum_{i \in I^S} \sum_{j \in J^S} s_{ij} x_{ij} + (k-2) \bar{y}_1 + \sum_{i=2}^{k} \bar{y}_i \leq 2k - 2$$

is a facet-defining inequality of P^{kk}.

Theorem 3.10 (Cánovas et al. 2002) *Consider three numbers, $k \geq 5$, $1 \leq a < k-3$ and $b = k-3-a$ and let S be the $k \times k$ adjacency matrix given by*

$$S = \begin{pmatrix} I_{a\times a} & \mathbf{0}_{a\times b} & \mathbf{0}_{a\times 1} & \mathbf{0}_{a\times 1} & \mathbf{1}_{a\times 1} \\ \mathbf{0}_{b\times a} & I_{b\times b} & \mathbf{1}_{b\times 1} & \mathbf{0}_{b\times 1} & \mathbf{1}_{b\times 1} \\ \mathbf{1}_{1\times a} & \mathbf{0}_{1\times b} & 1 & 0 & 0 \\ \mathbf{0}_{1\times a} & \mathbf{1}_{1\times b} & 0 & 1 & 0 \\ \mathbf{0}_{1\times a} & \mathbf{0}_{1\times b} & 1 & 1 & 1 \end{pmatrix}.$$

Then,

$$\sum_{i\in I^S}\sum_{j\in J^S} s_{ij}x_{ij} + \sum_{i\in I^S\setminus\{k-2,k-1\}} \bar{y}_i + a\bar{y}_{k-2} + b\bar{y}_{k-1} \leq 2k-3$$

is a facet-defining inequality of P^{kk}.

Theorem 3.11 (Cánovas et al. 2002) *Let B be the cyclic $(2k+1, 2)$ matrix, $k \geq 1$, and let S be the $(2k+2) \times (4k+2)$ adjacency matrix given by*

$$S = \begin{pmatrix} B_{(2k+1)\times(2k+1)} & I_{(2k+1)\times(2k+1)} \\ \mathbf{0}_{1\times(2k+1)} & \mathbf{1}_{1\times(2k+1)} \end{pmatrix}.$$

Then,

$$\sum_{i\in I^S}\sum_{j\in J^S} s_{ij}x_{ij} + \sum_{i=1}^{2k+1} 2\bar{y}_i + (k+1)\bar{y}_{2k+2} \leq 6k+3$$

is a facet-defining inequality of $P^{(2k+2)(4k+2)}$.

Other types of inequalities have been suggested. For instance, Myung and Tcha (1996) develop a family of inequalities that may cutoff feasible solutions but not optimal ones. In particular, they propose a method for generating inequalities for a constrained KUFLP which considers its feasible domain and the objective function value, as well. For the sake of brevity, details are omitted here.

3.5.3 *Lifting Procedures*

The procedures that transform a valid inequality (facet) of a polyhedron $P^{m^*n^*}$ into a valid inequality (facet) of an higher polyhedron P^{mn}, $m \geq m^*$ or $n \geq n^*$, are called lifting procedures. Such results invite the study of small polyhedra. The following result indicates how to lift all the facets in the previous section.

Theorem 3.12 (Cho et al. 1983b) *Let*

$$\sum_{i\in P}\sum_{j\in D}\pi_{ij}x_{ij}+\sum_{i\in P}\mu_i\bar{y}_i\leq\pi_0 \tag{3.68}$$

be a facet-defining inequality of $P^{m^*n^*}$. Then, (3.68) is also a facet-defining inequality of P^{mn} for $m\geq m^*$, $n\geq n^*$.

Cho et al. (1983b) also give a constructive procedure for obtaining facets of P^{mn} from cyclic adjacency matrices which do not define facets themselves.

Theorem 3.13 (Cho et al. 1983b) *Consider $P\subseteq I$, $D\subseteq J$, such that $|P|=|D|=q$, $q\geq 3$. Consider the facet-defining inequality of P^{qq} given by*

$$\sum_{i\in P}\sum_{j\in D_i}x_{ij}+\sum_{i\in P}\bar{y}_i\leq 2q-2$$

where the sets D_i are all the different subsets of D with $|D_i|=q-1$. Suppose we add $|S|+|T|$ facilities of I to P in such a way that each facility in S covers $q-1$ destinations and each facility in T covers all the q destinations. Let $|S|=s$ and $|T|=t$. Then,

$$\sum_{i\in I\cup S\cup T}\sum_{j\in D_i}\pi_{ij}x_{ij}+\sum_{i\in I\cup S\cup T}\mu_i\bar{y}_i\leq(2q+s-2)(q-1)+t(q-2)$$

is a facet-defining inequality of $P^{(q+s+t)q}$, where

I. $\pi_{ij}=\mu_i=q-1,\ \ i\in P\cup S, j\in D_i$,
II. $\pi_{ij}=\mu_i=q-2,\ \ i\in T, j\in D_i$.

3.6 Conclusions

Fixed-Charge Facility Location Problems capture the main issues arising in fixed-charge location, so they are an excellent workbench for reviewing relevant aspects in this field. This was the aim of this chapter where we have covered a broad range of possibilities related to the modeling and the solution process of FLPs. Indeed the problems studied in this chapter can be seen as simplifications of more realistic models that take into account additional issues. We have studied deterministic static problems, without taking uncertainty into account (see, for instance, Lin 2009; Albareda-Sambola et al. 2011, 2013; Gao 2012) or temporal aspects (see, for instance, Albareda-Sambola et al. 2009a, 2010, 2012). Also, the way we have considered capacity constraints on the facilities may seem simplistic, since modular capacities (incurring their corresponding costs) can be more realistic (see, for instance, Gouveia and Saldanha da Gama 2006; Gourdin and Klopfenstein

2008; Correia et al. 2010). FLPs can be extended in various ways: One can consider more involved objective functions or multiple objectives (Fernández and Puerto 2003; Boland et al. 2006; Wu et al. 2006; Zanjirani Farahani et al. 2010), problems combining FLP decisions with network design (Melkote and Daskin 2011; Contreras et al. 2012), additional constraints (Albareda-Sambola et al. 2009b; Gendron and Semet 2009; Marín 2011), or the possibility of installing several facilities at the same site (Ghiani et al. 2002), to mention just a few possibilities. Some of these extensions are addressed in other chapters of this book.

Acknowledgements This work was partly supported by the Spanish Ministry of *Economía y Competitividad* through grant MTM2012-36163-C06:04-05 and ERDF funds.

References

Aardal K (1998) Reformulation of capacitated facility location problems: how redundant information can help. Ann Oper Res 82:289–308

Ahuja RK, Orlin JB, Pallottino S, Scaparra MP, Scutellà MG (2004) A multi-exchange heuristic for the single-source capacitated facility location problem. Manag Sci 50:749–760

Akinc U, Khumawala BM (1977) An efficient branch and bound algorithm for the capacitated warehouse location problem. Manag Sci 23:585–594

Albareda-Sambola, M, Fernández E, Hinojosa Y, Puerto J (2009a) The multi-period sequential coverage facility location problem. Comput Oper Res 36:1356–1375

Albareda-Sambola M, Fernández E, Laporte G (2009b) The capacity and distance constrained plant location problem. Comput Oper Res 36(2): 597–611

Albareda-Sambola M, Fernández E, Hinojosa Y, Puerto J (2010) The single period coverage facility location problem: lagrangean heuristic and column generation approaches. TOP 18:43–61

Albareda-Sambola M, Fernández E, Saldanha da Gama F (2011) The facility location problem with Bernoulli demands. Omega 39:335–345

Albareda-Sambola M, Fernández E, Nickel S (2012) Multiperiod location-routing with decoupled time scales. Eur J Oper Res 217:248–258

Albareda-Sambola M, Alonso-Ayuso A, Escudero LF, Fernández E, Pizarro C (2013) Fix-and-relax-coordination for a multi-period location-allocation problem under uncertainty. Comput Oper Res 40:2878–2892

Babayev DA (1974) Comments on a note of Frieze. Math Program 7:249–252

Baïou M, Barahona F (2009a) On the integrality of some facility location polytopes. SIAM J Discret Math 23:665–679

Baïou M, Barahona F (2009b) A polyhedral study of a two-level facility model. IBM Research Report RC24886 (W0910–176) October 28

Balcik B, Beamon M (2008) Facility location in humanitarian relief. Int J Logist Res Appl Leading J Supply Chain Manag 11:101–121

Balinski M (1966) On finding integer solutions to linear programs. In: Proceedings of IBM scientific symposium on combinatorial problems, IBM Data processing division, White Plains, New York

Barahona F, Chudak FA (2005) Near-optimal solutions to large-scale facility location problems. Discret Optim 2:35–50

Barceló J, Casanovas (1984) A heuristic algorithm for the capacitated plant location problem. Eur J Oper Res 15:212–226

Barceló J, Hallefjord Å, Fernández E, Jörnsten K (1990) Lagrangean relaxation and constraint generation procedures for capacitated plant location problems. OR Spektrum 12:79–88

Barceló J, Fernández, E, Jörnsten K (1991) Computational results from a new lagrangean relaxation algorithm for the capacitated plant location problem. Eur J Oper Res 53:38–45
Beasley JE (1988) An algorithm for solving large capacitated warehouse location problems. Eur J Oper Res 33:314–325
Beasley JE (1993) Lagrangean heuristics for location problems. Eur J Oper Res 65:383–399
Beltran-Royo C, Vial J-P, Alonso-Ayuso A (2012) Semi-lagrangian relaxation applied to the uncapacitated facility location problem. Comput Optim Appl 51:387–409
Benders JF (1962) Partitioning procedures for solving mixed-variables programming problems. Numer Math 4:238–252
Bilde O, Krarup J (1977) Sharp lower bounds and efficient algorithms for the simple plant location problem. Ann Discret Math 1:79–97
Boland N, Domínguez-Marín P, Nickel S, Puerto J (2006) Exact procedures for solving the discrete ordered median problem. Comput Oper Res 33:3270–3300
Cánovas L, Landete M, Marín A (2000) New facets for the set packing polytope. Oper Res Lett 27:153–161
Cánovas L, Landete M, Marín A (2001) Extreme points of discrete location polyhedra. TOP 9:115–138
Cánovas L, Landete M, Marín A (2002) On the facets of the simple plant location problem. Discret Appl Math 124:27–53
Cánovas L, Landete M, Marín A (2003) Facet obtaining procedures for set packing problems. SIAM J Discret Math 16:127–155
Chen X, Chen Z, Zang W (2012) Total dual integrality in some facility location problems. SIAM J Discret Math 26:1022–1030
Cho DC, Johnson EL, Padberg MW, Rao MR (1983a) On the uncapacitated plant location problem I: valid inequalities and facets. Math Oper Res 8:579–589
Cho DC, Padberg MW, Rao MR (1983b) On the uncapacitated plant location problem II: facets and lifting theorems. Math Oper Res 8:590–612
Christofides N, Beasley JE (1983) Extensions to a lagrangean relaxation approach for the capacitated warehouse location problem. Eur J Oper Res 12:19–28
Chudak FA, Shmoys DB (1999) Improved approximation algorithms for a capacitated facility location problem. In: Proceedings of the 10th annual ACM-SIAM symposium on discrete algorithms, pp 875–886
Contreras I, Fernández E (2014) Hub location as the minimization of a supermodular set function. Oper Res 62:557–570
Contreras I, Fernández E, Reinelt G (2012) The center facility location/network design problem with budget constraint. Omega 40:847–860
Cornuéjols GP, Thizy JM (1982) Some facets of the simple plant location polytope. Math Program 23:50–74
Cornuéjols GP, Fisher M, Nemhauser GL (1977) On the uncapacitated location problem. Ann Discret Math 1:163–177
Cornuéjols GP, Nemhauser GL, Wolsey LA (1990) The uncapacitated facility location problem. In: Mirchandani PB, Francis RL (eds) Discrete location theory. Wiley-Interscience, New York
Cornuéjols GP, Sridharan R, Thizy JM (1991) A comparison of heuristics and relaxations for the capacitated plant location problem. Eur J Oper Res 50:280–297
Correia I, Gouveia LE, Saldanha da Gama F (2010) Discretized formulations for capacitated location problems with modular distribution costs. Eur J Oper Res 204:237–244
Cortinhal MJ, Captivo ME (2003) Genetic algorithms for the single source capacitated plant location problem. In: Resende MGC, Pinho de Sousa J, Viana A (eds) Metaheuristics: computer decision-making. Kluwer Academic Publishers, Boston, pp 187–216
Daskin MS, Coullard CR, Shen ZM (2002) An inventory-location model: formulation, solution algorithm and computational results. Ann Oper Res 110:83–106
Davis PS, Ray TL (1969) Branch and bound algorithm for the capacitated facilities location problem. Nav Res Logist Q 16:331–343

Delmaire H, Díaz JA, Fernández E, Ortega M (1999b) Comparing new heuristics for the pure integer capacitated plant location problem. Investig Oper 8:217–242

Delmaire H, Díaz JA, Fernández E, Ortega M (1999a) Reactive GRASP and tabu search based heuristics for the single source capacitated plant location problem. Inform Syst Oper Res (INFOR) 37:194–225

Díaz JA, Fernández E (2002) A branch and price algorithm for the single source capacitated plant location problem. J Oper Res Soc 53:728–748

Drezner Z, Hamacher HW (eds) (2002) Facility location: applications and theory. Springer, New York

Efroymson MA, Ray TL (1966) A branch and bound algorithm for plant location. Oper Res 14:361–368

Elloumi S, Labbé M, Pochet Y (2004) A new formulation and resolution method for the p-center problem. INFORMS J Comput 16:84–94

Ellwein LB, Gray P (1971) Solving fixed charge location-allocation problems with capacity and configuration constraints. AIIE T 3:290–299

Erkenkotter D (1978) A dual-based procedure for uncapacitated facility location. Oper Res 26:992–1009

Escudero LF, Landete M, Marín A (2009) A branch-and-cut algorithm for the winner determination problem. Decis Support Syst 46:649–659

Fernández E, Puerto J (2003) Multiobjective solution of the uncapacitated plant location problem. Eur J Oper Res 145:509–529

Filho VJMF, Galvão RD (1998) A tabu search heuristic for the concentrator location problem. Locat Sci 6:189–209

Fisher ML, Nemhauser GL, Wolsey LA (1978) An analysis of approximations for maximizing submodular set functions -II. Math Program Stud 8:73–87

Fisher ML, Jaikumar R, Van Wassenhove LN (1986) A multiplier adjustment method for the generalized assignment problem. Manag Sci 32:1095–1103

Frieze AM (1974) A cost function property for plant location problems. Math Program 7:245–248

Gao Y (2012) Uncertain models for single facility location problems on networks. Appl Math Model 36:2592–2599

Garey MR, Johnson DS (1979) Computers and intractabiliy: a guide to the theory of NP-completeness. Freeman, San Francisco

Gendron B, Semet F (2009) Formulations and relaxations for a multi-echelon capacitated location-distribution problem. Comput Oper Res 36:1335–1355

Geoffrion AM, McBride R (1978) Lagrangean relaxation applied to capacitated facility location problems. AIIE T 10:40–47

Ghiani G, Guerriero F, Musmanno R (2002) The capacitated plant location problem with multiple facilities in the same site. Comput Oper Res 29:1903–1912

Ghiani G, Laganà, Manni E, Triki C (2012) Capacitated location of collection sites in an urban waste management system. Waste Manag 32:1291–1296

Goldengorin B, Ghosh D, Sierksma G (2004) Branch and peg algorithms for the simple plant location problem. Comput Oper Res 31:241–255

Gourdin É, Klopfenstein O (2008) Multi-period capacitated location with modular equipments. Comput Oper Res 35:661–682

Gouveia LE, Saldanha da Gama F (2006) On the capacitated concentrator location problem: a reformulation by discretization. Comput Oper Res 33:1242–1258

Guignard M (1980) Fractional vertices, cuts and facets of the simple plant location problem. Math Program Stud 12:150–162

Guignard, M (1988) A Lagrangean dual ascent algorithm for simple plant location problems. Eur J Oper Res 35:193–200

Guignard M, Kim S (1983) A strong Lagrangean relaxation for capacitated plant location problems. Working Paper 56, Decision Sciences Department, The Wharton School, University of Pennsylvania

Guignard M, Opaswongkarn K (1990) Lagrangean dual ascent algorithms for computing bounds in capacitated location problems. Eur J Oper Res 46:73–83
Guignard M, Spielberg K (1977) Algorithms for exploiting the structure of the simple plant location problem. Ann Discret Math 1:247–271
Hamacher HW, Labbé M, Nickel S, Sonneborn T (2004) Adapting polyhedral properties from facility to hub location problems. Discret Appl Math 145:104–116
Hindi KS, Pieńkosz K (1999) Efficient solution of large scale, single-source, capacitated plant location problem. J Oper Res Soc 50:268–274
Holmberg K, Rönnqvist M, Yuan D (1999) An exact algorithm for the capacitated facility location problems with single sourcing. Eur J Oper Res 113:554–559
Jacobsen SK (1983) Heuristics for the capacitated plant location model. Eur J Oper Res 12:253–261
Janacek J, Buzna L (2008) An acceleration of Erlenkotter-Körkel's algorithms for the uncapacitated facility location problem. Ann Oper Res 164:97–109
Jiaa H, Ordoñez F, Dessouky M (2007) A modeling framework for facility location of medical services for large-scale emergencies. IIE T 39:41–55
Khumawala BM (1972) An efficient branch and bound algorithm for the warehouse location problem. Manag Sci 18:718–731
Klincewicz JG, Luss H (1986) A Lagrangean relaxation heuristic for the capacitated facility location with single-source constraints. J Oper Res Soc 37:495–500
Klose A, Drexl A (2005) Facility location models for distribution system design. Eur J Oper Res 162:4–29
Klose A, Görtz S (2007) A branch-and-price algorithm for the capacitated facility location problem. Eur J Oper Res 179:1109–1125
Körkel M (1989) On the exact solution of large-scale simple plant location problems. Eur J Oper Res 39:157–173
Krarup J, Pruzan PM (1983) The simple plant location problem: survey and synthesis. Eur J Oper Res 12:36–81
Kuehn AA, Hamburger MJ (1963) A heuristic program for locating warehouses. Manag Sci 9:645–666
Labbé M, Peeters D, Thisse JF (1995) Location on networks. In: Ball MO, Magnanti TL, Monma CL, Nemhauser GL (eds) Network routing, handbooks in operations research and management science, vol 8. North-Holland, Amsterdam, pp 551–624
Letchford AN, Miller SJ (2012) Fast bounding procedures for large instances of the simple plant location problem. Comput Oper Res 39:985–990
Letchford AN, Miller SJ (2014) An aggressive reduction scheme for the simple plant location problem. Eur J Oper Res 234:674–682
Leung J, Magnanti TL (1989) Valid inequalities and facets of the capacitated plant location problem. Math Program 44:271–291
Lin CKY (2009) Stochastic single-source capacitated facility location model with service level requirements. Int J Prod Econ 117:439–451
Louveaux FV, Peeters D (1992) A dual-based procedure for stochastic facility location. Oper Res 40:564–573
Magnanti TL, Wong RT (1990) Decomposition methods for facility location problems. In: Mirchandani PB, Francis RL (eds) Discrete location theory. Wiley-Interscience, New York
Manne AS (1964) Plant location under economies-of-scale – decentralization and computations. Manag Sci 11:213–235
Marín A (2011) The discrete facility location problem with balanced allocation of customers. Eur J Oper Res 210:27–38
Melkote S, Daskin MS (2001) Capacitated facility location/network design problems. Eur J Oper Res 129:481–495
Melo T, Nickel S, Saldanha da Gama F (2009) Facility location and supply chain management: a review. Eur J Oper Res 196:401–412

Mladenović N, Brimberg J, Hansen P (2006) A note on duality gap in the simple plant location problem. Eur J Oper Res 174:11–22
Myung YS, Tcha DW (1996) Feasible region reduction cuts for the simple plant location problem. J Oper Res Soc Jpn 39:614–622
Nagy G, Salhi S (2007) Location-routing: issues, models and methods. Eur J Oper Res 177:649–672
Nauss RM (1978) An improved algorithm for the capacitated facility location problem. J Oper Res Soc 29:1195–1201
Neebe AW, Rao MR (1983) An algorithm for the fixed-charge assigning users to sources problem. J Oper Res Soc 34:1107–1113
Nemhauser GL, Wolsey LA (1981) Maximizing submodular functions: formulations and analysis of algorithms. Ann Discret Math 11:279–301
Nemhauser GL, Wolsey LA (1988) Integer and combinatorial optimization. Wiley, New York
Nemhauser GL, Wolsey LA, Fisher ML (1978) An analysis of approximations for maximizing submodular set functions I. Math Program 14:265–294
Owen SH, Daskin MS (1998) Strategic facility location: a review. Eur J Oper Res 111:423–447
Pirkul H (1987) Efficient algorithms for the capacitated concentrator location problems. Comput Oper Res 14:197–208
ReVelle CS, Laporte G (1996) The plant location problem: new models and research prospects. Oper Res 44:864–874
Sá G (1969) Branch and bound and approximate solutions to the capacitated plant-location problem. Oper Res 17:1005–1016
Sankaran JK (2007) On solving large instances of the capacitated facility location problem. Eur J Oper Res 178:663–676
Sharma RRK, Berry V (2007) Developing new formulations and relaxations of single stage capacitated warehouse location problem (SSCWLP): empirical investigation for assessing relative strengths and computational effort. Eur J Oper Res 177:803–812
Shetty B (1990) Approximate solutions to large scale capacitated facility location problems. Appl Math Comput 39:159–175
Shmoys DB, Tardos E, Aardal K (1997) Approximation algorithms for facility location problems. In: Proceedings of the 29th annual ACM symposium on theory of computing (STOC). ACM Press, New York, pp 265–274
Singh KN (2008) The uncapacitated facility location problem: some applications in scheduling and routing. Int J Oper Res 5:36–43
Spielberg K (1969a) Plant location with generalized search origin. Manag Sci 16:165–178
Spielberg K (1969b) Algorithms for the simple plant location problem with some side conditions. Oper Res 17:85–111
Sridharan R (1993) A lagrangian heuristic for the capacitated plant location problem with single source constraints. Eur J Oper Res 66:305–312
Sridharan R (1995) The capacitated plant location problem. Eur J Oper Res 87:203–213
Stollsteimer JF (1963) A working model for plant numbers and locations. J Farm Econ 45:631–645
Van Roy TJ (1986) A cross decomposition algorithm for capacitated facility location. Oper Res 34:145–163
Verter V (2011) Uncapacitated and capacitated facility location problems. In: Eiselt HA, Marianov V (eds) Principles of location science. Springer, New York, pp 25–37
Wentges P (1996) Accelerating Benders decomposition for the capacitated facility location problem. Math Method Oper Res 44:267–290
Wu LY, Zhang XS, Zhang JL (2006) Capacitated facility location problem with general setup cost. Comput Oper Res 33:1226–1241
Zanjirani Farahani R, SteadieSeifi M, Asgari N (2010) Multiple criteria facility location problems: a survey. Appl Math Model 34:1689–1709
Zhen Y, Chu F, Chen H (2012) A cut-and-solve based algorithm for the single-source capacitated facility location problem. Eur J Oper Res 221:521–532

Chapter 4
p-Center Problems

Hatice Calik, Martine Labbé, and Hande Yaman

Abstract A p-center is a minimax solution that consists in a set of p points that minimizes the maximum distance between a demand point and a closest point belonging to that set. We present different variants of that problem. We review special polynomial cases, determine the complexity of the problems and present mixed integer linear programming formulations, exact algorithms and heuristics. Several extensions are also reviewed.

Keywords p-Center • Location in public sector • Minimax facility location

4.1 Introduction

Minimizing the total or average distance that potential users have to travel to reach a facility may not be the right decision criterion for placing a public facility. Total- or average distance minimization tends to favor clients who are clustered in population centers to the detriment of clients who are spatially dispersed. Discrimination of this kind with regard to accessibility may have a negative impact on remote clients in the case of an emergency service (ambulances, fire brigades, police stations, etc.). As a result, decision makers may want to consider a criterion focusing on clients who are the poorest served.

H. Calik (✉)
Department of Industrial Engineering, Bilkent University, Bilkent, 06800 Ankara, Turkey

HEC Montréal, Canada Research Chair in Logistics and Transportation, Montréal, QC, Canada H3T 2A7
e-mail: calik@bilkent.edu.tr

M. Labbé
Department of Computer Science, Université Libre de Bruxelles, CP 212, Boulevard du Triomphe, 1050 Brussels, Belgium
e-mail: mlabbe@ulb.ac.be

H. Yaman
Department of Industrial Engineering, Bilkent University, Bilkent, 06800 Ankara, Turkey
e-mail: hyaman@bilkent.edu.tr

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_4

The center problem, defined as finding a vertex whose distance to all the other vertices of a graph is minimum, has been known for a long time in graph theory (see, for instance, Berge 1967).

Hakimi (1964) introduced the absolute center problem to locate a police station or a hospital such that the maximum distance of the station to a set of communities connected by a highway system is minimized. Given a graph $G = (V, E)$ with $V = \{v_1, \ldots, v_n\}$, weight w_j for node $v_j \in V$ and length ℓ_{ij} for edge $\{i, j\} \in E$ connecting nodes v_i and v_j, the aim of the *absolute center problem* is to find a point x on the nodes or edges such that $\max_{j=1,\ldots,n} w_j d(v_j, x)$ is minimized, where $d(v_j, x)$ is the length of the shortest path between node v_j and point x (referred to as distance between v_j and x). The optimal value is called the *absolute radius* of graph G. If x is limited to the nodes of G, then we obtain the *center* of graph G and the optimal value is the *radius* of G. The center of G is not necessarily an absolute center of G. In other words, the absolute radius can be smaller than the radius. To see this, consider a very simple example with two nodes of weight 1 and an edge connecting them with length 1. In this example, the absolute radius is 0.5 whereas the radius is 1.

Hakimi (1964) proposed a solution method to compute the absolute center of a graph and motivated further studies of this problem by casting it as a game. Two people, X and Y, are playing a game on a graph G. First player X chooses a point x in G, then player Y chooses a point y in G and X pays $d(x, y)$ units to Y. When X chooses point x, Y chooses a point farthest from x to maximize his gain. Hence, player X computes the absolute radius of graph G to minimize his loss.

In the conclusion of his subsequent paper on median and covering problems, Hakimi (1965) mentions the generalization of the absolute center problem to the p-center problem. Given a set $X_p = \{x_1, \ldots, x_p\}$ of p points in G, the distance $d(X_p, v_j)$ between X_p and node v_j is computed as $\min_{i=1,\ldots,p} d(x_i, v_j)$. The *$p$-center problem* is to find a set X_p of p points in G such that $\max_{j=1,\ldots,n} w_j d(v_j, X_p)$ is minimized.

As defined above, the p-center problem is a network location problem. The literature contains several variants. In this chapter, we refer to the following variants:

- *vertex-restricted p-center problem*: X_p is restricted to be a subset of the node set;
- *unweighted p-center problem*: all node weights are equal;
- *discrete p-center problem*: the graph $G = (J \cup I, E)$ is bipartite and complete with I denoting the set of possible facility locations and J denoting the set of demand points.

One can find a discussion of several theoretical results and exact methods for the p-center problem on general and tree networks in Tansel (2011). A large scale review of the exact and heuristic methods proposed for the p-center and capacitated p-center problems is provided by Calik (2013).

This chapter is organized as follows. We review some polynomial cases, identify the complexity of the problems in general and present some approximation results in Sect. 4.2. Section 4.3 is devoted to the mixed integer linear programming models and algorithms for solving p-center problems. Heuristics are discussed in Sect. 4.4 and some extensions of the p-center problem are considered in Sect. 4.5. Section 4.6 concludes the chapter.

4.2 Polynomial Cases, Complexity and Approximation Results

An algorithm to compute an absolute center of a graph was proposed by Hakimi (1964). The idea is to compute, for each edge, an optimal point assuming that the center is restricted to be on that edge. Such an optimal point is called a local center of that edge. Then the algorithm finds the best local center. Hence, the overall complexity is equal to the number of edges multiplied by the complexity of computing a local center of an edge.

The computation of a local absolute center is based on the observation that the objective function is piecewise linear on each edge and that local minima correspond to *intersection points* and vertices (see Minieka 1970). A point x on edge $\{v_k, v_m\} \in E$ qualifies as an intersection point if there exist two distinct nodes $v_i, v_j \in V$ such that x is the unique point on $\{v_k, v_m\}$ for which $d(v_i, x) = d(v_i, v_k) + d(v_k, x) = d(x, v_j) = d(x, v_m) + d(v_m, v_j)$.

It follows from this definition that the number of intersection points on an edge is bounded by $O(n^2)$. Nevertheless, Kariv and Hakimi (1979) observed that at most $n + 1$ such points can be local minima of the objective function. The resulting algorithm proposed by Kariv and Hakimi (1979) solves the absolute center problem in $O(|E|n + n^2 logn)$ time.

An algorithm for finding an absolute center in the weighted case can be derived along the same lines. First, a solution can also be found in the set of local centers, i.e., solutions to the problems in which the solution is restricted to be on an edge. Then, the objective function remains piecewise linear on each edge but the slopes of the linear pieces depend on the vertex weights w_j. Kariv and Hakimi (1979) showed that, on an edge, at most $3n - 2$ intersections points can determine a local minima. A point x on an edge $\{v_k, v_m\}$ is now an intersection point if there exist two distinct nodes $v_i, v_j \in V$ such that x is the unique point on $\{v_k, v_m\}$ for which $w_i d(v_i, x) = w_i(d(v_i, v_k) + d(v_k, x)) = w_j d(x, v_j) = w_j(d(x, v_m) + d(v_m, v_j))$. The complexity of the resulting algorithm proposed by Kariv and Hakimi (1979) is $O(|E|nlogn)$.

Goldman (1972) proposed an $O(n^2)$ algorithm to find an absolute center of a tree in the unweighted case. The algorithm checks whether an edge contains an absolute center and if not, searches the two subtrees obtained by deleting this edge. Handler (1973) proposed an $O(n)$ algorithm exploiting the fact that the midpoint of a longest

path of the tree is an absolute center and that the distance is a convex function along any path of the tree. Given any node v_i, the algorithm first determines the vertex v_j whose distance to v_i is maximum, then determines the node v_k whose distance to v_j is maximum. The path linking v_j and v_k is a longest one and the absolute center is its midpoint.

Kariv and Hakimi (1979) provided an $O(nlogn)$ algorithm for the weighted center problem on a tree, which was improved to $O(n)$ by Megiddo (1983).

For an arbitrary graph G and $p \geq 2$, Kariv and Hakimi (1979) proved that the p-center problem is NP-hard even on a planar graph where the maximum degree is 3 and all node weights and edge lengths are equal to 1. The result is also true for the vertex-restricted problem. The authors show that the problem with $p \geq 2$ can be solved in $O(n^2 logn)$ time when G is a tree.

Hochbaum and Shmoys (1985) developed a two-approximation algorithm for the unweighted discrete problem with $I = J$ and edge lengths satisfying the triangle inequality. The algorithm runs in $O(|E|log|E|)$ time. Hsu and Nemhauser (1979) proved that it is NP-hard to find an approximation with a better guarantee. Dyer and Frieze (1985) gave an $O(np)$ algorithm with a guarantee of $\min\{3, 1 + \alpha\}$, where α is the ratio of the largest weight and the minimum weight. In the unweighted case, this guarantee is 2.

4.3 Exact Methods for p-Center Problems

We first observe that the different variants of the p-center problem can be transformed into a discrete p-center problem and solved as such.

In the case of the vertex-restricted p-center problem, the set I of possible locations and the set J of demand points are both equal to the set of vertices V.

The weighted and unweighted absolute p-center problems enjoy the same property as their single facility counterpart: an optimal solution can always be found in the set of vertices and intersection points. This follows from the fact that each point x_i of an optimal solution X_p must be a local minimizer of the function given by the maximum (possibly weighted) distance to the vertices that are allocated to x_i, i.e., which are closer to that x_i than to any other point of X_p. To transform an absolute p-center problem into a discrete p-center problem one thus simply sets $I = V \cup P$, where P denotes the set of intersection points, and $J = V$.

The remainder of this section is now devoted to models and algorithms for solving the discrete p-center problem.

Several methods based on solving finite series of an auxiliary problem called the set covering problem are developed. The set covering problem is a kind of covering problem (see Chap. 5), which is closely related to the p-center problem. Given a zero-one matrix $A = [a_{ji}]$, the set covering problem consists of finding a set of columns at minimum cost that covers the rows of the matrix A. In order to minimize the number of facilities required to serve all customers within a given radius value

r, one can solve a set covering problem with unit column costs by constructing A as follows:

$$a_{ji} = \begin{cases} 1, \text{ if } d(j,i) \leq r, \\ 0, \text{ otherwise} \end{cases} \quad \forall j \in J, i \in I.$$

If the optimal value of the set covering problem is greater than p, then the optimal value of the p-center problem needs to be greater than r; if it is less than or equal to p, then it means that the optimal value of the p-center problem is less than or equal to r.

The first set covering based approach was proposed by Minieka (1970). Let $r_1 < r_2 < \ldots < r_K$ be an ordering of the distinct distance values in the distance matrix $D = [d_{ji}] : d_{ji} = d(j,i), i \in I, j \in J$ and $R = \{r_1, r_2, \ldots, r_K\}$. The method by Minieka (1970) solves the set covering problem for a smaller value in R not yet considered at each step by updating the matrix A. The algorithm terminates when the optimal value of the set covering problem is greater than p. Since the number of different distance values in D is at most $|I|.|J|$, the algorithm converges to an optimal solution in a finite number of steps.

Garfinkel et al. (1977) improved the set covering based approach by Minieka (1970) by first finding a heuristic solution, then, reducing the search space of the radius values and eliminating some of the intersection points. They also reduce the size of the set covering matrix by using standard matrix reductions and heuristic techniques. For the selection of the radius values at the next step, they proposed using a bisection or binary search strategy instead of moving to the next smaller radius value.

The first mixed integer programming (MIP) formulation for the discrete p-center problem was proposed by Daskin (2013). The following decision variables are defined: $y_i = 1$ if a facility is placed at node $i \in I$ and 0 otherwise, $x_{ij} = 1$ if $j \in J$ is assigned to a facility placed at $i \in I$ and 0 otherwise. The formulation by Daskin can be stated as follows:

$$\text{Minimize} \quad z \tag{4.1}$$

$$\text{subject to} \quad \sum_{i \in I} d_{ji} x_{ij} \leq z \qquad \forall j \in J, \tag{4.2}$$

$$\sum_{i \in I} x_{ij} = 1 \qquad \forall j \in J, \tag{4.3}$$

$$x_{ij} \leq y_i \qquad \forall i \in I, j \in J, \tag{4.4}$$

$$\sum_{i \in I} y_i \leq p, \tag{4.5}$$

$$y_i \in \{0,1\} \qquad \forall i \in I, \tag{4.6}$$

$$x_{ij} \in \{0,1\} \qquad \forall i \in I, j \in J. \tag{4.7}$$

The objective function (4.1) together with (4.2) ensure that the objective value is no less than the maximum of the distances between demand points and their facilities. Constraints (4.3) establish the assignment of each demand point to exactly one facility. Constraints (4.4) avoid assignment of demand points to locations with no facility. Constraint (4.5) restricts the number of facilities to p. Constraints (4.6) and (4.7) are the binary restrictions.

Daskin (2013) also proposed a set covering based algorithm, in which the radius value of the set covering problem is selected from an interval of real numbers between pre-determined lower and upper bounds. At each step of the algorithm, the interval is halved and one of the segments is removed depending on whether the objective value of the set covering problem is greater than p or less than or equal to p.

Ilhan and Pınar (2001) proposed a two-phase extension of the algorithm developed by Daskin (2013). In the first phase, they solve the linear programming (LP) relaxation of the feasibility problem defined by (4.5), (4.6), and

$$\sum_{i \in I} a_{ji} y_i \geq 1, \quad \forall j \in J, \tag{4.8}$$

iteratively for fixed r values to obtain a relatively tight lower bound for the p-center problem. In the second phase, they restrict the interval of the radius values with the lower bound obtained in the first phase and solve the integer programming (IP) version of the same feasibility problem iteratively to obtain the optimal value of the p-center problem.

Elloumi et al. (2004) proposed a new IP formulation for the p-center problem. This formulation utilizes the fact that the optimal value of the p-center problem is restricted to a finite set of distance values. They introduced additional binary variables z^k, $k = 2, \ldots, K$, with $z^k = 0$ if all demand points can be covered by p facilities within a radius value of r_{k-1} and $z^k = 1$ otherwise. The formulation is given below:

$$\text{Minimize} \quad r_1 + \sum_{k=2}^{K} (r_k - r_{k-1}) z^k \tag{4.9}$$

subject to (4.5), (4.6),

$$\sum_{i \in I} y_i \geq 1, \tag{4.10}$$

$$z^k + \sum_{i : d_{ji} < r_k} y_i \geq 1 \qquad \forall j \in J, k = 2, \ldots, K, \tag{4.11}$$

$$z^k \in \{0, 1\} \qquad k = 2, \ldots, K. \tag{4.12}$$

Constraint (4.10) eliminates the solutions with no open facility. Constraints (4.11) and the objective function (4.9) ensure that all demand points are served by a facility within the smallest possible distance.

A semi-relaxation of this formulation, which is obtained by removing the binary restriction on the y variables, provides the best known lower bound for the p-center problem. This lower bound can be obtained by solving a finite series of LP problems, which are the LP relaxations of the set covering problems. Elloumi et al. (2004) also provided an exact algorithm that combines the important properties of the algorithms of Minieka (1970) and Ilhan and Pınar (2001). Their algorithm uses the two-phase idea and a binary search strategy similar to the algorithm by Ilhan and Pınar (2001), but restricts the set of radius values to solve the set covering problems with the finite radius set R as in Minieka (1970).

Calik and Tansel (2013) developed new IP formulations and a new exact algorithm based on the decomposition of their models for solving the p-center problem. They associated a binary variable u_k with r_k, for each $k \in \{1, \ldots, K\}$. In particular, u_k is equal to 1 if r_k is selected as the optimal value and 0 otherwise. Initially, they proposed the following formulation:

$$\text{Minimize} \quad \sum_{k=1}^{K} r_k u_k \tag{4.13}$$

$$\text{subject to} \quad (4.5), (4.6),$$

$$\sum_{i:d_{ji} \le r_k} y_i \ge u_k \qquad \forall j \in J, k = 1, \ldots, K, \tag{4.14}$$

$$\sum_{k=1}^{K} u_k = 1, \tag{4.15}$$

$$u_k \in \{0, 1\} \qquad k = 1, \ldots, K. \tag{4.16}$$

Constraint (4.15) sets exactly one of the variables u_k to 1 and the corresponding r_k value is selected as the optimal value according to the objective function (4.13). Constraints (4.14) ensure that each customer is served within the selected radius by at least one facility. Constraints (4.16) are binary restrictions. The authors proposed a tightened formulation by using a relationship between their formulation and the formulation proposed by Elloumi et al. (2004). In this formulation, constraints (4.14) are replaced with constraints (4.17) given below:

$$\sum_{i:d(i,j) \le r_k} y_i \ge \sum_{q=1}^{k} u_q, \quad \forall j \in J, k = 1, \ldots, K. \tag{4.17}$$

The semi relaxations of these formulations, in which the binary restriction of the y-variables are removed, provide the tight lower bound obtained by Elloumi

et al. (2004). The algorithm developed by Calik and Tansel (2013) solves their formulations for restricted sets of radius values iteratively to converge to an optimal solution. They proposed several selection strategies for a two-element specialization of their algorithm. They also utilize the matrix reduction rules known for the set covering problem in their restricted formulations when solving large problems.

In the recent studies, instances from the OR-Library (Beasley 1990) and TSPLIB (Reinelt 1991) have been used for making computational experiments. The data for the uncapacitated p-median problem found in the OR-Library consists of 40 instances with $n = 100 - 900$ and $p = 5 - (n/3)$. This data was used in the experiments conducted by Ilhan and Pınar (2001), Elloumi et al. (2004), and Calik and Tansel (2013). In addition to these instances, Elloumi et al. (2004) used the instances u1060, rl1323 and u1817 ($n = 1060, 1323$, and 1817, respectively) and Calik and Tansel (2013) used the instances u1817, d15112, and pcb3038 ($n = 1817, 2500$, and 3038, respectively) from the TSPLIB.

4.4 Heuristics

Mladenović et al. (2003) introduced the first meta-heuristic approaches for finding approximate solutions to the p-center problem. They proposed a multistart local search algorithm (M-I), a chain substitution Tabu Search (TS) algorithm, and a variable neighborhood search (VNS) algorithm and conducted large scale experiments on 40 p-median instances from the OR-Library and instances with up to 3,038 nodes from TSPLIB. These experiments reveal that their algorithms outperform the algorithm proposed by Hochbaum and Shmoys (1985). Among the three heuristics proposed, TS and VNS algorithms outperform M-I algorithm, VNS performs the best on the average in terms of both the solution quality and solution time; however, TS provides slightly better results for the instances with smaller p values.

Pullan (2008) proposed a memetic genetic algorithm (PBS) for the vertex-restricted p-center problem, which combines a population based meta-heuristic with a local search algorithm. By using the phenotype crossover and directed mutation tools of the genetic algorithm, a wide range of elite starting solutions are generated and then, these solutions are improved to local optimality by using a local search algorithm. From the computational experiments using the instances previously tackled by Mladenović et al. (2003), an improvement in the CPU times and in the objective value of some problems is observed when PBS is compared with the VNS algorithm. The PBS algorithm can be executed also in a parallel processing mode. The experiments conducted by increasing the number of parallel processors utilized in the algorithm provide better CPU times.

Salhi and Al-Khedhairi (2010) obtained tight lower and upper bounds by using a three-level meta-heuristic and integrated these bounds with the algorithm by Daskin (2013) to solve the vertex-restricted p-center problem. In the first and second levels of the algorithm, a variable neighborhood strategy is utilized with distinct neighborhood structures. In the third level, a perturbation mechanism is introduced

to avoid sticking at local optima. The computational experiments conducted on the 40 p-median instances of the OR-Library revealed that the utilization of these bounds decreases the solution times of Daskin's algorithm.

Other than the meta-heuristic algorithms, Martinich (1988) proposed a vertex closing approach for the vertex-restricted p-center problem on complete networks with distance values that satisfy the triangle inequality. Initially, the algorithm places a facility on each node and considers the problem of finding $n - p$ facilities to close so that the maximum of the distances between the nodes and their facilities is minimized. In this study, the optimal solutions were characterized with the embedded sub-graphs of the original graph. From this analysis, initial lower and upper bounds were obtained, two polynomial time algorithms were proposed and procedures to verify the optimality of the solutions for several special cases were developed. In terms of the number of instances solved to optimality, they outperform the algorithm by Hochbaum and Shmoys (1985).

Bozkaya and Tansel (1998) showed that there exists a spanning tree of any connected network such that the optimal absolute p-center of this tree is optimal also for the network under consideration. They conducted experiments on two classes of spanning trees to observe how often these trees provide the optimal solution. They concluded that these two classes of spanning trees do not always include the optimizing tree, but they do in most of the instances.

Mihelič and Robič (2005) solved the vertex-restricted p-center problem by introducing a heuristic algorithm based on solving a finite series of minimum dominating set problems. Given a graph $G = (V, E)$, the minimum dominating set problem aims to find a node subset $S \subset V$ of minimum cardinality so that any node in $V \setminus S$ is adjacent to some node in S. They assumed that the underlying network is complete and the distance values satisfy the triangle inequality. The computational experiments performed on 40 standard test instances indicate that their algorithm performs much better than the other polynomial time heuristics found in the literature and competes with the best known non-polynomial time algorithms.

4.5 Variants

In this section, we briefly discuss some extensions of the p-center problem.

4.5.1 The Capacitated p-Center Problem

One first variant concerns problems with capacitated facilities. There are few studies on this variant. Bar-Ilan et al. (1993) introduced a ten-approximation algorithm for the special case of unit demands. The guarantee was improved to 6 by Khuller and

Sussmann (2000). If multiple centers can be located at the same location, then the guarantee is further improved to 5.

Jaeger and Goldberg (1994) proposed a polynomial time algorithm for the capacitated p-center problem when the graph is a tree, capacities are equal, and multiple facilities can be located at the same location. In this work, the demand of a node can be split among different facilities.

Ozsoy and Pinar (2006) proposed an exact algorithm to solve the capacitated p-center problem. The idea is to see if the all nodes can be assigned within a given distance and update lower and upper bounds on the radius using this information. In the subproblem solved to see whether it is possible to assign all nodes within a given distance, the objective is to minimize the number of facilities required.

In addition to the subproblem solved by Ozsoy and Pinar (2006) to obtain bounds on the optimal radius, Albareda-Sambola et al. (2010) proposed a second approach where they solved the problem of maximizing the demand covered within a given distance using at most p facilities. They used bounds from the Lagrangian relaxation of the two subproblems to eliminate some radius values and concluded that the first approach for finding the minimum number of required facilities is a better approach. Based on this conclusion, they proposed an exact algorithm using binary search over possible values of the optimal radius.

A very large-scale neighborhood heuristic was developed by Scapparra et al. (2004). Two types of exchanges were considered. In a cyclic exchange, one takes a sequence of nodes that are served by different facilities and replaces the facility of each node with the facility of the next node in the sequence (the facility of the last node in the sequence becomes the facility of the first node). In a path exchange, we again take a sequence of nodes served by different facilities and replace the facility of each node with the facility of the next node. The facility of the last node is replaced by a facility different from the facilities of the nodes in the sequence. A relocation step that moves the facilities to better locations with respect to the set of nodes they are serving is also added to the algorithm.

Three data sets were used in the last three papers mentioned. The first data set contains 20 instances of the capacitated p-median problem from the OR-Library (Beasley 1990), with 50 and 100 nodes. The second data set is from Lorena and Senne (2004) and is also for the capacitated p-median problem. Here there are six instances with the number of nodes ranging from 100 to 402. Finally, Scapparra et al. (2004) provided a data set with 8 instances containing 100 and 150 nodes. Additional instances of the p-median problem were used by Albareda-Sambola et al. (2010). These authors also compared their approach with the one of Ozsoy and Pinar (2006).

4.5.2 The Conditional p-Center Problem

The second variant is the conditional p-center problem. In this variant, there are q existing facilities and additional p facilities are to be located so that the

maximum distance between a node and its facility (among $p + q$ facilities) is minimized. Minieka (1980) introduced the conditional 1-center problem. Drezner (1989) showed that the conditional p-center problem can be solved by solving $O(logn)$ p-center problems. Suppose that the nodes are ranked in non-increasing order of their distances to their facilities (using the existing q facilities). Then there exists a node s such that the optimal value of the conditional p-center problem is equal to the maximum of the optimal value of the p-center problem solved for the first s nodes and the distance of the $s + 1$st node to its facility using the existing q facilities. The algorithm tries to find the best s using bisection.

Berman and Simchi-Levi (1990) solved the conditional p-center problem by solving a $p + 1$ center problem. They add a dummy demand node and a dummy possible location. The distance from a demand node to the dummy location is the distance of that node to its facility considering the existing facilities. The distance of the dummy demand node to the dummy location is zero and its distance to the other possible locations is a very large number. As a result, an optimal solution to the $p + 1$-center problem includes the dummy facility location and opens p other facilities. Berman and Drezner (2008) improved this approach and showed that the conditional p-center problem can be solved by solving a p-center problem where the distance between a node and a potential facility is set to the minimum of this distance and the distance between this node and the closest existing facility.

4.5.3 The Continuous p-Center Problem

The next variant is the continuous p-center problem. When demand points are continuously distributed over the whole graph, a set X_p of p points of the graph minimizing the largest distance from a demand point to a closest point of X_p is called a continuous p-center.

In the single facility case, i.e., when $p = 1$, the problem can still be solved by choosing a best solution among all the local continuous centers, i.e., solutions to continuous center problem in which the location is restricted to an edge. On an edge, the objective function is again piecewise linear with $O(|E|)$ breakpoints. Based on these facts, $O(|E|^2 log(|E|)$ algorithms were proposed by Hansen et al. (1991) and Tamir (1988).

On a tree, the absolute center coincides with the unweighted absolute center.

For the continuous p-center problem, Tamir (1987) identified a finite set of rational numbers containing the optimal solution value. Hence, a continuous p-center can be found by solving a finite number of continuous set covering problems, i.e., problems in which one looks for the smallest set of facilities needed to cover all points of the graph (vertices and interior points to edges) within a given maximum distance.

4.5.4 The p-Center Problem with Uncertain Parameters

Finally, we consider the variants with uncertain parameters. Averbakh and Berman (1997) studied the minmax regret version of the problem where the node weights are uncertain within given intervals. They showed that the robust version of the problem can be reduced to the resolution of $n + 1$ deterministic problems. Averbakh (1997) showed that the robust 1-center problem is strongly NP-hard on general networks when there is uncertainty in edge lengths. Averbakh and Berman (2000) developed polynomial time algorithms for the robust weighted 1-center problem with uncertainty in both node weights and edge lengths on a tree network.

4.6 Conclusions

We conclude this chapter with some future research directions. The majority of the solution methods proposed for the p-center problem are based on either the set covering or the dominating set problems. Well known optimization methods such as the cutting plane, branch-and-cut, Benders decomposition, or dynamic programming are rarely used. Recently, Calik (2013) provided a Benders decomposition method to solve the vertex restricted p-center problem and developed a branch-and-cut method for the capacitated p-center problem with multiple allocation. The experimental study conducted by Calik (2013) revealed that the utilization of a branch-and-cut method enables obtaining optimal solutions of large instances in small CPU time. The multiple allocation variant, which was previously studied by Jaeger and Goldberg (1994) on trees, is also an open research area for the capacitated p-center problem.

Although there are many studies for the p-center problem on trees, the capacitated version is not extensively investigated. The only study on this problem considers facilities with identical capacities and allows multi centers and multiple allocation. Hence investigating the capacitated p-center problem on tree networks with non-identical capacities, without multi centers and/or with single allocation might be a worthwhile undertaking.

Another variant of the p-center problem that has recently attracted the attention of the researchers is the fault tolerant p-center problem. This is a generalization of the p-center problem in which each customer is assigned to α different facilities. The idea is to make back-up services available in case of a failure of some facilities. The fault tolerance can also be taken into account for the capacitated p-center problem. Among the existing studies for the fault tolerant p-center and capacitated p-center problems, Krumke (1995), Khuller et al. (2000), and Chechik and Peleg (2012) focus on approximation algorithms and a recent study by Chen and Chen (2013) presents two exact algorithms. Therefore, developing different exact approaches and meta-heuristic algorithms for this problem might appeal to the researchers.

Acknowledgements The research of the second author is supported by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office and the research of the third author is supported by the Turkish Academy of Sciences.

References

Albareda-Sambola M, Díaz JA, Fernández E (2010) Lagrangean duals and exact solution to the capacitated p-center problem. Eur J Oper Res 201:71–81

Averbakh I (1997) On the complexity of a class of robust location problems. Working paper. Western Washington University, Bellingham

Averbakh I, Berman O (1997) Minimax regret p-center location on a network with demand uncertainty. Locat Sci 5:247–254

Averbakh I, Berman O (2000) Algorithms for the robust 1-center problem on a tree. Eur J Oper Res 123:292–302

Bar-Ilan J, Kortsarz G, Peleg D (1993) How to allocate network centers. J Algorithm 15:385–415

Beasley JE (1990) OR-Library: distributing test problems by electronic mail. J Oper Res Soc 41:1069–1072

Berge B (1967) Théorie des graphes et ses applications. Dunod, Paris

Berman O, Drezner Z (2008) A new formulation for the conditional p-median and p-center problems. Oper Res Lett 36:481–483

Berman O, Simchi-Levi D (1990) Conditional location problems on networks. Transp Sci 24:77–78

Bozkaya B, Tansel B (1998) A spanning tree approach to the absolute p-center problem. Locat Sci 6:83–107

Calik H (2013) Exact solution methodologies for the p-center problem under single and multiple allocation strategies. Ph.D. thesis, Bilkent University, Ankara

Calik H, Tansel BC (2013) Double bound method for solving the p-center location problem. Comput Oper Res 40:2991–2999

Chechik S, Peleg D (2012) The fault tolerant capacitated k-center problem. In: Structural information and communication complexity. Springer, Berlin/Heidelberg

Chen D, Chen R (2013) Optimal algorithms for the α-neighbor p-center problem. Eur J Oper Res 225:36–43

Daskin MS (2013) Network and discrete location: models, algorithms, and applications, 2nd edn. Wiley, Hoboken

Drezner Z (1989) Conditional p-center problems. Transp Sci 23:51–53

Dyer ME, Frieze AM (1985) A simple heuristic for the p-center problem. Oper Res Lett 3:285–288

Elloumi S, Labbé M, Pochet Y (2004) A new formulation and resolution method for the p-center problem. INFORMS J Comput 16:84–94

Garfinkel R, Neebe A, Rao M (1977) The m-center problem: minimax facility location. Manag Sci 23:1133–1142

Goldman AJ (1972) Minimax location of a facility in a network. Transp Sci 6:407–418

Hakimi SL (1964) Optimum locations of switching centers and the absolute centers and medians of a graph. Oper Res 12:450–459

Hakimi SL (1965) Optimum distribution of switching centers in a communication network and some related graph theoretic problems. Oper Res 13:462–475

Handler GY (1973) Minimax location of a facility in an undirected tree network. Transp Sci 7:287–293

Hansen P, Labbé M, Nicloas B (1991) The continuous center set of a network. Discrete Appl Math 30:181–195

Hochbaum DS, Shmoys DB (1985) A best possible heuristic for the k-center problem. Math Oper Res 10:180–184

Hsu W-L, Nemhauser GL (1979) Easy and hard bottleneck location problems. Discrete Appl Math 1:209–215

Ilhan T, Pınar MÇ (2001) An efficient exact algorithm for the vertex p-center problem. Technical report, Department of Industrial Engineering, Bilkent University. http://www.ie.bilkent.edu.tr/~mustafap/pubs

Jaeger M, Goldberg J (1994) A polynomial algorithm for the equal capacity p-center problem on trees. Transp Sci 28:167–175

Kariv O, Hakimi SL (1979) An algorithmic approach to network location problems. I: the p-centers. SIAM J Appl Math 37:513–538

Khuller S, Sussmann YJ (2000) The capacitated k-center problem. SIAM J Discrete Math 13:403–418

Khuller S, Pless R, Sussmann YJ (2000) Fault tolerant K-center problems. Theor Comput Sci 242:237–245

Krumke OS (1995) On a generalization of the p-center problem. Inform Process Lett 56:67–71

Lorena LAN, Senne ELF (2004) A column generation approach to capacitated p-median problems. Comput Oper Res 31:863–876

Martinich JS (1988) A vertex-closing approach to the p-center problem. Nav Res Log 35:185–201

Megiddo N (1983) Linear-time algorithms for linear programming in R^3 and related problems. SIAM J Comput 12:759–776

Mihelič J, Robič B (2005) Solving the k-center problem efficiently with a dominating set algorithm. J Comput Inform Technol 13:225–233

Minieka E (1970) The m-center problem. SIAM Rev 12:138–139

Minieka E (1980) Conditional centers and medians on a graph. Networks 10:265–272

Mladenović N, Labbé M, Hansen P (2003) Solving the p-center problem with tabu search and variable neighborhood search. Networks 42:48–64

Ozsoy FA, Pinar MC (2006) An exact algorithm for the capacitated vertex p-center problem. Comput Oper Res 33:1420–1436

Pullan W (2008) A memetic genetic algorithm for the vertex p-center problem. Evol Comput 16:417–436

Reinelt G (1991) TSPLIB - a traveling salesman problem library. ORSA J Comput 3:376–384

Salhi S, Al-Khedhairi A (2010) Integrating heuristic information into exact methods: the case of the vertex p-centre problem. J Oper Res Soc 61:1619–1631

Scapparra MP, Pallotino S, Scutella MG (2004) Large-scale local search heuristics for the capacitated vertex p-center problem. Networks 43:241–255

Tamir A (1987) On the solution value of the continuous p-center location problem on a graph. Math Oper Res 12:340–349

Tamir A (1988) Improved complexity bounds for center location problems on networks fy using dynamic data structures. SIAM J Discrete Math 1:377–396

Tansel BÇ (2011) Discrete center problems. In: Eiselt HA, Marianov V (eds) Foundations of location analysis. Springer, New York, pp 79–106

Chapter 5
Covering Location Problems

Sergio García and Alfredo Marín

Abstract When deciding where to locate facilities (e.g., emergency points where an ambulance will wait for a call) that provide a service, it happens quite often that a customer (e.g., a person) can receive this service only if he/she is under a certain distance to the closest facility (e.g., the ambulance can arrive in less than 7 min at this person's home). The problems that share this property receive the name of covering problems and have many applications (analysis of markets, archaeology, crew scheduling, emergency services, metallurgy, nature reserve selection, etc.). This chapter surveys the Set Covering Problem, the Maximal Covering Location Problem, and related problems and introduces a general model that has as particular cases the main covering location models. The main theoretical results in this topic as well as exact and heuristic algorithms are reviewed. A Lagrangian approach to solve the general model is detailed and, although the emphasis is on discrete models, some information on continuous covering is provided at the end of the chapter.

Keywords Covering • Discrete optimization • Location

5.1 Introduction

When deciding where to locate facilities (e.g., emergency points where an ambulance will wait for a call) that provide a service, it happens quite often that a customer (e.g., a person) can receive this service only if he/she is under a certain distance to the closest facility (e.g., the ambulance can arrive in less than 7 min at this person's home). The problems that have this property receive the name of *covering problems* and, when the previous condition holds, it is said that the customer is covered.

S. García (✉)
School of Mathematics, University of Edinburgh, Edinburgh, UK
e-mail: sergio.garcia-quiles@ed.ac.uk

A. Marín
Departamento de Estadística e Investigación Operativa, Universidad de Murcia, Murcia, Spain
e-mail: amarin@um.es

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_5

The first mentions to covering problems in literature can be found in Berge (1957) where the problem of finding a minimum cover on a graph is introduced and a theorem that provides an algorithm to find a minimum cover using a matching is stated and in Hakimi (1965) where it must be decided on the minimum number of police patrols required to protect a highway network. However, the problem was mathematically formulated for the first time in the Location area in Toregas et al. (1971), although out of a Location context it had already been formulated in Roth (1969).

In general, there are two types of covering problems: *set covering* and *maximal covering*. In a set covering problem (Toregas et al. 1971), the total cost of locating a set of facilities so that every customer is covered must be minimized. Particularly, if all the facilities have the same location cost, this is equivalent to minimize the total number of facilities to be located. A quick analysis of a solution to the set covering problem will usually show that with just a few facilities it is possible to cover an important percentage of the demand and that only by locating a high number full coverage can be achieved. Since locating as many facilities as needed may not be possible (e.g., due to budget constraints), a natural variant is to maximize the number of customers that are covered (or, equivalently, minimize the non-covered customers) by locating a fixed number of facilities. This problem is the maximal covering problem which was introduced in Church and ReVelle (1974).

According to Balas and Padberg (1976), the set covering problem is one of the three special structures in pure integer programming with the most wide-spread applications, together with set partitioning and the traveling salesman problem. Just to mention a few, set covering models have been applied in the following areas: analysis of markets (Storbeck 1988), archaeology (Bell and Church 1985), crew scheduling (Ceria et al. 1998), deployment of emergency services (Toregas et al. 1971; Eaton et al. 1986), mail advertising (Dwyer and Evans 1981), metallurgy (Vasko et al. 1989), nature reserve selection (Church et al. 1996) and Steiner matrices (Feo and Resende 1989).

Due to its importance and the rich literature on this topic, it is not surprising that reviews have been published regularly. The first one is Christofides and Korman (1975), a comparison of five computational methods for the set covering problem. Later, we have Chung (1986) which examines several applications of the maximal covering model to problems that do not belong to the Location field, and ReVelle (1989), a review focused on emergency service. Broader reviews are Schilling et al. (1993), an exhaustive survey on covering models in Location reviewing 96 papers, and Caprara et al. (2000), a comparison of recent algorithms (exact and heuristic) for the set covering problem. Plastria (2002) is an exhaustive review of continuous covering models and it is a perfect complement to this chapter. More recently, we have Berman et al. (2010) which considers some of the latest trends by reviewing gradual coverage, cooperative coverage, and variable radius coverage models, and Snyder (2011) which reviews the seminal covering models plus some extensions. Finally, the most recent survey is Farahani et al. (2012), an exhaustive list of models reviewing more than 150 papers that study covering problems in the area of facility

location. More focused as a detailed tutorial than as a proper survey, Daskin (1995) is an excellent introduction to the basic properties of covering models.

At this point, it must be said that there are many different models involving covering and that the goal of this chapter is not to cover them all but to provide an insight on the main models and results on the topic. Particularly, we focus on discrete models because they have received most of the attention in literature. The rest of this chapter is organized as follows: the main models from the literature are obtained in Sect. 5.2 as particular cases of a general model. Section 5.3 summarizes the main theoretical results on two of the main models (Set Covering and Maximal Covering Location). Then, we survey exact (Sect. 5.4) and heuristic (Sect. 5.5) solution methods. Since Lagrangian relaxation technique is widely used for approaching covering models, we extend it to the general model described in Sect. 5.6. Finally, although the focus of this chapter is on discrete models, some information on continuous covering is provided in Sect. 5.7 for the sake of completeness.

5.2 Models

We will use a general covering model to present as particular cases the main covering location problems in the literature as well as several other basic location problems which can be also considered sophisticated extensions of covering models.

Let $J = \{1, \ldots, n\}$ be the set of customers (also called demand points) and let $I = \{1, \ldots, m\}$ be the set of potential centers (facilities). Since many applications of covering models come from Location, we will use indistinctively "sites" for customers and potential centers. For each pair $(i, j) \in I \times J$, a known constant $a_{ij} \in \{0, 1\}$ represents whether demand point j can be covered (value one) or not (value zero) by a center installed at site i. These constants can be obtained with different procedures depending on the model under consideration as we will see below.

Associated to each $i \in I$, a fixed cost $f_i \geq 0$ has to be paid for opening a center at site i. In some models it is possible to open more than one center at the same site. In this case we assume that the cost of the centers to be opened in $i \in I$ is equal (i.e., f_i is the opening cost for all centers to be opened at site i). Each demand point $j \in J$ must be covered by at least $b_j \in \mathbb{Z}_0^+$ facilities, where $b_j = 0$ if site j does not need to be covered. Besides, a maximum number of $p \in \mathbb{Z}^+$ facilities can be opened (note that when the fixed costs of the centers are zero, this maximum number is always reached by some optimal solution).

Non-negative integer variables y_i represent the number of facilities to be opened at site $i \in I$. These are the main location variables and they will be explicitly present in all the particular cases that are obtained from the general model. The maximum number of facilities that can be opened at site i is given by the constant $e_i \in \mathbb{Z}^+$. Particularly, if $e_i = 1$, then y_i is a binary variable that takes value one if a facility is located at site i (and zero otherwise).

A second family of (binary) variables is w_{jk}. Here, j belongs to the set of demand points J while k belongs to an index set $K = \{1, \ldots, h\}$, whose meaning will depend on the particular model that is considered. Associated to variables w_{jk}, fixed costs $g_{jk} \in \mathbb{R}$ are given. These costs g_{jk} can be negative, representing in this case the profit from w_{jk} taking value one. In order to avoid unnecessary complicating constraints in the basic model, without loss of generality, we assume that $g_{j1} \leq g_{j2} \leq \ldots \leq g_{jh}$ for each $j \in J$. Whenever this condition does not hold, it will be explicitly stated.

The mathematical Integer Programming formulation for our general covering model is:

$$\text{(COV) Minimize} \quad \sum_{i \in I} f_i y_i + \sum_{j \in J} \sum_{k \in K} g_{jk} w_{jk} \tag{5.1}$$

$$\text{subject to} \quad \sum_{i \in I} y_i \leq p, \tag{5.2}$$

$$\sum_{i \in I} a_{ij} y_i = b_j + \sum_{k \in K} w_{jk} \qquad \forall j \in J, \tag{5.3}$$

$$y_i \in \{0, 1, \ldots, e_i\} \qquad \forall i \in I, \tag{5.4}$$

$$w_{jk} \in \{0, 1\} \qquad \forall j \in J, \forall k \in K. \tag{5.5}$$

The objective function (5.1) has two parts. The first sum returns the total fixed cost of opening y_i facilities at site $i \in I$. The second sum returns the total cost (or profit, if negative) provided by the w-variables that take value one. Constraint (5.2) limits the number of centers to p. Note that all the centers installed at the same site contribute to the sum.

The main constraints in the model are (5.3). For each demand point $j \in J$, the left-hand side of (5.3) measures the number of open facilities which are covering j. This number must be at least equal to the lower bound b_j on the right-hand side, while the sum of w_{jk} variables measures the slack in the coverage of j, i.e., the number of centers which are covering j besides the minimum number b_j. Due to the condition that we imposed on the g-values, the w-variables taking value one will be in the first positions, that is, constraints $w_{jk} \geq w_{j,k+1}$, $j \in J, k \in \{1, \ldots, h-1\}$ are satisfied without including them explicitly in the formulation. In such a way, a cost g_{j1} will be paid if demand point j is covered by at least $b_j + 1$ centers; additional cost g_{j2} will be paid if demand point j is covered by at least $b_j + 2$ centers and so on.

Constraints (5.4) are the integrality constraints for y-variables and impose that at most e_i centers can be installed at site i. Constraints (5.5) state that variables w are binary.

Therefore, model (COV) forces to cover each demand point j with a minimum of b_j facilities by using at most p facilities while minimizing the location cost of the facilities plus an additional cost (or, instead, minus an additional benefit) associated to the number of facilities which over-cover customers. By giving particular values

to the constants in (COV), different models from the literature (and, particularly, all the classical models) are obtained. The details are given next.

Set Covering Problem: In the Set Covering Problem (SCP) we have that, under the context of emergency center location of Toregas et al. (1971), $a_{ij} = 1$ if the response time or distance d_{ij} from a center located at $i \in I$ when an emergency happens at $j \in J$ is under a certain given threshold s (i.e., $a_{ij} = 1$ if and only if $d_{ij} \leq s$). There is no maximum number of centers to be located (i.e., $p = m$) and all demand points must be covered at least once ($b_j = 1 \; \forall j \in J$). The only costs in the objective function are $f_i = 1 \; \forall i \in I$ because the goal is just to minimize the number of open centers. Therefore, variables w_{jk} can be removed from the model by replacing the equalities in (5.3) by inequalities "$\geq$" (equivalently, take $h = m - 1$ and $g_{jk} = 0$ for all $j \in J, k \in K$ in (COV)). In the SCP, opening more than one facility at the same site is not optimal. Thus, $e_i = 1 \; \forall i \in I$. Given the special importance of this model, its classical formulation is explicitly shown:

$$\begin{aligned} \text{(SCP) Minimize} \quad & \sum_{i \in I} y_i & & \\ \text{subject to} \quad & \sum_{i \in I} a_{ij} y_i \geq 1 & \forall j \in J, & \qquad (5.6) \\ & y_i \in \{0, 1\} & \forall i \in I. & \end{aligned}$$

As an optimization problem, the SCP is a classical problem. The particular case where $I = J$ is the set of nodes of an undirected graph and $a_{ij} = 1$ if and only if edge (i, j) exists, usually called *Node Covering Problem*, has been deeply studied during the last century. The interested reader can consult the survey by Balinski (1965). Other interesting seminal papers are Norman and Rabin (1959) and Hohn (1955), where the mathematical problem is identified in the context of electronic circuits when analyzing a general way of designing a contact network satisfying given requirements and employing a minimum number of contacts. Surprisingly, although the SCP is an NP-complete problem (Garey and Johnson 1979), it happens often that the linear relaxation already provides an integer solution. Another important property that must be remarked is that the SCP has usually many different optimal solutions, i.e., sets of centers with the same minimum cardinality which cover all the demand points.

Weighted Set Covering Problem: The Weighted SCP (WSCP) is a generalization of the SCP where the opening costs f_i can be different from one.

Redundant Covering Location Problem: The Redundant Covering Location Problem (RCLP) was approached in Daskin and Stern (1981) as an extension of the SCP where the aim is to choose, among the optimal solutions to the SCP, the one which maximizes the number of demand points covered at least twice. Each site can only shelter one center. Again, $a_{ij} = 1$ if and only if $d_{ij} \leq s$, $p = m$, $b_j = 1 \; \forall j \in J$ (because the demand points must be covered at least

once), and $e_i = 1\ \forall i \in I$. Since we are also interested in knowing whether each demand point $j \in J$ is covered or not by a second center (disregarding the number of additional facilities which cover j), only variables w_{j1} would be necessary if equalities (5.3) were replaced by inequalities (5.6) as in the SCP discussed above. Alternatively, the RCLP can be obtained as a particular case of (COV) by taking $h = m - 1$, $g_{jk} = 0\ \forall j \in J, k \geq 2$, and $g_{j1} = -1\ \forall j \in J$. In order to prioritize the minimization of the number of open facilities, we define $f_i = n + 1\ \forall i \in I$ as a cost large enough.

Hierarchical Covering Location Problem (HCLP): An objective function which allows the simultaneous minimization of the number of facilities that are opened and the maximization of the number of previously existing facilities that are kept (within the minimum total number of facilities) was introduced in Plane and Hendrick (1977) in a paper devoted to the location of fire stations. Values a_{ij} are equal to one if and only if focal point i can be served by a pumper company at location j in less than the response time specified for site i. They found a major difficulty when using the SCP: this model does not differentiate between those sites that have existing fire stations and those that require the construction of a station. This drawback was fixed by modifying the objective function of the SCP as follows: consider a partition of the set of facilities $I = I_0 \cup I_1$, where I_0 is the set of existing facilities and I_1 is the set of potential new facilities. Then, define $f_i = 1\ \forall i \in I_1$ and $f_i = 1 - \varepsilon > 0\ \forall i \in I_0$ with ε a small positive amount. This way, the slightly lower cost of the already existing centers makes them more interesting when minimizing the total cost.

Maximal Covering Location Problem: The Maximal (or Maximum) Covering Location Problem (MCLP) was introduced in Church and ReVelle (1974) and, as it has been explained in the previous section, it entails an important change with regard to the goal of the previous models listed in this section because, since now the number of facilities to be located is limited to a given value $p < m$, we do not require to cover all the demand but to maximize the covered demand. Then, $h = p$ and $b_j = 0\ \forall j \in J$. Again, $e_i = 1\ \forall i \in I$ and values a_{ij} are defined as usual. Since we need to know whether a demand point is covered or not without minding about the number of different facilities that cover it, we avoid that variables y_i and variables w_{jk} with $k \neq 1$ contribute to the objective function (5.1) by fixing their corresponding coefficients to zero, i.e., $f_i = 0\ \forall i \in I$ and $g_{jk} = 0\ \forall j \in J, \forall k \geq 2$. Besides, we set $g_{j1} = -1$ in order to maximize the number of demand points covered by the open facilities.

An alternative to this model that was proposed in Church and ReVelle (1974) is to combine mandatory covering of some demand points (assume these points are indexed by means of $J_1 \subset J$) and maximization of the coverage of the remaining points (those in $J \setminus J_1$). This situation can also be approached by means of model (COV) by taking $h = p, b_j = 1\ \forall j \in J_1, b_j = 0\ \forall j \in J \setminus J_1$, $e_i = 1\ \forall i \in I$, and $f_i = 0\ \forall i \in I$. The g-coefficients are defined as follows: $g_{j1} = -1\ \forall j \in J \setminus J_1$, $g_{jk} = 0\ \forall j \in J \setminus J_1, \forall k \geq 2$, and $g_{jk} = 0\ \forall j \in J_1$, $\forall k \in K$. We call this model MCLP'.

Backup Set Covering Problems: Several models can be grouped under this name. The common idea is to cover the demand points with more than one facility in order to guarantee the coverage in case of either failure or overflow in one or some of the centers (in this sense, the RCLP can be considered a backup problem). There are two natural goals: minimization of the number of open facilities and maximization of the backup coverage. Sometimes this problem has been approached from the point of view of multiobjective optimization as, for example, in Storbeck and Vohra (1988) and model BACOP1 in Hogan and ReVelle (1986). Some other times, both objectives are combined into a unique function as in model BACOP2 in Hogan and ReVelle (1986). Details are provided next.

Coverage of all demand points is not mandatory, and each site can host several facilities. Demands t_j are associated to points $j \in J$. A maximum number of p facilities can be opened ($h = p$). Values a_{ij} are obtained as in most of the previous models. A parameter $0 < \beta < 1$ measures the relative importance of covering once or twice each demand point: the smaller β is, the more importance is given to cover each point twice. The goal here is to maximize the demand covered by the facilities and also the demand covered twice, using β to give each objective its relative importance. Taking this into account, we define $f_i = 0\ \forall i \in I$, $e_i = p\ \forall i \in I$, $g_{jk} = 0\ \forall j \in J, \forall k \geq 3$ and $b_j = 0\ \forall j \in J$. Variables w_{j1} are used to represent whether customer j is covered or not and variables w_{j2} are used to check whether j is covered twice or not. We define $g_{j1} = -\beta t_j$ and $g_{j2} = -(1-\beta)t_j$. Model (COV) is valid when $\beta \geq 1/2$. When $\beta < 1/2$, constraints $w_{j1} \geq w_{j2}\ \forall j \in J$ must be included to preserve the correct definition of the w-variables.

Batta and Mannur (1990) propose a different criterion for coverage which can also be viewed as a particular case of (COV). Recently, Curtin et al. (2010) developed a backup coverage model in order to locate police patrols, where a priority t_j of crime incident in $j \in J$ is known, the number of police patrols is limited to p and a_{ij} takes value one if, and only if, a patrol located at i can cover a crime incident located at j. The model is called PPAC and is a particular case of (COV) obtained by defining $f_i = 0\ \forall i \in I$, $h = p$, $g_{jk} = -t_j\ \forall k$, $b_j = 0\ \forall j \in J$, and $e_i = 1\ \forall i \in I$.

Maximum Expected Covering Location Problem: Several covering location models are based on probabilistic principles. One of the most important is the Maximum Expected Covering Location Problem (MECLP) (Daskin 1983), where each facility has a probability of $0 < q < 1$ of being busy or failing, independently of any circumstance of the system. Therefore, a demand point covered by ℓ facilities has a probability $1-q^{\ell}$ of receiving service. In this model, demands t_j associated to the demand points are also known, and the goal is to locate at most p facilities in such a way that the total expected demand (the sum of the demands of the points times their probability of being serviced) is maximized. Apart from PPAC, this is the first model considered here where all the w-variables really make sense, since it is necessary to know how many facilities are covering each demand point in a given feasible solution. When

variable w_{jk} takes value one, this can be then be re-interpreted as *demand point j is covered at least k times*. Thus, in order to obtain the right total in the objective function (5.1), we define $g_{jk} = -t_j(1-q)q^{k-1}\ \forall j \in J,\ \forall k \in K$. This way, we have that $\sum_{k=1}^{\ell} g_{jk} = -t_j(1-q^{\ell})$ which is the correct contribution of j to objective function when j is covered by ℓ facilities and $w_{jk} \leq w_{j,k+1}\ \forall k$. But this last inequality is satisfied implicitly because $q^k \geq q^{k+1}$ means that coefficients $\{g_{jk}\}_k$ are sorted in increasing order for every demand point j. Finally, we define $f_i = 0\ \forall i \in I$ and $b_j = 0\ \forall j \in J$. It is also natural in this problem to assume that a site can host more than one facility because it could lead to better solutions which is why we define $e_i = p\ \forall i \in I$.

Some of the strong assumptions of this model (e.g., servers are independent, servers have the same failure probabilities) have been relaxed several times in the literature. See, for example, Batta et al. (1989) and Galvão et al. (2005).

Probabilistic Location Set Covering Problem: In order to examine the relationships between the number of facilities being located and their reliability, ReVelle and Hogan (1989a) proposed a Probabilistic Location Set Covering Problem (PLSCP) whose main (and almost unique) difference with the SCP is that values b_j can be greater than one and they are obtained in such a way that the reliability of coverage of each point $j \in J$ is guaranteed to be at least equal to a threshold value α. Particularly, b_j is calculated as the minimum integer number such that

$$\left(\frac{F_j}{b_j}\right)^{b_j} \leq 1-\alpha,$$

where F_j is an average busy fraction associated with point j. Optionally, in this model e_i can take values greater than one since this can lead to better solutions.

Maximum Availability Location Problem: Suppose now that a profit u_j associated with each demand point $j \in J$ is obtained only if at least ℓ_j facilities cover it. The total number of facilities is limited, a site can host more than one facility and there is no facility opening cost. The Maximum Availability Location Problem (MALP), first described in ReVelle and Hogan (1989b), is a particular case of (COV) obtained by defining $f_i = 0\ \forall i \in I$, $e_i = p\ \forall i \in I$, $b_j = 0\ \forall j \in J$, and $g_{jk} = 0\ \forall j \in J, \forall k \neq \ell_j$, whereas $g_{j\ell_j} = -u_j\ \forall j \in J$. Since now the g-values are not sorted in increasing order, constraints $w_{jk} \geq w_{j,k+1}\ \forall j \in J,\ \forall k < h$, must be included.

Covering Problem: The so-called Covering Problem (CP) in Kolen and Tamir (1990) is that of minimizing the costs of opening some facilities plus the penalty costs associated to uncovered demand points. It is obtained from (COV) by defining $p = m$, $e_i = 1\ \forall i \in I$, $b_j = 0\ \forall j \in J$, $g_{jk} = 0\ \forall j \in J, \forall k \geq 2$ and $g_{j1} = -u_j\ \forall j \in J$ where u_j is the penalty for not covering demand point j.

A constant $-\sum_{j\in J} g_{j1}$ must be added to the objective to get the right optimal value. This way, when variable w_{j1} takes value one, j is covered and the penalty cost $-g_{j1}$ is removed from the objective function.

Minimum Cost Maximal Covering Problem (MCMCP): This is the name for the model introduced in Broin and Lowe (1986) whose only difference with regard to CP is that the total number of facilities is limited. They gave a dynamic programming algorithm for solving MCMCP in $O(p^2 n \min\{m^2, n^2\})$ time when the matrix $A = (a_{ij})$ is totally balanced.

p-Median Problem: Studied in detail in Chap. 2, the p-Median Problem (pMP) consists in, given a set of n demand points, choosing p of them to locate facilities and allocating each demand point to one of these facilities (which receive the name of medians) in such a way that the total cost be minimum, where the cost of allocating j to i is the distance d_{ij} between the two points (assuming $d_{ii} = 0\ \forall i$ and $d_{ij} > 0$ in all other cases).

Instead of using the classical formulation for pMP, an artificial set J can be designed in order to get it as a particular case of (COV): for each demand point j, a vector $D_j = (D_{1j}, \ldots, D_{G_j j})$ which is obtained by sorting in increasing order the values in $\{d_{1j}, \ldots, d_{nj}\}$ (removing multiplicities):

$$0 = D_{1j} < D_{2j} < \ldots < D_{G_j j} = \max_{1\le i\le n} \{d_{ij}\}.$$

Then define $J = \{(\ell, j) : j \in \{1, \ldots, n\},\ \ell \in \{2, \ldots, G_j\}\}$ and $a_{i,(\ell,j)} = 1$ if and only if $d_{ij} < D_{\ell j}$. Besides, we set $f_i = 0\ \forall i \in I$, $e_i = 1\ \forall i \in I$, $b_{(j,\ell)} = 0\ \forall (\ell, j) \in J$, and $h = p$. Coefficients $g_{(\ell,j)1}$ are defined with value $D_{\ell-1,j} - D_{\ell j}$ and $g_{(\ell,j)k} = 0\ \forall k \ge 2$.

With this approach, constraints (5.3) force variables $w_{(j,\ell)1}$ to take value zero if there is no open facility at a distance less than $D_{\ell j}$ from demand point j and the allocation cost of j is increased from $D_{\ell-1,j}$ to $D_{\ell j}$, as desired. A constant $\sum_{j=1}^{n} D_{G_j j}$ must be added to the objective function to get the right optimal value. This formulation has been very successfully used in García et al. (2011), where a column-and-row generation algorithm is developed to solve very large instances.

Uncapacitated Facility Location Problem: The problem considered in Chap. 3 (UFLP) and pMP differ in the number of centers which in UFLP is not fixed beforehand, but there is a fixed cost f_i for opening a facility at site i. Therefore, a straightforward modification of these parameters will allow to obtain UFLP as a particular case of (COV). This particular formulation was first proposed in Cornuéjols et al. (1980) and later in Kolen and Tamir (1990).

Table 5.1 summarizes the information about covering models in the literature which have been shown in this chapter to be particular cases of (COV).

Table 5.1 Covering location models derived from (COV)

Model	f	h	g	p	b	e
SCP	1	$m-1$	0	m	1	1
WSCP	f	$m-1$	0	m	1	1
RCLP	$n+1$	$m-1$	$(-1,0,\ldots,0)$	m	1	1
HCLP	$(1,\ldots,1,1-\varepsilon,\ldots,1-\varepsilon)$	$m-1$	0	m	1	1
MCLP	0	p	$(-1,0,\ldots,0)$	p	0	1
MCLP'	0	p	$(-1,\ldots,-1,0,\ldots,0)$	p	0	1
BACOP2[a]	0	p	$(-\beta t_j,-(1-\beta)t_j,0,\ldots,0)$	p	0	p
PPAC	0	p	$-t_j$	p	0	1
MECLP	0	p	$-t_j(1-q)q^{k-1}$	p	0	p
PLSCP	1	$m-b_j$	0	m	$\min\{b_j \in \mathbb{Z} \,/\, \left(\frac{F_j}{b_j}\right)^{b_j} \leq 1-\alpha\}$	1
MALP[b]	0	p	$(0,\ldots,0,-u_j,0,\ldots,0)$	p	0	p
CP	f	m	$(-u_j,0,\ldots,0)$	m	0	1
MCMCP	f	p	$(-u_j,0,\ldots,0)$	p	0	1
pMP	0	p	$(D_{\ell-1,j}-D_{\ell j},0,\ldots,0)$	p	0	1
UFLP	f	m	$(D_{\ell-1,j}-D_{\ell j},0,\ldots,0)$	m	0	1

[a] Constraint (5.5) must be added if $\beta < 1/2$
[b] Constraint (5.5) must be added

5.3 Theoretical Results

The Set Covering Problem is an NP-hard model (Garey and Johnson 1979). As a consequence, much effort has been put into understanding better the structure of this model in order to develop solving algorithms (which are reviewed later in this chapter). This knowledge can be divided mainly into three categories: preprocessing, relation with other problems, and polyhedral analysis.

When solving SCP, all the setup costs f_i can be assumed to be positive because if $f_i \leq 0$ for a certain facility i, then we can fix $y_i = 1$, remove this variable from the model and delete any inequality (5.6) that includes y_i. As explained in some early papers (Roth 1969; Lemke et al. 1971; Toregas and Revelle 1972, 1973), it is trivial that if a demand point j can be covered only by a certain facility i_1 (that is, $\{i \in I : a_{ij} = 1\} = \{i_1\}$), then we can fix $y_{i_1} = 1$. We have also some dominance rules: constraint (5.6) for a demand point j_1 can be removed if there is another demand point j_2 such that $\{i \in I : a_{ij_2} = 1\} \subset \{i \in I : a_{ij_1} = 1\}$, that is, if all the facilities covering demand point j_2 can cover also j_1. Similarly, a facility i_1 which covers a set of demand points which can be all covered by a cheaper facility i_2 will never be used: if $f_{i_1} \geq f_{i_2}$ and $\{j \in J : a_{i_1 j} = 1\} \subset \{j \in J : a_{i_2 j} = 1\}$, then we can fix $y_{i_1} = 0$. Sometimes, it is possible to use several facilities to cover all the demand points covered by another facility (Lorena and Lopes 1994): if we assume that the y-columns are sorted in increasing order in cost (with those columns with equal cost sorted in decreasing order in the number of rows that they cover), and we define $\beta_j = \min\{i \in I : a_{ij} = 1\}\ \forall j$ and $H_i = \cup_{j \in J}\{\beta_j : a_{ij} = 1\}\ \forall i$, then we can fix $y_i = 0$ if $\sum_{\ell \in H_i} f_\ell < f_i$. Applying these tests cyclically (i.e., not just once) can lead to substantial reductions in the size of the formulation.

The SCP formulation can be further improved by studying the polyhedral structure of its polytope. Balas (1980) uses disjunctions based on conditional bounds to obtain strong cuts in the form of cover constraints. Particularly, the inequalities introduced in Bellmore and Ratliff (1971) are generalized. Given an inequality of the form $\sum_{j \in J} \alpha_j y_j \geq \beta$, with $\alpha_j \in \{0, 1\}\ \forall j$ and β a positive integer, some necessary and sufficient conditions using the bipartite incidence graph of the matrix defining the SCP polytope are given in Cornuéjols and Sassano (1989) for this inequality to be a facet. Sassano (1989) studies the properties of this polytope and presents two sequential lifting procedures to obtain valid inequalities and facets. Particularly, it is shown that the SCP polytope is full dimensional if and only if every demand point can be covered by at least two different facilities. It is also characterized when an inequality of the form $\sum_{i \in J_0} y_i \geq 1$ with $J_0 \subset J$ is a facet. When the polytope is full-dimensional, then the trivial inequality $y_j \leq 1$ is shown to be always a facet, and the trivial inequality $y_j \geq 0$ is a facet if and only if every demand point can be covered by at least two different facilities different from j. Some deeper results on facets and lifting can be found in Nobili and Sassano (1989). Balas and Ng (1989a) characterize facet-defining inequalities for the SCP polytope with right-hand side 2 and coefficients 0, 1 or 2. In Balas and Ng (1989b) it is shown that each of these facets can be obtained using a lifting procedure from an inequality with only three

non-zero coefficients that is valid in a lower dimensional polytope. Sánchez-García et al. (1998) do a similar study for the case of facets with coefficients in $\{0, 1, 2, 3\}$ and right-hand side equal to 3.

The connection of SCP to other classical problems has also been studied in the literature. Balas and Padberg (1976) show how to turn a set partitioning problem into a set covering. In Krarup and Pruzan (1983) it is discussed how SCP can be transformed into a set packing, set partitioning or simple plant location problem. Reciprocal results are given to turn a set partitioning or simple plant location problem into a set covering problem.

Less theoretical results can be found for the Maximal Covering Location Problem, which is known to be NP-hard (Megiddo et al. 1983). In the literature, MCLP has been formulated using other classical models. For example, Church and ReVelle (1976) show the equivalence between MCLP and a certain p-median problem where the distances in this second problem are defined as

$$d'_{ij} = \begin{cases} 0, & \text{if } d_{ij} \leq s, \\ 1, & \text{if } d_{ij} > s, \end{cases}$$

with d_{ij} the distances from the original problem and s is the maximum distance that a demand point can be from the facility that covers it. Another different reformulation is given in Klastorin (1979) where the problem is formulated as a generalized assignment problem by adding some artificial variables.

The Maximal Expected Coverage Location Problem and the Backup Coverage Location Problem are shown in Church and Weaver (1986) to be special cases of the vector assignment p-median problem. Techniques developed for this latter model are used to solve instances of the first two problems. The Capacitated Set Covering Problem and the Capacitated Maximal Covering Location Problem are formulated in Current and Storbeck (1988) as a capacitated plant location problem and a capacitated p-median problem, respectively.

Several technical results on covering problems with special emphasis on trees and matrices in standard greedy form can be found in Kolen and Tamir (1990).

5.4 Solution Methods

The first exact algorithms for the Set Covering Problem were almost purely enumerative: Lemke et al. (1971) develop a branch-and-bound method that exploits the structure of the SCP formulation and solutions. Later, Etcheberry (1977) uses a branch-and-bound strategy where the branching is done on constraints and not on variables. The lower bounds of the tree are calculated using Lagrangian relaxation instead of the simplex method.

Using cutting planes from conditional bounds, the algorithm proposed in Balas (1980) is exploited in Balas and Ho (1980). This method uses two sets of heuristics:

one to find good upper bounds (primal heuristics) and another to obtain lower bounds and cutting planes (dual heuristics). Subgradient optimization is applied to find better lower bounds. This last technique is also used in Beasley (1987) where a branch-and-bound method is proposed whose main elements are a dual ascent procedure and subgradient optimization. This algorithm is improved in Beasley and Jørnsten (1992) by incorporating the heuristic published in Beasley (1990) along with some other enhancements.

Of special interest is Neebe (1988) which solves the problem of calculating for every possible maximum distance the minimum number of facilities that cover all the nodes (instead of solving the set covering problem for a single maximum distance). This approach uses a chain of linear programming relaxations and, after every linear model, some tests are used to obtain an integer solution. Although these tests do not guarantee that an optimal integer solution will be found, the author claims to solve to optimality almost all the instances he considers (up to 100 nodes). Each of the auxiliary problems is solved with a modification of the procedure suggested in Lemke et al. (1971).

Fisher and Kedia (1990) propose an algorithm for a model which includes both set covering and set partitioning constraints. It is an exact branch-and-bound algorithm that uses greedy and 3-opt heuristics applied to the dual problem. Exploiting the use of bounds, Mannino and Sassano (1995) propose a lower bounding procedure and a branch-and-bound scheme to solve set covering problems that appear in Steiner triple systems (a certain matrix structure). Balas and Carrera (1996) develop a procedure applied to a Lagrangian dual problem at each node that combines subgradient optimization with primal and dual heuristics which tighten the upper and lower bounds. These strengthened bounds allow to fix some variables. In general, Lagrangian methods are the most extended and effective methods in the literature. More recently, Avella et al. (2009) propose a cutting plane algorithm where the separation algorithm is solved in an exact way on a subproblem defined by a subset of the original constraints and variables of the set covering problem formulation.

On the contrary, not many exact algorithms have been developed for the Maximal Covering Location Problem. Downs and Camm (1996) obtain a primal solution using the greedy heuristic of Church and ReVelle (1974). They use complementary slackness conditions for the maximal covering problem formulation to obtain a dual feasible solution. This solution is the starting vector of multipliers for the Lagrangian dual problem of MCLP which is solved with subgradient optimization. If an integer solution is not obtained, branch-and-bound is used.

5.5 Approximate Solution Methods

As it happens with any hard optimization problem, there are more heuristic algorithms than exact methods in the literature. Roth (1969), the first paper to formulate the Set Covering Problem, already proposes a probabilistic heuristic. A

random initial solution is selected and then refined using a set of predefined rules based on the concept of λ-optimal cover. This procedure is repeated many times with the hope of finding a good solution. Chvátal (1979) proposes a basic greedy heuristic that selects iteratively the facility with the largest number of nodes covered per unit cost. A bound is established for the worst value of the solution provided by the heuristic. Feo and Resende (1989) develop a probabilistic heuristic for set covering problems arising in Steiner triple systems. It is a non-deterministic variation of a previous deterministic heuristic where randomization is introduced to escape from local minima.

Many more different metaheuristic techniques have been used to approach SCP: surrogate relaxation (Lorena and Lopes 1994), simulated annealing (Jacobs and Brusco 1995; Brusco et al. 1999), genetic algorithms (Al-Sultan et al. 1996; Beasley and Chu 1996). However, as with the exact case, subgradient methods are the most effective. Ceria et al. (1998) use a primal-dual subgradient Lagrangian algorithm to provide information for a later greedy heuristic to decide which variables to fix to one. Caprara et al. (1999) use variable pricing to update the subset of columns that define a core problem in their subgradient optimization heuristic. This is a difference with respect to Ceria et al. (1998) where the core set is not modified. They also improve the way in which the step-size and ascent direction definitions are usually done in subgradient optimization in order to speed up convergence.

For the Maximal Covering Location Problem and similar problems, we can find several heuristics. Already in Church and ReVelle (1974) where the problem is introduced, a greedy heuristic is provided. Later, Daskin (1983) describes a heuristic for the Maximum Expected Covering Location Problem which finds good solutions for all values of q (the probability of a facility not working). It starts with all the facilities located at the node that covers the maximum demand and then considers single node substitutions. For each of the new solutions, it is analyzed if there is an interval where the current best solution is improved. By iterating this procedure, interval [0,1] is partitioned and a heuristic solution is given for each of the resulting subintervals. In MCLP, Galvão and ReVelle (1996) develop a Lagrangian heuristic that uses a vertex interchange heuristic to improve upper bounds. In Galvão et al. (2000), heuristics based on Lagrangian and surrogate relaxations are compared. Here, the relaxed surrogate problem is a binary knapsack problem whose linear relaxation is solved in the heuristic. The authors show that, when the initial set of multipliers is obtained using a dual descent procedure, the performance of the two methods is similar.

Eaton et al. (1986) deal with a hierarchical covering problem where sites with multiple cover are maximized while the number of vehicles is minimized in an application to ambulance deployment in Santo Domingo. Although they proposed two formulations, no solver was available at that moment in the Ministry of Health of Dominican Republic and they then developed a heuristic that minimizes the number of facilities, maximizes multiple coverage and minimizes response time. In their algorithm, they create a cover matrix, then order coverage zones in a list and remove dominated sites iteratively.

A further reason for using heuristics is that aggregation is used to reduce the size of the problem so that larger size instances can be tackled. Daskin et al. (1989) study the effect of node aggregation for MCLP. Three aggregation schemes are tested based on relative demands on the disaggregate nodes, distances between the disaggregate nodes and a mix of both. The first and the third methods are shown to perform much better than the second. In Current and Schilling (1990) three rules are proposed to reduce the aggregation error in SCP and MCLP.

5.6 Lagrangian Relaxation

Among the many different methods that have been developed in the literature for covering models, we highlight here Lagrangian Relaxation (LR) for several reasons. First, LR can be used as a heuristic method but can additionally provide good lower bounds which can be embedded into a branch-and-bound framework to develop an exact method. Second, as shown in Sects. 5.4 and 5.5, LR has been widely used in covering problems. Third, it can be designed for the general model (COV) and then used on any particular case without loss of accuracy. And, finally, LR usually produces very good results in a reasonable amount of computational time. Readers not familiarized with this technique are referred to Guignard (2003).

In what follows, we apply LR to model (COV) by making the natural choice of relaxing constraints (5.3). Since the non-relaxed linear constraints (5.2) and $y_i \leq e_i \ \forall i \in I$ give rise to a totally unimodular coefficients matrix, lower bounds produced by means of LR will not be greater than lower bounds produced by the usual linear relaxation. A Lagrangian multiplier $v_j \in \mathbb{R}$ associated to each constraint in (5.3), unrestricted in sign, will be used. So, a family of Lagrangian relaxed subproblems is obtained with objective functions

$$\sum_{i\in I} f_i y_i + \sum_{j\in J}\sum_{k\in K} g_{jk} w_{jk} + \sum_{j\in J} v_j \left(\sum_{i\in I} a_{ij} y_i - b_j - \sum_{k\in K} w_{jk} \right) =$$

$$\sum_{i\in I} \left(f_i + \sum_{j\in J} v_j a_{ij} \right) y_i + \sum_{j\in J}\sum_{k\in K} \left(g_{jk} - v_j \right) w_{jk} - \sum_{j\in J} v_j b_j .$$

By solving

$$\begin{array}{ll} \text{(COVLR}(v)\text{)} & \min \sum_{i\in I} \left(f_i + \sum_{j\in J} v_j a_{ij} \right) y_i + \sum_{j\in J}\sum_{k\in K} \left(g_{jk} - v_j \right) w_{jk} \\ & \text{s.t.} \qquad\qquad (5.2), (5.4), (5.5), \end{array}$$

and then adding constant $-\sum_{j\in J} v_j b_j$, we will get a lower bound on the objective value of (COV) when the set of multipliers is $v = (v_1, \ldots, v_n)$.

Let now $(y^*(v), w^*(v))$ be an optimal solution to (COVLR(v)). Problem (COVLR(v)) splits into

$$\begin{array}{ll} \text{(COVLRy}(v)) & \min \sum_{i\in I}\left(f_i + \sum_{j\in J} v_j a_{ij}\right) y_i \\ & \text{s.t.} \qquad (5.2), (5.4), \end{array}$$

and

$$\begin{array}{ll} \text{(COVLRw}(v)) & \min \sum_{j\in J}\sum_{k\in K}\left(g_{jk} - v_j\right) w_{jk} \\ & \text{s.t.} \qquad (5.5). \end{array}$$

(COVLRw(v)) can be easily solved by inspection:

$$w^*_{jk}(v) = 1 \;\Leftrightarrow g_{jk} \le v_j \quad \forall j \in J, \forall k \in K.$$

If, as in most of the models that we considered, g_{jk}-values are sorted in increasing order for each $j \in J$, and assuming that $v_j \in (g_{j\ell_j}, g_{j,\ell_j+1}]$, then the optimal solution to (COVLRw(v)) will look like as follows:

$$w^*_{j1}(v) = \ldots = w^*_{j\ell_j}(v) = 1,\ w^*_{j,\ell_j+1}(v) = \ldots = w^*_{jh}(v) = 0.$$

The corresponding optimal value will be $v(\text{COVLRw}(v)) = \sum_{j\in J}(\sum_{k=1}^{\ell_j} g_{jk} - \ell_j v_j)$.

Regarding (COVLRy(v)), we define $f'_i := f_i + \sum_{j\in J} v_j a_{ij}\ \forall i \in I$ and we sort these values in increasing order:

$$f'_{(1)} \le \ldots \le f'_{(t)} \le 0 \le f'_{(t+1)} \le \ldots \le f'_{(n)}.$$

An optimal solution to (COVLRy(v)) is recursively obtained by taking

$$y^*_{(i)}(v) = \begin{cases} e_{(i)} & \text{if } \sum_{\ell=1}^{i-1} y^*_{(\ell)}(v) \le p - e_{(i)}, \\ p - \sum_{\ell=1}^{i-1} y^*_{(\ell)}(v) & \text{if } \sum_{\ell=1}^{i-1} y^*_{(\ell)}(v) > p - e_{(i)}, \end{cases}$$

$i = 1, \ldots t$, and $y^*_{(i)}(v) = 0,\ i = t+1, \ldots, n$. Assuming that $\sum_{\ell=1}^{i'} e_{(\ell)} \le p < \sum_{\ell=1}^{i'+1} e_{(\ell)}$, with $i' \le t$, we then have that

$$\text{v(COVLRy}(v)) = \sum_{i=1}^{i'-1} e_{(i)}\left(f_{(i)} + \sum_{j\in J} v_j a_{(i)j}\right)$$

$$+\left(p-\sum_{i=1}^{i'} e_{(i)}\right)\left(f_{(i')}+\sum_{j\in J} v_j a_{(i')j}\right).$$

A suitable set of Lagrangian multipliers v must be chosen so that v(COVLR(v)) provides a good lower bound on the optimal value of (COV). This can be achieved by means of ascent procedures which iteratively modify v, like subgradient algorithms or tailored dual ascent algorithms. Good feasible solutions (and the corresponding upper bounds) can be generated from good sets of multipliers as follows. Consider any optimal solution to the relaxed problem given by $(y^*(v), w^*(v))$. We relax the notation by calling simply y^* the optimal values of the y-variables. Once these have been determined, the best values which the w-variables can take are obtained by solving for each $j \in J$ the subproblem

$$\begin{aligned}
(\mathrm{COV})_j \text{ Minimize } & \sum_{k\in K} g_{jk} w_{jk} \\
\text{subject to } & \sum_{k\in K} w_{jk} = \sum_{i\in I} a_{ij} y_i^* - b_j, \\
& w_{jk} \in \{0,1\} \qquad \forall k \in K.
\end{aligned}$$

If $\sum_{i\in I} a_{ij} y_i^* - b_j < 0$, the subproblem is infeasible. Otherwise, assuming that $\sum_{i\in I} a_{ij} y_i^* - b_j \leq h$ (note that in general h is taken large enough) and sorting g-values in increasing order, the optimal solution to $(\mathrm{COV})_j$ can be obtained just by making the first $\sum_{i\in I} a_{ij} y_i^* - b_j$ w-variables equal to one, that is,

$$\mathrm{v(COV)}_j = \sum_{k=1}^{\sum_{i\in I} a_{ij} y_i^* - b_j} g_{jk}.$$

5.7 Continuous Covering Location Problems

When speaking about continuous covering, it means that the set of candidates where facilities can be located is not discrete but a whole (continuous) space. Because of the nature of these problems, most of them are in the plane or, if height/depth is relevant, in the 3D-space. Besides, most of the applications locate one single facility because these models are already difficult enough.

Analogous to the discrete Set Covering Problem, the continuous Minimal Covering Circle Problem (MCCP) consists in finding the smallest circle in the plane that contains all the points of a given set which need to be covered. The center of this circle is the optimal site. This is a very old problem which according to Plastria (2002) was studied in the nineteenth century, but may have been introduced even earlier. One of the main properties of the solution to MCCP is that there are always

at least two demand points on the border of the minimal circle. Although several algorithms to solve this problem have been proposed over time, the best known is the method published in Elzinga and Hearn (1972) for the case of Euclidean distances.

When the radius of the circle is fixed, it may be not large enough to cover all the demand points and, as in the discrete Maximal Covering Location Problem, the objective is now to cover as much demand as possible. These maximal covering problems have usually multiple solutions, maybe even a region of optimal solutions, and this region may not even be convex (see Plastria 2002). However, it can be proved that there is an optimal solution which is either a demand point or an intersection point of two circles centered at demand points (see Drezner 1981 and Chazelle and Lee 1986 for details on algorithms). There is a similar property when the facilities can be located on any part of a network (Church and Meadows 1979). Church (1984) shows an analogous property for planar maximal covering problems with Euclidean or rectilinear distances.

More recently, Drezner et al. (2004) studied a gradual covering problem with Euclidean distances where a finite set of points needs to be covered with one single facility. If the facility can be located anywhere on the plane and the total cost of non-covered points is minimized, then the solution is in the convex hull of the demand points.

5.8 Conclusions

In this chapter we have provided an overview on covering problems with a special emphasis on discrete models. Instead of providing a list of the many covering models that can be found in the literature, we have focused on detailing those that are considered to be more relevant because of the attention received in the literature in the last decades. Moreover, we show that many of the models discussed in this review can be seen as particular cases of a general covering model that we introduce here. As far as we know, this is the first attempt to develop such a unified approach for the study of set covering problems.

Having set covering problems received so much attention in the literature, it seems that the number of theoretical results is too small. These results reduce basically to some preprocessing rules and to the study of some facets. And none of them has been used to develop an algorithm that can be considered to be a major breakthrough in the area. Therefore, future research should try to make better use of these results or obtain new theoretical properties for these problems. Particularly, developing exact methods for covering models that are not the SCP seems highly desirable.

Acknowledgements Research supported by Fundación Séneca, project 08716/PI/08, ERDF funds and Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (I+D+I), project MTM2012-36163-C06-04.

References

Al-Sultan KS, Hussain MF, Nizami JS (1996) A genetic algorithm for the set covering problem. J Oper Res Soc 47:702–709

Avella P, Boccia M, Vasilyev I (2009) Computational experience with general cutting planes for the set covering problem. Oper Res Lett 37:16–20

Balas E (1980) Cutting planes from conditional bounds: a new approach to set covering. Math Program 12:19–36

Balas E, Carrera MC (1996) A dynamic subgradient-based branch-and-bound procedure for set covering. Oper Res 44:875–890

Balas E, Ho A (1980) Set covering algorithms using cutting planes, heuristics and subgradient optimization: a computational study. Math Program 12:37–60

Balas E, Ng SM (1989a) On the set covering polytope: I. All the facets with coefficients in $\{0, 1, 2\}$. Math Program 43:57–69

Balas E, Ng SM (1989b) On the set covering polytope: II. Lifting the facets with coefficients in $\{0, 1, 2\}$. Math Program 45:1–20

Balas E, Padberg MW (1976) Set partitioning: a survey. SIAM Rev 18:710–760

Balinski ML (1965) Integer programming: methods, uses, computations. Manag Sci 12:253–313

Batta R, Mannur NR (1990) Covering-location models for emergency situations that require multiple response units. Manag Sci 36:16–23

Batta R, Dolan JM, Krishnamurthy NN (1989) The maximal expected covering location problem: revisited. Transp Sci 23:277–287

Beasley JE (1987) An algorithm for the set covering problem. Eur J Oper Res 31:85–93

Beasley JE (1990) A Lagrangian heuristic for set-covering problems. Nav Res Log 37:151–164

Beasley JE, Chu PC (1996) A genetic algorithm for the set covering problem. Eur J Oper Res 94:392–404

Beasley JE, Jørnsten K (1992) Enhancing an algorithm for set covering problems. Eur J Oper Res 58:293–300

Bell T, Church RL (1985) Location-allocation theory in archaeological settlement pattern research: some preliminary applications. World Archaeol 16:354–371

Bellmore M, Ratliff HD (1971) Set covering and involutory bases. Manag Sci 18:194–206

Berge C (1957) Two theorems in graph theory. Proc Natl Acad Sci USA 43:842–844

Berman O, Drezner Z, Krass D (2010) Generalized coverage: new developments in covering location models. Comput Oper Res 37:1675–1687

Broin MW, Lowe TJ (1986) A dynamic programming algorithm for covering problems with (greedy) totally balanced constraint matrices. SIAM J Algebr Discrete Method 7:348–357

Brusco MJ, Jacobs LW, Thompson GM (1999) A morphing procedure to supplement a simulated annealing heuristic for cost- and coverage-correlated set-covering problems. Ann Oper Res 86:611–627

Caprara A, Fischetti M, Toth P (1999) A heuristic method for the set covering problem. Oper Res 47:730–743

Caprara A, Toth P, Fischetti M (2000) Algorithms for the set covering problem. Ann Oper Res 98:353–371

Ceria S, Nobili P, Sassano A (1998) A Lagrangian-based heuristic for large-scale set covering problems. Math Program 81:215–228

Chazelle BM, Lee DT (1986) On a circle placement problem. Computing 36:1–16

Christofides N, Korman S (1975) A computational survey of methods for the set covering problem. Manag Sci 21:591–599

Chung C (1986) Recent applications of the maximal covering location planning (M.C.L.P.) model. J Oper Res Soc 37:735–746

Church RL (1984) The planar maximal covering location problem. J Reg Sci 24:185–201

Church RL, Meadows ME (1979) Location modeling utilizing maximum service distance criteria. Geogr Anal 11:358–373

Church RL, ReVelle C (1974) The maximal covering location problem. Pap Reg Sci Assoc 32:101–118
Church RL, ReVelle C (1976). Theoretical and computational links between the *p*-median, location set-covering, and the maximal covering location problem. Geogr Anal 8:406–415
Church RL, Weaver JR (1986) Theoretical links between median and coverage location problems. Ann Oper Res 6:1–19
Church RL, Stoms DM, Davis FW (1996) Reserve selection as a maximal covering location problem. Biol Conserv 76:105–112
Chvátal V (1979) A greedy heuristic for the set-covering problem. Math Oper Res 4:233–235
Cornuéjols G, Sassano A (1989). On the 0,1 facets of the set covering polytope. Math Program 43:45–55
Cornuéjols G, Nemhauser GL, Wolsey LA (1980) A canonical representation of simple plant location problems and its applications. SIAM J Algebr Discrete Method 1:261–272
Current JR, Schilling DA (1990) Analysis of errors due to demand data aggregation in the set covering and maximal covering location problems. Geogr Anal 22:116–126
Current JR, Storbeck JE (1988) Capacitated covering problems. Environ Plan B 15:153–163
Curtin KM, Hayslett-McCall K, Qiu F (2010) Determining optimal police patrol areas with maximal covering and backup covering location models. Netw Span Econ 10:125–145
Daskin MS (1983) A maximum expected covering location model: formulation, properties and heuristic solution. Transp Sci 17:48–70
Daskin MS (1995) Covering problems. In: Networks and discrete location. Models, algorithms and applications. Wiley, New York, pp 92–153
Daskin MS, Stern EH (1981) A hierarchical objective set covering model for emergency medical service vehicle deployment. Transp Sci 15:137–152
Daskin MS, Haghani AE, Khanal M, Malandraki C (1989) Aggregation effects in maximum covering problems. Ann Oper Res 18:115–140
Downs BT, Camm JD (1996) An exact algorithm for the maximal covering problem. Nav Res Log 43:435–461
Drezner Z (1981) On a modified one-center problem. Manag Sci 27:848–851
Drezner Z, Wesolowsky GO, Drezner T (2004) The gradual covering problem. Nav Res Log 51:841–855
Dwyer FP, Evans JR (1981) A branch and bound algorithm for the list selection problem in direct mail advertising. Manag Sci 27:658–667
Eaton DJ, Sánchez HML, Lantigua RR, Morgan J (1986) Determining ambulance deployment in Santo Domingo, Dominican Republic. J Oper Res Soc 37:113–126
Elzinga D, Hearn D (1972) Geometric solutions for some minimax location problems. Transp Sci 6:379–394
Etcheberry J (1977) The set-covering problem: a new implicit enumeration algorithm. Oper Res 25:760–772
Farahani RZ, Asgari N, Heidari N, Hosseininia M, Goh M (2012) Covering problems in facility location: a review. Comput Ind Eng 62:368–407
Feo TA, Resende MGC (1989) A probabilistic heuristic for a computationally difficult set covering problem. Oper Res Lett 8:67–71
Fisher ML, Kedia P (1990) Optimal solutions of set covering/partitioning problems using dual heuristics. Manag Sci 36:674–688
Galvão RD, ReVelle C (1996) A Lagrangean heuristic for the maximal covering location problem. Eur J Oper Res 88:114–123
Galvão RD, Espejo LGA, Boffey B (2000) A comparison of Lagrangean and surrogate relaxations for the maximal covering location problem. Eur J Oper Res 124:377–389
Galvão RD, Chiyoshia FY, Morabito R (2005) Towards unified formulations and extensions of two classical probabilistic location models. Comput Oper Res 32:15–33
García S, Labbé M, Marín A (2011) Solving large p-median problems with a radius formulation. INFORMS J Comput 23:546–556

Garey MR, Johnson DS (1979) Computers and intractibility: a guide to the theory of NP-completeness. WH Freeman and Co., San Francisco
Guignard M (2003) Lagrangean relaxation. TOP 11:151–200
Hakimi SL (1965) Optimum distribution of switching centers in a communication network and some related graph theoretic problems. Oper Res 13:462–475
Hogan K, ReVelle C (1986) Concepts and applications of backup coverage. Manag Sci 32:1434–1444
Hohn F (1955) Mathematical aspects of switching. Am Math Mon 62:75–90
Jacobs LW, Brusco MJ (1995) Note: a local-search heuristic for large set-covering problems. Nav Res Log 42:1129–1140
Klastorin TD (1979) On the maximal covering location problem and the generalized assignment problem. Manag Sci 25:107–112
Kolen A, Tamir A (1990) Covering problems. In: Mirchandani PB, Francis RL (eds) Discrete location theory. Wiley, New York, pp 263–304
Krarup J, Pruzan PM (1983) The simple plant location problem: survey and synthesis. Eur J Oper Res 12:36–81
Lemke CE, Salkin HM, Spielberg K (1971) Set covering by single branch enumeration with linear programming subproblems. Oper Res 19:998–1022
Lorena LAN, Lopes FB (1994) A surrogate heuristic for set covering problems. Eur J Oper Res 79:138–150
Mannino C, Sassano A (1995) Solving hard set covering problems. Oper Res Lett 18:1–5
Megiddo N, Zemel E, Hakimi SL (1983) The maximum coverage location problem. SIAM J Algebr Discrete Method 4:253–261
Neebe AW (1988) A procedure for locating emergency-service facilities for all possible response distances. J Oper Res Soc 39:743–748
Nobili P, Sassano A (1989) Facets and lifting procedures for the set covering polytope. Math Program 45:111–137
Norman RZ, Rabin MO (1959) An algorithm for a minimum cover of a graph. Proc Am Math Soc 10:315–319
Plane DR, Hendrick TE (1977) Mathematical programming and the location of fire companies for the Denver fire. Oper Res 25:563–578
Plastria F (2002) Continuous covering location problems. In: Hamacher HW, Drezner Z (eds) Facility location: applications and theory. Springer, New York, pp 37–79
ReVelle C (1989) Review, extension and prediction in emergency service siting models. Eur J Oper Res 40:58–69
ReVelle C, Hogan K (1989a) The maximum reliability location problem and α-reliable p-center problem: derivatives of the probabilistic location set covering problem. Ann Oper Res 18:155–174
ReVelle C, Hogan K (1989b) The maximum availability location problem. Transp Sci 23:192–200
Roth R (1969) Computer solutions to minimum-cover problems. Oper Res 17:455–465
Sánchez-García M, Sobrón MI, Vitoriano B (1998) On the set covering polytope: facets with coefficients in $\{0, 1, 2, 3\}$. Ann Oper Res 81:343–356
Sassano A (1989) On the facial structure of the set covering polytope. Math Program 44:181–202
Schilling DA, Jayaraman V, Barkhi R (1993) A review of covering problems in facility location. Locat Sci 1:25–55
Snyder LV (2011) Covering problems. In: Eiselt HA, Marianov V (eds) Foundations of location analysis. Springer, Berlin, pp 109–135
Storbeck JE (1988) The spatial structuring of central places. Geogr Anal 20:93–110
Storbeck JE, Vohra RV (1988) A simple trade-off model for maximal and multiple coverage. Geogr Anal 20:220–230
Toregas C, Revelle C (1972) Optimal location under time or distance constraints. Pap Reg Sci Assoc 28:133–144

Toregas C, Revelle C (1973) Binary logic solutions to a class of location problem. Geogr Anal 5:145–175
Toregas C, Swain A, ReVelle C, Bergman L (1971) The location of emergency service facilities. Oper Res 19:1363–1373
Vasko FJ, Wolf FE, Stott KL (1989) A set covering approach to metallurgical grade assignment. Eur J Oper Res 38:27–34

Chapter 6
Anti-covering Problems

Emilio Carrizosa and Boglárka G.-Tóth

Abstract In covering location models, one seeks the location of facilities optimizing the weight of individuals covered, i.e., those at the distance from the facilities below a threshold value. Attractive facilities are wished to be close to the individuals, and thus the covering is to be maximized, while for repulsive facilities the covering is to be minimized. On top of such individual-facility interactions, facility-facility interactions are relevant, since they may repel each other. This chapter is focused on models for locating facilities using covering criteria, taking into account that facilities are repulsive from each other. Contrary to the usual approach, in which individuals are assumed to be concentrated at a finite set of points, we assume the individuals to be continuously distributed in a planar region. The problem is formulated as a global optimization problem, and a branch and bound algorithm is proposed.

Keywords Big square small square • Covering problems • Global optimization • Regional demand • Repulsive facilities

6.1 Introduction

Locational Analysis addresses decision problems involving the location of facilities which interact with a set of individuals, and, eventually interact among them. For *attractive* facilities, such as schools, libraries, emergency services or supermarkets, individuals wish the facilities to be as close as possible to them. Such *pull* models (facilities are pulled towards demand) do not properly model *repulsive* facility location problems (Alonso et al. 1998; Carrizosa and Plastria 1998; Erkut and Neuman 1989; Fliege 2001; Plastria and Carrizosa 1999), like, for instance, the location of a polluting plant, wished to be as far as possible from the individuals.

E. Carrizosa (✉)
Universidad de Sevilla, Instituto de Matemáticas de la Universidad de Sevilla, Sevilla, Spain
e-mail: ecarrizosa@us.es

B.G.-Tóth
Faculty of Mathematics, Department of Differential Equations, Budapest University of Technology and Economics, 1111 Budapest, Egry József u. 1., Hungary
e-mail: bog@math.bme.hu

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_6

For such undesirable facilities, a *push* model, pushing facilities away from the sites affected by facilities nearness, is more suitable: the location for the facilities is then sought maximizing a certain *non-increasing* function of the distances from the individuals to the facilities. For both desirable and undesirable facilities, interactions may be measured as a function of the individual-facility distance (or time), or, as studied here, via *coverage*; see e.g. Kolen and Tamir (1990), Li et al. (2011), Murray et al. (2009), Schilling et al. (1993) for extensive reviews on covering models and solution approaches. It is important to stress here that, independently of the nature of the facility, either attractive or repulsive, the very same models for covering function apply (Farhan and Murray 2006), the difference being algorithmic: such covering is to be maximized for desirable facilities and minimized for undesirable facilities.

On top of individual-facility interactions, facility-facility interactions are also likely to be relevant. Such interactions may be critical when facilities are obnoxious, and risk or damage to population scales nonlinearly (e.g., with hazardous materials deposits or dangerous plants which may suffer chain reactions) and thus negative impacts are to be dispersed. Facility-facility interactions are also important in models for locating facilities which, although they are perceived as attractive by the users, they are perceived as repelling by other facilities competing for the very same market. In these models, locating the facilities far away from each other avoids cannibalization and optimizes competitive market advantage (Christaller 1966; Curtin and Church 2006; Lei and Church 2013).

Although the models described are general, the algorithmic approach presented here is restricted to the *planar* case (Drezner and Wesolowsky 1994; Plastria 2002; Plastria and Carrizosa 1999): facilities are identified with points in the plane, and interact with the remaining facilities and with individuals, also identified with points in the plane. Interactions are measured via distances in the plane. See Plastria (1992) for an excellent review of planar distances and planar location models and e.g. Berman et al. (1996), Berman and Huang (2008), Berman and Wang (2011) for covering models for which interactions are not measured via planar distances, but network distances instead, typically shortest-path distances.

Contrary to most papers in the literature, affected individuals are not assumed here to be concentrated at a finite number of points, and, instead, an arbitrary distribution (in particular, a continuous distribution) on their location is given. This way we can directly address models in which affected individuals are densely spread on a region, but we also address models in which uncertainties exist about the exact location of the individuals, due to their mobility (Carrizosa et al. 1998b).

Regional models are not so common in the location literature, since, even when individuals are assumed to be continuously distributed, a discretization process is usually done, and such continuous distribution is replaced by a discrete one, by e.g. replacing all points in each district by its centroid, or other central point, see e.g. Francis and Lowe (2011), Francis et al. (2008, 2000, 2002), Murray and O'Kelly (2002), Plastria (2001), Tong and Church (2012). Nevertheless, discretization is well known not to perform well in applications, this issue being especially relevant in covering models, since significant discrepancies may exist between what is modeled as covered and what is actually covered, see e.g. Current and Schilling (1990),

Daskin et al. (1989), Kim and Murray (2008), Murray (2005), Murray and Wei (2013), Tong (2012), Tong and Murray (2009). For this reason, some papers are found in which the regional aspect is directly handled. See for instance Blanquero and Carrizosa (2013), Carrizosa et al. (1995, 1998c), Fekete et al. (2005), Yao and Murray (2014) for single-facility Weber problems with regional demand, Murat et al. (2010) for a heuristic method for the extension to p facilities, and Tong (2012), Tong and Murray (2009) for discrete covering problems, in which the individuals are identified with objects (polygons) in the plane, which can be considered as fully or partially covered.

The remainder of the chapter is structured as follows. In Sect. 6.2, a rather general p-facility covering model for continuously distributed demand is described; how to address the optimization problem is presented in Sect. 6.3, and illustrated in Sect. 6.4. Conclusions and future lines of research are outlined in Sect. 6.5.

6.2 Regional Covering Model

Location models are specific in the way the interactions are modeled. Two types of interactions take place, namely, individual-facility interactions and facility-facility interactions. Depending on the specific problem, just one or the two types of interactions may be relevant; see e.g. Erkut and Neuman (1989).

Since these two types of interactions have different nature, they are discussed separately in what follows.

6.2.1 Individual-Facility Interactions

For a given individual location a and any facility location x, let $c(a, x) \in [0, 1]$ denote how much a is covered (affected) by the facility at x. In its general form, $c(\cdot, \cdot)$ may be any function $\varphi : \mathbb{R}^+ \longrightarrow [0, 1]$, which is non-increasing in the (Euclidean) distance $\|x - a\|$ separating a and x,

$$c(a, x) = \varphi(\|x - a\|), \tag{6.1}$$

so that, the lower the distance, the higher the coverage. This assumption, yet sensible, may not be sound for specific problems of locating undesirable facilities; for instance, Karkazis and Papadimitriou (1992) addresses the problem of locating a polluting plant whose pollutant is discharged by means of high stacks, and thus maximal interaction (damage) takes place at a non-negligible distance of the facility.

We remark that we are using the Euclidean distance, but this is not the only choice of distance function $\|\cdot\|$ found in the literature in covering models: see e.g. Fernández et al. (2000) for a proposal of (weighted) ℓ_p norms and Plastria (2002) for a thorough discussion on planar distances.

The basic form of φ is an all-or-nothing function, already suggested in Church and ReVelle (1974), see also e.g. Drezner and Wesolowsky (1994),

$$c(a,x) = \varphi(\|x-a\|) = \begin{cases} 1, \text{ if } \|x-a\| \leq R \\ 0, \text{ otherwise,} \end{cases} \tag{6.2}$$

where the threshold value R is called the *range* (Christaller 1966) or *coverage standard*. For an attractive facility, R represents the highest distance a user is willing to overcome to utilize a facility, whereas for undesirable facilities, R represents the distance of the boundary of the zone within which the facility would have a negative impact (Farhan and Murray 2006). Extensions of (6.2) abound in the literature, leading to so-called *gradual covering* models (Berman et al. 2009c, 2003; Drezner et al. 2004). For instance the all-or-nothing function above is replaced by a piecewise constant function modeling different levels of coverage in Berman and Krass (2002), by a piecewise linear function in Berman et al. (2003), Berman and Wang (2011), Drezner et al. (2004), or by more general nonlinear functions, such as the logistic model

$$c(a,x) = \varphi(\|x-a\|) = \frac{1}{1+exp(\alpha_a + \beta_a \|x-a\|)}, \tag{6.3}$$

in Fernández et al. (2000), see also Berman et al. (2010, 2003), Karasakal and Karasakal (2004). Observe that in some of the papers cited above the coverage functions c are introduced for attractive facilities, and thus maximization, instead of minimization, is pursued. However, the models for c are the very same.

Expressions above for c, as (6.2), are adequate just for the single-facility case. When several facilities are to be located, the covering model (6.1) can be extended in several ways, by first defining, for each facility $i = 1, 2, \ldots, p$, the function φ_i converting distances into coverage. In the simplest and most popular model in the literature, for a p-tuple of facility locations $\mathbf{x} = (x_1, \ldots, x_p)$, covering c of an individual location a by $\mathbf{x}$ is given by

$$c(a,\mathbf{x}) = \max_{1 \leq i \leq p} c_i(a, x_i). \tag{6.4}$$

In the particular form of individual covering c_i given by (6.2) using φ_i instead of φ and R_i instead of R, one considers the individual location a to be covered by the p-tuple of facility locations $\mathbf{x} = (x_1, \ldots, x_p)$ if it is covered by at least one of the p facilities, i.e., if at least one facility i is at a distance smaller than its threshold value R_i.

Multifacility covering functions other than (6.4) can be found in the literature, see Berman et al. (2010) for an updated review. One may consider fuzzy operators

to aggregate the covering functions c_i, yielding, for example, the proposal of Hwang et al. (2004),

$$c(a, \mathbf{x}) = 1 - \prod_{1 \le i \le p} (1 - c_i(a, x_i)), \tag{6.5}$$

which, if each c_i has the form (6.2) is identical to (6.4). Alternatively, realizing that the max operator used in (6.4) is nothing but taking one of the ordered values of $c_i(a, x_i)$, further extensions are natural:

$$c(a, \mathbf{x}) = \max_{(\lambda_1, \ldots, \lambda_p) \in \Lambda} \sum_{i=1}^{p} \lambda_i c_i(a, x_i) \tag{6.6}$$

for a given Λ. Taking as Λ the set

$$\Lambda = \left\{ (\lambda_1, \ldots, \lambda_p) : \sum_{i=1}^{p} \lambda_i = 1, \ \lambda_i \ge 0 \quad \forall i \right\},$$

one recovers (6.4); taking

$$\Lambda = \left\{ (\lambda_1, \ldots, \lambda_p) : \sum_{i=1}^{p} \lambda_i = 1, \ \frac{1}{r} \ge \lambda_i \ge 0 \quad \forall i \right\},$$

for some integer $r \in \{1, 2, \ldots, p\}$, one obtains as coverage the weighted sum of the r highest covers. These covering models belong to the class of so-called ordered covering models (Berman et al. 2009c), in which a weighted sum of the ordered values of the covers are considered.

Another class of models is given by the so-called cooperative cover model, discussed in Berman et al. (2009a):

$$c(a, \mathbf{x}) = \begin{cases} 1, \text{ if } \sum_{i=1}^{p} \lambda_i c_i(a, x_i) \ge \tau \\ 0, \text{ otherwise} \end{cases} \tag{6.7}$$

for some positive fixed scalars λ_i and threshold value τ. Assuming that each facility covering function c_i follows the all-or-nothing model (6.2), model (6.7) means that we may consider an individual to be covered if the weighted sum of 1-facility covers yields a value above a threshold limit τ.

Summing up, the different proposals in the literature can be considered as particular cases of a general model of the form

$$c(a, \mathbf{x}) = \Psi\left(c_1(a, x_1), c_2(a, x_2), \ldots, c_p(a, x_p)\right), \tag{6.8}$$

where Ψ should take values in $[0, 1]$ and should be componentwise non-decreasing, so that the higher each individual-facility cover, the higher the cover of individual location a by the p facilities.

So far we have modeled the interaction between an affected individual at a and the facilities at $\mathbf{x} = (x_1, \ldots, x_p)$. Now we address the problem of defining a global individuals-facilities covering measure $C(\mathbf{x})$.

If the main concern is how much the highest coverage is, a worst-case performance measure is suitable:

$$C(\mathbf{x}) = \sup_{a \in A} c(a, \mathbf{x}). \tag{6.9}$$

Under (6.9) as criterion, searching locations $\mathbf{x}$ for the facilities such that $C(\mathbf{x}) \leq \alpha$ means that no individual at all suffers a coverage of more than α.

The (safe) worst-case approach (6.9) may be unfeasible for densely populated regions, and, instead of searching locations not affecting individuals, the *average* coverage may be a suitable choice. Formally, assume that affected individuals are distributed along the plane, following a distribution given by a probability measure μ on a set $A \subset \mathbb{R}^2$, and the individuals-facilities coverages are aggregated into one single measure, namely, the *expected coverage*, given by

$$C(\mathbf{x}) = \int_A c(a, \mathbf{x})\, d\mu(a). \tag{6.10}$$

Assuming, as in (6.10), an arbitrary probability measure μ for the distribution of affected individual locations gives us full freedom to accommodate different important models. Obviously, for a finite set A of affected individual locations, $A = \{a_1, \ldots, a_n\}$, denoting $\mu_a = \mu(\{a\})$, we recover the basic covering model,

$$C(\mathbf{x}) = \sum_{a \in A} \mu_a c(a, \mathbf{x}), \tag{6.11}$$

in which the covering is given by the weighted sum of the covers of the different points a. However, we can consider absolutely continuous distributions, in which μ has associated a probability density function f in the plane, and now (6.10) becomes

$$C(\mathbf{x}) = \int_A c(a, \mathbf{x})\, f(a)\, da. \tag{6.12}$$

Several types of density functions f are worthy to be considered. One can take, for instance, f as the uniform density on a region $A \subset \mathbb{R}^2$ (a polygon, a disc), and thus f is given as

$$f(a) = \begin{cases} \frac{1}{ar(A)}, & \text{if } a \in A \\ 0, & \text{otherwise,} \end{cases} \tag{6.13}$$

where $ar(A)$ denotes the area of the region A; assuming a uniform density of individuals along the full region A under study seems to be rather unrealistic; instead, one may better split the region A into smaller and more homogeneous subregions A_j (e.g. polygons), give a weight ω_j to each A_j, and assume a uniform distribution f_j for each A_j :

$$f(a) = \sum_{j=1}^{r} \omega_j f_j(a), \tag{6.14}$$

where each f_j is uniform on A_j, and thus its expression is given in (6.13).

Let us particularize (6.14) for the all-or-nothing case in which the covering function is given by (6.4), and each c_i is given by (6.2), i.e., $c(a, \mathbf{x})$ takes the value 1 if at least one facility i is at a distance from a below the threshold R_i, and takes the value 0 otherwise. Then, for any $\mathbf{x}$, $C(\mathbf{x})$ takes the form

$$\begin{aligned} C(\mathbf{x}) &= \int c(a, \mathbf{x}) f(a) da \\ &= \sum_{j=1}^{r} \omega_j \frac{1}{ar(A_j)} \int_{A_j} c(a, \mathbf{x}) da \\ &= \sum_{j=1}^{r} \omega_j \frac{1}{ar(A_j)} ar(A_j \cap \cup_{i=1}^{r} B_i(x_i)), \end{aligned} \tag{6.15}$$

where, for each $i = 1, \ldots, p$, $B_i(x_i)$ gives the set of points covered by facility i, i.e., the disc centered at x_i and radius R_i. Hence, the problem is reduced to calculating areas of intersections of discs $B_i(x_i)$ with the subregions A_j. Such calculation, although cumbersome in general, are supported in GIS, see Kim and Murray (2008), Murray et al. (2009), Tong and Murray (2009).

Needless to say, the density f does not need to be piecewise constant, and one can take, for instance, a mixture of bivariate gaussians, $f(a) = \sum_{j=1}^{r} \omega_j f_j(a)$, where each f_j is a bivariate gaussian density centered at some u_j and with covariance matrix S_j,

$$f_j(a) = \frac{1}{2\pi \sqrt{|S_j|}} e^{-\frac{1}{2}(a-u_j)^\top S_j^{-1}(a-u_j)}, \tag{6.16}$$

or, more generally, a radial basis function (RBF) density,

$$f_j(a) = g_j(\|a - u_j\|) \tag{6.17}$$

for some decreasing function g_j, so that the density is the highest at some knot point u_j and decreasing in all directions.

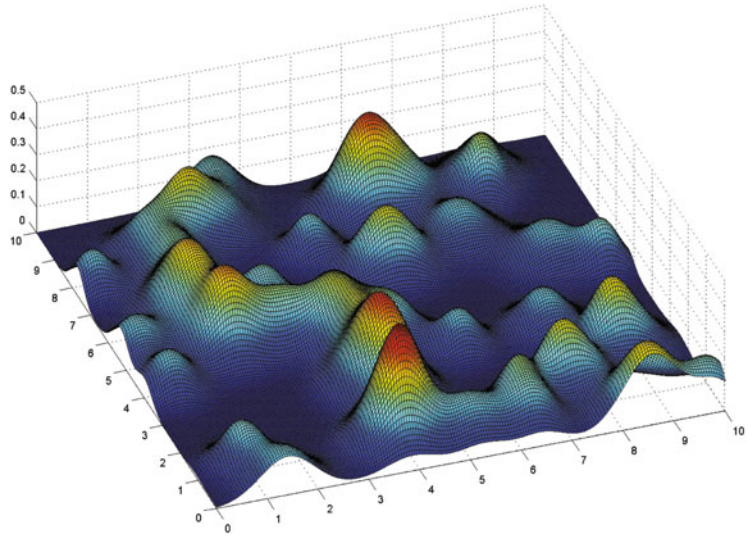


Fig. 6.1 Pdf of a mixture of 50 bivariate gaussians

A model like (6.16), or in general (6.17), may be rather promising when the only information provided for the region is just a set $u_1, \ldots, u_r$ of points, aggregating the actual coordinates of affected individuals around, and then a *kernel density estimation* process (Bowman and Foster 1993; Wand and Jones 1993, 1995) is done. For instance, Fig. 6.1 represents the probability density function (pdf) of the form (6.16) with 50 knots.

6.2.2 Facility-Facility Interactions

The facility-facility interactions may be defined similarly. As in (6.1), the effect caused by facility at x_i on facility at x_j is measured by the scalar $c_{ij}^F(x_i, x_j)$,

$$c_{ij}^F(x_i, x_j) = \varphi_{ij}^F(\|x_i - x_j\|) \tag{6.18}$$

for some non-increasing function φ_{ij}^F. All pairwise facility-facility effects are aggregated into one single facility-facility interactions measure $C^F(\mathbf{x})$, which, similarly to (6.8), is assumed to take the form

$$C^F(\mathbf{x}) = \Psi^F\left((c_{ij}^F(x_i, x_j))_{i \neq j}\right)$$

for some componentwise non-decreasing Ψ^F. The simplest case is given by

$$\Psi^F\left((c_{ij}^F(x_i,x_j))_{i\neq j}\right) = \max_{i\neq j} c_{ij}^F(x_i,x_j), \tag{6.19}$$

and thus $C^F(\mathbf{x})$ is calculated as the highest facility-facility interaction, i.e., the one of the closest pairs of facilities. Hence, under (6.19),

$$\begin{aligned} C^F(\mathbf{x}) &\leq \delta \text{ if and only if} \\ c_{ij}^F(x_i,x_j) &\leq \delta \ \ \forall i,j,\ i\neq j, \text{ if and only if} \\ \|x_i - x_j\| &\geq (\varphi_{ij}^F)^{-1}(\delta) \ \ \forall i,j,\ i\neq j. \end{aligned}$$

Assuming all c_{ij}^F in (6.18) are modeled by means of the same φ_{ij}^F function, $\varphi_{ij}^F = \varphi^F$, we have

$$C^F(\mathbf{x}) \leq \delta \quad \text{if and only if} \quad \min_{\substack{i,j \\ i\neq j}} \|x_i - x_j\| \geq \gamma, \tag{6.20}$$

with $\gamma = \left(\varphi^F\right)^{-1}(\delta)$. See Lei and Church (2013) for a discussion and extension of (6.19) to so-called partial-sum criteria.

6.2.3 The Anti-covering Model

Depending on the specific problem under consideration, either one or the two covering criteria C, C^F are to be optimized. Pure repulsion among facilities naturally leads to a dispersion criterion (Erkut and Neuman 1991; Kuby 1987; Lei and Church 2013). By (6.20), minimizing C^F amounts to maximizing the minimal distance among facilities. This criterion alone yields a simple geometrical interpretation: a set of p non-overlapping circles (the location of the facilities) is sought so that their (common) radius is maximized (Mladenović et al. 2005).

When both C and C^F are relevant, one naturally faces a biobjective optimization problem in which both C and C^F are to be minimized,

$$\min_{\mathbf{x}\in\mathscr{S}} \left(C(\mathbf{x}), C^F(\mathbf{x})\right), \tag{6.21}$$

where $\mathscr{S} \subset (\mathbb{R}^2)^p$ is the feasible region, which is assumed to be a compact subset, and thus embedded in a box. Sensible examples for $\mathscr{S}$ may be $\mathscr{S} = S^p$, where S is a polygon in the plane, or $\mathscr{S} = \{\xi_1\} \times \{\xi_2\} \times \ldots \times \{\xi_k\} \times S^{p-k}$, where S is a polygon in the plane, and $\xi_1, \ldots, \xi_k$ are fixed points in the plane, corresponding to facilities already located.

One can address the problem of finding (an approximation to) the set of Pareto-optimal solutions to (6.21), as done for other problems in Blanquero and Carrizosa (2002), Romero-Morales et al. (1997). Alternatively, one can consider one of the criteria as constraint, and address instead the problem of minimizing the covering $C(\mathbf{x})$ keeping the facility-facility cover $C^F(\mathbf{x})$ below a threshold limit δ :

$$\begin{array}{ll} \text{minimize} & C(\mathbf{x}) \\ \text{subject to} & C^F(\mathbf{x}) \leq \delta \\ & \mathbf{x} \in \mathscr{S}. \end{array} \tag{6.22}$$

Assuming for C^F the model given by (6.18), problem (6.22) amounts to finding p points $x_1, \ldots, x_p$ so that they are at a distance at least $\left(\varphi^F\right)^{-1}(\delta)$ from each other and the covering C is minimized. This is the approach proposed e.g. in Berman and Huang (2008), in which undesirable facilities are located (on a network) so as no facilities are allowed to be closer than a pre-specified distance.

6.3 Computational Approach

While nowadays computational tools allow one to address *discrete* p-facility problems with a very large p, e.g. Avella and Boccia (2007), Avella et al. (2006), nonconvex continuous location problems, as those addressed here, can only be solved exactly for a very small number of facilities to be located. The most popular and most effective technique is a geometric branch and bound, which can already be found under the name of Big Square Small Square (BSSS) (Hansen et al. 1985), and later modified by a number of authors (Blanquero and Carrizosa 2008; Drezner and Suzuki 2004; Plastria 1992; Schöbel and Scholz 2010), coining names such as BTST (Big Triangle Small Triangle) or Big Cube Small Cube. See Drezner (2012) for a recent review of such variants. In our case the search space is the set of p rectangles for the p facilities, that gives a multi-dimensional interval, also called a box. The main steps of the branch and bound are as usual: a list of boxes is handled, each box being associated with a subproblem, namely, the covering location problem in which facilities are to be located within such box; at each step one box is selected from the list and divided into smaller boxes. Bounds on the optimum over the subboxes are calculated, so that boxes which are found not to contain the global optimum are removed, while the rest is saved for further processing. The branching and bounding rules are iterated until the gap between the underestimation and underestimation of the optimal value is smaller than the prescribed accuracy.

In our implementation, selection of the next box is done by the smallest lower bound, and the division rule is defined by halving both sides of the largest rectangle into four equal sized rectangles. An upper bound on the minimum is calculated evaluating the objective function at the midpoint of the selected box. In what follows, a bounding procedure, valid for arbitrary pdfs, is discussed.

A branch and bound can only be used as soon as increasingly tight bounds are built for $C(\mathbf{x})$ on a box $\mathbf{X} = (X_1, \ldots, X_p)$. Each X_i is a rectangle $X_i = ([a_i, b_i], [c_i, d_i])$ where the i-th facility is allowed to be located. One has then on a given box $\mathbf{X}$

$$\min_{\mathbf{x}\in\mathbf{X}} C(\mathbf{x}) = \min_{\mathbf{x}\in\mathbf{X}} \int_A c(a,\mathbf{x})d\mu(a) \geq \int_A \min_{\mathbf{x}\in\mathbf{X}} c(a,\mathbf{x})d\mu(a).$$

For the general function $c(a,\mathbf{x}) = \Psi(c_1(a,x_1), c_2(a,x_2), \ldots, c_p(a,x_p))$, as in (6.8), with Ψ non-decreasing function of $c_i(a,x_i)\ \forall i$, it can be derived further to

$$\int_A \min_{\mathbf{x}\in\mathbf{X}} c(a,\mathbf{x})d\mu(a) = \int_A \Psi\left(\min_{x_1\in X_1} c_1(a,x_1), \ldots, \min_{x_p\in X_p} c_p(a,x_p)\right) d\mu(a)$$
$$= \int_A \Psi\left(\min_{x_1\in X_1} \varphi_1(\|a-x_1\|), \ldots, \min_{x_p\in X_p} \varphi_p(\|a-x_p\|)\right) d\mu(a),$$

where, as in (6.1), $c_i(a,x_i) = \varphi_i(\|a-x_i\|)$ for a non-increasing function φ_i of the distance for all i. This leads to

$$\min_{\mathbf{x}\in\mathbf{X}} C(\mathbf{x}) \geq \int_A \Psi\left(\varphi_1(\max_{x_1\in X_1} \|a-x_1\|), \ldots, \varphi_p(\max_{x_p\in X_p} \|a-x_p\|)\right) d\mu(a)$$
$$= \int_A \Psi\left(\varphi_1(\max_{x_1\in \mathrm{ext}(X_1)} \|a-x_1\|), \ldots, \varphi_p(\max_{x_p\in \mathrm{ext}(X_p)} \|a-x_p\|)\right) d\mu(a),$$

where $\mathrm{ext}(X_i)$ denotes the set of vertices of the box X_i. For the particular case of an all-or-nothing covering function as given in (6.2), the above integral simplifies to

$$\int_{I(\mathbf{X})} d\mu(a),$$

where the set $I(\mathbf{X}) = \bigcup_{i=1}^p I_i(X_i)$ with $I_i(X_i) = \{a \in A | c_i(a,x_i) = 1\ \forall x_i \in \mathrm{ext}(X_i)\}$, i.e. $I_i(X_i)$ is the set of points a such that, for facility i, all points in X_i cover a (the gray region in Fig. 6.2). For an easier description of the set $I_i(X_i)$ one can consider its inscribed circle, $I_i^*(X_i)$ as shown in Fig. 6.2.

This leads to

$$\min_{\mathbf{x}\in\mathbf{X}} C(\mathbf{x}) \geq \int_{\bigcup_{i=1}^p I_i(X_i)} d\mu(a) \geq \sum_{i=1}^p \int_{I_i^*(X_i)} d\mu(a) - \sum_{\substack{i,j=1 \\ i<j}}^p \int_{I_i^*(X_i)\bigcap I_j^*(X_j)} d\mu(a).$$

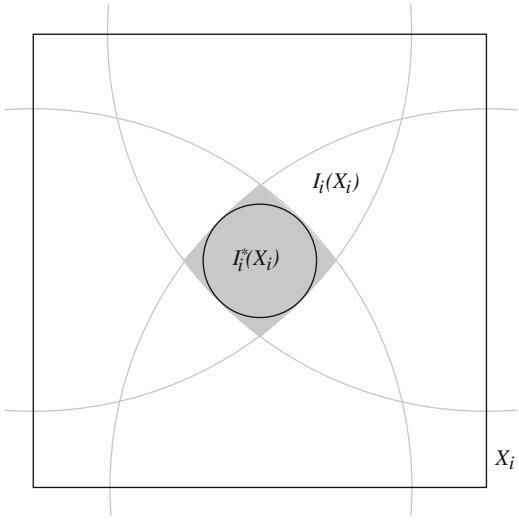


Fig. 6.2 Intersection of covered areas from $\text{ext}(X_i)$ giving the region which is covered by all points in the box. The integral is computed over the inscribed circle of this region, $I_i^*(X_i)$

In what follows, the so obtained bound will be denoted by $LB(\mathbf{X})$,

$$LB(\mathbf{X}) = \sum_{i=1}^{p} \int_{I_i^*(X_i)} d\mu(a) - \sum_{\substack{i,j=1 \\ i<j}}^{p} \int_{I_i^*(X_i) \bigcap I_j^*(X_j)} d\mu(a).$$

Notice, that the integral could be computed directly as $\int_A f(a) \min_{\mathbf{x} \in \mathbf{X}} c(a, \mathbf{x}) da$, but that is not practical for the all-or-nothing covering function. Numerical integrators take many sample points around discontinuities, that are introduced with $c(a, \mathbf{x})$, therefore taking a very long time for a single integration.

6.4 Numerical Examples

The branch and bound method outlines above was implemented in Fortran 90 (Intel©Fortran Compiler XE 12.0), using the integration tools of the IMSL Fortran Numerical Library. Executions were carried out on an Intel Core i7 computer with 8.00 Gb of RAM memory at 2.8 GHz, running Windows 7.

Two types of experiments were performed. First, a series of problems with randomly generated demand functions were solved for $p = 1$ and $p = 2$. The demand function was generated as a mixture of r bivariate gaussian distribution functions (6.16) with centers and weights uniformly generated in $[0, 10]^2$ and $[0.1, 0.1 + 1/(10r)]$, respectively. We set the covariance matrix to $w_i E$, that is the identity matrix scaled by the knot weight. The location of the facilities were sought in the square $[2, 8]^2$. Three parameters were considered, leading to different

problems: the radius R, the minimal distance γ in (6.20), and the number of knots r. As stopping criterion, the algorithm, stopped when the gap was smaller than 10^{-2}.

In order to reduce the random variability of the results, for each choice of radius R, minimal distance γ and number of knots r, three independent instances were generated and solved. The results presented in the tables correspond to the median out of the three values obtained.

In Table 6.1 running times in seconds are shown for the problem of locating one facility with a smaller and a larger radius ($R = 1.8$ and $R = 2.4$). It is not surprising that the computational time grows with the number of knots, as for all knots we need to do at least one integration.

Running times in seconds are reported in Table 6.2 for the problem of locating two facilities. Again, the values presented are the median value of the three runs performed. When at least two out of the three instances could not reach the desired accuracy in 8 h, the message "$> 8h$" is reported. The results clearly show that, the higher the number of knots or the radius, the higher the running times. The connection between the elapsed time and the minimal distance is not so evident. One can find cases where either smaller or higher minimal distance can be solved faster, so it looks rather problem dependent.

A second experiment was done in order to analyze the impact of the radius, displaying the Pareto frontier if one maximizes the radius and minimizes the coverage. In Fig. 6.3 the Pareto front is displayed for a problem with a mixture of 50 bivariate gaussian distributions setting minimal distance $\gamma = R$, and radii $R = 0.45, 0.6, \ldots, 1.65, 1.8$. The pdf of such mixture of gaussians was shown in

Table 6.1 Results for single-facility problems ($p = 1$) with different minimal distances

r	$R = 1.8$	$R = 2.4$
10	3.6	1.9
20	11.8	38.0
50	143.7	244.0
100	675.5	897.6

Table 6.2 Results for two-facility problems ($p = 2$) with different minimal distances

r	Minimal distance	$R = 1.2$	$R = 1.8$
10	R	110.5	186.1
	$1.5R$	182.8	124.7
	$2R$	178.1	83.4
20	R	114.0	2714.5
	$1.5R$	95.7	2593.5
	$2R$	86.4	2543.9
50	R	3926.2	12282.9
	$1.5R$	3754.7	18167.5
	$2R$	3675.1	$>8h$
100	R	20026.1	$>8h$
	$1.5R$	$>8h$	$>8h$
	$2R$	$>8h$	$>8h$

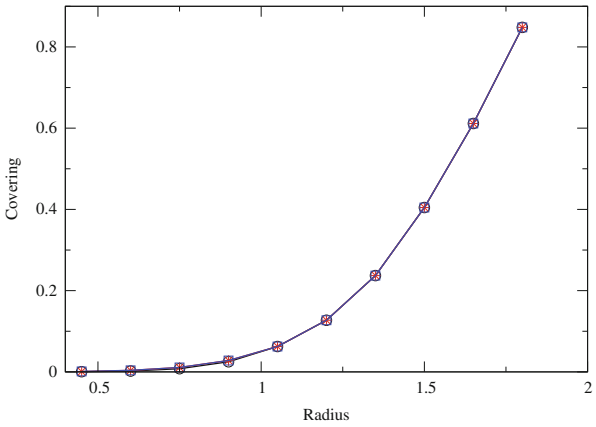


Fig. 6.3 Pareto frontier of the problem of maximizing the radius and minimizing covering

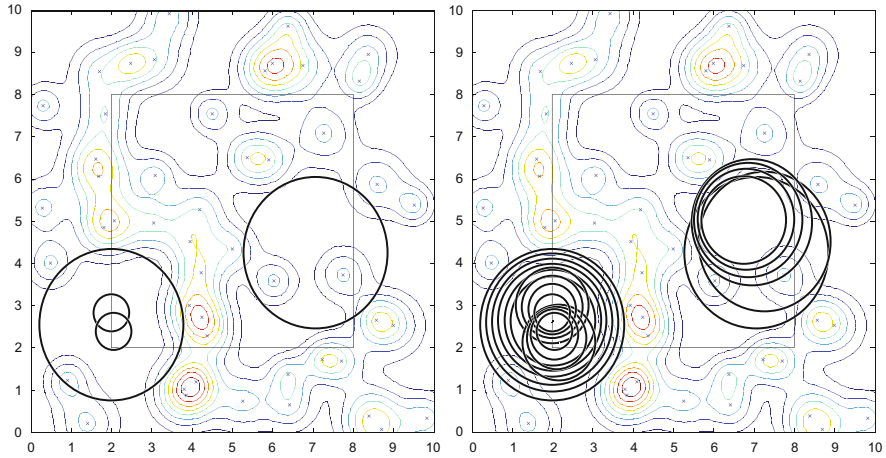


Fig. 6.4 Optimal covering for extreme radii (*left*) and all radii (*right*)

Fig. 6.1, while the solutions for the different radii are drawn in Fig. 6.4. In the latter, the demand function contours as well as the knots (with blue crosses) are shown. On the left, we focus on the optimal solution of the two extreme radii ($R = 0.45$ and $R = 1.8$). The optimal covered regions, i.e., the disc centered at the optimal facilities and radius R, are plotted. On the right, the optimal covered regions for all radii addressed are given.

6.5 Conclusions

While we have focused on purely repulsive facilities, the approach described here can be used to address location problems of semi-desirable facilities (Carrizosa and Plastria 1999; Blanquero and Carrizosa 2002; Romero-Morales et al. 1997; Plastria et al. 2013), in which, instead of having a set A of affected individuals, all negatively affected and wishing to have the facilities as far as possible, one has two separated sets, A^+ and A^-, identifying respectively the individuals feeling the facilities attractive, and thus want them as close as possible, and those feeling the facilities repulsive, and thus want them as far as possible. This would imply replacing the expected coverage function (6.10) by

$$C(\mathbf{x}) = -\int_{A^+} c^+(a,\mathbf{x})\, d\mu^+(a) + \int_{A^-} c^-(a,\mathbf{x})\, d\mu^-(a), \tag{6.23}$$

where c^+ and c^- are the covering models respectively for positively and negatively affected individuals. For finite probability measures μ^+ and μ^-, this model corresponds to minimizing a weighted sum of the points covered, where now the points in A^+ have a negative weight, already studied in Berman et al. (2009b) in a discrete setting. The planar version, including the regional case, remains unexplored. It calls for deriving new bounds for the branch and bound; but, as done here in the repulsive case, on can construct bounds after obtaining bounds for the covering functions $c(a,x)$. Whilst for c^- the key is that c^- is nonincreasing, monotonicity (in this case, decreasingness) can be used to bound $-c^+$. This approach is not new, since it already dates back to the seminal branch and bound BSSS (Hansen et al. 1985) but it deserves being tested.

The basic all-or-nothing cover function c in (6.2) is built assuming R fixed, and given R, the cover C is minimized. A dual problem consists of maximizing R so that the cover C remains below a threshold value. This so-called maxquantile problem (Plastria and Carrizosa 1999) would be solved by doing a binary search in the space of the values R, and solving, for each R, one problem as those solved in this chapter.

While affected individuals have been assumed to be (continuously) distributed in a planar region, facilities are considered here to have negligible size, so they are properly modeled as points. Adapting the branch and bound (in particular, the design of bounds) for the case of extensive facilities, e.g. Carrizosa et al. (1998a), deserves further study.

We have considered from the beginning the number of facilities p to be fixed. A related, somehow dual, problem is the problem of locating as many facilities as possible so that the coverage function C (or C^F, or both) remain(s) within a given interval. Such is the case of the so-called *anticovering* location problem, e.g. Chaudhry (2006), Moon and Chaudhry (1984), Murray and Church (1997), which, in its simplest version, seeks the highest number p^* of facilities such that no two are at a distance smaller than a threshold value R. Aggregation of the individual-facility cover functions $c(a,x)$ to $C(\mathbf{x})$ by any of the procedures described in Sect. 6.2 is

easily shown to be monotonic in the number p of facilities. The same holds for the aggregation of the facility-facility cover $c_{jk}^F(x_j, x_k)$ to $C^F(\mathbf{x})$. Hence, in order to find the highest p^* for which such covers remain within a given interval, one only needs to solve sequentially the problem for different values of p. The design of more direct and efficient procedures is definitely a promising research line.

Acknowledgements Research partially supported by research grants and projects ICT COST Action TD1207 (EU), MTM2012-36163 (Ministerio de Ciencia e Innovación, Spain), P11-FQM-7603, FQM329 (Junta de Andalucía, Spain), all with EU ERDF funds.

References

Alonso I, Carrizosa E, Conde E (1998) Maximin location: discretization not always works. TOP 6:313–319

Avella P, Boccia M (2007) A cutting plane algorithm for the capacitated facility location problem. Comput Optim Appl 43:39–65

Avella P, Sassano A, Vasil'ev I (2006) Computational study of large-scale p-median problems. Math Program 109:89–114

Berman O, Huang R (2008) The minimum weighted covering location problem with distance constraints. Comput Oper Res 35:356–372

Berman O, Krass D (2002) The generalized maximal covering location problem. Comput Oper Res 29:563–581

Berman O, Wang J (2011) The minmax regret gradual covering location problem on a network with incomplete information of demand weights. Eur J Oper Res 208:233–238

Berman O, Drezner Z, Wesolowsky GO (1996) Minimum covering criterion for obnoxious facility location on a network. Networks 28:1–5

Berman O, Krass D, Drezner Z (2003) The gradual covering decay location problem on a network. Eur J Oper Res 151:474–480

Berman O, Drezner Z, Krass D (2009a) Cooperative cover location problems: the planar case. IIE Trans 42:232–246

Berman O, Drezner Z, Wesolowsky GO (2009b) The maximal covering problem with some negative weights. Geogr Anal 41:30–42

Berman O, Kalcsics J, Krass D, Nickel S (2009c) The ordered gradual covering location problem on a network. Discrete Appl Math 157:3689–3707

Berman O, Drezner Z, Krass D (2010) Generalized coverage: new developments in covering location models. Comput Oper Res 37:1675–1687

Blanquero R, Carrizosa E (2002) A DC biobjective location model. J Global Optim 23:139–154

Blanquero R, Carrizosa E (2008) Continuous location problems and big triangle small triangle: constructing better bounds. J Global Optim 45:389–402

Blanquero R, Carrizosa E (2013) Solving the median problem with continuous demand on a network. Comput Optim Appl 56:723–734

Bowman A, Foster P (1993) Density based exploration of bivariate data. Stat Comput 3:171–177

Carrizosa E, Plastria F (1998) Locating an undesirable facility by generalized cutting planes. Math Oper Res 23:680–694

Carrizosa E, Plastria F (1999) Location of semi-obnoxious facilities. Stud Locat Anal 12:1–27

Carrizosa E, Conde E, Muñoz-Márquez M, Puerto J (1995) The generalized Weber problem with expected distances. RAIRO-Rech Oper 29:35–57

Carrizosa E, Muñoz-Márquez M, Puerto J (1998a) Location and shape of a rectangular facility in $\Re^n$ convexity properties. Math Program 83:277–290

Carrizosa E, Muñoz-Márquez M, Puerto J (1998b) A note on the optimal positioning of service units. Oper Res 46:155–156
Carrizosa E, Muñoz-Márquez M, Puerto J (1998c) The Weber problem with regional demand. Eur J Oper Res 104:358–365
Chaudhry SS (2006) A genetic algorithm approach to solving the anti-covering location problem. Expert Syst 23:251–257
Christaller W (1966) Central places in southern Germany. Prentice-Hall, London
Church R, ReVelle C (1974) The maximal covering location problem. Pap Reg Sci Assoc 32:101–118
Current JR, Schilling DA (1990) Analysis of errors due to demand data aggregation in the set covering and maximal covering location problems. Geogr Anal 22:116–126
Curtin KM, Church RL (2006) A family of location models for multiple-type discrete dispersion. Geogr Anal 38:248–270
Daskin MS, Haghani AE, Khanal M, Malandraki C (1989) Aggregation effects in maximum covering models. Ann Oper Res 18:113–139
Drezner Z (2012) Solving planar location problems by global optimization. Logist Res 6:17–23
Drezner Z, Suzuki A (2004) The big triangle small triangle method for the solution of nonconvex facility location problems. Oper Res 52:128–135
Drezner Z, Wesolowsky G (1994) Finding the circle or rectangle containing the minimum weight of points. Locat Sci 2:83–90
Drezner Z, Wesolowsky GO, Drezner T (2004) The gradual covering problem. Nav Res Log 51:841–855
Erkut E, Neuman S (1989) Analytical models for locating undesirable facilities. Eur J Oper Res 40:275–291
Erkut E, Neuman S (1991) Comparison of four models for dispersing facilities. INFOR 29:68–86
Farhan B, Murray AT (2006) Distance decay and coverage in facility location planning. Ann Reg Sci 40:279–295
Fekete SP, Mitchell JSB, Beurer K (2005) On the continuous fermat-weber problem. Oper Res 53:61–76
Fernández J, Fernández P, Pelegrín B (2000) A continuous location model for siting a non-noxious undesirable facility within a geographical region. Eur J Oper Res 121:259–274
Fliege J (2001) OLAF—a general modeling system to evaluate and optimize the location of an air polluting facility. OR Spectr 23:117–136
Francis RL, Lowe TJ (2011) Comparative error bound theory for three location models: continuous demand versus discrete demand. TOP 22:144–169
Francis RL, Lowe TJ, Tamir A (2000) Aggregation error bounds for a class of location models. Oper Res 48:294–307
Francis RL, Lowe TJ, Tamir A (2002) Demand point aggregation for location models. In: Drezner Z, Hamacher HW (eds) Facility location. Springer, Berlin/Heidelberg, pp 207–232
Francis RL, Lowe TJ, Rayco MB, Tamir A (2008) Aggregation error for location models: survey and analysis. Ann Oper Res 167:171–208
Hansen P, Peeters D, Richard D, Thisse JF (1985) The minisum and minimax location problems revisited. Oper Res 33:1251–1265
Hwang M, Chiang C, Liu Y (2004) Solving a fuzzy set-covering problem. Math Comput Model 40:861–865
Karasakal O, Karasakal EK (2004) A maximal covering location model in the presence of partial coverage. Comput Oper Res 31:1515–1526
Karkazis J, Papadimitriou C (1992) A branch-and-bound algorithm for the location of facilities causing atmospheric pollution. Eur J Oper Res 58:363–373
Kim K, Murray AT (2008) Enhancing spatial representation in primary and secondary coverage location modeling. J Reg Sci 48:745–768
Kolen A, Tamir A (1990) Covering problems. In: Mirchandani P, Francis R (eds) Discrete location theory. Wiley, New York

Kuby MJ (1987) Programming models for facility dispersion: the p-dispersion and maxisum dispersion problems. Geogr Anal 19:315–329
Lei TL, Church RL (2013) A unified model for dispersing facilities. Geogr Anal 45:401–418
Li X, Zhao Z, Zhu X, Wyatt T (2011) Covering models and optimization techniques for emergency response facility location and planning: a review. Math Method Oper Res 74:281–310
Mladenović N, Plastria F, Urošević D (2005) Reformulation descent applied to circle packing problems. Comput Oper Res 32:2419–2434
Moon ID, Chaudhry SS (1984) An analysis of network location problems with distance constraints. Manag Sci 30:290–307
Murat A, Verter V, Laporte G (2010) A continuous analysis framework for the solution of location-allocation problems with dense demand. Comput Oper Res 37:123–136
Murray AT (2005) Geography in coverage modeling: exploiting spatial structure to address complementary partial service of areas. Ann Assoc Am Geogr 95:761–772
Murray AT, Church RL (1997) Solving the anti-covering location problem using lagrangian relaxation. Comput Oper Res 24:127–140
Murray AT, O'Kelly ME (2002) Assessing representation error in point-based coverage modeling. J Geogr Syst 4:171–191
Murray AT, Wei R (2013) A computational approach for eliminating error in the solution of the location set covering problem. Eur J Oper Res 224:52–64
Murray AT, Tong D, Kim K (2009) Enhancing classic coverage location models. Int Reg Sci Rev 33:115–133
Plastria F (1992) Gbsss: the generalized big square small square method for planar single-facility location. Eur J Oper Res 62:163–174
Plastria F (2001) On the choice of aggregation points for continuous p-median problems: a case for the gravity centre. TOP 9:217–242
Plastria F (2002) Continuous covering location problems. In: Drezner Z, Hamacher HW (eds) Facility location. Springer, Berlin/Heidelberg, pp 39–83
Plastria F, Carrizosa E (1999) Undesirable facility location with minimal covering objectives. Eur J Oper Res 119:158–180
Plastria F, Gordillo J, Carrizosa E (2013) Locating a semi-obnoxious covering facility with repelling polygonal regions. Discrete Appl Math 161:2604–2623
Romero-Morales D, Carrizosa E, Conde E (1997) Semi-obnoxious location models: a global optimization approach. Eur J Oper Res 102:295–301
Schilling DA, Jayaraman V, Barkhi R (1993) A review of covering problems in facility location. Locat Sci 1:25–55
Schöbel A, Scholz D (2010) The big cube small cube solution method for multidimensional facility location problems. Comput Oper Res 37:115–122
Tong D (2012) Regional coverage maximization: a new model to account implicitly for complementary coverage. Geogr Anal 44:1–14
Tong D, Church RL (2012) Aggregation in continuous space coverage modeling. Int J Geogr Inf Sci 26:795–816
Tong D, Murray AT (2009) Maximising coverage of spatial demand for service. Pap Reg Sci 88:85–97
Wand MP, Jones MC (1993) Comparison of smoothing parameterizations in bivariate kernel density estimation. J Am Stat Assoc 88:520–528
Wand MP, Jones MC (1995) Kernel smoothing. Springer, New York
Yao J, Murray AT (2014) Serving regional demand in facility location. Pap Reg Sci 93:643–662

Part II
Advanced Concepts

Chapter 7
Location of Dimensional Facilities in a Continuous Space

Anita Schöbel

Abstract In many cases, the facilities to be located cannot be represented by isolated points, but may be modeled as dimensional structures. Examples for one-dimensional facilities are straight lines, line segments, or circles while boxes, strips, or balls are two-dimensional facilities. The goal of this chapter is to review the location of lines and circles in the plane and the location of hyperplanes and hyperspheres in higher dimensional spaces. We also discuss the location of some other dimensional facilities. We formulate the resulting location problems, point out some of their important properties and review the basic solution techniques and algorithmic approaches. Our focus lies on presenting a unified understanding of the common characteristics these problems have, and on reviewing the new findings obtained in this field within the last 10 years.

Keywords Circle location • Hyperplane location • Line location

7.1 Introduction

Within the locational context, the problem of locating a dimensional facility was first posed in Wesolowsky (1972, 1975) where the location of a line minimizing the sum of rectangular or Euclidean distances to a set of existing points was introduced. Since this time, the subject of locating lines and hyperplanes, circles, spheres, and other dimensional facilities has been intensively studied. Surveys are given in Martini and Schöbel (1998), Díaz-Bánez et al. (2004), an extensive list of papers dealing with the location of dimensional structures (most of them before 2000) is also given in Blanquero et al. (2009).

Within the last 10 years, many new results have been found and published. In this chapter, one goal is to review these new findings. More importantly, another goal is to present a unified understanding of the subject which is now possible since the field has become more mature. We hence not only present a list of problems

A. Schöbel (✉)
University of Göttingen, Göttingen, Germany
e-mail: schoebel@math.uni-goettingen.de

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_7

treated in the literature, but point out common characteristics and common solution techniques which are used for many different types of such location problems.

Applications in the location of dimensional facilities are various: These range from real-world applications in location theory and operations research to applications in robust statistics and computational geometry. Particular applications are mentioned at the beginning of the respective sections.

The chapter is organized as follows. We start with a general introduction into the topic in Sect. 7.2 where we introduce the basic notation, define the problems to be considered and mention the properties on which we will put some focus later on. We then discuss the two most extensively researched structures in dimensional facility location: The location of lines and hyperplanes in Sect. 7.3 and the location of circles and hyperspheres in Sect. 7.4. We finally review other interesting extensions and problem variations in Sect. 7.5. The chapter is ended by some conclusion in Sect. 7.6 summarizing the findings and pointing out lines for further research.

7.2 Location of Dimensional Facilities

The location of dimensional facilities is a natural generalization of locating one or more points. As in classical location problems we have given

- a finite set $V = \{v_1, \ldots, v_n\} \subseteq \mathbb{R}^D$ of existing facilities or existing points with positive weights $w_j > 0, j = 1, \ldots, n$, and
- a distance measure $d : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ evaluating the distance for each pair of points in $\mathbb{R}^D$. We mostly consider distances derived from norms or gauges.

We look for a new facility X which minimizes a function of the weighted distances to the existing points

$$\text{minimize } f(X) = g\begin{pmatrix} w_1 d(X, v_1) \\ w_2 d(X, v_2) \\ \vdots \\ w_n d(X, v_n) \end{pmatrix}, \tag{7.1}$$

where the most common functions used for f are the minsum (or median) function, i.e., $g_1(y_1, \ldots, y_n) = \sum_{j=1}^{n} y_j$ or the minmax (or center) function given as $g_{\max}(y_1, \ldots, y_n) = \max_{j=1,\ldots,n} y_j$. Also, other objective functions such as the centdian, or more general, ordered median objective functions g_λ (see Chap. 10) are possible.

If the new facility X is required to be a point, or a set of points, we are in the situation of classical continuous facility location, see Drezner et al. (2001). In this chapter, however, we assume that X is not a point but a dimensional structure such as a line, a circle, a hyperplane, a hypersphere, a polygonal line, etc. This, in turn, means that the distance $d(X, v)$ in (7.1) is the distance between a set X (which

represents the dimensional facility) and a point v. It is given by using the standard definition

$$d(X, v) = \min_{x \in X} d(x, v). \tag{7.2}$$

Note that in some applications $d(X, v)$ is defined as $\max_{x \in X} d(x, v)$, and that the average distance to all points in the set also is a reasonable definition; however, (7.2) is the most common model in this context.

We now specify the distances d we are mostly working with in this chapter. The most common distances in location theory are derived from norms, i.e., $d : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ is given as $d(x, y) := \|x - y\|$ for some norm $\|\cdot\|$. Moreover, distances derived from a gauge $\gamma : \mathbb{R}^D \to \mathbb{R}$ given through $d(x, y) = \gamma(y - x)$ have also been used in the location of dimensional facilities. Note that gauge-distances are no metrics since they are in general not symmetric, and that norms are special gauges. We also use the *vertical distance* and its generalizations, being neither a norm nor a gauge but giving insight into the problem, in particular for the location of lines and hyperplanes. For two points $x = (x^1, \ldots, x^D), y = (y^1, \ldots, y^D) \in \mathbb{R}^D$ the vertical distance is given as

$$d_{ver}(x, y) = \begin{cases} |x^D - y^D| & \text{if } x^i = y^i,\ i = 1, \ldots, D-1 \\ \infty & \text{otherwise.} \end{cases} \tag{7.3}$$

This distance leads to trivial location problems if X is required to be a point but yields interesting problems with applications mainly in statistics when locating lines or hyperplanes.

Figure 7.1 presents two examples on how distances are computed, and optimal dimensional structures may look like. In both examples we have given six existing points, all of them with unit weights. The left part of Fig. 7.1 shows a line minimizing the maximum vertical distance to the set of existing facilities. In the

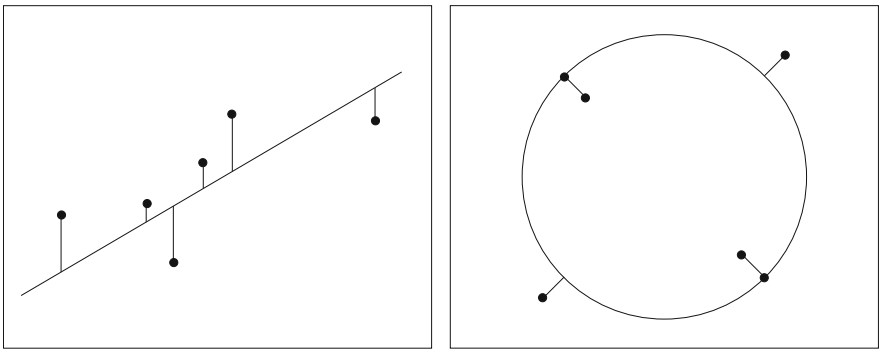


Fig. 7.1 Two illustrations for locating dimensional facilities. *Left*: A *line* minimizing the maximum vertical distance. *Right*: A *circle* minimizing the sum of Euclidean distances

right part a circle minimizing the sum of Euclidean distances to the existing facilities is depicted. The lengths of the thin lines in both examples correspond to the distances from the existing facilities to the line (or to the circle, respectively). Note that the distance between a facility $v \in X$ and X is zero—this happens twice in the right part of the figure where the minsum circle passes through two of the existing points.

In the following sections we discuss different types of dimensional facilities to be located. Most of the resulting optimization problems are multi-modal and neither convex nor concave. Hence, methods of global optimization are required. However, in many of these location problems it is possible to exploit one or more of the following properties showing that they have much more structure than just an arbitrary global optimization problem.

LP properties: Some of the problems become piecewise linear, sometimes even resulting in linear programming (LP) approaches which can be solved highly efficiently.

FDS properties: A finite dominating set (FDS) is a finite set of possible solutions from which it is known that it contains an optimal solution to the problem. This allows an enumeration approach by evaluating all possible elements of the FDS.

Halving properties: In many cases, any optimal facility to be located splits the sets of existing points into two sets of nearly equal weights. This allows to enhance enumeration approaches.

In our conclusion we provide a summary on these properties and give some general hints when they hold and why they are useful.

7.3 Locating Lines and Hyperplanes

Given a set of points $V \subseteq \mathbb{R}^D$ the hyperplane location problem is to find a hyperplane H minimizing the distances to the points in V. In this section we consider such hyperplane location problems for different types of distances and different objective functions.

Note that line location deals with finding a line in $\mathbb{R}^2$ minimizing the distances to a set of two-dimensional points and is included in our discussion as the special case $D = 2$.

7.3.1 Applications

The location of lines and hyperplanes has many applications within at least three different mathematical fields: Operations research, computational geometry, and statistics. Applications in *operations research* are various. The new facility to be located may be, e.g., a highway (see Díaz-Bánez et al. 2013), a train line (see Espejo

and Rodríguez-Chía 2011), a conveyor belt, or a mining shaft (e.g., Brimberg et al. 2002). Line location has also been mentioned in connection with the planning of pipelines, drainage or irrigation ditches, or in the field of plant layout (see Morris and Norback 1980).

In *computational geometry*, the width of a set is defined as the smallest possible distance between two parallel hyperplanes enclosing the set (Houle and Toussaint 1985). If the set is a polyhedron with extreme points $V = \{v_1, \ldots, v_n\}$ determining the width of this set is equivalent to finding a hyperplane minimizing the maximum distance to V. The relation between hyperplane location and transversal theory is mentioned in Sect. 7.3.4.1. In machine learning, a *support vector machine* is a hyperplane (if it exists) separating red from blue data points and maximizing the minimal distance to these points (see Bennet and Mangasarian 1992; Mangasarian 1999). If the set of red and blue points are not linearly separable, one may look for a hyperplane which minimizes the maximum distance to the points on the wrong side. This problem can again be solved as a restricted hyperplane location problem, see Carrizosa and Plastria (2008) and Plastria and Carrizosa (2012).

In *statistics*, classical linear regression asks for a hyperplane which minimizes the sum of squared vertical distances to a set of data points, while orthogonal regression (also called total least squares, see Golub and van Loan 1980) calls for a hyperplane minimizing the sum of squared Euclidean distances. However, these estimators are usually not considered as robust. This gives a reason for computing L_1-estimators minimizing the sum of absolute vertical (or orthogonal) differences, since the median of a set is considered more robust than its mean. We refer to Narula and Wellington (1982) for a survey on absolute errors regression. More general, many *robust estimators* can be found as optimal solutions to ordered hyperplane location problems, i.e., hyperplane location problems minimizing an ordered median objective function (see Chap. 10 for the definition of ordered median functions). Such problems are treated in Sect. 7.3.6. An example are *trimmed* estimators which neglect the k largest distances assuming that these belong to outliers. We list some of the most popular estimators and their corresponding hyperplane location problems in Table 7.1. For each of them we specify the distance function d which is used to measure the distance from the data points (i.e., the existing points) to the hyperplane, and the vector $\lambda \in \mathbb{R}^n$ which specifies the ordered

Table 7.1 Correspondence between line and hyperplane location problems and robust estimators

Estimator	Distance	Weights of ordered median function
Least squares	$d = d_{ver}^2$	$\lambda = (1, \ldots, 1)$
Total least squares	$d = \ell_2^2$	$\lambda = (1, \ldots, 1)$
Least trimmed squares	$d = d_{ver}^2$	$\lambda = (1, \ldots, 1, 0, \ldots, 0)$
Least absolute deviation	$d = d_{ver}$	$\lambda = (1, \ldots, 1)$
Least trimmed absolute deviation	$d = d_{ver}$	$\lambda = (1, \ldots, 1, 0, \ldots, 0)$
Least median of squares	$d = d_{ver}^2$	$\lambda = (0, \ldots, 0, 1, 0, \ldots, 0)$ (n odd) $\lambda = (0, \ldots, 0, 1, 1, 0, \ldots, 0)$ (n even)

median function g_λ used for modeling the respective estimator. More applications to classification and regression are pointed out in Bertsimas and Shioda (2007).

7.3.2 *Ingredients for Analyzing Hyperplane Location Problems*

7.3.2.1 Distances Between Points and Hyperplanes

A hyperplane is given by its normal vector $a = (a^1, \ldots, a^D) \in \mathbb{R}^D$ and a real number $b \in \mathbb{R}$:

$$H_{a,b} = \{x \in \mathbb{R}^D : a^t x + b = 0\}.$$

Given a distance $d : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$, the distance between a point $v \in \mathbb{R}^D$ and a hyperplane $H_{a,b}$ is given as $d(H_{a,b}, v) = \min\{d(x, v) : a^t x + b = 0\}$. For the vertical distance (see again the left part of Fig. 7.1) the following formula can easily be computed:

Lemma 7.1 (Schöbel 1999a)

$$d_{ver}(H_{a,b}, v) = \begin{cases} \frac{|a^t v + b|}{a^D} & \text{if } a^D \neq 0 \\ 0 & \text{if } a^D = 0 \text{ and } a^t v + b = 0 \\ \infty & \text{if } a^D = 0 \text{ and } a^t v + b \neq 0 \end{cases}$$

The second case and the third case comprise the case of a hyperplane which is vertical itself. Its distance to a point v is defined as infinity unless the hyperplane passes through v. If not all existing points lie in one common vertical hyperplane, this means that a vertical hyperplane can never be an optimal solution to the hyperplane location problem, hence without loss of generality we can assume the hyperplane $H_{a,b}$ to be non-vertical if the vertical distance is used.

If d is derived from a norm or a gauge $\gamma : \mathbb{R}^D \to \mathbb{R}$, the following formula for computing $d(H_{a,b}, v)$ has been derived in Plastria and Carrizosa (2001).

Lemma 7.2 (Plastria and Carrizosa 2001)

$$d(H_{a,b}, v) = \begin{cases} \frac{a^t v + b}{\gamma^\circ(a)} & \text{if } a^t v + b \geq 0 \\ \frac{-a^t v - b}{\gamma^\circ(-a)} & \text{if } a^t v + b < 0, \end{cases}$$

where $\gamma^\circ : \mathbb{R}^D \to \mathbb{R}$ is the dual (polar) norm common in convex analysis (e.g., Rockafellar 1970), i.e.,

$$\gamma^\circ(v) = \sup\{v^t x : \gamma(x) \leq 1\}.$$

Note that $d(H_{a,b}, v) = \frac{|a^t v + b|}{\gamma^\circ(a)}$ if γ is a norm.

7.3.2.2 Dual Interpretation

The following geometric interpretation is helpful when dealing with hyperplane location problems: A non-vertical hyperplane $H_{a,b}$ (with $a^D = 1$) may be interpreted as point $(a^1, \ldots, a^{D-1}, b)$ in $\mathbb{R}^D$. Vice versa, any point $v = (v^1, \ldots, v^D)$ may be interpreted as a hyperplane. Formally, we use the following transformation.

Definition 7.1

$$T_H(v^1, \ldots, v^D) := H_{v^1,\ldots,v^{D-1},1,v^D}$$
$$T_P(H_{a^1,\ldots,a^{D-1},1,b}) := (a^1, \ldots, a^{D-1}, b)$$

It can easily be verified that

$$d_{ver}(H_{a,b}, v) = d_{ver}(T_H(v), T_P(H_{a,b}))$$

for non-vertical hyperplanes with $a^D = 1$. In particular, we obtain

Lemma 7.3 *Let H be a non-vertical hyperplane and $v \in \mathbb{R}^D$ be a point. Then*

$$v \in H \iff T_P(H) \in T_H(v).$$

This means that $H_{a,b}$ passes through a point v if and only if $T_H(v)$ passes through $(a^1, \ldots, a^{D-1}, b)$.

In the resulting *dual space* the goal is to locate a point which minimizes the sum of distances to a set of given hyperplanes $\{T_H(v) : v \in V\}$. In the results of the next sections it will become clear that this is a helpful interpretation.

Figure 7.2 shows an example of the dual interpretation in $\mathbb{R}^2$. We consider five points (depicted in the left part of the figure), namely $v_1 = (0, \frac{1}{2})$, $v_2 = (0, 1)$,

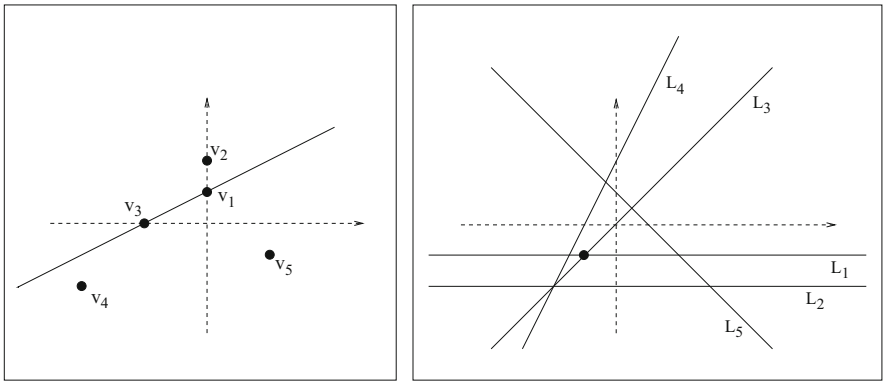


Fig. 7.2 *Left*: Five existing points and a line in primal space. *Right*: The same situation in dual space corresponds to five lines and one point

$v_3 = (-1, 0)$, $v_4 = (-2, -1)$ and $v_5 = (1, -\frac{1}{2})$. In the dual interpretation these points are transferred to the five lines in the right part of the figure.

$$L_1 = H_{0,1,\frac{1}{2}} = \{(x^1, x^2) : x^2 = -\frac{1}{2}\}$$
$$L_2 = H_{0,1,1} = \{(x^1, x^2) : x^2 = -1\}$$
$$L_3 = H_{-1,1,0} = \{(x^1, x^2) : x^2 = x^1\}$$
$$L_4 = H_{-2,1,-1} = \{(x^1, x^2) : x^2 = 2x^1 + 1\}$$
$$L_5 = H_{1,1,-\frac{1}{2}} = \{(x^1, x^2) : x^2 = -x^1 + \frac{1}{2}\}$$

It can also be seen that the line $H_{-\frac{1}{2},1,-\frac{1}{2}}$ through the two points v_1 and v_3 is transformed to the point $v = (-\frac{1}{2}, -\frac{1}{2})$ in dual space which lies on the intersection of L_1 and L_3. Furthermore, note that in the point $(-1, -1)$ in dual space three of the lines meet, namely, L_2, L_3, and L_4. Hence, this point corresponds to the line $H_{-1,1,-1} = \{(x^1, x^2) : x^2 = x^1 + 1\}$ which passes through the three points v_2, v_3, and v_4.

7.3.3 The Minsum Hyperplane Location Problem

Let us now start with the *minsum hyperplane location problem* defined as follows: Given a set of existing points $V = \{v_1, \ldots, v_n\} \subseteq \mathbb{R}^D$ with positive weights $w_j > 0$, $j = 1, \ldots, n$, find a hyperplane $H_{a,b}$ which minimizes

$$f_1(H_{a,b}) = \sum_{j=1}^{n} w_j d(H_{a,b}, v_j).$$

A hyperplane H minimizing $f_1(H)$ is called *minsum hyperplane w.r.t the distance* d. Let us assume throughout this section that there are $n > D$ affinely independent points, otherwise an optimal solution is the hyperplane containing all of them.

7.3.3.1 Minsum Hyperplane Location with Vertical Distance

We first look at the problem with vertical distance d_{ver}. As explained after Lemma 7.1 we may without loss of generality assume that $a^D = 1$. This simplifies the problem formulation to the question of finding $a^1, \ldots, a^{D-1}, b \in \mathbb{R}$ such that

$$f_1(a, b) = \sum_{j=1}^{n} w_j |v_j^t a + b| \tag{7.4}$$

is minimal (with $a^D = 1$). In order to get rid of the absolute values, we define

$$H^{>}_{a,b} := \{j \in \{1,\ldots,n\} : v_j^t a + b > 0\}$$
$$H^{<}_{a,b} := \{j \in \{1,\ldots,n\} : v_j^t a + b < 0\}$$
$$H^{=}_{a,b} := \{j \in \{1,\ldots,n\} : v_j^t a + b = 0\}.$$

We furthermore set

$$W^{>}_{a,b} := \sum_{j \in H^{>}_{a,b}} w_j, \quad W^{=}_{a,b} := \sum_{j \in H^{=}_{a,b}} w_j, \quad W^{<}_{a,b} := \sum_{j \in H^{<}_{a,b}} w_j$$

and let $W := \sum_{j=1}^{n} w_j$ be the sum of all weights. Since $f_1(a,b)$ is piecewise linear in b we receive:

Theorem 7.1 (Halving Property for Minsum Hyperplanes) *(Schöbel 1999a; Martini and Schöbel 1998) Let $H_{a,b}$ be a minsum hyperplane w.r.t the vertical distance d_{ver}. Then*

$$W^{>}_{a,b} \leq \frac{W}{2} \text{ and } W^{<}_{a,b} \leq \frac{W}{2} \tag{7.5}$$

Note that the halving property (7.5) is equivalent to

$$W^{>}_{a,b} \leq W^{<}_{a,b} + W^{=}_{a,b} \text{ and } W^{<}_{a,b} \leq W^{>}_{a,b} + W^{=}_{a,b}. \tag{7.6}$$

Looking again at (7.4), note that f_1 is not only piecewise linear in b but is also convex and piecewise linear in the D variables $a^1,\ldots,a^{D-1},b$. The latter yields the following *incidence property*.

Theorem 7.2 (FDS for Minsum Hyperplanes with Vertical Distance) *Let d_{ver} be the vertical distance and let $n \geq D$. Then there exists a minsum hyperplane w.r.t d_{ver} that passes through D affinely independent points.*

Sketch of Proof We can rewrite the objective function $f_1(H_{a,b})$ to

$$f_1(H_{a,b}) = \sum_{j \in H^{>}_{a,b}} w_j (v_j^t a + b) + \sum_{j : \in H^{<}_{a,b}} w_j (-v_j^t a - b) \tag{7.7}$$

which is easily seen to be linear as long as the signs of $v_j^t a + b$ do not change, i.e., on any polyhedral *cell* given by disjoint sets $H^{\geq}, H^{\leq}$ specifying which existing points should be below (or on) and above (or on) the hyperplane:

$$R(H^{\geq}, H^{\leq}) := \Big\{(a^1,\ldots,a^{D-1},b) : v_j^t a + b \geq 0 \text{ for all } j \in H^{\geq}$$
$$v_j^t a + b \leq 0 \text{ for all } j \in H^{\leq}\Big\}.$$

Note that these polyhedra can be constructed in dual space by using the arrangement of hyperplanes $T_H(v_j), j = 1, \ldots, n$, i.e., the right hand side of Fig. 7.2 shows exactly the polyhedra in dual space on which the objective function is linear. The fundamental theorem of linear programming then yields an optimal solution at a vertex of some of the cells $R(H^{\geq}, H^{\leq})$, i.e., a hyperplane satisfying $v_j^t a + b = 0$ for at least D indices from $\{1, \ldots, n\}$. □

Note that many papers mention this result. For $D = 2$, it was shown in Wesolowsky (1972), Morris and Norback (1983), Megiddo and Tamir (1983) and generalized to higher dimensions, e.g., in Schöbel (1999a).

In our example of Fig. 7.2 the depicted line is an optimal solution.

7.3.3.2 Minsum Hyperplane Location with Norm-Based Distance

We now turn our attention to the location of hyperplanes with respect to a norm $\|\cdot\|$. In this case, we can use Lemma 7.2 and obtain the following objective function

$$f_1(H_{a,b}) = \sum_{j=1}^{n} w_j \frac{|v^t a + b|}{\|a\|^\circ} \tag{7.8}$$

where $\|\cdot\|^\circ$ denotes the dual norm of $\|\cdot\|$. Still, the objective function is piecewise linear in b, hence the halving property holds again:

Theorem 7.3 (Halving Property for Minsum Hyperplanes) *(Schöbel 1999a; Martini and Schöbel 1998) Let d be a norm and $H_{a,b}$ be a minsum hyperplane w.r.t d. Then*

$$W_{a,b}^{+} \leq \frac{W}{2} \text{ and } W_{a,b}^{-} \leq \frac{W}{2}$$

We also receive the incidence property of Theorem 7.2.

Theorem 7.4 (FDS for Minsum Hyperplanes) *(Schöbel 1999a; Martini and Schöbel 1998, 1999) Let d be derived from a norm and let $n \geq D$. Then there exists a minsum hyperplane w.r.t d that passes through D affinely independent points. If and only if the norm is smooth, we have that* all *minsum hyperplanes pass through D affinely independent points.*

Sketch of Proof Different proofs for this property exist. Here, we use the cell structure of the proof of Theorem 7.2 for the vertical distance. The idea is to use piecewise quasiconcavity instead of piecewise linearity on these cells. Neglecting

vertical hyperplanes, we again look at the regions $R(H^{\leq}, H^{\geq})$ in dual space. On any such region we obtain that the objective function (7.8) can be rewritten as

$$f_1(H_{a,b}) = \sum_{j \in H^{>}_{a,b}} w_j \frac{v_j^t a + b}{\|a\|^\circ} + \sum_{j : \in H^{<}_{a,b}} w_j \frac{-v_j^t a - b}{\|a\|^\circ}$$

$$= \frac{1}{\|a\|^\circ} \left(\sum_{j \in H^{>}_{a,b}} w_j (v_j^t a + b) + \sum_{j : \in H^{<}_{a,b}} w_j (-v_j^t a - b) \right),$$

i.e., it is a positive linear function divided by a positive convex function and hence is quasiconcave. Consequently, it takes its minimum at a vertex of a region $R(H^{\leq}, H^{\geq})$, i.e., again at a hyperplane passing through D affinely independent existing points. □

Note that this theorem has been known for a long time for line location problems ($D = 2$) in the case of rectangular or Euclidean distances (Wesolowsky 1972, 1975; Morris and Norback 1980, 1983; Megiddo and Tamir 1983), and has been generalized to line location problems with arbitrary norms in Schöbel (1998, 1999a) and to D-dimensional hyperplane location problems with Euclidean distance in Korneenko and Martini (1990, 1993). The extension to hyperplanes with arbitrary norms is due to Schöbel (1999a) and Martini and Schöbel (1998).

7.3.3.3 Minsum Hyperplane Location with Gauges

For gauges the results of Theorems 7.4 and 7.3 do not hold any more. There exist counterexamples showing that optimal hyperplanes need not be halving, see, e.g., Schöbel (1999a). However, redefining the halving property by taking into account the non-symmetry on both sides of a hyperplane, the following similar result [based on formulation (7.6)] may be transferred to gauge distances.

Theorem 7.5 (Halving Property for Minsum Hyperplanes with Gauges) *(Plastria and Carrizosa 2001) Let d be a gauge and $H(a, b)$ be a minsum hyperplane w.r.t. d. Then we have*

$$\sum_{j \in H^{<}_{a,b}} \frac{w_j}{\gamma^\circ(a)} \leq \sum_{j \in H^{>}_{a,b} \cup H^{=}_{a,b}} \frac{w_j}{\gamma^\circ(a)}$$

$$\sum_{j \in H^{>}_{a,b}} \frac{w_j}{\gamma^\circ(-a)} \leq \sum_{j \in H^{<}_{a,b} \cup H^{=}_{a,b}} \frac{w_j}{\gamma^\circ(-a)}.$$

Also, for gauge-distances it does not hold that there always exists an optimal minsum hyperplane passing through D of the existing points, for a counterexample see again Schöbel (1999a). However, the following weaker result holds.

Theorem 7.6 (Incidence Property for Minsum Hyperplanes) *(Plastria and Carrizosa 2001) Let d be derived from a gauge and let $n \geq D$. Then there exists a minsum hyperplane w.r.t the distance d that passes through $D-1$ affinely independent points.*

Note that this incidence property does not define an FDS.

7.3.4 The Minmax Hyperplane Location Problem

We now turn our attention to the *minmax hyperplane location problem* in which we look for a hyperplane $H_{a,b}$ which minimizes

$$f_{\max}(H_{a,b}) = \max_{j=1,\ldots,n} w_j d(H_{a,b}, v_j).$$

A hyperplane H minimizing $f_{\max}(H)$ is called *minmax hyperplane w.r.t d*. Again, let us assume $n > D$. Since the main results for the location of minmax hyperplanes are similar for different types of distance functions, we need not distinguish between vertical, norm-based and gauge distances here. We start with a link to computational geometry.

7.3.4.1 Relation to Transversal Theory

Minmax location problems often rely on Helly's theorem (Helly 1923). For the location of hyperplanes, this result can only be applied for the vertical distance, since the sets $\{(a,b) : d(H_{a,b}, v) \leq \alpha\}$ are non-convex in general if $d \neq d_{ver}$. Instead, the following relation to transversal theory may be exploited.

Definition 7.2 Given a family of sets $\mathscr{M}$ in $\mathbb{R}^D$, a hyperplane H is called a *hyperplane transversal with respect to $\mathscr{M}$* if $M \cap H \neq \emptyset$ for all $M \in \mathscr{M}$.

Using this definition it is directly clear that $f_{\max}(H) \leq r$ if and only if H is a hyperplane transversal for the set $\mathscr{M} = \{M_j(r), j = 1, \ldots, n\}$ with

$$M_j(r) = \{x \in \mathbb{R}^D : w_j d(x, v_j) \leq r\}.$$

Instead of looking for a hyperplane minimizing the maximum distance to a set of existing points, we can hence equivalently look for the smallest possible $r \geq 0$ such that a hyperplane transversal for the sets $M_j(r), j = 1, \ldots, n$ exists. As an example, in Fig. 7.3 we search a line minimizing the maximum rectangular distance to the five given points, each of them with unit weight. Since it is a line transversal for the five sets $M_j(r)$, the depicted line l satisfies $f_{\max}(l) \leq r$.

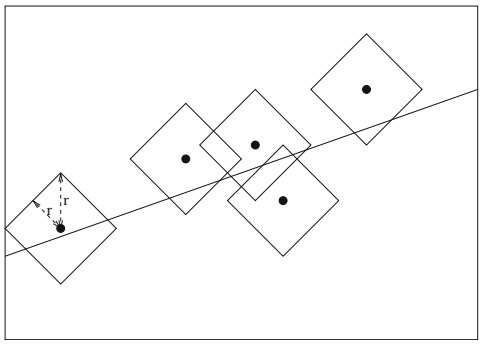


Fig. 7.3 A line transversal l to the five sets (each of them with radius r) exists, hence the objective function value of this line satisfies $f_{\max}(l) \leq r$

7.3.4.2 The Finite Dominating Set Property

The main result for minmax hyperplane location is the following *blockedness property*.

Theorem 7.7 (FDS for Minmax Hyperplanes) *(Schöbel 1999a; Martini and Schöbel 1998, 1999; Plastria and Carrizosa 2012) Let d be derived from a norm or a gauge and let $n \geq D+1$. Then there exists a minmax hyperplane w.r.t d that is at the same (maximum) distance from $D+1$ affinely independent points. If and only if the norm or the gauge is smooth, we have that* all *minmax hyperplanes are at maximum distance from $D+1$ affinely independent points.*

Sketch of Proof for Norms Similar to the proof for median hyperplanes we look at the case for vertical distances first. Here, the objective function is linear as long as the maximum distance does not change (if $n > 1$). We hence may use a type of farthest Voronoi diagram in the dual space, i.e., a partition of the dual space into (not necessarily connected) polyhedral cells

$$\begin{aligned} C(v_j) := \quad & \{(a,b) : d(H_{a,b}, v_j) \geq d(H_{a,b}, v) \text{ for all } v \in V\} \\ = \{(a^1, \ldots, a^{D-1}, b) & : |v_j^t a + b| \geq |v_i^t a + b| \text{ for all } i = 1, \ldots, n\} \end{aligned}$$

and it can be shown that an extreme point of such a cell is an optimal solution for the case of the vertical distance. Note that the cell structure does not change when we replace the vertical distance by a distance d derived from a norm, since we have

$$\begin{aligned} C'(v_j) := \quad & \{(a,b) : d(H_{a,b}, v_j) \geq d(H_{a,b}, v) \text{ for all } v \in V\} \\ = \{(a^1, \ldots, a^{D-1}, b) & : \frac{|v_j^t a + b|}{\gamma^\circ(a)} \geq \frac{|v_i^t a + b|}{\gamma^\circ(a)} \text{ for all } i = 1, \ldots, n\} \\ = C(v_j), & \end{aligned}$$

and using again that the objective function on these cells is quasiconcave, the result follows. □

Note that in contrast to minsum hyperplane location problems, the result also holds for gauges. This was shown for $D = 2$ in Schöbel (1999a) and for arbitrary D recently in Plastria and Carrizosa (2012). Using transversal theory, it can furthermore be extended to metrics (under some mild conditions of monotonicity), see Schöbel (1999a) for the case of $D = 2$.

A geometric point of view was taken in Nievergelt (2002) for the Euclidean case. He interprets the minmax hyperplane location problem as follows: locate two parallel hyperplanes such that the set of existing points lies completely between these two hyperplanes and minimize the distance between these parallel hyperplanes. He shows that in an optimal solution the two hyperplanes are *rigidly supported* by the points in V, i.e., there does not exist any other pair of parallel hyperplanes enclosing all points and passing through the same points of V. This property coincides with the blockedness property of Theorem 7.7. The algorithm proposed in Nievergelt (2002) uses projective shifts to improve a solution in a finite number of steps.

7.3.5 *Algorithms for Minsum and Minmax Hyperplane Location*

We describe the main approaches used for computing minsum hyperplanes.

7.3.5.1 Enumeration

Theorems 7.2, 7.4, and 7.7 specify a finite dominating set for both the minsum and the minmax hyperplane location problem. The trivial approach is to enumerate all candidates in the FDS. For the minsum case these are just the hyperplanes passing through D of the existing points. More effort is necessary to determine the hyperplanes being at maximum distance from $D + 1$ of the existing points for the minmax case. For $D = 2$ and norm-based distances these are parallel to one edge of the convex hull of the existing points (Schöbel 1999a).

7.3.5.2 Linear Programming for Hyperplane Location with Vertical and Block Norm Distances

For the vertical distance d_{ver} the hyperplane location problem can be formulated as a linear program. To this end, we define additional variables $d_j \geq 0$ which contain the distances $d(H, v_j)$, $j = 1, \ldots, n$. For the minsum problem we then obtain

$$\text{minimize} \sum_{j=1}^{n} w_j d_j \tag{7.9}$$

$$\text{subject to} \quad d_j \geq v_j^T a + b \text{ for } j = 1, \ldots, n \tag{7.10}$$

$$d_j \geq -v_j^T a - b \text{ for } j = 1, \ldots, n \tag{7.11}$$

$$d_j \geq 0 \text{ for } j = 1, \ldots, n \tag{7.12}$$

$$a^D = 1 \tag{7.13}$$

$$b, a^i \in \mathbb{R} \text{ for } i = 1, \ldots, D-1. \tag{7.14}$$

For the minmax problem, the objective (7.9) has to be replaced by the minmax objective function $f_{\max}$, i.e., by

$$\text{minimize} \max_{j=1,\ldots,n} w_j d_j,$$

which can be rewritten as linear program by using a bottleneck variable z and then replacing the objective by Minimize$z$ and adding $w_j d_j \leq z$ for $j = 1, \ldots, n$ as constraints. It is also possible to use other types of objective functions. For the minsum problem (see Zemel 1984) and for the minmax problem (see Megiddo 1984), the above LP formulation can be solved in O(n) time.

Now consider a block norm γ_B with unit ball $B = conv\{e_1, \ldots, e_G\}$, i.e., $e_g, g = 1, \ldots, G$ are the *fundamental directions* of the block norm. The idea is to solve the problem for each of the fundamental directions separately. To this end, we extend the vertical distance d_{ver} to a distance $d_t, t \in \mathbb{R}^D$ as follows.

$$d_t(u, v) := \begin{cases} |\alpha| & \text{if } u - v = \alpha t \text{ for some } \alpha \in \mathbb{R} \\ \infty & \text{otherwise.} \end{cases}$$

We then know the following result.

Lemma 7.4 (Schöbel 1999a) *Let H be a hyperplane and let d be derived from a block norm γ_B with fundamental directions $e_1, \ldots, e_G$. Then for any point $v \in \mathbb{R}^D$ there exists $\bar{g} \in \{1, \ldots, G\}$ such that*

$$d(H, v) = d_{e_{\bar{g}}}(H, v) = \min_{g=1,\ldots,G} d_{e_g}(H, v),$$

i.e., the fundamental direction $e_{\bar{g}}$ is independent of the point v.

This result allows to solve the problem with block norm distance in O(Gn) time in the planar case by iteratively solving the minmax hyperplane location problem with respect to distance d_{e_g}, $g = 1, \ldots, G$, and taking the best solution. Note that the G problems may be solved by transformation to the vertical distance as follows: Choose a linear (invertible) transformation T with $T(e_g) = (0, 0, \ldots, 0, 1)$. Transform all points $v'_j = T(v_j)$, $j = 1, \ldots, n$. We obtain that

$$d_{ver}(T(H), T(v)) = d_{e_g}(H, v)$$

for any hyperplane H and any point $v \in \mathbb{R}^D$, i.e., we have transformed the problem with distance d_{e_g} to a problem with vertical distance which can be solved by linear programming as above. Transforming an optimal hyperplane H' for the resulting problem back to $T^{-1}(H')$ gives an optimal solution to the problem with distance d_{e_g}. Details can be found in Schöbel (1999a, 1996).

7.3.5.3 Enhancing the Enumeration for Line Location with Euclidean Distance

For the Euclidean distance, the minsum straight line problem has received a lot of attention. Many of the ideas proposed here could also be used for other distance functions (see Schieweck and Schöbel 2012); nevertheless they have been investigated mainly for the Euclidean case. Algorithms rely on Theorems 7.3 and 7.4 and use the representation of the problem in the dual space.

The Euclidean minsum straight line problem with unit weights can be solved by sweeping along the so called *median trajectory* in the dual space (see Yamamoto et al. 1988). The median trajectory is the point-wise median of the lines $T_H(v_j)$, $j = 1, \ldots, n$, see Fig. 7.4 for the median trajectory in our example. The breakpoints on the median trajectory coincide with lines passing through two of the existing points and satisfying the halving property. Hence, the complexity of the approach depends on the number $h(n)$ of halving lines. In Yamamoto et al. (1988) the complexity of the approach is given as $O(\log^2(n)h(n))$ which can be improved to $O(\log(n)h(n))$ (see Schieweck and Schöbel 2012) by substituting the algorithm for dynamic convex hulls of Overmars and van Leeuwen (1981) by the newer $O(\log(n))$ algorithm of Brodal and Jacob (2002).

Note that the order of $h(n)$ is not known yet. It has been shown that the number of halving lines is in $O(n^{4/3})$ (see Dey 1998) yielding an $O(n^{4/3}\log(n))$ approach for the line location problem with Euclidean distance. The best known lower bound for the Euclidean minsum line location problem is $\Omega(n \log n)$ using reduction from

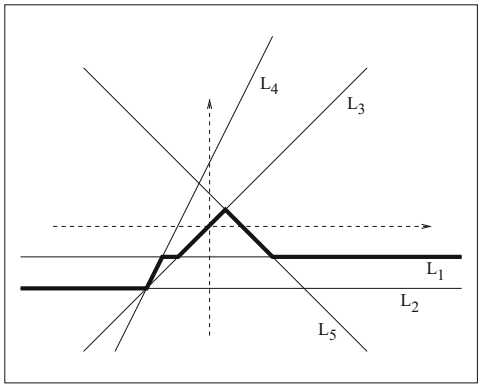


Fig. 7.4 The median trajectory for the example of Fig. 7.2

the uniform-gap on a circle problem (Yamamoto et al. 1988). We conclude that the question for an optimal algorithm for this problem is still open.

The Euclidean line location problem with arbitrary weights can be solved in $O(n^2)$, see Lee and Ching (1985).

For the Euclidean minmax line location problem the relation to transversal theory is exploited leading to an optimal $O(n \log n)$ algorithm for the case with arbitrary weights (Edelsbrunner 1985).

7.3.6 Ordered Median Line and Hyperplane Location Problem

A rather general objective function in location theory is the ordered median function (see Nickel and Puerto 2005, or Chap. 10). For tackling ordered median line location problems, one can combine the ideas of the preceding results on minsum and minmax location.

Theorem 7.8 (FDS for Ordered Line Location) *(See Lozano and Plastria 2009 for the Planar Euclidean Case) Let d be derived from a norm and let $n \geq 2$. Then there exists a solution l^* to the ordered line location problem w.r.t distance d that satisfies at least one of the following conditions:*

- *l^* passes through two of the existing points.*
- *l^* passes through one of the existing points and is at same weighted distance from two of the existing points.*
- *l^* is at the same weighted distance from three of the existing points.*
- *There exist two pairs of existing points $v_j, v_{j'} \in V$ and $v_k, v_{k'} \in V$ such that*

$$w_j d(l^*, v_j) = w_{j'} d(l^*, v_{j'}) \text{ and } w_k d(l^*, v_k) = w_{k'} d(l^*, v_{k'}),$$

i.e., l^* is at the same weighted distance from both points of each of the two pairs.

Sketch of Proof The theorem has been shown in Lozano and Plastria (2009) for the ordered Euclidean line location problem, but also holds for all distances derived from norms: Again, we look at the regions in dual space in which the order of the distances from the line to the existing points does not change, i.e., in which

$$d(H_{a,b}, v_j) = d(H_{a,b}, v_i)$$

does not hold for any $j \neq i$. These regions are hence bounded by the affine linear sets

$$\left\{(a,b) : \frac{w_j |a^t v_j + b|}{\gamma^\circ(a)} = \frac{w_i |a^t v_i + b|}{\gamma^\circ(a)}\right\} = \{(a,b) : w_j |a^t v_j + b| = w_i |a^t v_i + b|\}$$

in dual space and may be interpreted as the weighted bisectors of the lines $T_H(v_j)$ and $T_H(v_i)$. Taking the intersection of these regions with the regions $R(H^{\geq}, H^{\leqq})$ of the proof of Theorem 7.4, we obtain quasiconcavity on the resulting (smaller) cells. This yields that the extreme points of these new cells are a finite dominating set. □

This FDS allows an algorithm to solve the ordered line location problem in $O(n^4)$, see Lozano and Plastria (2009) for the Euclidean case. The problem of locating a hyperplane minimizing the Euclidean ordered median function has been investigated in Kapelushnik (2008) where its equivalence to searching within the levels of an arrangement is shown. The resulting algorithm runs in $O(n^{2D})$ where its complexity is reduced to $O(n^{D+\min\{D-1,K+1\}})$ if $K = |\{j = 1, \ldots, n : \lambda_j \neq 0\}|$.

A special case concerns the k-centrum line location problem, in which the sum of distances from the line to the k most distant points is minimized. It is also an ordered median problem and has been treated in Lozano et al. (2010). The methodology is similar to the approach of the general ordered median problem and exploits quasiconcavity of the objective function in the cells mentioned above. For smooth norms, it is shown that the resulting finite dominating set consists of lines either passing through two existing points or being at equal weighted distance from three of them. Based on this, an $O((k + \log n)n^3)$ algorithm is proposed for computing all t-centrum lines for $1 \leq t \leq k$. For unweighted points, Kapelushnik (2008) suggests an algorithm that finds a k-centrum line in the plane in time $O(n \log n + nk)$.

7.3.7 *Some Extensions of Line and Hyperplane Location Problems*

7.3.7.1 Obnoxious Line and Hyperplane Location

Instead of *minimizing* the distances to the existing points, one may also consider an obnoxious problem in which the new facility should be as far away from the existing points as possible. A rather general approach for obnoxious line location is presented in Lozano et al. (2013) in which a weighted ordered median function is maximized. More precisely, the problem treated is the following: Given a connected polygonal set S in the plane, the goal is to find a line which intersects S and maximizes the sum of ordered weighted Euclidean distances to the existing points. For such problems, the authors are again able to derive a finite dominating set which yields an $O(n^4)$ algorithm for the general Euclidean anti-ordered median case, and an $O(n^2)$ algorithm for the case of the Euclidean anti-median line. The case of locating an obnoxious plane (i.e., finding the widest empty slab through a set of existing points V) has been considered in Díaz-Bánez et al. (2006a). Also here, a finite dominating set could be identified leading to an algorithm in time $O(n^3)$.

7.3.7.2 Locating p Lines or Hyperplanes

As in point facility location it is also possible to study the problem of locating p lines or hyperplanes $H_1, \ldots, H_P$. In this setting, every existing point is served by its closest line. We may either minimize the sum of distances

$$f_1(H_1, \ldots, H_p) = \sum_{j=1}^{n} w_j \min_{q=1,\ldots,p} d(H_q, v_j) \tag{7.15}$$

or the maximum distance

$$f_{\max}(H_1, \ldots, H_p) = \max_{j=1,\ldots,n} w_j \min_{q=1,\ldots,p} d(H_q, v_j) \tag{7.16}$$

from the existing points to their closest lines. Minimizing the sum of distances is called *p-minsum-hyperplane location problem* and minimizing the maximum distance to a set of p hyperplanes is called *p-minmax-hyperplane location problem*. Locating p lines has important applications in statistics with latent classes, and also provides an alternative approach for clustering, called *projective clustering* (see, e.g., Har-Peled and Varadarajan 2002; Deshpande et al. 2006).

Both problems are known to be NP-hard for most reasonable distance measures (see Megiddo and Tamir 1982). However, since each of the p hyperplanes $H_1, \ldots, H_p$ to be located is a minsum (or minmax) hyperplane for the set of points

$$V_q = \{v \in \{v_1, \ldots, v_n\} : d(H_q, v) \leq d(H_{q'}, v) \text{ for all } q' = 1, \ldots, p\}$$

the results on the finite dominating sets of Theorems 7.4 and 7.7 still hold:

Theorem 7.9 *Given $p \in \mathbb{N}$ and a set of existing points V.*

- *If $n \geq D$ then there exists an optimal solution to the p-minsum-hyperplane location problem in which each hyperplane passes through D existing points.*
- *If $n \geq D + 1$ then there exists an optimal solution to the p-minmax-hyperplane location problem in which each of hyperplane is at maximum distance from $D+1$ existing points.*

Hence, enumeration approaches based on such an FDS are possible, however, the number of candidates to be enumerated is of order $O(n^D)$. Recently, such an enumeration approach for the p-minsum line location problem has been enhanced by computing lower bounds and using them to discard elements from the FDS, see Schieweck (2013). The idea is to cluster the demand points and find a line which minimizes the sum of distances to the resulting demand regions. This problem is not easier than the original problem, but since the number of demand regions is much smaller than n it can be solved quicker.

Based on the FDS, another approach is possible: The problem may be transformed to a p-median or p-center problem on a bipartite graph with $O(|FDS|)$ nodes.

The two node sets of the graph are given by the existing points V and by the potential hyperplanes in the FDS. Every node v from V is connected to every node H from the FDS where the edge (v, H) is weighted by the distance, the node v has from the hyperplane H. The goal is to serve all customers in V by installing p new locations in the FDS.

Finally, the problem of finding p lines in the plane is studied in Bertsimas and Shioda (2007) where it is formulated as an integer program. Binary variables $x_{j,q}$ determine to which of the $q = 1, \ldots, p$ lines the existing point v_j is assigned. Applying their basic formulation to the linear program (7.9)–(7.14) of Sect. 7.3.5 gives

$$
\begin{aligned}
\text{minimize} \quad & \sum_{j=1}^{n} w_j d_j \\
\text{subject to} \quad & d_j \geq v_j^T a_q + b_q - M(1 - x_{j,q}) \quad \text{for } j = 1, \ldots, n,\ q = 1, \ldots, p \\
& d_j \geq -v_j^T a_q - b_q - M(1 - x_{j,q}) \quad \text{for } j = 1, \ldots, n,\ q = 1, \ldots, p \\
& \sum_{q=1}^{p} x_{j,q} = 1 \quad \text{for } j = 1, \ldots, n \\
& x_{j,q} \in \{0, 1\} \quad \text{for } j = 1, \ldots, n,\ q = 1, \ldots, p \\
& d_j \geq 0 \quad \text{for } j = 1, \ldots, n \\
& a_q^D = 1 \quad \text{for } q = 1, \ldots, p \\
& b_q, a_q^i \in \mathbb{R} \text{ for } i = 1, \ldots, D-1,\ q = 1, \ldots, p.
\end{aligned}
$$

Solving the integer program in its basic form is not possible in reasonable time; in Bertsimas and Shioda (2007) clustering algorithms are performed in a preprocessing step. The above integer program can also be used for solving the minmax version of the problem, if $\sum$ is replaced by max in its objective function.

7.3.7.3 Restricted Line Location

Line location problems in which the line is not allowed to pass through a specified set $R \subseteq \mathbb{R}^2$ can be tackled by looking at the dual space and transforming the restriction to a forbidden set there. Since the problem is convex for vertical distances, techniques from location theory can be used, e.g., the boundary theorem saying that there exists a solution on the boundary of the restricted set whenever the restriction is not redundant (see Hamacher and Nickel 1995). The results may be generalized to block norms or to arbitrary norms, see Schöbel (1999b).

In some statistical applications it is preferable to restrict the slope of the line (or the norm of a) as done in types of RLAD approaches (Wang et al. 2006). Such

restrictions on the parameters of the hyperplane can again be treated and solved in dual space, see Krempasky (2012).

Another type of restriction is to force a subset of points of V to lie on, above or below the hyperplane. Also for such problems, finite dominating sets have been derived, see Schöbel (2003) for hyperplane location problems in which the hyperplane is forced to pass through a subset of points and Plastria and Carrizosa (2012) for the more general case of requiring a specified subset of points below or above the hyperplane.

7.3.7.4 Line Location with Polyhedra as Existing Facilities

There are also a few approaches considering the location of lines when the existing facilities are connected sets or polyhedra in $\mathbb{R}^2$. The minmax problem is equivalent to finding the thinnest strip transversal, i.e., a strip of minimal width which intersects each of the existing polyhedra. For m polyhedra with a total of n vertices, Robert (1991) and Robert and Toussaint (1994) solve the Euclidean problem by computing the upper and the lower envelope of the dual representation of the existing sets resulting in an $O(n \log n)$ approach in the unweighted case and in an $O(n^2 \log n)$ approach in the weighted case. For the minsum problem, the algorithm works by sweeping the dual arrangement and takes $O(mn \log m)$ time.

7.3.7.5 Line Location in $\mathbb{R}^D$

Locating a line in $\mathbb{R}^D$ turns out to be a difficult problem since all of the structure of line and hyperplane location problems gets lost. In Brimberg et al. (2002, 2003) some special cases are investigated for the case $D = 3$, such as locating a vertical line, or locating a line where the distance measure is given as the lengths of horizontal paths. If these lengths are measured with the rectangular distance, the problem can be reduced to two planar line location problems with vertical distance. For the general case of locating a minsum line in $\mathbb{R}^3$, global optimization methods such as Big-Cube-Small-Cube (Schöbel and Scholz 2010) have been successfully used, see Blanquero et al. (2011). The case of locating a minmax line in $\mathbb{R}^D$ is known in computational geometry as smallest enclosing cylinder problem. It has been mainly researched in $\mathbb{R}^3$ (Schömer et al. 2000; Chan 2000).

7.4 Locating Circles and Spheres

We now turn our attention to the location of hyperspheres. Again, we have given a set of existing points $V \subseteq \mathbb{R}^D$ with positive weights $w_j > 0, j = 1, \ldots, n$. The *hypersphere location problem* is to find the center point and the radius of a hypersphere S which minimizes the distances from its surface to the points in V.

The problem is interesting not only for Euclidean circles and spheres but also for all unit balls derived from a norm. In this section we consider such hypersphere location problems for different types of norms and different objective functions.

Note that circle location deals with finding a circle in $\mathbb{R}^2$ minimizing the distances from its circumference to a set of points in the plane. For circle location, more and stronger results are known than for general hypersphere location; it will hence be treated separately where appropriate.

7.4.1 Applications

Hyperspheres and circles are mathematical objects which are well-known for hundreds of years. The Rhind Mathematical Papyrus, written around 1650 BC by Egyptian mathematicians, already contains a method for approximating the surface area of a circle, see Robins and Shute (1987). The problem of fitting a circle or a sphere to a set of data points has also been mentioned in the fourth century BC by notes of Aristotle on the earth's sphericity, see Dicks (1985).

Also nowadays, the location of circles and spheres has applications in different fields. The Euclidean version of the problem is of major interest in measurement science, where it is used as a model for the out-of-roundness problem which occurs in quality control and consists of deciding whether or not the roundness of a manufactured part is in the normal range (see, e.g., Farago and Curtis 1994; Ventura and Yeralan 1989; Yeralan and Ventura 1988). To this end, measurements are taken along the boundary of the manufactured part. In order to evaluate the roundness of the part, a circle is searched which fits the measurements. Mathematical models for different variants of the out-of-roundness problem are studied for instance in Le and Lee (1991), Swanson et al. (1995), and Sun (2009).

Circle and hypersphere location problems have also applications in other disciplines, e.g., in particle physics (Moura and Kitney 1992; Crawford 1983) when fitting a circular trajectory to a large number of electrically charged particles within uniform magnetic fields, or in archeology where minmax circles are used to estimate the diameter of an ancient shard (Chernov and Sapirstein 2008). In Suzuki (2005), the construction of ring roads is mentioned as an application. Many further applications are collected in Nievergelt (2010). They include

- the analysis of the design and layout of structures in archeology,
- the analysis of megalithic monuments in history,
- the identification of the shape of planetary surfaces in astronomy,
- computer graphics and vision,
- calibration of microwave devices in electrical engineering,
- measurement of the efficiency of turbines in mechanical engineering,
- monitoring of deformations in structural engineering, or
- the identification of particles in accelerators in particle physics.

There is also a relation to equity problems (see Gluchshenko 2008; Drezner and Drezner 2007) of point facility location and to a problem in computational geometry which is to find an annulus of smallest width. These relations are specified in Sect. 7.4.4.1.

In statistics, the problem is also of interest. As Nievergelt (2002) points out, many attempts have been made of transferring total least squares algorithms from hyperplane location problems to hypersphere location problems (e.g., Kasa 1976; Moura and Kitney 1992; Crawford 1983; Rorres and Romano 1997; Späth 1997, 1998; Coope 1993; Gander et al. 1994; Nievergelt 2004).

7.4.2 *Distances Between Points and Hyperspheres*

Let d be a distance derived from some norm $\|\cdot\|$, i.e., $d(x, y) = \|y - x\|$. A circle or a sphere with respect to the norm $\|\cdot\|$ is given by its center point $x = (x^1, \ldots, x^D) \in \mathbb{R}^D$ and its radius $r > 0$:

$$S_{x,r} = \{y \in \mathbb{R}^D : d(x, y) = r\}.$$

The distance between a sphere $S = S_{x,r}$ and a point $v \in \mathbb{R}^D$ is defined as

$$d(S, v) = \min_{y \in S} d(y, v)$$

and can be computed as

$$d(S_{x,r}, v) = |d(x, v) - r|.$$

The following properties of the distance can easily be shown.

Lemma 7.5 (Körner et al. 2012; Körner 2011) *Given a distance d derived from a norm, and a point $v \in \mathbb{R}^D$, the following hold:*

- *$d(S_{x,r}, v)$ is convex and piecewise linear in r,*
- *$d(S_{x,r}, v)$ is locally convex in (x, r) if v is a point outside the sphere, and*
- *$d(S_{x,r}, v)$ is concave in (x, r) if v is inside the sphere.*

Before analyzing minsum or minmax circles or hyperspheres, let us remark that even the special case with only $n = 3$ existing points in the plane ($D = 2$) is a surprisingly interesting problem. Within a wider context it has recently been studied in Alonso et al. (2012a,b). Here, the circumcircle of a set of three points is investigated (which is the optimal minmax or minsum circle for the three points). Dependent on the norm considered, such a circumcircle need not exist, and need not be unique. Among other results on covering problems, the work focuses on a complete description of possible locations of the center points of such circumcircles.

7.4.3 The Minsum Hypersphere Location Problem

We start with the minsum hypersphere location problem. Given a distance d derived from a norm, the goal is to find a hypersphere $S = S_{x,r}$ which minimizes

$$f_1(S_{x,r}) = \sum_{j=1}^{n} w_j d(S_{x,r}, v_j) = \sum_{j=1}^{n} w_j |d(x, v_j) - r|. \tag{7.17}$$

The location of a Euclidean circle in the plane has been defined and treated in Drezner et al. (2002). This has then been generalized to the location of a norm-circle in the plane in Brimberg et al. (2009b), and later to the location of a hypersphere with respect to any norm in $\mathbb{R}^D$ (Körner et al. 2012). The Euclidean case in dimension d has been also extensively analyzed in Nievergelt (2010).

We start by presenting some general properties of minsum hypersphere location problems. In contrast to hyperplanes, it is not obvious in which cases a minsum hypersphere exists, since a hypersphere can degenerate to a point (for $r = 0$) and to a hyperplane (for $r \to \infty$). The following results are known.

Lemma 7.6 (Brimberg et al. 2011a; Körner et al. 2012)

- *No hypersphere with $r = 0$ can be a minsum hypersphere.*
- *For any smooth norm there exist instances for which no minsum hypersphere exists.*
- *For any elliptic norm and any block norm a minsum hypersphere exists for all instances with $n \geq D + 1$.*

Since no optimal solution degenerates to a point, we need not bother with existence results if we restrict r to an upper bound and solve the problem then.

Let us now discuss the halving property. To this end, we define the set of points outside, on, and inside the hypersphere

$$S^{>}_{x,r} := \{j \in \{1, \ldots, n\} : d(x, v_j) > r\}$$
$$S^{<}_{x,r} := \{j \in \{1, \ldots, n\} : d(x, v_j) < r\}$$
$$S^{=}_{x,r} := \{j \in \{1, \ldots, n\} : d(x, v_j) = r\}$$

and let

$$W^{>}_{x,r} := \sum_{j \in S^{>}_{x,r}} w_j, \quad W^{=}_{x,r} := \sum_{j \in S^{=}_{x,r}} w_j, \quad W^{<}_{x,r} := \sum_{j \in S^{<}_{x,r}} w_j.$$

As before, let $W = \sum_{j=1}^{n} w_j$ be the sum of all weights.

Theorem 7.10 (Halving Property for Minsum Hyperspheres) *(Brimberg et al. 2011a; Körner et al. 2012) Let $S_{x,r}$ be a minsum hypersphere w.r.t to any distance derived from a norm. Then*

$$W^{>}_{x,r} \leq \frac{W}{2} \text{ and } W^{<}_{x,r} \leq \frac{W}{2} \tag{7.18}$$

Sketch of Proof If we increase the radius from r to $r + \epsilon$ the distance to points in $S^{>}_{x,r}$ decreases by ϵ, and the distance to points in $S^{<}_{x,r}$ increases by ϵ. This means, if $W^{>}_{x,r} > \frac{W}{2}$ we can improve the objective function by increasing the radius. (Analogously, if $W^{<}_{x,r} > \frac{W}{2}$ we can improve the objective function by reducing the radius.) □

While the halving property can be nicely generalized, this is unfortunately not true for the determination of a finite dominating set. The generalization of Theorem 7.4 would be that there always exists an optimal Euclidean circle passing through three of the existing points. However, this turned out to be wrong, even in the unweighted case (see Fig. 7.1 for a counter-example). For most distances it is not even guaranteed that there exists an optimal circle passing through two points. The only incidence property that can be shown is the following.

Lemma 7.7 *Let d be any distance derived from a norm. Then there exists a minsum hypersphere w.r.t d which passes through at least one point $v \in V$.*

Sketch of Proof Let $S_{x,r}$ be a hypersphere. Fix its center point x and assume without loss of generality that the existing points are ordered such that $d(x, v_1) \leq d(x, v_2) \leq \ldots \leq d(x, v_n)$. Then the objective function $f'(r) := f_1(S_{x,r})$ in (7.17) is piecewise linear in r on the intervals $I_j := \{r : d(x, v_j) \leq r \leq d(x, v_{j+1}\}$, $j = 1, \ldots, n-1$, and hence takes a minimum at a boundary point, i.e., there exists an optimal radius $r = d(x, v_j)$ for some $v_j \in V$. □

The proof uses that the radius of an optimal circle is the median of the distances $d(x, v_1), \ldots, d(x, v_n)$ which was already recognized in Drezner et al. (2002).

Not much more can be said in the general case. The only (again, weak) property into this direction we are aware of is the following:

Lemma 7.8 (Körner et al. 2012) *Let $S = S_{x,r}$ be a minsum hypersphere with radius $r < \infty$. Then S intersects the convex hull of the existing points in at least two points, i.e., $|S \cap conv(V)| \geq 2$.*

Furthermore, if $|S \cap conv(V)| < \infty$, then $S \cap conv(V) \subseteq V$.

7.4.3.1 Location of a Euclidean Minsum Circle

For the Euclidean distance and the planar case $D = 2$ it is possible to strengthen the incidence property of Corollary 7.7.

Theorem 7.11 (Brimberg et al. 2009b) *Let d be the Euclidean distance, and consider the planar case, i.e., let $D = 2$. Then there exists a minsum circle which passes through two points of V.*

The result has been shown by looking at the second derivatives of the objective function (in an appropriately defined neighborhood) which reveal that a circle passing through exactly one or none of the existing points cannot be a local minimum.

An algorithmic consequence of Theorem 7.11 is that there exists an optimal circle with center point x being on a bisector of two of the existing points, hence a line search along the bisectors is possible. Using Theorem 7.10 a large amount of bisectors may be excluded beforehand. Figure 7.5 shows the Euclidean bisectors for five existing facilities where the relevant parts (which contain center points of circles having the halving property) are marked in bold.

Another approach was followed in Drezner and Brimberg (2014): Here the unweighted case is shown to be an ordered median *point* location problem with weights $\lambda = (-1, \ldots, -1, 1, \ldots, 1)$ with equal number of -1's and 1's if n is even, and with weights $\lambda = (-1, \ldots, -1, 0, 1, \ldots, 1)$ with equal number of -1's and 1's if n is odd. The resulting ordered median point location problem was then solved using the Big-Triangle-Small-Triangle method (Drezner and Suzuki 2004) with the d.c. bounding technique proposed in Brimberg and Nickel (2009).

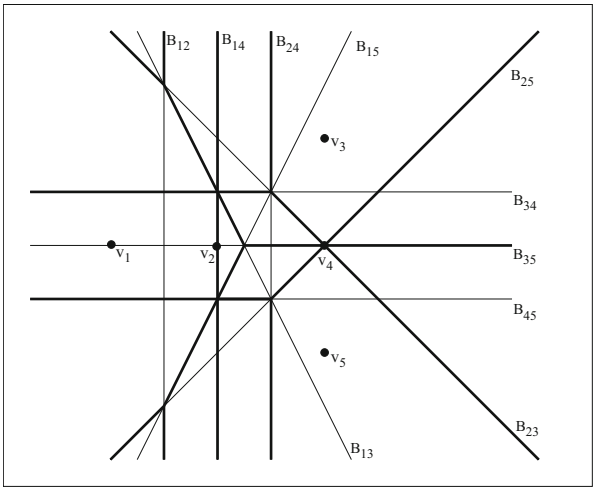


Fig. 7.5 The Euclidean bisectors for five existing points. The notation B_{ij} indicates that the corresponding line is the bisector for points v_i and v_j. The parts of the bisectors which may contain a center point of a minsum circle are marked in *bold*

7.4.3.2 Location of Minsum Circles and Hyperspheres with Block Norms

If d is derived from a block norm, a finite dominating set can be constructed for the center point of the minsum circle. To this end, graph all fundamental directions $\{e_1, \ldots, e_G\} \subseteq \mathbb{R}^2$ of the block norm through any of the existing points $v \in V$ and add the bisectors for all pairs of existing points in V. The intersection points of these lines are a finite dominating set which can be tested within $O(n^3)$ time, see Körner (2011) and Brimberg et al. (2011a).

Using that the block norm of a point y is given as

$$\|y\| = \min\{\sum_{g=1}^{G} \alpha_g : y = \sum_{g=1}^{G} \alpha_g e_g, \alpha_g \geq 0 \text{ for } g = 1, \ldots, G\}$$

the problem can in the case of block norms alternatively be formulated as the following linear program with $nG + 2n + D + 1$ variables, see Brimberg et al. (2011a) for the planar case and Körner et al. (2012) for the case of hyperspheres.

$$\begin{aligned}
\text{minimize} \quad & \sum_{j=1}^{n} w_j \left(z_j^+ + z_j^-\right) \\
\text{subject to} \quad & \sum_{g=1}^{G} \alpha_{g,j} = r + z_j^+ - z_j^- \text{ for } j = 1, \ldots, n \\
& \sum_{g=1}^{G} \alpha_{g,j} e_g = x - v_j \text{ for } j = 1, \ldots, n \\
& z_j^+, z_j^- \geq 0 \text{ for } j = 1, \ldots, n \\
& \alpha_{g,j} \geq 0 \text{ for } g = 1, \ldots, G, j = 1, \ldots, n \\
& r \geq 0 \\
& x \in \mathbb{R}^D.
\end{aligned}$$

7.4.4 The Minmax Hypersphere Location Problem

We now turn our attention to the location of a minmax hypersphere, i.e., we look for a hypersphere which minimizes the maximum weighted distance from its surface to the set V of existing points. Given a distance d derived from a norm, the goal hence is to find a hypersphere $S = S_{x,r}$ which minimizes

$$f_{\max}(S_{x,r}) = max_{j=1}^{n} w_j d(S_{x,r}, v_j) = \sum_{j=1}^{n} w_j |d(x, v_j) - r|. \tag{7.19}$$

Note that the problem of locating a Euclidean minmax circle in the plane is older than the corresponding Euclidean minsum circle problem; a finite dominating set has already been identified in Rivlin (1979). Its rectangular version is due to Gluchshenko et al. (2009). In $\mathbb{R}^D$ the Euclidean minmax hypersphere location problem has been analyzed mainly in the Euclidean case, see Nievergelt (2002).

7.4.4.1 Relation to Minimal Covering Annulus Problem and Equity Problem

The problem of locating a minmax circle has a nice geometric interpretation. For equally weighted points it may be interpreted as finding an annulus of minimal width covering all existing points. This problem has been studied in computational geometry, hence results on minmax circle location have been obtained independently in location theory and in computational geometry.

In location science the minmax hypersphere location problem has an interesting application as a point location problem. Namely, the (unweighted) center point x of an optimal hypersphere $S_{x,r}$ minimizes the difference

$$\max_{j=1,\dots,n} d(x, v_j) - \min_{j=1,\dots,n} d(x, v_j),$$

i.e., it minimizes the *range* to the set V. We conclude that minmax hypersphere location problems can be interpreted as ordered median point location problems. Therefore, the point x may be interpreted as a fair location for a service facility as used in equity problems, see Gluchshenko (2008) for further results.

7.4.4.2 Location of a Euclidean Minmax Circle

Let us start with the Euclidean case in dimension $D = 2$: In this case, the problem has been discussed extensively in the literature, mainly in computational geometry under the name of finding an annulus of smallest width. In contrast to the Euclidean minsum circle problem, where an FDS could not be found, the following result shows that an FDS for the (Euclidean) minmax hypersphere exists.

Theorem 7.12 (FDS for the Euclidean Minmax Circle) *(e.g., Rivlin 1979; Brimberg et al. 2009a) Let $D = 2$ and let C be a minmax circle with finite radius. Let $h := \max_{j=1,\dots,n} w_j d(C, v_j)$. Then there exist four points having distance h to the circle C, two of them inside the circle and two of them outside the circle.*

The theorem was shown for the unweighted case independently in many papers, among others in Rivlin (1979), Ebara et al. (1989), García-López et al. (1998) and it was generalized to the weighted case in Brimberg et al. (2009a). The result can be interpreted in different ways:

- In the geometric interpretation, the result means that the annulus of minimal width covering all points has two points on its inner circumference and two points on its outer circumference (Rivlin 1979).
- It also shows that the center point of a minimax circle is either a vertex of the (nearest neighbor) Voronoi diagram or of the farthest neighbor Voronoi diagram or lies at an intersection point of both diagrams (Le and Lee 1991; García-López et al. 1998).

For the unweighted problem, Ebara et al. (1989) use this result and present an enumeration algorithm with runtime in $O(n^2)$. If the points in V are given in an angular order, García-López et al. (1998) present an algorithm which runs in $O(n \log n)$ and which can even be improved to $O(n)$ if the points in V are the vertices of a convex polygon. This is in particular helpful for solving the out-of-roundness problem (see Sect. 7.4.1), since the measurements are taken along the manufactured part in angular order in this case. A gradient search heuristic is provided in Drezner et al. (2002) and global optimization methods were used in Drezner and Drezner (2007) who use the Big-Triangle-Small-Triangle method (based on Drezner and Suzuki 2004) for its solution. Randomized and approximation algorithms are also possible, see Agarwal et al. (2004, 1999).

More references on the computation of Euclidean minmax circles can be found in García-López et al. (1998) and in Brimberg et al. (2009a).

7.4.4.3 Location of a Minmax Circle with Rectangular Distance

Gluchshenko (2008) and Gluchshenko et al. (2009) consider the minimal annulus problem for the rectangular distance. This means, the circle to be located is a diamond, and the distances from the given points to the circle are measured in the rectangular norm. The following is an important result.

Theorem 7.13 (FDS for the Rectangular Minmax Circle) *(Gluchshenko et al. 2009) There exists a minmax circle whose center point is a center point of a smallest enclosing square.*

This means the set of all center points of smallest enclosing squares (which can be determined easily) is an FDS. Based on this, Gluchshenko et al. (2009) develop an optimal $O(n \log n)$ algorithm for finding a minmax circle with respect to the rectangular norm.

Recently, the problem in which the annulus may also be rotated has been considered in Mukherjee et al. (2013) where an $O(n^2 \log n)$ algorithm has been proposed.

7.4.4.4 Location of a Euclidean Minmax Hypersphere

The problem of finding a minmax hypersphere in dimension $D \geq 3$ was considered in García-López et al. (1998). The authors give necessary and sufficient conditions for a point to be the center point of a *locally* minimal hypersphere with respect to $f_{\max}$. Independently, also Nievergelt (2002) considers the problem of locating a hypersphere in $\mathbb{R}^D$ with Euclidean distance. Analogously to his approach for minmax hyperplanes, he interprets the problem as the location of two concentric hyperspheres with minimal distance which enclose the set V of existing points. This results in a generalization of Theorem 7.12 to higher dimensions.

Theorem 7.14 (FDS for the Euclidean Minmax Hypersphere) *(Nievergelt 2002) There exists a Euclidean minmax hypersphere S which is rigidly supported by the point set V, i.e., there does not exist any other pair of concentric hyperspheres enclosing all points of V and passing through the same points of V as S.*

Based on this property, Nievergelt (2002) derives a finite algorithm finding a minmax hypersphere with respect to the Euclidean distance. A linear time $(1+\epsilon)$ factor approximation algorithm for finding a Euclidean minmax hypersphere is given in Chan (2000).

7.4.5 Some Extensions of Circle Location Problems

7.4.5.1 Minimizing the Sum of Squared Distances

An earlier variant of the hypersphere location problem minimizes the sum of *squared* distances of the existing points to the circle, i.e., it considers

$$f_2^2(S_{x,r}) = \sum_{j=1}^{n} w_j \left(d(S_{x,r}, v_j)\right)^2$$

as objective function. In Drezner et al. (2002) it is shown that the least squares objective is equivalent to minimizing the variance of the distances. The problem is (like the minsum and minmax problem) non-convex; heuristic solution approaches are suggested. In Drezner and Drezner (2007) the Big-Triangle-Small-Triangle global optimization algorithm is successfully applied.

Minimizing the sum of squared distances from the points in V to a circle has been also considered within statistics in Kasa (1976), Crawford (1983), Moura and Kitney (1992), Coope (1993), Gander et al. (1994), Rorres and Romano (1997), Späth (1997, 1998), and Nievergelt (2004).

7.4.5.2 Locating Euclidean Concentric Circles

In a recent paper, Drezner and Brimberg (2014) introduce the following interesting extension of the circle location problem: They look for p concentric circles with different radii $r_1, \ldots, r_p$ which minimize the distances to a given set of points. In their paper they assume a partition of V into sets $V_1, \ldots, V_p$ and require that each point in V_i is served by the circle with radius r_i. This means the variables to be determined are the center point $x \in \mathbb{R}^2$ and the radii $r_1, \ldots, r_p$ of the p circles. The model is considered for the least squares objective function, the minsum, and the minmax objective function. Using that

$$d(S_{x,r_j}, v_j) = |d(x, v_j) - r|$$

the objective functions which are considered are given as

$$f_2^2(x, r_1, \ldots, r_p) = \sum_{q=1}^{p} \sum_{v_j \in V_q} w_j \left(d(x, v_j) - r\right)^2$$

$$f_1(x, r_1, \ldots, r_p) = \sum_{q=1}^{p} \sum_{v_j \in V_q} w_j |d(x, v_j) - r|$$

$$f_{\max}(x, r_1, \ldots, r_p) = \max_{q=1,\ldots,p} \max_{v_j \in V_q} w_j |d(x, v_j) - r|.$$

Drezner and Brimberg (2014) solve the problem by global optimization methods, using a reformulation of the circle location problem as an ordered median point location problem (see the location of a Euclidean minsum circle in Sect. 7.4.3) and applying the Big-Triangle-Small-Triangle method (Drezner and Suzuki 2004).

7.4.5.3 Location of a Circle with Fixed Radius

The location of a circle with fixed radius is considered in Brimberg et al. (2009a). In this case, it can be shown that considering every triple of points separately yields an optimal solution, i.e., a finite dominating set can be derived by solving $\binom{n}{3}$ smaller optimization problems.

7.4.5.4 Generalized Circle Location: Locating the Unit Ball of One Norm Measuring Distances with Respect to Another Norm

The circle location problem treated so far is to translate and scale a circle $S = \{x \in \mathbb{R}^2 : \|x\| \leq 1\}$ (derived from norm $\|\cdot\|$) in such a way that the distances to the set V are minimized, where the distances are measured with respect to the same norm $\|\cdot\|$.

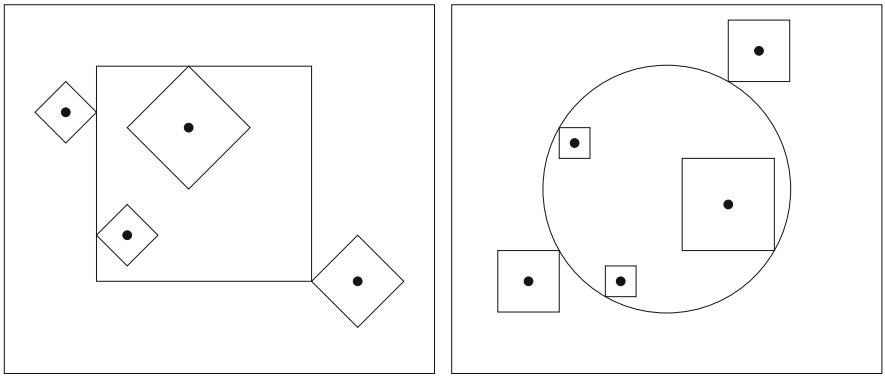


Fig. 7.6 Locating a unit ball of norm k_1 with respect to another norm k_2. *Left*: The unit circle of the maximum-norm is to be located, distances are measured w.r.t the rectangular norm. *Right*: The Euclidean circle is to be located, distances are measured w.r.t the maximum norm

In Körner et al. (2009, 2011) this problem is studied for two *different* norms under the name *generalized circle location*.

More precisely, given two norms k_1 and k_2 and a set of points V in the plane with positive weights $w_j > 0$, the goal of generalized circle location is to locate and scale the unit ball of norm k_1 such that the sum of weighted distances between its circumference and the given points is minimized, where distances are measured by the other norm k_2. Figure 7.6 shows two possible situations. In the left part of the figure, the new facility is the scaled and translated unit circle of the $k_1 := \|\cdot\|_{\max}$ norm and the distances to the four given points are measured by the $k_2 := \|\cdot\|_1$ norm. In the right part, $k_1 := \|\cdot\|_2$ and $k_2 := \|\cdot\|_{\max}$.

In Körner et al. (2011), properties of minsum generalized circle location are investigated, and it is shown that not much of the properties for minsum circle location still hold. There is neither an easy formula for computing the distance between a point and such a generalized circle, nor does any of the incidence criteria hold. In fact, there are examples in which no optimal circle passes through any of the existing points. However, if both norms k_1 and k_2 are block norms, a finite dominating set can still be identified (see Körner et al. 2009). The problem of locating a general circle is interesting for many special cases, e.g. if a box should be located. Such cases have been studied in Brimberg et al. (2011b).

7.5 Locating Other Types of Dimensional Facilities

7.5.1 Locating Line Segments

The *line segment location problem* looks for a line segment with specified length which minimizes the distances to the set V of existing points.

Location of line segments has been considered in Imai et al. (1992), Agarwal et al. (1993), Efrat and Sharir (1996) for the Euclidean minmax problem, and in Schöbel (1997) for the minsum problem with vertical distances. In both the cases it is possible to determine a finite dominating set; the latter case can be transformed to a restricted line location problem.

Recently, locating line segments received new interest within the following problem: A line segment and a point facility are to be located simultaneously. In this setting, the line segment can be used to speed up traveling in the plane in which a new point facility should be built. The problem has been treated in the plane, using rectangular distances in Espejo and Rodríguez-Chía (2011, 2012) where a characterization of optimal solutions was used to derive an algorithm. This could be improved in Díaz-Bánez et al. (2013) to an $O(n^3)$ approach. These approaches are based on a finite dominating set which can be obtained by reduction of the location problem to a finite number of simpler optimization problems.

7.5.1.1 The Widest Empty 1-Corner Corridor in the Plane

An *empty corridor* in the plane is an open region bounded by two parallel polygonal chains that does not contain any of the existing points $V = \{v_1, \ldots, v_n\}$, and that partitions the existing points into two non-empty parts. This can be interpreted as an obnoxious dimensional location problem: locate a polygonal chain maximizing the minimum distance to the existing facilities. Empty corridors have been of interest in computational geometry (see e.g., Janardan and Preparata 1996). An empty corridor is called a *1-corner empty corridor* if each of the two bounding polygonal chains has exactly one corner point. The problem in which the angle at the corner point is given and fixed has been studied in Cheng (1996). Recently, Díaz-Bánez et al. (2006b) considered the problem of locating a widest 1-corner corridor using techniques of facility location: they were able to derive a finite dominating set consisting of locally widest 1-corner corridors among which a solution may be chosen. Their approach needs $O(n^4 \log n)$ time. It was further improved to $O(n^3 \log^2 n)$ time in Das et al. (2009).

7.5.1.2 Two-Dimensional Facilities

Covering problems are the most common problems in which the location of full-dimensional facilities is considered. There exist, e.g., many papers about covering points by a circle (i.e., locating one point x such that all given points are in a given threshold distance from x), by a set of circles, or even by a set of aligned circles (occurring when the center points of the circles to be located are forced to lie on a common straight line), or circles satisfying other restrictions. Covering problems are not reviewed here, we refer to Plastria (2001) or to Chap. 5.

However, also the location of a two-dimensional facility X such that the minsum or minmax objective function is minimized, has been considered in the literature. If

there exists a location for X such that all existing points are covered, this location is clearly an optimal solution with objective value zero both for the minsum and for the minmax problem. If it is not possible to cover all points, the minsum and the minmax problem usually have different solutions.

A paper dealing with the location of a two-dimensional facility is Brimberg and Wesolowsky (2000) where the rectangular distance is considered and special cases could be transformed to classical point location problems. In the context of facility layout the location of a rectangular office with minsum and minmax objective function has been studied in Savas et al. (2002), Kelachankuttu et al. (2007) and Sarkar et al. (2007). In these papers, already existing offices are treated as barriers. Various problem variations for the location of an axis-parallel rectangle (with fixed circumference, with fixed area, with fixed aspect ratio, or with fixed shape and size) have been considered in Brimberg et al. (2011b). For most cases, a finite dominating set could be derived.

The location of a two-dimensional ball

$$B_x = \{y \in \mathbb{R}^2 : d(x, y) \leq r\}$$

with given and fixed radius r has been considered in Brimberg et al. (2013a) both for the minsum and the minmax objective function. Note that the distance between B_x and v

$$d(B_x, v) = \min_{y \in B_x} d(y, v)$$

is measured as the closest distance to any point in B, and not only to points on its circumference $S_{x,r}$. This means that

$$d(B_x, v) = \begin{cases} 0 & \text{if } v \in B_x \\ d(S_{x,r}, v) & \text{otherwise.} \end{cases}$$

Hence, Lemma 7.5 yields that $d(B_x, v)$ is a convex function and consequently, the resulting optimization problems are much easier to solve than the circle location problems of Sects. 7.4.3 and 7.4.4. We remark that the location of a full-dimensional ball has the following interesting interpretation as a point location problem with *partial coverage*:

Assume that we are looking for a new facility $x \in \mathbb{R}^2$ for which we know that little or no service cost (or inconvenience) is associated with existing points that are within an acceptable travel distance r from x. Thus, costs will be associated only to those existing points that are further away from the facility than this threshold distance r. If we assume that these costs are proportional to the distance in excess of r, the resulting problem is equivalent to the location of a ball with radius r, and its center point is the optimal location x we are looking for. This has been pointed out in Brimberg et al. (2013a) where the behavior of the optimal solution with respect to the threshold distance r is studied.

Line location with the partial coverage objective function is equivalent to locating a strip of given width and has recently been considered in Brimberg et al. (2013b).

7.5.1.3 A General Approach Based on dc-Programming

Blanquero et al. (2009) deal with the location of a variety of dimensional facilities such as segments, arcs of circumferences, arbitrary convex sets, their complements, or their boundaries. The idea is to fix the shape of the dimensional facility and to look for a shift vector and an angle of rotation. The objective they follow is very general, including most objective functions used in location theory, and allows also to model obnoxious or semi-obnoxious location problems as follows: The set of existing facilities is split into a subset V^+ for which the new facility is attractive and a subset V^- for which the new facility has negative effects. The distance from the new facility to an existing point should be small when the point is in V^+ and large when it is in V^-. In order to combine the distances within the same set V^+ and V^- Blanquero et al. (2009) propose to evaluate the norm (or the gauge) of the resulting single distances.

Using that the Euclidean distance $d(S, v)$ between a point and a set can be written as difference of convex functions, Blanquero et al. (2009) solve the model by d.c.-programming methods, outer approximation and branch and bound.

7.6 Conclusions

For the location of dimensional facilities we can draw the following conclusions.

- The location of a one-dimensional facility (i.e., a point) and a two-dimensional facility of convex shape with respect to a norm are convex problems if distances are measured by norms.
- In contrast, the location of a one-dimensional facility with respect to a norm is a non-convex problem which usually has many locally optimal solutions. Only the vertical distance leads to convex hyperplane location problems (if also the objective function g is convex).
- However, many of the investigated problems of locating a one-dimensional facility are piecewise quasiconcave on a cell structure in dual space. This leads to a finite dominating set. Another possibility for deriving an FDS is via Helly-type theorems.
- When distances are measured w.r.t a block norm, problems are often piecewise linear and can hence be solved by linear programming methods.
- The halving property holds when the problem is linear with respect to one of its variables.

The main properties pointed out in this chapter are summarized in Table 7.2. They have the following algorithmic consequences.

Table 7.2 Summary of properties for some of the considered location problems

Problem	FDS	Halving	LP
Minsum hyperplane with $d = d_{ver}$	Yes	Yes	Yes
Minsum hyperplane with norm	Yes	Yes	No
Minsum hyperplane with block norm	Yes	Yes	Yes
Minsum hyperplane with gauges	No	(Yes)	No
Minmax hyperplane with norm	Yes	No	No
Minmax hyperplane with block norm	Yes	No	Yes
Minmax hyperplane with gauges	Yes	No	No
Ordered minsum hyperplane with norm	Yes	Yes	No
Minsum line in $\mathbb{R}^3$	No	No	No
Line may not pass through a polyhedral set	Yes	No	No
Minsum/minmax p-line with norm	Yes	No	No
Minsum hypersphere with norm	No	Yes	No
Minsum hypersphere with block norm	Yes	Yes	Yes
Minmax hypersphere with Euclidean norm	Yes	No	No
Minmax circle with rectangular norm	Yes	No	Yes

The FDS property gives the straightforward possibility of enumerating the candidate set. Also for the location of p facilities the FDS property is still helpful, although the number of candidates increases to $O(|FDS|^p)$. As demonstrated for the p-minsum line location problem in Sect. 7.3.7, an FDS also allows to transfer the problem of locating p facilities to a p-location problem on a bipartite graph with $O(|FDS|)$ nodes. It is ongoing work to test such approaches numerically.

Enumeration may be enhanced by the halving property which can be used to directly discard candidates. Such discarding tests are also useful in other approaches, even if no FDS is known, since the halving property allows to discard whole regions when searching for an optimal solution. An example is the search along bisectors which can be reduced to the relevant parts in the Euclidean minsum circle location problem. Also in geometric branch & bound approaches such as Big-Square-Small-Square (Plastria 1992), Big-Triangle-Small-Triangle (Drezner and Suzuki 2004), or Big-Cube-Small-Cube (Schöbel and Scholz 2010), discarding tests motivated by the halving property may be interesting.

Using linear programming methods is an efficient way of solving facility location problems, in particular if the number of variables needed for the linear program is not too large. This is the case for block norms with not too many fundamental directions.

While many questions in the location of lines and hyperplanes seem to be solved, there are still questions remaining in the location of hyperspheres. These concern, on one hand, general properties about the location of hyperspheres with other than the minsum objective function and with arbitrary norms or gauges. On the other hand, there are also many special cases waiting to be investigated, in particular if the sphere is defined with respect to another norm as the distance function.

Concerning the location of new types of dimensional structures, researchers should look for shapes which are of interest for other disciplines or for applications. Similarly, identifying additional restrictions and particularities arising in applications in operations research, statistics, and computational geometry and including them in the models is a future challenge.

Acknowledgements I want to thank Robert Schieweck for providing useful hints on line and hyperplane location problems.

References

Agarwal P, Efrat A, Sharir M, Toledo S (1993) Computing a segment center for a planar point set. J Algorithm 15:314–323

Agarwal P, Aronov B, Peled S, Sharir M (1999) Approximation and exact algorithms for minimum-width annuli and shells. In: Proceedings of the 15th ACM symposium on computational geometry, pp 380–389

Agarwal P, Peled SH, Varadarajan K (2004) Approximation extent measures of points. J ACM 51:605–635

Alonso J, Martini H, Spirova M (2012a) Minimal enclosing discs, circumcircles, and circumcenters in normed planes (part I). Comput Geom Theor Appl 45:258–274

Alonso J, Martini H, Spirova M (2012b) Minimal enclosing discs, circumcircles, and circumcenters in normed planes (part II). Comput Geom Theor Appl 45:350–369

Bennet K, Mangasarian O (1992) Robust linear programming discrimination of two linearly inseparable sets. Optim Method Softw 1:23–34

Bertsimas D, Shioda R (2007) Classification and regression via integer optimization. Oper Res 55:252–271

Blanquero R, Carrizosa E, Hansen P (2009) Locating objects in the plane using global optimization techniques. Math Oper Res 34:837–858

Blanquero R, Carrizosa E, Schöbel A, Scholz D (2011) Location of a line in the three-dimensional space. Eur J Oper Res 215:14–20

Brimberg J, Nickel S (2009) Constructing a DC decomposition for ordered median problems. J Global Optim 45:187–201

Brimberg J, Wesolowsky G (2000) Note: facility location with closest rectangular distances. Nav Res Log 47:77–84

Brimberg J, Juel H, Schöbel A (2002) Linear facility location in three dimensions - models and solution methods. Oper Res 50:1050–1057

Brimberg J, Juel H, Schöbel A (2003) Properties of 3-dimensional line location models. Ann Oper Res 122:71–85

Brimberg J, Juel H, Schöbel A (2009a) Locating a circle on the plane using the minimax criterion. Stud Locat Anal 17:46–60

Brimberg J, Juel H, Schöbel A (2009b) Locating a minisum circle in the plane. Discrete Appl Math 157:901–912

Brimberg J, Juel H, Körner MC, Schöbel A (2011a) Locating a general minisum 'circle' on the plane. 4OR-Q J Oper Res 9:351–370

Brimberg J, Juel H, Körner MC, Schöbel A (2011b) Locating an axis-parallel rectangle on a Manhattan plane. TOP 22:185–207

Brimberg J, Juel H, Körner MC, Schöbel A (2013a) On models for continuous facility location with partial coverage. J Oper Res Soc. doi:JORS.2013.142

Brimberg J, Schieweck R, Schöbel A (2013b) Locating a median line with partial coverage distance. Preprint 32, Institut für Numerische und Angewandte Mathematik, Universität Göttingen. http://num.math.uni-goettingen.de/preprints/files/2013-32.pdf

Brodal GS, Jacob R (2002) Dynamic planar convex hull. In: Proceedings of the 43rd annual IEEE symposium on foundations of computer science, pp 617–626

Carrizosa E, Plastria F (2008) Optimal expected-distance separating halfspace. Math Oper Res 33:662–677

Chan TM (2000) Approximating the diameter, width, smalat enclosing cylinder, and minimum-width annulus. In: Proceedings of the 16th annual symposium on computational geometry SCG '00. ACM, New York, pp 300–309

Cheng SW (1996) Widest empty L-shaped corridor. Inform Process Lett 58:277–283

Chernov N, Sapirstein P (2008) Fitting circles to data with correlated noise. Comput Stat Data Anal 52:5328–5337

Coope I (1993) Circle fitting by linear and nonlinear least squares. J Optim Theory Appl 76:381–388

Crawford J (1983) A non-iterative method for fitting circular arcs to measured points. Nucl Instrum Methods 211:223–225

Das G, Mukhopadhyay D, Nandy S (2009) Improved algorithm for the widest empty 1-corner corridor. Inform Process Lett 109:1060–1065

Deshpande A, Rademacher L, Vempala S, Wang G (2006) Matrix approximation and projective clustering via volume sampling. In: Proceedings of the 17th annual ACM-SIAM symposium on discrete algorithms. ACM, New York, pp 1117–1126

Dey T (1998) Improved bounds for planar k-sets and related problems. Discrete Comput Geom 19:373–382

Díaz-Bánez JM, Mesa J, Schöbel A (2004) Continuous location of dimensional structures. Eur J Oper Res 152:22–44

Díaz-Bánez JM, López MA, Sellarès JA (2006a) Locating an obnoxious plane. Eur J Oper Res 173:556–564

Díaz-Bánez JM, López MA, Sellarès JA (2006b) On finding a widest empty 1-corner corridor. Inform Process Lett 98:199–205

Díaz-Bánez J, Korman M, Pérez-Lantero P, Ventura I (2013) The 1-median and 1-highway problem. Eur J Oper Res 225:552–557

Dicks DR (1985) Early Greek astronomy to aristotle (Aspects of Greek and Roman life series). Cornell University Press, Ithaca

Drezner Z, Brimberg J (2014) Fitting concentric circles to measurements. Math Method Oper Res 29:119–133

Drezner T, Drezner Z (2007) Equity models in planar location. Comput Manag Sci 4:1–16

Drezner Z, Suzuki A (2004) The big triangle small triangle method for the solution of non-convex facility location problems. Oper Res 52:128–135

Drezner Z, Klamroth K, Schöbel A, Wesolowsky G (2001) The weber problem, chap 1. In: Drezner Z, Hamacher H (eds) Facility location - applications and theory. Springer, Berlin/Heidelberg, pp 1–36

Drezner Z, Steiner S, Wesolowsky G (2002) On the circle closest to a set of points. Comput Oper Res 29:637–650

Ebara H, Fukuyama N, Nakano H, Nakanishi Y (1989) Roundness algorithms using the voronoi diagrams. In: Proceedings of the 1st Canadian conference on computational geometry, p 41

Edelsbrunner H (1985) Finding transversals for sets of simple geometric figures. Theor Comput Sci 35:55–69

Efrat A, Sharir M (1996) A near-linear algorithm for the planar segment-center problem. Discrete Comput Geom 16:239–257

Espejo I, Rodríguez-Chía A (2011) Simultaneous location of a service facility and a rapid transit line. Comput Oper Res 38:525–538

Espejo I, Rodríguez-Chía A (2012) Simultaneous location of a service facility and a rapid transit line. Comput Oper Res 39:2899–2903

Farago F, Curtis M (1994) Handbook of dimensional measurement, 3rd edn. Industrial Press, New York
Gander W, Golub G, Strebel R (1994) Least-squares fitting of circles and ellipses. BIT 34:558–578
García-López J, Ramos P, Snoeyink J (1998) Fitting a set of points by a circle. Discrete Comput Geom 20:389–402
Gluchshenko O (2008) Annulus and center location problems. Ph.D. thesis, Technische Universität Kaiserslautern
Gluchshenko ON, Hamacher HW, Tamir A (2009) An optimal $o(n \log n)$ algorithm for finding an enclosing planar rectilinear annulus of minimum width. Oper Res Lett 37:168–170
Golub G, van Loan C (1980) An analysis of the total least squares problem. SIAM J Numer Anal 17:883–893
Hamacher H, Nickel S (1995) Restricted planar location problems and applications. Nav Res Log 42:967–992
Har-Peled S, Varadarajan K (2002) Projective clustering in high dimensions using core-sets. In: Proceedings of the 18th annual symposium on computational geometry. ACM, New York, pp 312–318
Helly E (1923) Über Mengen konvexer Körper mit gemeinschaftlichen Punkten. Jahrb Dtsch Math Ver 32:175–176
Houle M, Toussaint G (1985) Computing the width of a set. In: Proceedings of the 1st ACM symposium on computational geometry, pp 1–7
Imai H, Lee D, Yang CD (1992) 1-Segment center problems. ORSA J Comput 4:426–434
Janardan R, Preparata F (1996) Widest-corridos problems. Nord J Comput 1:231–245
Kapelushnik L (2008) Computing the k-centrum and the ordered median hyperplane. Master's thesis, School of Computer Science, Tel-Aviv University
Kasa I (1976) A circle fitting procedure and its error analysis. IEEE T Instrum Meas 25:8–14
Kelachankuttu H, Batta R, Nagi R (2007) Contour line construction for a new rectangular facility in an existing layout with rectangular departments. Eur J Oper Res 180:149–162
Korneenko N, Martini H (1990) Approximating finite weighted point sets by hyperplanes. In: SWAT90. Lecture notes in computer science, vol 447. Springer, pp 276–286
Korneenko N, Martini H (1993) Hyperplane approximation and related topics. In: Pach J (ed) New trends in discrete and computational geometry. Springer, New York, pp 135–162
Körner MC (2011) Minisum hyperspheres. Springer, New York
Körner MC, Brimberg J, Juel H, Schöbel A (2009) General circle location. In: Proceedings of the 21st Canadian conference on computational geometry, pp 111–114
Körner MC, Brimberg J, Juel H, Schöbel A (2011) Geometric fit of a point set by generalized circles. J Global Optim 51:115–132
Körner MC, Martini H, Schöbel A (2012) Minisum hyperspheres in normed spaces. Discrete Appl Math 16:2221–2233
Krempasky T (2012) Locating median lines and hyperplanes with a restriction on the slope. Ph.D. thesis, Universität Göttingen
Le V, Lee D (1991) Out-of-roundness problem revisited. IEEE Trans Pattern Anal 13:217–223
Lee D, Ching Y (1985) The power of geometric duality revisited. Inform Process Lett 21:117–122
Lozano AJ, Plastria F (2009) The ordered median Euclidean straight-line location problem. Stud Locat Anal 17:29–43
Lozano AJ, Mesa J, Plastria F (2010) The k-centrum straight-line location problem. J Math Model Algorithms 9:1–17
Lozano AJ, Mesa J, Plastria F (2013) Location of weighted anti-ordered median straight lines with euclidean distances. Discrete Appl Math. doi:10.1016/j.dam.2013.04.016
Mangasarian O (1999) Arbitrary-norm separating plane. Oper Res Lett 24:15–23
Martini H, Schöbel A (1998) Median hyperplanes in normed spaces—a survey. Discrete Appl Math 89:181–195
Martini H, Schöbel A (1999) A characterization of smooth norms. Geom Dedicata 77:173–183
Megiddo N (1984) Linear programming in linear time when the dimension is fixed. J ACM 31:114–127

Megiddo N, Tamir A (1982) On the complexity of locating linear facilities in the plane. Oper Res Lett 1:194–197
Megiddo N, Tamir A (1983) Finding least-distance lines. SIAM J Algebr Discrete Method 4:207–211
Morris J, Norback J (1980) A simple approach to linear facility location. Transp Sci 14:1–8
Morris J, Norback J (1983) Linear facility location - solving extensions of the basic problem. Eur J Oper Res 12:90–94
Moura L, Kitney R (1992) A direct method for least-squares circle fitting. Comput Phys Commun 64:57–63
Mukherjee J, Sinha Mahapatra PR, Karmakar A, Das S (2013) Minimum-width rectangular annulus. Theor Comput Sci 508:74–80
Narula SC, Wellington JF (1982) The minimum sum of absolute errors regression: a state of the art survey. Int Stat Rev 50:317–326
Nickel S, Puerto J (2005) Location theory: a unified approach. Springer, Berlin/Heidelberg
Nievergelt Y (2002) A finite algorithm to fit geometrically all midrange lines, circles, planes, spheres, hyperplanes, and hyperspheres. Numer Math 91:257–303
Nievergelt Y (2004) Perturbation analysis for circles, spheres, and generalized hyperspheres fitted to data by geometric total least-squares. Math Comput 73:169–180
Nievergelt Y (2010) Median spheres: theory, algorithms, applications. Numer Math 114:573–606
Overmars MH, van Leeuwen J (1981) Maintenance of configurations in the plane. J Comput Syst Sci 23:166–204
Plastria F (1992) GBSSS: the generalized big square small square method for planar single-facility location. Eur J Oper Res 62:163–174
Plastria F (2001) Continuous covering location problems. In: Drezner Z, Hamacher H (eds) Facility location - applications and theory. Springer, Berlin/Heidelberg, pp 1–36
Plastria F, Carrizosa E (2001) Gauge-distances and median hyperplanes. J Optim Theory Appl 110:173–182
Plastria F, Carrizosa E (2012) Minmax-distance approximation and separation problems: geometrical properties. Math Program 132:153–177
Rivlin T (1979) Approximation by circles. Computing 21:1–17
Robert JM (1991) Linear approximation and line transversals. Ph.D. thesis, School of Computer Sciences, McGill University, Montreal
Robert JM, Toussaint G (1994) Linear approximation of simple objects. Comput Geom Theor Appl 4:27–52
Robins G, Shute C (1987) The Rhind mathematical papyrus. An ancient Egyptian text. British Museum
Rockafellar R (1970) Convex analysis. Princeton Landmarks, Princeton
Rorres C, Romano D (1997) Finding the center of a circular starting line in an Ancient Greek stadium. SIAM Rev 39:745–754
Sarkar A, Batta R, Nagi R (2007) Placing a finite size facility with a center objective on a rectangular plane with barriers. Eur J Oper Res 179:1160–1176
Savas S, Batta R, Nagi R (2002) Finite-size facility placement in the presence of barriers to rectilinear travel. Oper Res 50:1018–1031
Schieweck R (2013) Lower bounds for line location problems via demand regions. Technical report 28. Institut für Numerische und Angewandte Mathematik, Universität of Göttingen, Göttingen
Schieweck R, Schöbel A (2012) Properties and algorithms for line location with extensions. In: Proceedings of the 28th European workshop on computational geometry, Assisi, Italy, pp 185–188
Schöbel A (1996) Locating least-distant lines with block norms. Stud Locat Anal 10:139–150
Schöbel A (1997) Locating line segments with vertical distances. Stud Locat Anal 11:143–158
Schöbel A (1998) Locating least distant lines in the plane. Eur J Oper Res 106:152–159
Schöbel A (1999a) Locating lines and hyperplanes—theory and algorithms. Applied optimization series, vol 25. Kluwer, Dordrecht

Schöbel A (1999b) Solving restricted line location problems via a dual interpretation. Discrete Appl Math 93:109–125
Schöbel A (2003) Anchored hyperplane location problems. Discrete Comput Geom 29:229–238
Schöbel A, Scholz D (2010) The Big Cube Small Cube solution method for multidimensional facility location problems. Comput Oper Res 37:115–122
Schömer E, Sellen J, Teichmann M, Yap C (2000) Smallest enclosing cylinders. Algorithmica 27:170–186
Späth H (1997) Least squares fitting of ellipses and hyperbolas. Comput Stat 12:329–341
Späth H (1998) Least-squares fitting with spheres. J Optim Theory Appl 96:191–199
Sun T (2009) Applying particle swarm optimization algorithm to roundness measurement. Expert Syst Appl 36:3428–3438
Suzuki T (2005) Optimal location of orbital routes in a circular city. Presented at ISOLDE X—10th international symposium on locational decisions, Sevilla and Islantilla, 2–8 June 20005
Swanson K, Lee DT, Wu V (1995) An optimal algorithm for roundness determination on convex polygons. Comput Geom Theor Appl 5:225–235
Ventura J, Yeralan S (1989) The minmax center estimation problem. Eur J Oper Res 41:64–72
Wang L, Gordon MD, Zhu J (2006) Regularized least absolute deviations regression and an efficient algorithm for parameter tuning. In: Proceedings of the 6th international conference on data mining. IEEE, New York, pp 690–700
Wesolowsky G (1972) Rectangular distance location under the minimax optimality criterion. Transp Sci 6:103–113
Wesolowsky G (1975) Location of the median line for weighted points. Environ Plan A 7:163–170
Yamamoto P, Kato K, Imai K, Imai H (1988) Algorithms for vertical and orthogonal L_1 linear approximation of points. In: Proceedings of the 4th annual symposium on computational geometry, pp 352–361
Yeralan S, Ventura J (1988) Computerized roundness inspection. Int J Prod Res 26:1921–1935
Zemel E (1984) An O(n) algorithm for the linear multiple choice knapsack problem and related problems. Inform Process Lett 18:123–128

Chapter 8
Facility Location Under Uncertainty

Isabel Correia and Francisco Saldanha da Gama

Abstract In this chapter, we cover some essential knowledge on facility location under uncertainty. We put a major emphasis on modeling aspects related with discrete facility location problems. Different modeling frameworks are discussed. In particular, we distinguish between robust optimization, stochastic programming and chance-constrained models. We also discuss relevant aspects such as solution techniques, multi-stage stochastic programming models, scenario generation, and extensions of basic problems.

Keywords Chance constraints • Robust optimization • Stochastic programming

8.1 Introduction

Many facility location problems involve strategic decisions that must hold for some considerable time. During this time, changes may occur in the underlying conditions. For instance, we may observe an unexpected disruption in the network due to some failure, or we may realize that the values of some parameters (e.g., demand levels) vary in an unpredictable manner. In such cases, it may be desirable to account for uncertainty in advance. This can be accomplished by embedding uncertainty in the models, leading to solutions that somehow anticipate it.

The review papers by Louveaux (1993) and Snyder (2006) show that much work has been done within this topic. The different sources of uncertainty we may observe in a facility location problem have led to the development of different research branches. One of them regards unexpected disruptions in the network structures (e.g., in the facilities or in the transportation channels) and is addressed in detail in Chap. 24. Another important research branch concerns congestion models. In

I. Correia (✉)
Departamento de Matemática, Centro de Matemática e Aplicações, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
e-mail: isc@fct.unl.pt

F. Saldanha da Gama
Departamento de Estatística e Investigação Operacional, Centro de Investigação Operacional, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal
e-mail: fsgama@fc.ul.pt

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_8

this case, the customers' requests for service have a probabilistic behavior and a facility or equipment may be busy when a new request arrives. This is the topic addressed in Chap. 17. In the current chapter, we focus on aspects emerging from uncertainty associated with the parameters of a facility location problem. We show how uncertainty can be embedded in the models built for supporting the decision making process. For illustrative purposes, we consider several well-known facility location problems. We focus on discrete models. This is motivated by the practical relevance that such models have acquired in the recent decades due to many successful applications of facility location theory to areas such as logistics, transportation and routing (see Chap. 1).

In the following sections we assume that the reader is familiar with basic concepts from robust and stochastic optimization. Important references in these fields include Birge and Louveaux (2011) and Shapiro et al. (2009) (for stochastic programming) and Kouvelis and Yu (1997) and Ben-Tal et al. (2009) (for robust optimization).

The remainder of this chapter is organized as follows. In the next section, we discuss general aspects related with uncertainty. In Sect. 8.3, we address robust facility location problems. In Sect. 8.4, we focus on stochastic programming models. Section 8.5 is devoted to chance-constrained problems. In Sect. 8.6 we discuss some challenges and give suggestions for further reading. The chapter ends with a short conclusion.

8.2 Uncertainty Issues

Basic information underlying a facility location problem includes demand levels, travel time or cost for supplying the customers, location of the customers, presence or absence of the customers, and price for the commodities. Uncertainty may occur in one or several of these parameters.

One crucial aspect when dealing with uncertainty regards its representation. First, uncertain parameters may be discrete or continuous. Second, if probabilistic information is available, the uncertain parameters can be represented through random variables. In this case, using the well-known characterization proposed by Rosenhead et al. (1972), we say that we are making a decision under risk and we can resort to stochastic programming models and methods for dealing with the problem. If this is not the case, we are making a decision under uncertainty and a robustness measure is usually considered for evaluating the performance of the system. It is important to note that the existence of a probabilistic description for the uncertainty does not prevent the use of some robustness measures, as it will be detailed in the next section.

We call "scenario" a complete realization of all the uncertain parameters. This notion is independent of whether or not probabilistic information is available. Nevertheless, if uncertain parameters can be represented by random variables, some probability can be associated with each scenario. Depending on the problem, we may have a finite or infinite number of scenarios. As it will be discussed later, this fact has impact on the models and techniques that can be used.

One important feature that influences the type of model to be considered, regards the attitude of the decision maker towards risk. Two attitudes are typically considered: risk neutral and risk averse. In the first case, the decision maker does not take risk into account when making a decision and a linear function is a correct representation of the utility associated with the decision maker. When a probability can be associated with each scenario, a risk neutral decision maker looks for the decision which minimizes the expected cost (or maximizes the expected return or utility). A risk averse decision maker can be associated with a concave utility function (when utility is measured on the vertical axis and monetary value is measured on the horizontal axis). In this case, the decision maker wants to avoid unnecessary risk and the expected value of the future assets is no longer an appropriate objective. Such decision maker may look, for instance, for the solution minimizing the maximum cost across all scenarios.

Finally, in some classes of problems, there is another aspect that influences the mathematical model to be considered: the identification of the *ex ante* and *ex post* decisions. In the first case, we have the here-and-now decisions, i.e., the decisions that must be implemented before uncertainty is revealed; in the second case, we have the decisions to be implemented after uncertainty is disclosed. The latter set of decisions is often used as a reaction to the values observed for the uncertain parameters. In a facility location problem, the location of the facilities is often an *ex ante* decision. This is a consequence of the strategic nature of such decisions in many problems, which imposes their fully implementation before uncertainty is revealed. Regarding the allocation or distribution decisions, they will depend on the specific problem addressed whether they will be *ex ante* or *ex post* decisions. In the following sections we address both situations.

8.3 Robust Facility Location Problems

We start by assuming that uncertainty is appropriately captured by a finite set of scenarios. As mentioned above, each scenario fully determines the value of all the uncertain parameters. If no probabilistic information is available, one possibility for measuring the performance of a system is to use a robustness measure. Two classical objectives are often considered: minmax cost and minmax regret.

For illustrative purposes, we consider a well-known facility location problem: the p-median problem. In this problem, we have a set of demand nodes, J, each of which to be served by one out of p new facilities to be located. The potential

locations for the facilities coincide with the locations of the demand nodes. In its discrete version, the p-median problem can be formulated as follows:

$$\text{Minimize} \quad \sum_{i \in J} \sum_{j \in J} d_j a_{ij} x_{ij} \tag{8.1}$$

$$\text{subject to} \quad \sum_{i \in J} x_{ij} = 1, \quad j \in J \tag{8.2}$$

$$x_{ij} \leq x_{ii}, \quad i \in J,\ j \in J \tag{8.3}$$

$$\sum_{i \in J} x_{ii} = p \tag{8.4}$$

$$x_{ij} \in \{0, 1\}, \quad i \in J,\ j \in J. \tag{8.5}$$

In this formulation, a_{ij} represents the distance or travel time between demand nodes i and j ($i, j \in J$) and d_j is the demand or weight of node j ($j \in J$); x_{ij} is a binary variable equal to 1 if node $j \in J$ is allocated to node $i \in J$ and 0 otherwise; $x_{ii} = 1$ indicates that a facility is located at i. The goal is to minimize the total weighted distance or travel time.

In a p-median problem, uncertainty can occur in the demands (or weights) or in the distances (or travel times). Denote by Ω the finite set of scenarios and by $\omega \in \Omega$ one particular scenario (that fully determines the uncertain parameters). Suppose that the location of the facilities is an *ex ante* decision and the allocation of the customers to the operating facilities is an *ex post* decision. In order to capture uncertainty, we need to consider binary location variables y_i indicating whether a facility is located at $i \in J$, and scenario-indexed binary allocation variables $x_{ij\omega}$ indicating whether demand node $j \in J$ is allocated to facility $i \in J$ in scenario $\omega \in \Omega$. The minmax p-median problem can be formulated as follows:

$$\text{Minimize} \quad v \tag{8.6}$$

$$\text{subject to} \quad \sum_{i \in J} \sum_{j \in J} d_{j\omega} a_{ij\omega} x_{ij\omega} \leq v, \quad \omega \in \Omega \tag{8.7}$$

$$\sum_{i \in J} x_{ij\omega} = 1, \quad j \in J,\ \omega \in \Omega \tag{8.8}$$

$$x_{ij\omega} \leq y_i, \quad i \in J,\ j \in J,\ \omega \in \Omega \tag{8.9}$$

$$\sum_{i \in J} y_i = p \tag{8.10}$$

$$x_{ij\omega} \in \{0, 1\}, \quad i \in J,\ j \in J,\ \omega \in \Omega \tag{8.11}$$

$$y_i \in \{0, 1\}, \quad i \in J. \tag{8.12}$$

In this model, $d_{j\omega}$ represents the demand of node $j \in J$ under scenario $\omega \in \Omega$, and $a_{ij\omega}$ represents the distance (or travel time) between nodes $i \in J$ and $j \in J$ under scenario $\omega \in \Omega$. The minmax objective arises from the combination of (8.6) and (8.7).

The solution provided by the previous model tends to be overly conservative. It reflects a complete aversion of the decision maker towards risk. In fact, by planning for the worst case scenario (the maximum weighted distance occurring across all scenarios), the decision maker may be planning for a scenario which turns out to be very unlikely. A better compromise can be achieved by considering the minmax regret[1] criterion, in which the decision maker chooses the decision that minimizes the maximum regret across all scenarios. The corresponding model is obtained by replacing (8.7) with

$$\sum_{i \in J} \sum_{j \in J} d_{j\omega} a_{ij\omega} x_{ij\omega} - v_{\omega}^{*} \leq v, \quad \omega \in \Omega, \tag{8.13}$$

where v_{ω}^{*} is the optimal value of problem (8.1)–(8.5) solved for scenario $\omega \in \Omega$. Serra and Marianov (1998) consider the above minmax regret model after scaling the demands. In particular, for each scenario, they divide each demand by the total demand under that scenario. The authors also note the well-known fact that when the optimal objective function differs significantly across the different scenarios, the relative regret is a more appropriate robustness measure (see, for instance, Kouvelis and Yu 1997). In this case, (8.13) should be replaced by

$$\frac{\sum_{i \in J} \sum_{j \in J} d_{j\omega} a_{ij\omega} x_{ij\omega} - v_{\omega}^{*}}{v_{\omega}^{*}} \leq v, \quad \omega \in \Omega. \tag{8.14}$$

For this problem, the same authors propose a heuristic approach.

A different problem is addressed by Serra et al. (1996). They consider a firm that wishes to locate p facilities in a competitive environment. The goal is to maximize the minimum market captured in a region where competitors are already operating. The criterion considered corresponds to the "maximization" version of the minmax "cost" criterion discussed above. Uncertainty is assumed for the demand and for the location of the competitors. Again, a heuristic approach is proposed for tackling the problem.

If the allocation of customers to facilities is also an *ex ante* decision, the models above can be easily adapted. In this case, the scenario index should be removed from the allocation variables, i.e., the allocation variables become those introduced in model (8.1)–(8.5). Furthermore, the location variables y_i are no longer necessary, as variables x_{ii} ($i \in J$) can play their role.

[1] In each scenario, the regret of a solution is the difference between the cost of the solution if the scenario occurs and the optimal cost that can be achieved under that scenario (see Kouvelis and Yu 1997 for further details).

The above models work with a finite set of scenarios. In practice, however, this is not always a correct representation for the uncertainty. In many situations, an uncertain parameter can lie in some infinite set. A popular way of capturing uncertainty in these cases is via intervals. In the general context of robust optimization, two types of uncertainty sets are often considered: box and ellipsoidal uncertainty sets (see Ben-Tal et al. 2009, for further details). In the first case, uncertainty is defined by a set of linear constraints; in the second case, quadratic expressions involving the uncertain parameters are used. We illustrate the use of box uncertainty sets considering the uncapacitated facility location problem (UFLP), whose well-known formulation is the following:

$$\text{Minimize} \quad \sum_{i \in I} f_i y_i + \sum_{i \in I} \sum_{j \in J} c_{ij} d_j x_{ij} \tag{8.15}$$

$$\text{subject to} \quad \sum_{i \in I} x_{ij} = 1, \quad j \in J \tag{8.16}$$

$$x_{ij} \leq y_i, \quad i \in I,\ j \in J \tag{8.17}$$

$$y_i \in \{0, 1\}, \quad i \in I \tag{8.18}$$

$$x_{ij} \geq 0, \quad i \in I,\ j \in J. \tag{8.19}$$

In this model, I denotes the set of potential locations for the facilities, J is the set of customers, f_i represents the setup cost for facility $i \in I$, c_{ij} corresponds to the unitary cost for supplying the demand of customer $j \in J$ from facility $i \in I$ and d_j gives the demand of customer $j \in J$. The binary variable y_i indicates whether a facility is installed at $i \in I$, and the continuous variable x_{ij} represents the fraction of the demand of customer $j \in J$ that is supplied from facility $i \in I$.

We consider now a common source of uncertainty in a facility location problem: the demand. Under box uncertainty, each demand level, d_j ($j \in J$), lies in an interval $\mathfrak{B}_j = [\overline{d}_j - \epsilon \Delta_j, \overline{d}_j + \epsilon \Delta_j]$ with $0 \leq \epsilon \leq 1$. The parameter ϵ measures the uncertainty "magnitude"; $\overline{d}_j$ denotes a reference value for the demand of customer $j \in J$, and is commonly referred to as the nominal value for the unknown parameter. Finally, Δ_j is a scaling factor.

A particular case of box uncertainty that we consider for illustrative purposes arises when $\Delta_j = \overline{d}_j$ ($j \in J$), which leads to the intervals $\mathfrak{B}_j = [\overline{d}_j(1 - \epsilon), \overline{d}_j(1 + \epsilon)]$ ($j \in J$). Given these intervals, we can formulate the so-called robust counterpart of model (8.15)–(8.19). Considering an auxiliary variable v, we can rewrite the objective function of the problem as

$$\text{Minimize} \quad v, \tag{8.20}$$

and the following constraint is added to the problem:

$$\sum_{i \in I} f_i y_i + \sum_{i \in I} \sum_{j \in J} c_{ij} d_j x_{ij} \leq v. \tag{8.21}$$

By considering an augmented constraint for (8.21), namely

$$\sum_{i \in I} f_i y_i + \max_{d_j \in \mathfrak{B}_j,\, j \in J} \left\{ \sum_{i \in I} \sum_{j \in J} c_{ij} d_j x_{ij} \right\} \leq v, \tag{8.22}$$

the robust counterpart of (8.21) becomes

$$\sum_{i \in I} f_i y_i + \sum_{i \in I} \sum_{j \in J} c_{ij} \left[\overline{d}_j (1 + \epsilon) \right] x_{ij} \leq v. \tag{8.23}$$

The robust counterpart of (8.15)–(8.19) consists of minimizing (8.20) subject to (8.16)–(8.19), and (8.23).

A drawback of box uncertainty is that it comprises the possibility of having all the uncertain parameters taking simultaneously their worst values. This is often not realistic. Accordingly, other type of uncertainty sets may be more appropriate, leading to less conservative solutions. Ellipsoidal uncertainty arises as an alternative in such cases. Baron et al. (2011) study the use of box and ellipsoidal uncertainty in a facility location problem with a time varying uncertain demand. The location of the facilities and their operating capacity are *ex ante* decisions that should hold for the entire planning horizon, during which the demands must be satisfied. The goal is to maximize the overall profit. Nikoofal and Sadjadi (2010) avoid the most conservative solutions arising from considering box uncertainty by imposing a maximum total scaled variation for the uncertainty parameters. The authors address a p-median problem with interval uncertainty associated with the distances (or travel times). In particular, for each pair $(i, j), i, j \in J$, they assume that a_{ij} can take any value within an interval $[\underline{a}_{ij}, \overline{a}_{ij}]$ previously defined. Additionally, the choices for the values a_{ij} are restricted by the relation $\sum_{i,j \in J,\, i<j} (a_{ij} - \underline{a}_{ij})(\overline{a}_{ij} - \underline{a}_{ij}) \leq L$, where L denotes a maximum level previously imposed for the total scaled variation. This type of relation avoids the situation in which all (or many) parameters take their extreme values simultaneously.

In all problems discussed above, no probabilities were associated with the scenarios. However, in some situations, a probability π_ω can, in fact, be associated with each scenario $\omega \in \Omega$. A well-known robustness measure in this case, is the expected cost, which is equivalent to the expected regret (see Snyder 2006). Current et al. (1997) study a facility location problem consisting of locating a set of p facilities here-and-now, together with the possibility of locating an extra set of facilities during a planning horizon previously defined. The number of facilities to locate during the planning horizon is an outcome of the problem. The authors

compare the solutions obtained using the minmax regret and the expected regret criteria.

When probabilities can be associated with the scenarios, an alternative robustness measure proposed by Snyder and Daskin (2006) is "α-robustness". The idea is to look for a solution minimizing the expected cost/distance but such that the relative regret in each scenario is less or equal than α. In the case of the p-median problem, assuming *ex ante* location decisions and *ex post* allocation of customers to the operating facilities, we obtain the following model:

$$\text{Minimize} \quad \sum_{\omega\in\Omega}\sum_{i\in J}\sum_{j\in J}\pi_\omega d_{j\omega}a_{ij\omega}x_{ij\omega} \tag{8.24}$$

$$\text{subject to} \quad (8.8)\text{–}(8.12)$$

$$\sum_{i\in J}\sum_{j\in J} d_{j\omega}a_{ij\omega}x_{ij\omega} \leq (1+\alpha)v^*_\omega, \quad \omega\in\Omega. \tag{8.25}$$

As pointed out by Snyder and Daskin (2006), this model generalizes the well-known models proposed by Weaver and Church (1983) and Mirchandani et al. (1985). Snyder and Daskin (2006) also apply these ideas to the UFLP. They analyze the complexity of both problems (the α-robustness p-median problem and the α-robustness UFLP) and develop Lagrangian relaxation based approaches in order to compute lower and upper bounds for the problems. The final gaps are closed using branch-and-bound procedures.

All the robustness measures discussed and illustrated above involve all scenarios. When the number of scenarios is too high, the large-scale models obtained may become intractable. In this case, restricting the scenario set may be unavoidable. This was done by Daskin et al. (1997) that introduced the α-reliable minmax regret p-median problem. The authors seek to minimize the maximum regret over a subset of scenarios. This subset is referred to as the reliability set. It is built from the original set in such a way that the total probability associated with its scenarios is at least some pre-specified value α. As pointed out by Baron et al. (2011), this idea has a purpose similar to the use of ellipsoid uncertainty: the exclusion of low-probability (typically extreme) scenarios. An extension of the above robustness measure was introduced by Chen et al. (2006) who introduced the α-reliable mean-excess regret. This measure weights the maximum regret over the reliability set and the conditional expectation of the regret over the scenarios not included in the reliability set.

A different robustness concept was introduced by Carrizosa and Nickel (2003) within the context of continuous facility location, although the concept can be extended to network or discrete problems. In that paper, nominal values are assumed to have been estimated for the (uncertain) weights of a set of nodes. A maximum value is preset for the weighted distance between a single facility to be located and the demand nodes. The robustness of a location is then defined as the minimum deviation of the vector of weights with respect to the nominal vector that turns that location an infeasible solution. The goal of the problem is to find the most robust

location. This yields a non-linear fractional model that the authors tackle by existing methods and by ad-hoc procedures they propose in the paper.

One final aspect worth mentioning in this section regards the relevance of using a model like the ones described above, instead of a "simplified" deterministic model. When probabilities can be associated with the scenarios, we can measure this relevance by using the expected value of perfect information (EVPI). The EVPI indicates how much the decision maker would be willing to pay for getting perfect information. Suppose we have an expected cost minimization problem. In this case, the EVPI is obtained by computing the difference between the weighted sum of the optimal values for all scenarios (using the probabilities as weights) and the minimum expected cost. The reader should refer to Kouvelis and Yu (1997) for further details.

8.4 Stochastic Facility Location Problems

A facility location problem under uncertainty, can often be casted within a stochastic programming modeling framework if uncertainty can be described by some probability distribution. In this case, we say that we are dealing with a stochastic facility location problem.

We start by considering the UFLP (8.15)–(8.19). In practice, several parameters in this model may be uncertain. This is the case of the distribution costs and of the demands. Let us assume that uncertainty can be measured probabilistically. In particular, denote by Ξ the random vector containing all the random parameters (e.g., $\Xi = \left((c_{ij})_{i\in I,\, j\in J}, (d_j)_{j\in J}\right)$). Furthermore, suppose that we know the joint probability distribution of Ξ. Assuming *ex ante* location decisions, the model to be adopted will depend on the *ex post* decisions, namely on the moment in time where allocation or distribution decisions are to be implemented. If we have *ex post* allocation decisions, the following stochastic uncapacitated facility location problem with recourse can be considered:

$$\text{Minimize} \quad \sum_{i\in I} f_i y_i + \mathrm{Q}(y) \tag{8.26}$$

$$\text{subject to} \quad y_i \in \{0,1\}, \quad i \in I, \tag{8.27}$$

with $\mathrm{Q}(y) = \mathbb{E}_{\Xi}\left[Q(y,\xi)\right]$, and $Q(y,\xi)$ denoting the optimal value of the following problem:

$$\text{Minimize} \quad \sum_{i\in I}\sum_{j\in J} c_{ij} d_j x_{ij} \tag{8.28}$$

$$\text{subject to} \quad \sum_{i \in I} x_{ij} = 1, \quad j \in J \tag{8.29}$$

$$x_{ij} \leq y_i, \quad i \in I,\ j \in J \tag{8.30}$$

$$x_{ij} \geq 0, \quad i \in I,\ j \in J. \tag{8.31}$$

Model (8.28)–(8.31) is defined for every realization, ξ, of Ξ, i.e., for every realization of costs and demands. Accordingly, the allocation decisions x_{ij} ($i \in I$, $j \in J$), which do not appear in the first-stage problem, can change with different realizations of the random vector. For this reason, they are referred to as recourse decisions. Regarding the variables associated with the location of the facilities, y_i, they correspond to *ex ante* (first-stage) decisions and thus, they must hold for all possible realizations of the random variables. The expectation defining the recourse function $Q(y)$, implicitly conveys a neutral attitude of the decision maker towards risk. Later in this section, we discuss another possible attitude and the corresponding consequences from a modeling point of view. It is also important to emphasize that constraints (8.30) and (8.31) together assure that at least one facility is installed. Finally, it should be noted that we are dealing with a problem that has relatively complete recourse, i.e., for every first-stage feasible solution, y_i ($i \in I$) there is at least one second-stage feasible completion (solution), x_{ij} ($i \in I$, $j \in J$) for every possible realization of the random quantities.

If we have a finite set of scenarios, say Ω, we can go farther with the above model. In order to do so, we consider scenario-indexed parameters and variables. Denote by $c_{ij\omega}$ the cost for supplying customer $j \in J$ from facility $i \in I$ under scenario $\omega \in \Omega$, and let $d_{j\omega}$ be the demand of customer $j \in J$ under scenario $\omega \in \Omega$. If $x_{ij\omega}$ is the fraction of the demand of customer $j \in J$ satisfied from facility $i \in I$ under scenario $\omega \in \Omega$, then we can consider the following extensive form of the deterministic equivalent:

$$\text{Minimize} \quad \sum_{i \in I} f_i y_i + \sum_{\omega \in \Omega} \pi_\omega \left(\sum_{i \in I} \sum_{j \in J} c_{ij\omega} d_{j\omega} x_{ij\omega} \right) \tag{8.32}$$

$$\text{subject to} \quad (8.28)$$

$$\sum_{i \in I} x_{ij\omega} = 1, \quad j \in J,\ \omega \in \Omega \tag{8.33}$$

$$x_{ij\omega} \leq y_i, \quad i \in I,\ j \in J,\ \omega \in \Omega \tag{8.34}$$

$$x_{ij\omega} \geq 0, \quad i \in I,\ j \in J,\ \omega \in \Omega. \tag{8.35}$$

In the above model, the non-anticipativity principle[2] is implicitly considered: each first-stage decision variable has the same value for all scenarios.

So far, no capacities have been considered for the facilities. When they exist, several adjustments are required. Denote by q_i the capacity of a facility established at $i \in I$. A model for the capacitated stochastic facility location problem is obtained if we replace (8.30) with

$$\sum_{j \in J} d_j x_{ij} \leq q_i y_i, \quad i \in I. \tag{8.36}$$

With the inclusion of these constraints, it may happen that for some first-stage feasible solution, no feasible completion exists in the second stage for one or several realizations of the random vector, i.e., the problem no longer has relatively complete recourse. This feasibility issue adds an extra difficulty to this stochastic programming problem. Infeasibility in the second stage is often an indication of an undesirable first-stage solution. A natural way for addressing this issue is to penalize the non-satisfied demand, which makes sense from a practical point of view. In fact, such penalties correspond, for instance, to costs associated with opportunity losses. Denote by ψ_j the demand of customer $j \in J$ which is not supplied and denote by μ_j the corresponding unitary penalty cost. Note that ψ_j is also a random variable as it depends on the occurring realization of the random vector Ξ. We can still consider the first stage problem (8.26)–(8.27). However, the second stage problem becomes the following:

$$\text{Minimize} \quad \sum_{i \in I} \sum_{j \in J} c_{ij} d_j x_{ij} + \sum_{j \in J} \mu_j \psi_j \tag{8.37}$$

$$\text{subject to} \quad (8.31), (8.36)$$

$$d_j \sum_{i \in I} x_{ij} + \psi_j = d_j, \quad j \in J \tag{8.38}$$

$$\psi_j \geq 0, \quad j \in J. \tag{8.39}$$

Again, if a finite set of scenarios exists, we can consider scenario-indexed recourse variables and parameters, and we can write the deterministic equivalent in its extensive form.

In the capacitated model just described, capacities are exogenous. Louveaux (1986) considers a stochastic facility location problem with endogenous capacities. In particular, capacity decisions are *ex ante* decisions, i.e., the capacities of the facilities must be decided in advance before uncertainty is disclosed. A unitary cost g_i is considered for the capacity to be installed at location $i \in I$. Additionally,

[2] A decision should depend only on the information available at the time it is made (see Rockafellar and Wets 1991).

the author considers the existence of variable production costs at the facilities as well as revenues associated with demand satisfaction. Denote by r_j the unitary revenue obtained from customer $j \in J$. Additionally, assume that c_{ij} ($i \in I$, $j \in J$) includes the production costs. A new decision variable z_i ($i \in I$) must be considered, representing the capacity to be installed at location $i \in I$. Now, it may not be rewarding to satisfy all the demand; the trade-off between revenues and costs will decide the best service level for each customer. The capacitated model formulated above, can be easily adapted to the new conditions, leading to the model proposed by Louveaux (1986):

$$\text{Minimize} \quad \sum_{i \in I} f_i y_i + \sum_{i \in I} g_i z_i + \mathrm{Q}(y, z) \tag{8.40}$$

$$\text{subject to} \quad (8.27)$$

$$z_i \geq 0, \quad i \in I, \tag{8.41}$$

with $\mathrm{Q}(y, z) = \mathbb{E}_{\Xi}[Q(y, z, \xi)]$, and $Q(y, z, \xi)$ denoting the optimal value of the following problem:

$$\text{Minimize} \quad \sum_{i \in I} \sum_{j \in J} \left(c_{ij} - r_j\right) d_j x_{ij} \tag{8.42}$$

$$\text{subject to} \quad \sum_{i \in I} x_{ij} \leq 1, \quad j \in J \tag{8.43}$$

$$(8.30), (8.31)$$

$$\sum_{j \in J} d_j x_{ij} \leq z_i, \quad i \in I. \tag{8.44}$$

Louveaux and Peeters (1992) consider a finite set of scenarios for this problem and propose a dual-based procedure for the extensive form of the deterministic equivalent.

A different type of models emerge when the distribution decisions (represented by x-variables) become first-stage decisions. In this case, penalties are paid in the second stage for excess and shortage inventory. In addition to the notation already presented, we denote by ϕ_j the excess inventory of customer $j \in J$ and by λ_j the corresponding unitary cost. Assuming deterministic distribution costs (as they are associated with an *ex ante* decision), we can formulate the stochastic facility location problem as follows:

$$\text{Minimize} \quad \sum_{i \in I} f_i y_i + \sum_{i \in I} \sum_{j \in J} c_{ij} x_{ij} + \mathrm{Q}(x) \tag{8.45}$$

$$\text{subject to} \quad (8.27), (8.30), (8.31),$$

with $\mathrm{Q}(x) = \mathbb{E}_{\Xi}[Q(x,\xi)]$, and $Q(x,\xi)$ denoting the optimal value of the following problem:

$$\text{Minimize} \quad \sum_{j\in J} \lambda_j \phi_j + \sum_{j\in J} \mu_j \psi_j \tag{8.46}$$

$$\text{subject to} \quad d_j \sum_{i\in I} x_{ij} + \psi_j - \phi_j = d_j, \qquad j \in J \tag{8.47}$$

$$\psi_j, \phi_j \geq 0, \qquad j \in J. \tag{8.48}$$

Capacities can be easily included in the above model leading to the so-called stochastic transportation-location problem which has been investigated by several authors (e.g., França and Luna 1982; Holmberg and Tuy 1999).

So far in this section, we have assumed that the allocation and distribution decisions are made simultaneously, either after or before uncertainty is disclosed. In some problems, these decisions can be made separately. We now consider the situation in which the allocation of the customers to the facilities is a here-and-now decision but the quantities to ship from the facilities to the customers are to be decided after uncertainty is revealed. This situation is motivated, for instance, by logistics applications, when a contract has to be previously signed, determining a priori the distribution channels but leaving the distribution decisions dependent on the observed values of the stochastic parameters. Such case can also occur in companies providing some service and that need to define a priori groups of customers that will be allocated to some server or facility. In this case, we need to explicitly consider allocation decision variables. In particular, we denote by w_{ij} the binary variable equal to 1 if customer $j \in J$ is allocated to facility $i \in I$ and 0 otherwise. The single-allocation version of the problem was introduced by Laporte et al. (1994) and has the following formulation:

$$\text{Minimize} \quad \sum_{i\in I} f_i y_i + \sum_{i\in I}\sum_{j\in J} b_{ij} w_{ij} + \mathrm{Q}(w) \tag{8.49}$$

$$\text{subject to} \quad w_{ij} \leq y_i, \quad i \in I,\ j \in J \tag{8.50}$$

$$\sum_{i\in I} w_{ij} \leq 1, \quad j \in J \tag{8.51}$$

$$y_i, w_{ij} \in \{0,1\}, \quad i \in I,\ j \in J, \tag{8.52}$$

with $\mathrm{Q}(w) = \mathbb{E}_{\Xi}[Q(w,\xi)]$, and $Q(w,\xi)$ denoting the optimal value of the following problem:

$$\text{Minimize} \quad \sum_{i\in I}\sum_{j\in J} \left(c_{ij} - r_j\right) d_j x_{ij} \tag{8.53}$$

$$\text{subject to} \quad x_{ij} \leq w_{ij}, \quad i \in I,\ j \in J \tag{8.54}$$

$$\sum_{j \in J} d_j x_{ij} \leq q_i, \quad i \in I \tag{8.55}$$

$$x_{ij} \geq 0, \quad i \in I,\ j \in J. \tag{8.56}$$

In the above model, b_{ij} is a fixed cost for allocating customer $j \in J$ to facility $i \in I$. The other notation was already introduced before. Note that in this problem, facilities are capacitated. Laporte et al. (1994) consider a finite set of scenarios and solve the extensive form of the deterministic equivalent using the integer L-shaped method previously proposed by Laporte and Louveaux (1993).

In line with the idea of allocating the customers before uncertainty is disclosed, Albareda-Sambola et al. (2011) consider Bernoulli demands, which represent a possible request for some service. This is an example of a problem in which the presence or absence of customers is itself a source of uncertainty. The problem, which we revisit below, is important to show that finding a deterministic equivalent is not always straightforward (or even possible) as the models above could indicate.

In the problem studied by Albareda-Sambola et al. (2011), there is a limited capacity for the facilities in terms of the number of customers that can be served. In particular, for each facility $i \in I$, there is a maximum number of customers, q_i, that can be served from the facility. Due to the uncertainty in the demand, it makes sense to allocate (a priori) to some facility more customers than the service capacity. In the end, it may turn out that a facility has a number of requests for service above its capacity. In this case, outsourcing is considered and the corresponding costs incurred. An important assumption in many logistics systems that the authors also consider is that, for each facility $i \in I$, there should be a minimum number of customers ℓ_i allocated to it to justify its establishment. The problem can be conceptually formulated as follows.

$$\text{Minimize} \quad \sum_{i \in I} f_i y_i + \mathbb{E}_{\Xi}\,[\text{Service cost} + \text{Outsourcing cost}] \tag{8.57}$$

$$\text{subject to} \quad \sum_{i \in I} x_{ij} = 1, \quad j \in J \tag{8.58}$$

$$x_{ij} \leq y_i, \quad i \in I,\ j \in J \tag{8.59}$$

$$\ell_i y_i \leq \sum_{j \in J} x_{ij}, \quad i \in I \tag{8.60}$$

$$y_i, x_{ij} \in \{0, 1\}, \quad i \in I,\ j \in J. \tag{8.61}$$

Denote by ξ_j the demand of customer $j \in J$ that is assumed to follow a Bernoulli distribution with parameter p_j. For each first-stage solution, denote by z_i the number of customers assigned to facility $i \in I$ (i.e., $z_i = \sum_{j \in J} x_{ij}$) and denote by η_i the random variable representing the number of customers that request the service (referred to as demand customers) among those assigned to facility $i \in I$ (i.e.,

$\eta_i = \sum_{j \in J} \xi_j x_{ij}$). Note that the probability distribution of η_i is quite involved as it depends on the actual values of x_{ij} ($j \in J$). Denote by $\mathbb{P}_x(\eta_i = s)$ the probability that η_i is equal to s ($s = 0, \ldots, z_i$).

Albareda-Sambola et al. (2011), consider two possible outsourcing actions. We use one of them to illustrate the difficulties that may arise in formulating a deterministic equivalent. In particular, we consider the so-called customer outsourcing. In this case, when the number of customers allocated to some facility $i \in I$ requesting the service (demand customers) exceeds q_i, $\eta_i - q_i$ customers have to be served directly from an external source. A FIFO policy is assumed for deciding which customers to serve from the facility and which ones to outsource. The cost for supplying each outsourced costumer is denoted by g_i and depends on the facility to which the customer was originally assigned. Denote by $\mathbb{P}_i(s)$ the conditional probability of serving a demand customer assigned to facility $i \in I$ given that the total number of demand customers assigned to facility $i \in I$ is s (i.e., $\eta_i = s$). We have

$$\mathbb{P}_i(s) = \frac{\min\{q_i, s\}}{s} = \begin{cases} 1 & \text{if } s \leq q_i \\ q_i/s & \text{otherwise} \end{cases} \tag{8.62}$$

Due to the fact that the expected value is additive, the recourse function can be written as the sum of the expected service cost plus the expected outsourcing cost. These terms can be computed as follows:

$$\begin{aligned} \mathbb{E}_\xi(\text{service cost}) &= \sum_{i \in I} \sum_{s=0}^{z_i} \mathbb{P}_x(\eta_i = s) \times \mathbb{E}(\text{Service cost} | \eta_i = s) \\ &= \sum_{i \in I} \sum_{s=0}^{z_i} \left[\mathbb{P}_x(\eta_i = s) \sum_{j \in J} \mathbb{P}(\xi_j = 1 | \eta_i = s) \mathbb{P}_i(s) c_{ij} x_{ij} \right], \end{aligned} \tag{8.63}$$

$$\begin{aligned} \mathbb{E}_\xi(\text{Outsourcing cost}) &= \sum_{i \in I} \sum_{s=0}^{z_i} \mathbb{P}_x(\eta_i = s) \times \mathbb{E}_\xi(\text{outsourcing cost} | \eta_i = s) \\ &= \sum_{i \in I} g_i \left(\sum_{s=q_i+1}^{z_i} \mathbb{P}_x(\eta_i = s)(s - q_i) \right). \end{aligned} \tag{8.64}$$

A close look into the above expressions reveals that even for tiny instances, a deterministic equivalent formulated from these expressions becomes intractable. In fact, the number of scenarios is huge even for a small number of customers (note that a scenario is defined not only by the set of customers which request the service but also by the order of the requests when calling for service). Nevertheless, for the homogeneous case, i.e., $p_j = p$, $j \in J$, it is possible to go farther and derive a tractable deterministic equivalent, as we show next.

When all the customers have the same probability of requesting the service, then η_i follows a binomial distribution with parameters z_i and p. Thus, $\mathbb{P}_x(\eta_i = s) = \binom{z_i}{s} p^s (1-p)^{z_i - s}$, $s = 0, \ldots, z_i$. We denote by ζ_{tps} the probability that a binomial random variable with parameters t and p takes the value s. In the homogeneous case, it is straightforward to show that $\mathbb{P}(\xi_j = 1 | \eta_i = s) = s/t$ and consequently (taking also (8.62) into account) that $\mathbb{P}(\xi_j = 1 | \eta_i = s)\mathbb{P}_i(s) = \min\{q_i, s\}/t$, which does not depend on x. Accordingly, the service cost (8.63) can be written as

$$\sum_{i \in I} \sum_{j \in J} \left(c_{ij} x_{ij} \sum_{s=0}^{z_i} \zeta_{z_i ps} \frac{\min\{q_i, s\}}{t} \right).$$

A deterministic equivalent can now be obtained by considering discretized location and allocation variables accounting for the number of customers allocated to a facility. In particular, define y_i^t as a binary variable equal to 1 is a facility is located at $i \in I$ with t customers allocated to it ($t = \ell_i, \ldots, |J|$) and 0 otherwise. Define, also, x_{ij}^t as a binary variable equal to 1 if customer $j \in J$ is allocated to facility $i \in I$ which has t customers allocated to it ($t = \ell_i, \ldots, |J|$). Finally, we can formulate a deterministic equivalent problem:

$$\begin{aligned}
\text{Minimize} \quad & \sum_{i \in I} \sum_{t=\ell_i}^{|J|} y_i^t g_i \left[\sum_{s=q_i+1}^{t} \zeta_{tps}(s - q_i) \right] \\
& + \sum_{i \in I} \sum_{j \in J} \left(c_{ij} \sum_{t=\ell_i}^{|J|} x_{ij}^t \left[\sum_{s=0}^{t} \zeta_{tps} \frac{\min\{q_i, s\}}{t} \right] \right) && (8.65) \\
\text{subject to} \quad & \sum_{i \in I} \sum_{t=\ell_i}^{|J|} x_{ij}^t = 1, \quad j \in J && (8.66) \\
& \sum_{j \in J} x_{ij}^t = t y_i^t, \quad i \in I,\ t = \ell_i, \ldots, |J| && (8.67) \\
& \sum_{t=\ell_i}^{|J|} y_i^t \leq 1, \quad i \in I && (8.68) \\
& y_i^t \in \{0, 1\}, \quad i \in I,\ t = \ell_i, \ldots, |J| && (8.69) \\
& x_{ij}^t \in \{0, 1\}, \quad i \in I,\ j \in J,\ t = \ell_i, \ldots, |J|. && (8.70)
\end{aligned}$$

Albareda-Sambola et al. (2011) show that using a general solver, instances of the problem with a realistic size can be solved within an acceptable CPU time using the model above. The authors also explore the advantages of the homogenous case for the alternative outsourcing action they consider. The reader should refer to their paper for further details.

Recently, Hinojosa et al. (2014) considered a problem with location decisions made at a tactical or operational level, i.e., location decisions are *ex post* decisions. The multi-product problem considered in this paper arises in the context of logistics systems. Like in some of the above problems, the available distribution channels correspond to a decision made before demand is known and result from some contract or option. Furthermore, due to the limited capacity at the facilities, the distribution channels contracted in advance may turn out to be insufficient for covering the demand that occurs. In this case, a penalty is incurred (corresponding, e.g., to a "last minute" and thus more expensive contract, to an outsourcing action, or simply to an opportunity loss cost). The location decisions correspond to the "activation" of existing equipments or facilities from where the commodities will be shipped to the customers. Accordingly, this becomes a decision that can be made only after demand is revealed. The authors formulate the extensive form of the deterministic equivalent and solve it for instances with a realistic size using a general solver.

In all of the above models, the recourse function is the expected value of the second-stage problem. As mentioned before, this conveys a neutral attitude of the decision maker towards risk. Location decisions are often strategic and involve significant investments. Accordingly, a risk-averse attitude towards risk cannot be disregarded as a possibility to be considered. One way of capturing such attitude consists of applying a Markowicz type of approach in which the recourse function is expanded to include a variability measure. Taking, as an example, model (8.26)–(8.31) this consists of defining

$$\mathrm{Q}(y) = \mathbb{E}_{\Xi}\left[Q(y,\xi)\right] - \lambda \mathrm{Var}_{\Xi}\left[Q(y,\xi)\right]. \tag{8.71}$$

Such a modeling framework in facility location is far from new (see Jucker and Carlson 1976). Nevertheless, this type of approach has a clear disadvantage: it often results in a non-linear large-scale mixed-integer model. Different possibilities for overcoming this drawback are discussed by Louveaux (1993).

Stochastic discrete facility location problems have attracted much attention in the recent years. Some papers not mentioned so far include those by Ravi and Sinha (2006), Lin (2009), Wang et al. (2011) and Kiya and Davoudpour (2012).

In the context of logistics with particular emphasis to logistics network design, we can also observe an increasing attention paid to stochastic facility location problems (see Chap. 16 for further details). We can refer, among others, to Aghezzaf (2005), Listeş and Dekker (2005), Mo and Harrison (2005), Romauch and Hartl (2005), Pan and Nagi (2010), Fonseca et al. (2010), and Nickel et al. (2012).

Recently, Alumur et al. (2012) explored the possibility of using a robustness measure within a stochastic programming modeling framework. The authors apply the idea to a hub location problem. Uncertainty is associated with two sets of parameters. In both cases, uncertainty can be captured by a finite set of scenarios. For one set of parameters, probabilistic information is assumed to be known which is not the case for the other set. The authors propose a so-called robust-stochastic model: for each scenario associated with the parameters that have no probabilistic

information associated to them, a stochastic program is formulated, capturing the uncertainty associated with the other set of parameters (those for which probabilistic information exists). Then, a minmax regret formulation is proposed for the overall problem. The reader should refer to the paper for further details.

As in the preceding section, when using a stochastic programming approach, it is important to evaluate its relevance compared to a more simplified deterministic approach. Although no robust measure exists for asserting such relevance, two measures are often used to give an indication of such relevance: the EVPI and the value of the stochastic solution (VSS). The EVPI is computed as described in Sect. 8.3. To obtain it, we have to solve the distributional problem (i.e., to find the optimal value for each scenario). In many cases this is cumbersome, namely when the number of scenarios is high or even infinity. The VSS emerges as an alternative and can be obtained in two steps: (i) the expected value problem is solved. This is the deterministic problem obtained when the random variables are replaced by their expectation; (ii) the stochastic problem is considered and the difference between its optimal value and the value of the solution obtained in (i) is computed. This difference gives the VSS (the reader should refer to Birge and Louveaux 2011 for further details).

8.5 Chance-Constrained Facility Location Problems

One important class of optimization problems under uncertainty includes chance-constrained problems. The idea is that one or several constraints of the problem are not required to always hold. Instead, the decision maker is satisfied if they hold with some given probability. This type of constraints may be of relevance when dealing with reliability issues.

In the particular case of a facility location problem, if demand is uncertain but still the decision maker wants to plan for satisfying all the demand whatever it may turn out to be, the resulting solution may call for an operational capacity much above the demand level that turns out being observed. In such situation, one alternative is to plan for assuring a certain service level, i.e., assuring that with some pre-specified probability, the overall demand does not exceed the capacity of the operating facilities.

In order to exemplify these modeling capability, we consider the classical single-source capacitated facility location problem. Assume that fixed costs are associated with the location of the facilities and also with the allocation of customers to the facilities. Additionally, assume that facility $i \in I$ has capacity q_i, and that demands d_j ($j \in J$) are stochastic. We can formulate a capacitated facility location problem with minimum service level as follows:

$$\text{Minimize} \quad \sum_{i \in I} f_i y_i + \sum_{i \in I} \sum_{j \in J} c_{ij} x_{ij} \tag{8.72}$$

subject to (8.16)–(8.18)

$$\mathbb{P}\left[\sum_{j\in J} d_j x_{ij} \leq q_i y_i\right] \geq \alpha_i, \quad i \in I \tag{8.73}$$

$$x_{ij} \in \{0,1\}, \quad i \in I,\ j \in J. \tag{8.74}$$

In this model, α_i is the minimum probability of having the demand assigned to facility $i \in I$ not exceeding the capacity of the facility. Typically, high values are assumed for α_i (e.g., 0.90 or 0.95).

One important feature in a model like the one above, is the possibility of obtaining a deterministic equivalent formulation with the probabilistic constraints being replaced by deterministic ones. Unfortunately, this is not always a straightforward task. One successful example for the problem above is due to Lin (2009). The author considers independent demands following a Poisson or a Gaussian distribution. For illustrative purposes, we detail the procedure in the former case.

If the demands d_j are independent and follow a Poisson distribution $P(\lambda_j)$, $j \in J$, then the total demand assigned to facility $i \in I$, i.e., $\sum_{j\in J} d_j x_{ij}$ follows a Poisson distribution $P(\mu_i)$ with $\mu_i = \sum_{j\in J} \lambda_j x_{ij}$. Accordingly, (8.73) becomes equivalent to

$$\sum_{\ell=0}^{q_i y_i} e^{-\mu_i} \frac{\mu_i^\ell}{\ell!} \geq \alpha_i, \quad i \in I \tag{8.75}$$

which, in turn, has a deterministic equivalent of the form

$$\sum_{j\in J} \lambda_j x_{ij} \leq \nu_i y_i, \quad i \in I. \tag{8.76}$$

In this model, $\nu_i = \mathbb{E}[\Upsilon]$, where Υ is a random variable following a Poisson distribution with an expectation that is equal to the largest value assuring that $\mathbb{P}(\Upsilon \leq q_i) \geq \alpha_i$. As detailed by Lin (2009), the value ν_i can be easily obtained by a search method in which the mean of Υ is changed until $P(\Upsilon \leq q_i)$ is approximately equal to α_i ($i \in I$). After replacing the probabilistic constraints (8.73) by (8.76) the resulting problem becomes a single-source capacitated facility location problem that can be tackled by any of the available methods for such problem. Lin (2009) also explores the possibility of having independent demands following a Gaussian distribution. In this case, the deterministic equivalent of the probabilistic constraints yields a non-convex feasible region. The author proposes a relaxation for the problem that is then embedded into a heuristic approach.

A well-known facility location problem with chance constraints is the covering-location problem proposed by ReVelle and Hogan (1989). The authors assume that a server may be busy when a customer requests to be served. Let us denote by π the probability that this occurs. In a discrete covering-location problem, we have

a set of potential locations for the facilities (see Chap. 5). A customer is said to be covered if a facility is established within a maximum distance or travel time specified in advance. Accordingly, for each customer, we can find the subset of potential locations for the facilities which cover the customer. The goal is to cover all the demand minimizing the number of facilities installed. The "classical" covering constraints are

$$\sum_{i \in I_j} y_i \geq 1, \quad j \in J, \tag{8.77}$$

where I_j denotes the set of locations covering customer $j \in J$. The probabilistic version of these constraints is the following:

$$\mathbb{P}\left[\text{At least one location is available for serving customer } j\right] \geq \alpha, \quad j \in J. \tag{8.78}$$

These constraints have as a deterministic equivalent,

$$\sum_{i \in I_j} y_i \geq \beta, \tag{8.79}$$

with $\beta = \lceil \ln(1-\alpha)/\ln \pi \rceil$. In fact, the probability that no location among those covering customer $j \in J$ is available to serve the customer immediately is given by $\pi^{\sum_{i \in I_j} y_i}$. Thus, the probability that at least one location covering customer $j \in J$ can serve it immediately is given by $1 - \pi^{\sum_{i \in I_j} y_i}$ which, together with (8.78) leads to the deterministic equivalent just presented.

8.6 Challenges and Further Reading

Despite all the existing work on facility location problems under uncertainty, many challenges still exist. In this section, we provide the reader with some notes on relevant issues not discussed in the previous pages, and we give suggestions for further reading.

8.6.1 Multi-Stage Stochastic Programming Models

In all the stochastic facility location problems discussed above, it was assumed that there is a single moment for uncertainty to be revealed. However, in many situations, this is not the case. Instead, we may observe uncertainty being progressively revealed in more than one occasion. When this happens, the two-stage stochastic programming modeling framework discussed in Sect. 8.4 is no longer appropriate,

and a multi-stage setting is required. Nickel et al. (2012) address one such case by considering a multi-period facility location problem with service level and investment decisions. The demand as well as the rates of return for the investments are uncertain. Uncertainty is captured via a scenario tree. In addition to minimizing the overall cost, the problem seeks to minimize the downside risk.[3] The deterministic equivalent problem is formulated in its extensive form and solved using a general solver. Other works addressing multi-stage stochastic facility location problems include the one by Hernández et al. (2012) which considers a multi-period problem with stochastic demands. The problem consists of determining the locations and dimensions of a preset number of new jails in Chile and determine when and where to expand the existing capacity. The goal is to minimize the total expected costs of the system. A large-scale model is obtained and solved approximately using a heuristic combining branch-and-fix coordination (Alonso-Ayuso et al. 2003) and branch-and-bound. Albareda-Sambola et al. (2013), propose a so-called fix-and-relax coordination approximation procedure for tackling a multi-period facility location problem with uncertainty in the costs and in the customers' requests for service.

Taking the previous works into account, one might think that a stochastic multi-period facility location problem necessarily leads to a multi-stage stochastic programming problem. However, this is not true. In some cases, the strategic multi-period decisions can be seen as first-stage decisions in a two-stage stochastic programming modeling framework. For instance, we may decide here-and-now how the location of the facilities will occur during the entire planning horizon. In the second stage problem, the operational decisions will be made, which can adapt to the different realizations of the uncertainty. Works exploring this possibility include those by Ahmed and Garcia (2004) and Aghezzaf (2005).

8.6.2 Solution Methods

Most facility location problems under uncertainty are NP-hard since they generalize well-known NP-hard problems. In particular, this is true for the discrete problems that have been discussed in this chapter. In these cases, either the size of an instance to be solved is such that the resulting model is manageable by a general solver or one must resort to techniques from combinatorial optimization, such as heuristics and relaxation-based approaches.

Regarding robust facility location problems, the minmax structure often considered makes them harder to solve than the corresponding minsum deterministic problems. The reader can refer to Snyder (2006) for a deeper discussion on this issue. That paper presents a sketch of the procedure typically followed for tackling minmax regret problems. Although some general procedures have been proposed

[3] Measure of how much the return on investment is below a target initially imposed.

for minmax problems (e.g., Mausser and Laguna (1998), for minmax regret linear problems with interval uncertainty) in most cases, specially tailored procedures, exact or approximate, must be developed for efficiently tackling the problems. Analytic results and polynomial time algorithms have also been proposed but only for problems with some underlying structure, such as a network.

As far as stochastic discrete facility location problems are concerned, again, they are often difficult to solve to optimality. Even when the number of scenarios is finite and an extensive form of the deterministic equivalent can be obtained, we often end up with a large-scale mixed-integer linear programming problem not manageable by a general solver. In this case, specific approaches, exact or heuristic, have to be developed for tackling the problems. Laporte et al. (1994) make use of the integer L-shaped method proposed by Laporte and Louveaux (1993) for solving a two-stage stochastic facility location problem with first-stage binary variables. In the context of logistics systems, Alonso-Ayuso et al. (2003) introduce the so-called branch-and-fix coordination scheme, which they consider for solving a stochastic facility location problem. The technique proposed can be used for solving general two-stage stochastic programming problems with binary first-stage variables and both binary and continuous variables in the second stage.

A general approach for multi-stage stochastic mixed-integer linear programming problems was proposed by Escudero et al. (2009, 2010). In those papers, the branch-and-fix coordination scheme proposed by Alonso-Ayuso et al. (2003) was extended in order to solve multi-stage problems with integer variables. As mentioned above, Hernández et al. (2012) embed such approach within a heuristic procedure.

When exact approaches fail to solve the problems, we must resort to approximate procedures. One particular difficulty in stochastic programming arises when the number of scenarios is too large or even infinite. In this case, one possibility is to use a sampling approach. The sample average approximation approach (SAA) introduced by Kleywegt et al. (2001) is one such example which has become quite popular. Applications of this approach to stochastic facility location were proposed by Kiya and Davoudpour (2012), Romauch and Hartl (2005) and Santoso et al. (2005). Sampling approaches have also been proposed for general chance-constrained problems by Luedtke and Ahmed (2008) and Pagnoncelli et al. (2009). The application to facility location problems is still to be explored.

Other algorithms for stochastic programming problems include the generation of cutting planes introduced by Guan et al. (2009) for multi-stage problems, and the dual decomposition based approaches developed by Carrøe and Schultz (1999) and Escudero et al. (2012). To the best of our knowledge, the first type of approach was never applied to stochastic facility location. However, there are several papers proposing dual decomposition based algorithms for problems that include location decisions, namely those by Schütz et al. (2008, 2009). The latter work combines dual decomposition with SAA. In this type of method, the non-anticipativity constraints are explicitly considered in the model and dualized, which allows a scenario-decoupling for the relaxed problem.

8.6.3 Scenario Generation

In this chapter it has often been assumed that uncertainty can be represented by a set of scenarios. In particular, it has been assumed that each scenario fully determines all the uncertain parameters. In practice, defining the scenarios is itself a relevant problem.

In some situations, scenarios are associated with driving forces (e.g., the political conditions in a specific region, economic trends or some technological developments) which, in turn, influence the input of the model that supports the decision making process. In this case, it is up to the decision maker to understand these driving forces and the way they influence the input of the model. This understanding leads to a complete definition of the scenarios. The reader should refer to Kouvelis and Yu (1997) for a deeper discussion on this matter.

In other situations, namely in the context of stochastic programming, scenario generation may be important either to instantiate large deterministic equivalent models or to restrict the set of scenarios in a sampling approach used within a solution procedure. The reader should refer to Dupačová et al. (2003), Høyland and Wallace (2001), Di Domenica et al. (2007) and the references therein for further details.

In the case of facility location problems, a short discussion on scenario generation is presented by Kouvelis and Yu (1997) who discuss the issue in the context of a network with uncertain node weights. Assuming a small set of possible values for the demand of each node, one possibility is to take as a scenario each element of the cartesian product of the sets for all nodes. Nevertheless, this is strongly discouraged since the number of scenarios easily leads to intractable models. Instead, the authors highlight that in many location problems the driving forces mentioned above are the key element inducing uncertainty and thus should be identified and taken into account. Typically, these forces induce high correlation between different parameters. If a small number of such factors is identified, the number of scenarios associated with them should be manageable.

8.6.4 Other Notes

One important research topic in facility location under uncertainty regards location-inventory problems. These are problems in which location decisions are combined with inventory management: uncertainty can hardly be disregarded in a realistic modeling framework. This type of problems that was introduced by Daskin et al. (2002) and extended by Snyder et al. (2007) is of great relevance in complex systems such as those arising in logistics. The reader should refer to Chap. 16 for further details.

Another area with great potential is stochastic location-routing. One such problem was solved by Albareda-Sambola et al. (2007). This is a complex and challenging topic.

Finally, this chapter could not come to an end before a brief reference to continuous and network facility location problems under uncertainty. We did not focus on this type of problems although some significant work has been done and much progress achieved. The reader can refer to Snyder (2006) for a review of the fundamental literature addressing these problems. Some recent works on network facility location under uncertainty include those by Conde (2007), Berman and Drezner (2008), Berman and Wang (2010), Sonmez and Lim (2012), Lim and Sonmez (2013), López-de-los-Mozos et al. (2013), Lu (2013), and Lu and Sheu (2013). Recent references on continuous problems include Blanquero et al. (2011) and Drezner et al. (2012).

8.7 Conclusions

In this chapter we have covered several essential aspects related with discrete facility location under uncertainty. Despite the extensive work reported, the existing literature can still be considered scarce in comparison with the literature devoted to deterministic models. However the relevance of facility location in areas where uncertainty if often unavoidable, such as logistics, routing and transportation, has led to an increased interest in the topic addressed in this chapter. In order to better support many decision making processes, it is important to embed uncertainty in the optimization models and, by doing so, to obtain solutions which can anticipate it. The existing literature shows that despite the advances we have observed, dealing with uncertainty in facility location problems remains a challenging and promising research field.

References

Aghezzaf E (2005) Capacity planning and warehouse location in supply chains with uncertain demands. J Oper Res Soc 56:453–462

Ahmed S, Garcia R (2004) Dynamic capacity acquisition and assignment under uncertainty. Ann Oper Res 124:267–283

Albareda-Sambola M, Fernández E, Laporte G (2007) Heuristic and lower bound for a stochastic location-routing problem. Eur J Oper Res 179:940–955

Albareda-Sambola M, Fernández E, Saldanha da Gama F (2011) The facility location problem with Bernoulli demands. Omega 39:335–345

Albareda-Sambola M, Alonso-Ayuso A, Escudero LF, Fernández E, Pizarro C (2013) Fix-and-relax coordination for a multi-period location-allocation problem under uncertainty. Comput Oper Res 40:2878–2892

Alonso-Ayuso A, Escudero LF, Garín A, Ortuño MT, Pérez G (2003) An approach for strategic supply chain planning under uncertainty based on stochastic 0-1 programming. J Global Optim 26:97–124

Alumur SA, Nickel S, Saldanha da Gama F (2012) Hub location under uncertainty. Transp Res B Methodol 46:529–543

Baron O, Milner J, Naseraldin H (2011) Facility location: a robust optimization approach. Prod Oper Manag 20:772–785

Ben-Tal A, El Ghaoui L, Nemirovski A (2009) Robust optimization. Princeton University Press, Princeton/Oxford

Berman O, Drezner Z (2008) The p-median problem under uncertainty. Eur J Oper Res 189:19–30

Berman O, Wang J (2010) The network p-median problem with discrete probabilistic demand weights. Comput Oper Res 37:1455–1463

Birge JR, Louveaux F (2011) Introduction to stochastic programming, 2nd edn. Springer, New York

Blanquero R, Carrizosa E, Hendrix EMT (2011) Locating a competitive facility in the plane with a robustness criterion. Eur J Oper Res 215:21–24

Carrizosa E, Nickel S (2003) Robust facility location. Math Method Oper Res 58:331–349

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

Chen G, Daskin MS, Shen Z-JM, Uryasev S (2006) The α-reliable mean-excess regret model for stochastic facility location modeling. Nav Res Log 53:617–626

Conde E (2007) Minmax regret location-allocation problem on a network under uncertainty. Eur J Oper Res 179:1025–1039

Current J, Ratick S, ReVelle CS (1997) Dynamic facility location when the total number of facilities is uncertain: a decision analysis approach. Eur J Oper Res 110:597–609

Daskin MS, Hesse SM, ReVelle CS (1997) α-Reliable p-minimax regret: a new model for strategic facility location modeling. Locat Sci 5:227–246

Daskin MS, Coullard CR, Shen Z-JM (2002) An inventory-location model: formulation, solution algorithm and computational results. Ann Oper Res 110:83–106

Di Domenica N, Mitra G, Valente P, Birbilis G (2007) Stochastic programming and scenario generation within a simulation framework: an information systems perspective. Decis Support Syst 42:2197–2218

Drezner Z, Nickel S, Ziegler H-P (2012) Stochastic analysis of ordered median problems. J Oper Res Soc 63:1578–1588

Dupačová J, Gröwe-Kuska N, Römisch W (2003) Scenario reduction in stochastic programming. Math Program A 95:493–511

Escudero LF, Garín MA, Merino M, Pérez G (2009) BFC-MSMIP: an exact branch-and-fix coordination approach for solving multistage stochastic mixed 0-1 problems. TOP 17:96–122

Escudero LF, Garín MA, Merino M, Pérez G (2010) On BFC-MSMIP strategies for scenario cluster partitioning, and twin node family branching selection and bounding for multistage stochastic mixed integer programming. Comput Oper Res 37:738–753

Escudero LF, Garín MA, Pérez G, Unzueta A (2012) Lagrangian decomposition for large-scale two-stage stochastic mixed 0-1 problems. TOP 20:347–374

Fonseca MC, García-Sánchez A, Ortega-Mier M, Saldanha da Gama F (2010) A stochastic bi-objective location model for strategic reverse logistics. TOP 18:158–184

França PM, Luna HPL (1982) Solving stochastic transportation-location problems by generalized Benders decomposition. Transp Sci 16:113–126

Guan Y, Ahmed S, Nemhauser GL (2009) Cutting planes for multistage stochastic integer programs. Oper Res 57:287–298

Hernández P, Alonso-Ayuso A, Bravo F, Escudero LF, Guignard M, Marianov V, Weintraub A (2012) A branch-and-cluster coordination scheme for selecting prison facility sites under uncertainty. Comput Oper Res 39:2232–2241

Hinojosa Y, Puerto J, Saldanha da Gama F (2014) A two-stage stochastic transportation problem with fixed handling costs and a priori selection of the distribution channels. TOP. doi:10.1007/s11750-014-0321-4

Holmberg K, Tuy H (1999) A production-transportation problem with stochastic demand and concave production costs. Math Program 85:157–179

Høyland K, Wallace SW (2001) Generating scenario trees for multistage decision problems. Manag Sci 47:295–307

Jucker JV, Carlson C (1976) The simple plant-location problem under uncertainty. Oper Res 24:1045–1055

Kiya F, Davoudpour H (2012) Stochastic programming approach to re-designing a warehouse network under uncertainty. Transp Res E-LOG 48:919–936

Kleywegt A, Shapiro A, Homem-de-Mello T (2001) The sample average approximation method for stochastic discrete optimization. SIAM J Optim 12:479–502

Kouvelis P, Yu G (1997) Robust discrete optimization. Kluwer Academic Publishers, Dordrecht

Laporte G, Louveaux FV (1993) The integer L-shaped method for stochastic integer programs with complete recourse. Oper Res Lett 13:133–142

Laporte G, Louveaux FV, Van hamme L (1994) Exact solution to a location problem with stochastic demands. Transp Sci 28:95–103

Lim GJ, Sonmez AD (2013) γ-Robust facility relocation problem. Eur J Oper Res 229:67–74

Lin CKY (2009) Stochastic single-source capacitated facility location model with service level requirements. Int J Prod Econ 117:439–451

Listeş O, Dekker R (2005) A stochastic approach to a case study for product recovery network design. Eur J Oper Res 160:268–287

López-de-los-Mozos MC, Puerto J, Rodríguez-Chía AM (2013) Robust mean absolute deviation problems on networks with linear vertex weights. Networks 61:76–85

Louveaux FV (1986) Discrete stochastic location models. Ann Oper Res 6:23–34

Louveaux FV (1993) Stochastic location analysis. Locat Sci 1:127–154

Louveaux FV, Peeters D (1992) A dual-based procedure for stochastic facility location. Oper Res 40:564–573

Lu C-C (2013) Robust weighted vertex p-center model considering uncertain data: an application to emergency management. Eur J Oper Res 230:113–121

Lu C-C, Sheu J-B (2013) Robust vertex p-center model for locating urgent relief distribution centers. Comput Oper Res 40:2128–2137

Luedtke J, Ahmed S (2008) A sample approximation approach for optimization with probabilistic constraints. SIAM J Optim 19:674–699

Mausser HE, Laguna M (1998) A new mixed integer formulation for the maximum regret problem. Int Trans Oper Res 5:389–403

Mirchandani PB, Oudjit A, Wong RT (1985) 'Multidimensional' extensions and a nested dual approach for the m-median problem. Eur J Oper Res 21:121–137

Mo Y, Harrison TP (2005) A conceptual framework for robust supply chain design under demand uncertainty. In: Geunes J, Pardalos PM (eds) Supply chain optimization. Springer, New York, pp 243–263

Nickel S, Saldanha da Gama F, Ziegler H-P (2012) A multi-stage stochastic supply network design problem with financial decisions and risk management. Omega 40:511–524

Nikoofal ME, Sadjadi SJ (2010) A robust optimization model for p-median problem with uncertain edge lengths. Int J Adv Manuf Technol 50:391–397

Pagnoncelli BK, Ahmed S, Shapiro A (2009) Sample average approximation method for chance constrained programming: theory and applications. J Optim Theory Appl 142:399–416

Pan F, Nagi R (2010) Robust supply chain design under uncertain demand in agile manufacturing. Comput Oper Res 37:668–683

Ravi R, Sinha A (2006) Hedging uncertainty: approximation algorithms for stochastic optimization problems. Math. Program A 108:97–114

ReVelle CS, Hogan K (1989) The maximum availability location problem. Transp Sci 23:192–200

Rockafeller R, Wets RJ-B (1991) Scenario and policy aggregation in optimisation under uncertainty. Math Oper Res 16:119–147

Romauch M, Hartl RF (2005) Dynamic facility location with stochastic demands. In: Lupanov OB, Kasim-Zade OM, Chaskin AV, Steinhöfel K (eds) Stochastic algorithms: foundations and applications. Springer, Berlin Heidelberg, pp 180–189

Rosenhead J, Elton M, Gupta SK (1972) Robustness and optimality as criteria for strategic decisions. Oper Res Q 23:413–431

Santoso T, Ahmed S, Goetschalckx M, Shapiro A (2005) A stochastic programming approach for supply chain network design under uncertainty. Eur J Oper Res 167:96–115

Schütz P, Stougie L, Tomasgard A (2008) Stochastic facility location with general long-run costs and convex short-run costs. Comput Oper Res 35:2988–3000

Schütz P, Tomasgard A, Ahmed S (2009) Supply chain design under uncertainty using sample average approximation and dual decomposition. Eur J Oper Res 199:409–419

Serra D, Marianov V (1998) The p-median problem in a changing network: the case of Barcelona. Locat Sci 6:383–394

Serra D, Ratick S, ReVelle CS (1996) The maximum capture problem with uncertainty. Environ Plan 23:49–59

Shapiro A, Dentcheva D, Ruszczyński A (2009) Lectures on stochastic programming: modeling and theory. MPS-SIAM series on optimization. SIAM-MPS, Philadelphia

Snyder L (2006) Facility location under uncertainty: a review. IIE Trans 38:537–554

Snyder L, Daskin MS (2006) Stochastic p-robust location problems. IIE Trans 38:971–985

Snyder L, Daskin MS, Teo C-P (2007) The stochastic location model with risk pooling. Eur J Oper Res 179:1221–1238

Sonmez AD, Lim GJ (2012) A decomposition approach for facility location and relocation problem with uncertain number of future facilities. Eur J Oper Res 218:327–338

Wang X, Xu D, Zhao XD (2011) A primal-dual approximation algorithm for stochastic facility location problem with service installation costs. Front Math China 6:957–964

Weaver JR, Church RL (1983) Computational procedures for location problems on stochastic networks. Transp Sci 17:168–180

Chapter 9
Location Problems with Multiple Criteria

Stefan Nickel, Justo Puerto, and Antonio M. Rodríguez-Chía

Abstract This chapter analyzes multicriteria continuous, network, and discrete location problems. In the continuous framework, we provide a complete description of the set of weak Pareto, Pareto, and strict Pareto locations for a general Q-criteria location problem based on the characterization of three criteria problems. In the network case, the set of Pareto locations is characterized for general networks as well as for tree networks using the concavity and convexity properties of the distance function on the edges. In the discrete setting, the entire set of Pareto locations is characterized using rational generating functions of integer points in polytopes. Moreover, we describe algorithms to obtain the solutions sets (the different Pareto locations) using the above characterizations. We also include a detailed complexity analysis. A number of references has been cited throughout the chapter to avoid the inclusion of unnecessary technical details and also to be useful for a deeper analysis.

Keywords Level curves • Networks • Pareto locations • Pareto-optimal • Rational functions • Tree networks

9.1 Introduction

Very often, locational decisions involve the investment of a significant amount of money. It will be therefore very probable that a locational decision is made by a group of Q decision makers (DM). In turn, it is very likely that each DM will choose a median function to evaluate the quality of a new location, but the weights assigned

S. Nickel (✉)
Institute for Operations Research, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Fraunhofer Institute for Industrial Mathematics (ITWM), Kaiserslautern, Germany
e-mail: stefan.nickel@kit.edu

J. Puerto
IMUS, Universidad de Sevilla, Sevilla, Spain
e-mail: puerto@us.es

A.M. Rodríguez-Chía
Universidad de Cádiz, Cádiz, Spain
e-mail: antonio.rodriguezchia@uca.es

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_9

to clients may differ a lot. The same scenario occurs if one location for different types of goods has to be found.

Multicriteria analysis of location problems has received considerable attention within the scope of continuous, network, and discrete models in the last years. For an overview of general methods as well as for a more bibliographic overview of the related location literature the reader is referred to Ehrgott (2005) and Nickel et al. (2005a). Presently, there are several problems that are accepted as classical ones: the point-objective problem (see, e.g., Wendell and Hurter 1973; Hansen et al. 1980; Carrizosa et al. 1993), the continuous multicriteria min-sum facility location problem (see, e.g., Hamacher and Nickel 1996; Puerto and Fernández 1999), the network multicriteria median location problem (see, for instance, Hamacher et al. 1999; Wendell et al. 1977) and the multicriteria discrete location problem (see, e.g., Fernández and Puerto 2003), among others.

In contrast to problems with only one objective, we do not have a natural ordering in higher dimensional objective spaces. Therefore, in multicriteria optimization one has to decide which concept of "optimality" to choose.

The goal in a multicriteria location problem is to optimize simultaneously a set of objective functions $(f^1, \ldots, f^Q)$. Therefore, the formulation of the problem is:

$$v - \min_{x \in X \subseteq \mathbb{R}^d} (f^1(x), \ldots, f^Q(x)), \tag{9.1}$$

where $v-$ min stands for vectorial optimization. Observe that we get points in a Q-dimensional objective space where we do not have the canonical order of $\mathbb{R}$ anymore. Accordingly, for this type of problems, different concepts of solution have been proposed in the literature (the reader is referred to Ehrgott (2005) as a general reference in multicriteria optimization). A point $x \in \mathbb{R}^d$ is called a Pareto location (or Pareto-optimal) if there exists no $y \in \mathbb{R}^d$ such that $f^q(y) \leq f^q(x) \quad \forall q \in \mathcal{Q} := \{1, \ldots, Q\}$ and $f^p(y) < f^p(x)$ for some $p \in \mathcal{Q}$. We denote the set of Pareto solutions by $\mathcal{X}^*_{\text{Par}}\left(f^1, \ldots, f^Q\right)$ or simply by $\mathcal{X}^*_{\text{Par}}$ if this is possible without causing confusion. If $f^q(x) \leq f^q(x')\ \forall\, q \in \mathcal{Q}$ and $\exists q \in \mathcal{Q} : f^q(x) < f^q(x')$ we say that x dominates x' in the decision space and $f(x)$ dominates $f(x')$ in the objective space.

Alternative solution concepts are weak Pareto-optimality and strict Pareto-optimality. A point $x \in \mathbb{R}^d$ is called a weak Pareto location (or weakly Pareto-optimal) if there exists no $y \in \mathbb{R}^d$, such that $f^q(y) < f^q(x)\ \forall\, q \in \mathcal{Q}$. We denote the set of weak Pareto solutions by $\mathcal{X}^*_{\text{w-Par}}\left(f^1, \ldots, f^Q\right)$ or simply by $\mathcal{X}^*_{\text{w-Par}}$ if this is possible without causing confusion. A point $x \in \mathbb{R}^d$ is called a strict Pareto location (or strictly Pareto-optimal) if there exists no $y \in \mathbb{R}^d$, $y \neq x$, such that $f^q(y) \leq f^q(x)\, \forall\, q \in \mathcal{Q}$. Analogously, the set of strict Pareto solutions is denoted by $\mathcal{X}^*_{\text{s-Par}}\left(f^1, \ldots, f^Q\right)$, or simply by $\mathcal{X}^*_{\text{s-Par}}$ if this is possible without causing confusion. Note that $\mathcal{X}^*_{\text{s-Par}} \subseteq \mathcal{X}^*_{\text{Par}} \subseteq \mathcal{X}^*_{\text{w-Par}}$ and in case we are considering strictly convex functions these three sets coincide. Finally, we recall that Warburton (1983) proved the connectedness of the set $\mathcal{X}^*_{\text{Par}}$ when the functions are convex.

In our proofs we use the concept of level sets. For a function $f : \mathbb{R}^d \to \mathbb{R}$ the level set for a value $\rho \in \mathbb{R}$ is given by $L_{\leq}(f, \rho) := \{x \in \mathbb{R}^d : f(x) \leq \rho\}$ (the strict level set is $L_{<}(f, \rho) := \{x \in \mathbb{R}^d : f(x) < \rho\}$) and the level curve for a value $\rho \in \mathbb{R}$ is given by $L_{=}(f, \rho) := \{x \in \mathbb{R}^d : f(x) = \rho\}$. For a function $f^i(\cdot)$ we use the notation

$$\mathscr{X}^*(f^i) := \arg\min_{x \in \mathbb{R}^d} f^i(x).$$

For two points x and y we denote the segment defined by x and y as $\overline{xy}$.

In this chapter we focus on some fundamental results in the continuous, network and discrete cases. We will describe in some detail a complete geometric characterization for the planar 1-facility case, an optimal time algorithm for the 1-facility network problem as well as the computation of the entire set of Pareto-optimal solutions of the discrete multicriteria p-median problem. Although we are concentrating on the median case we will give some outlook to extensions.

9.2 1-Facility Planar/Continuous Location Problems

In this section we study Problem (9.1) where $f^1(\cdot), \ldots, f^Q(\cdot)$ are convex, inf-compact functions, defined in $\mathbb{R}^2$, which represent different criteria or scenarios. Recall that a real function $f(\cdot)$ is said to be inf-compact if its lower level sets $\{x \in \mathbb{R}^d : f(x) \leq \rho\}$ are compact for any $\rho \in \mathbb{R}$. The next result states a useful characterization of the different solution sets defined in the previous section using level sets and level curves which will be used later.

Theorem 9.1 *The following characterizations hold:*

$$x \in \mathscr{X}^*_{\text{w-Par}}\left(f^1, \ldots, f^Q\right) \Leftrightarrow \bigcap_{q=1}^{Q} L_{<}(f^q, f^q(x)) = \emptyset \tag{9.2}$$

$$x \in \mathscr{X}^*_{\text{Par}}\left(f^1, \ldots, f^Q\right) \Leftrightarrow \bigcap_{q=1}^{Q} L_{\leq}(f^q, f^q(x)) = \bigcap_{q=1}^{Q} L_{=}(f^q, f^q(x)) \tag{9.3}$$

$$x \in \mathscr{X}^*_{\text{s-Par}}\left(f^1, \ldots, f^Q\right) \Leftrightarrow \bigcap_{q=1}^{Q} L_{\leq}(f^q, f^q(x)) = \{x\}. \tag{9.4}$$

Proof If $x \notin \mathscr{X}^*_{\text{w-Par}}\left(f^1,\ldots,f^Q\right)$, there exists $z \in \mathbb{R}^2$ such that $f^q(z) < f^q(x)$ for each $q \in \mathcal{Q}$, that means,

$$z \in \bigcap_{q=1}^{Q} L_<(f^q, f^q(x)).$$

Hence, we obtain that

$$\bigcap_{q=1}^{Q} L_<(f^q, f^q(x)) \neq \emptyset.$$

Since the implications above can be reversed the proof is concluded. The remaining results can be proved analogously. □

Remark 9.1 For the case $Q = 2$ the previous result states that the set $\mathscr{X}^*_{\text{w-Par}}(f^1, f^2)$ coincides with tangential cusps between the level curves of functions $f^1(\cdot)$ and $f^2(\cdot)$ union with $\mathscr{X}^*(f^1) \cup \mathscr{X}^*(f^2)$ (see Example 9.1).

Corollary 9.1 *If $f^1,\ldots,f^Q$ are strictly convex functions then*

$$\mathscr{X}^*_{\text{w-Par}}(f^1,\ldots,f^Q) = \mathscr{X}^*_{\text{Par}}\left(f^1,\ldots,f^Q\right) = \mathscr{X}^*_{\text{s-Par}}\left(f^1,\ldots,f^Q\right).$$

Example 9.1 (See Fig. 9.1) Let us consider the points $a_1 = (0,0)$, $a_2 = (8,3)$, $a_3 = (-3,5)$ and the functions $f^1(x) = \|x-a_1\|_1$, $f^2(x) = \|x-a_2\|_\infty$, $f^3(x) = \|x-a_3\|_1$. By Theorem 9.1, $\mathscr{X}^*_{\text{w-Par}}(f^1, f^2)$ is the rectilinear thick path joining a_1 and a_2 and $\mathscr{X}^*_{\text{w-Par}}(f^1, f^3)$ is the dark rectangle with a_1 and a_3 as opposite vertices.

In what follows, since we are dealing with general convex, inf-compact functions, we will focus on providing information about the geometrical structure of $\mathscr{X}^*_{\text{w-Par}}(f^1, f^2, f^3)$. This characterization will allow us to obtain a geometrical description of $\mathscr{X}^*_{\text{Par}}\left(f^1, f^2, f^3\right)$ and $\mathscr{X}^*_{\text{s-Par}}\left(f^1, f^2, f^3\right)$ in the next section for an important family of functions. Actually, we will characterize $\mathscr{X}^*_{\text{w-Par}}(f^1, f^2, f^3)$ as a kind of hull delimited by the chains of bicriteria solutions of any pair of functions f^p, f^q $p,q = 1,2,3$. This result enables us to obtain the set $\mathscr{X}^*_{\text{w-Par}}\left(f^1,\ldots,f^Q\right)$ by union of three-criteria solution sets already characterized. In order to do that, let

$$C_\infty(\mathbb{R}_0^+, \mathbb{R}^2) := \left\{\varphi \mid \varphi : \mathbb{R}_0^+ \to \mathbb{R}^2, \varphi \text{ continuous}, \lim_{t\to\infty} \|\varphi(t)\|_2 = \infty\right\},$$

where $\|x\|_2$ is the Euclidean norm of the point x. $C_\infty(\mathbb{R}_0^+, \mathbb{R}^2)$ is the set of continuous curves, which map the set of non-negative numbers $\mathbb{R}_0^+ := [0,\infty)$ into the two-dimensional space $\mathbb{R}^2$ and whose image $\varphi(\mathbb{R}_0^+)$ is unbounded in $\mathbb{R}^2$. These curves are introduced to characterize the geometrical locus of the points surrounded by weak-Pareto and Pareto chains.

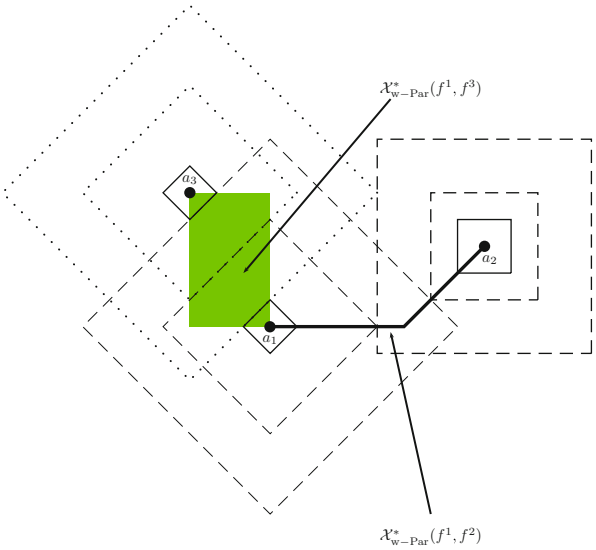


Fig. 9.1 Illustration of Example 9.1

For a set $S \subseteq \mathbb{R}^2$ we define the enclosure of S by

$$\begin{aligned}\text{encl}\,(S) := \big\{x \in \mathbb{R}^2 \,:\, &\exists\, \varepsilon > 0 \text{ with } B(x,\varepsilon) \cap S = \emptyset\,,\ \exists\, t_\varphi \in [0,\infty) \text{ with}\\ &\varphi(t_\varphi) \in S \text{ for all } \varphi \in C_\infty(\mathbb{R}_0^+, \mathbb{R}^2) \text{ with } \varphi(0) = x\big\}\,,\end{aligned}$$

where $B(x,\varepsilon) = \{y \in \mathbb{R}^2 \,:\, \|y - x\|_2 \le \varepsilon\}$. Note that $S \cap \text{encl}\,(S) = \emptyset$. Informally, encl (S) contains all the points which are surrounded by S, but do not belong themselves to S.

We denote the union of the bicriteria chains of weak-Pareto solutions by

$$\mathcal{X}^{\text{gen}}_{\text{w-Par}}\left(f^1, f^2, f^3\right) := \bigcup_{p=1}^{2} \bigcup_{q=p+1}^{3} \mathcal{X}^*_{\text{w-Par}}\left(f^p, f^q\right).$$

We use "gen" since this set will generate the set $\mathcal{X}^*_{\text{w-Par}}\left(f^1, f^2, f^3\right)$. The next theorem provides useful geometric information to build $\mathcal{X}^*_{\text{w-Par}}\left(f^1, f^2, f^3\right)$. Its proof can be found in Rodríguez-Chía and Puerto (2002).

Theorem 9.2

$$\mathcal{X}^*_{\text{w-Par}}(f^1, f^2, f^3) = \text{encl}\left(\mathcal{X}^{\text{gen}}_{\text{w-Par}}\left(f^1, f^2, f^3\right)\right) \cup \mathcal{X}^{\text{gen}}_{\text{w-Par}}\left(f^1, f^2, f^3\right).$$

Remark 9.2 It is worth noting that the region $encl\Big(\mathscr{X}^{\text{gen}}_{\text{w-Par}}\left(f^1, f^2, f^3\right)\Big)$ is well-defined because the set $\mathscr{X}^{\text{gen}}_{\text{w-Par}}\left(f^1, f^2, f^3\right)$ is connected (see Warburton 1983).

As an illustration of the above result we present the following example.

Example 9.2 Let us consider three points $a_1 = (0, 0)$, $a_2 = (3, -1)$ and $a_3 = (3, 3)$, and the functions $f^1(\cdot)$, $f^2(\cdot)$ and $f^3(\cdot)$ such that,

$$L_{\leq}(f^1, 1) = \left\{(x_1, x_2) \ : \ \frac{x_1^2}{4} + \frac{x_2^2}{9} \leq 1\right\}$$

$$L_{\leq}(f^2, 1) = \left\{(x_1, x_2) \ : \ (x_1 - 3)^2 + (x_2 + 1)^2 \leq 1\right\}$$

$$L_{\leq}(f^3, 1) = \left\{(x_1, x_2) \ : \ \frac{(x_1 - 3)^2}{9} + \frac{(x_2 - 3)^2}{4} \leq 1\right\}.$$

We can see that these three functions are convex functions. Therefore by the previous result we obtain the geometrical characterization of the set $\mathscr{X}^*_{\text{w-Par}}(f^1, f^2, f^3)$; this set is the shadowed region in Fig. 9.2.

Now we are in the right position to show the main result about the geometrical structure of $\mathscr{X}^*_{\text{w-Par}}(f^1, \ldots, f^Q)$.

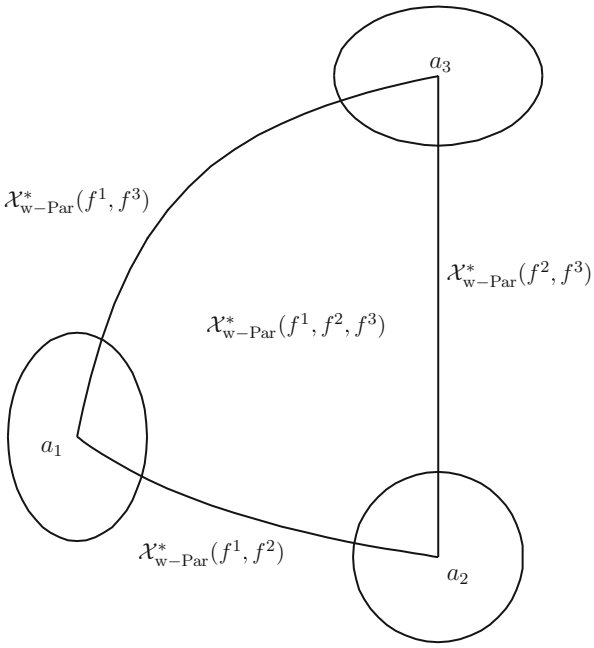


Fig. 9.2 Illustration of Example 9.2

Theorem 9.3

$$\mathscr{X}^*_{\text{w-Par}}(f^1,\ldots,f^Q) = \bigcup_{\substack{p,q,r\in\mathscr{Q}\\ p<q<r}} \mathscr{X}^*_{\text{w-Par}}(f^p,f^q,f^r).$$

Proof By Theorem 9.1, $x \in \mathscr{X}^*_{\text{w-Par}}(f^1,\ldots,f^Q)$ if and only if $\bigcap_{q\in\mathscr{Q}} L_<(f^q, f^q(x)) = \emptyset$. Furthermore, by Helly's theorem (see Rockafellar 1970), this intersection is empty if and only if there exist $p,q,r \in \mathscr{Q}$ $(p<q<r)$ such that $L_<(f^p,f^p(x)) \cap L_<(f^q,f^q(x)) \cap L_<(f^r,f^r(x)) = \emptyset$ and this is equivalent to $x \in \mathscr{X}^*_{\text{w-Par}}(f^p,f^q,f^r)$. Since in any case we have that

$$\bigcup_{\substack{p,q,r\in\mathscr{Q}\\ p<q<r}} \mathscr{X}^*_{\text{w-Par}}(f^p,f^q,f^r) \subset \mathscr{X}^*_{\text{w-Par}}(f^1,\ldots,f^Q),$$

the result follows. □

Remark 9.3 This result extends previous characterizations in the literature:

- Taking $f^i(x) = \|x-a_i\|$ with $a_i \in \mathbb{R}^2$ for $i=1,\ldots,Q$ and $\|\cdot\|$ being a strictly convex norm or a norm derived from a scalar product, we get Proposition 1.3, Theorem 4.3 and Corollary 4.1 in Durier and Michelot (1986). The set of weakly efficient locations is the convex hull of the points a_i with $i=1,\ldots,Q$. In Example 9.3, we illustrate this result.
- Taking $f^i(x) = \|x-a_i\|$ with $a_i \in \mathbb{R}^2$ for $i=1,\ldots,Q$ and $\|\cdot\|$ being a polyhedral gauge we get Theorem 6.1 in Durier (1990), where the set of weakly efficient locations is the union of elementary convex sets, (see Durier and Michelot 1985 for a definition). In Example 9.4, we illustrate this result.
- Taking $f^i(x) = \max_{j\in\mathscr{M}} w^i_j\|x-a_j\|$ with $a_j \in \mathbb{R}^2$, $w^i_j > 0$ for $i=1,\ldots,Q$, $j \in \mathscr{M} := \{1,\ldots,m\}$ and $\|\cdot\|$ being the ℓ_∞-norm, we get Theorem 6.1 in Hamacher and Nickel (1996), where the set of weakly efficient locations is the union of the sets of weakly efficient locations for all pairs of functions. In Example 9.5, we illustrate the use of this result.

Example 9.3 (See Fig. 9.3) Let us consider the points $a_1=(0,0)$, $a_2=(5,-10)$, $a_3=(10,0)$ and the functions $f^i(x)=\|x-a_i\|_2$ for $i=1,2,3$. By Theorem 9.2, $\mathscr{X}^*_{\text{w-Par}}(f^1,f^2,f^3)$ is the dark region, which in this case is the convex hull of a_1, a_2 and a_3.

Example 9.4 (See Fig. 9.4) Let us consider the points $a_1=(0,0)$, $a_2=(8,3)$, $a_3=(-3,5)$ and the functions $f^1(x)=\|x-a_1\|_1$, $f^2(x)=\|x-a_2\|_\infty$ and $f^3(x)=\|x-a_3\|_1$. By Theorem 9.1, $\mathscr{X}^*_{\text{w-Par}}(f^1,f^2)$ is the thick path joining a_1 and a_2, $\mathscr{X}^*_{\text{w-Par}}(f^2,f^3)$ is the thick path joining a_2 and a_3, and $\mathscr{X}^*_{\text{w-Par}}(f^1,f^3)$ is the dark rectangle with a_1 and a_3 as opposite extreme points. Therefore, by Theorem 9.2, $\mathscr{X}^*_{\text{w-Par}}(f^1,f^2,f^3)$ is the dark region surrounded by the union of

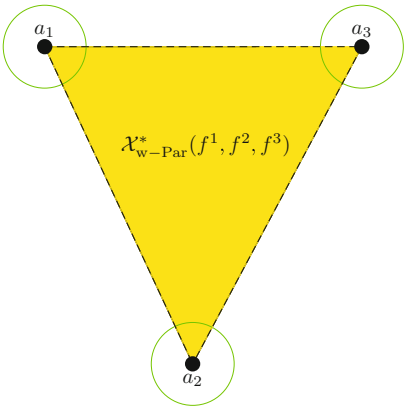


Fig. 9.3 Illustration of Example 9.3

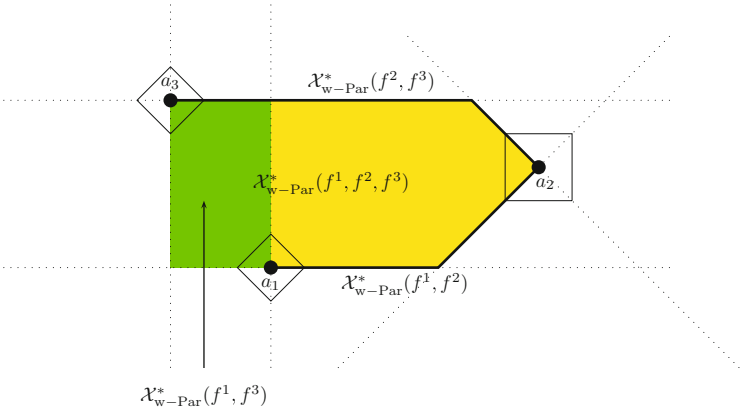


Fig. 9.4 Illustration of Example 9.4

the three previous sets. Note that this region is the union of two full dimensional elementary convex sets.

Example 9.5 (See Fig. 9.5) Let us consider the points $a_1 = (4, 16)$, $a_2 = (10, 5)$, $a_3 = (25, 12)$ and the functions $f^i(x) = \|x - a_i\|_\infty$ for $i = 1, 2, 3$. By Theorem 9.1, $\mathcal{X}^*_{\text{w-Par}}(f^1, f^2) = R_1$, $\mathcal{X}^*_{\text{w-Par}}(f^1, f^3) = R_2 \cup R_4$, $\mathcal{X}^*_{\text{w-Par}}(f^2, f^3) = R_3 \cup R_4$. By Theorem 9.2, $\mathcal{X}^*_{\text{w-Par}}(f^1, f^2, f^3) = R_1 \cup R_2 \cup R_3 \cup R_4$. Note that in this example $\mathcal{X}^*_{\text{w-Par}}(f^1, f^2, f^3) = \mathcal{X}^*_{\text{w-Par}}(f^1, f^2) \cup \mathcal{X}^*_{\text{w-Par}}(f^1, f^3) \cup \mathcal{X}^*_{\text{w-Par}}(f^2, f^3)$.

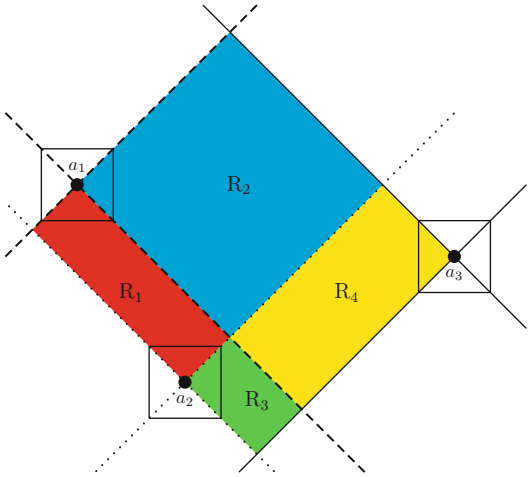


Fig. 9.5 Illustration of Example 9.5

9.2.1 *Polyhedral Planar Minisum Location Problems*

Consider a set of demand points $A := \{a_1, \ldots, a_M\} \subseteq \mathbb{R}^2$. Let $B_i \subset \mathbb{R}^2$, for $i \in \mathscr{M} := \{1, 2, \ldots, M\}$, be a compact, convex set containing the origin in its interior. The gauge with respect to B_i is defined as $\gamma_i : \mathbb{R}^2 \to \mathbb{R}$, $\gamma_i(x) := \inf\{r > 0 : x \in rB_i\}$. Taking this definition into account, the planar minisum location problem is

$$\min_{x \in \mathbb{R}^2} \sum_{i=1}^{M} w_i \gamma_i(x - a_i),$$

where w_i is a nonnegative weight associated with the demand point a_i ($i \in \mathscr{M}$).

In this section we study the particular case where the functions $f^1, \ldots, f^Q$ are minisum location objective functions and the distances are measured with polyhedral gauges, i.e., the unit balls associated with these gauges are convex polytopes. This type of objective function is not strictly convex and for this reason, the three solutions sets (Pareto, weak Pareto and strict Pareto locations) do not coincide. Therefore, in this section we focus on the characterization of the Pareto locations and how it can be extended to the remaining solution sets.

The polar set B_i^o of B_i is given by $B_i^o := \{p \in \mathbb{R}^2 : \langle p, x\rangle \leq 1 \; \forall x \in B_i\}$ and the normal cone to B_i at x is given by $N(B_i, x) := \{p \in \mathbb{R}^2 : \langle p, y - x\rangle \leq 0 \forall y \in B_i\}$, where $\langle \cdot, \cdot \rangle$ denotes the scalar product. In case of polyhedral gauges (i.e., B_i is a polytope), the set of extreme points of B_i is denoted by $\text{Ext}(B_i) := \{e_1^i, \ldots, e_{G_i}^i\}$. The maximal number of extreme points is denoted by $G_{\max} := \max\{G_i : i \in \mathscr{M}\}$. We define fundamental directions $d_1^i, \ldots, d_{G_i}^i$ as the half-lines determined by 0 and $e_1^i, \ldots, e_{G_i}^i$ (see Fig. 9.6).

Let $\pi = (p_i)_{i \in \mathscr{M}}$ be a family of elements of $\mathbb{R}^2$ such that $p_i \in B_i^o$ for each $i \in \mathscr{M}$ and let $C_\pi = \bigcap_{i \in \mathscr{M}} (a_i + N(B_i^o, p_i))$. According to Durier and Michelot

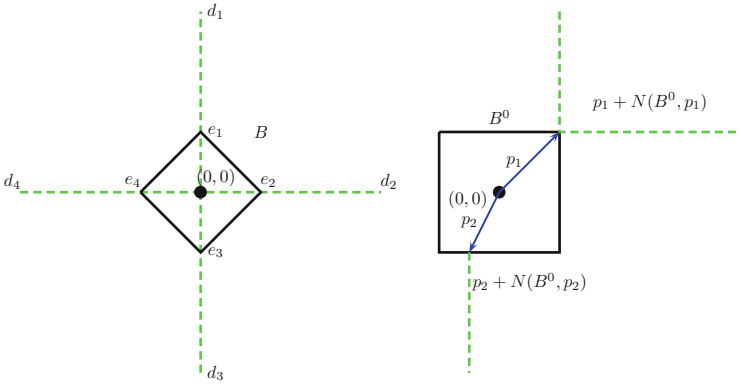


Fig. 9.6 Illustration of the unit ball for the ℓ_1-norm, its dual ball and two normal cones of this dual ball

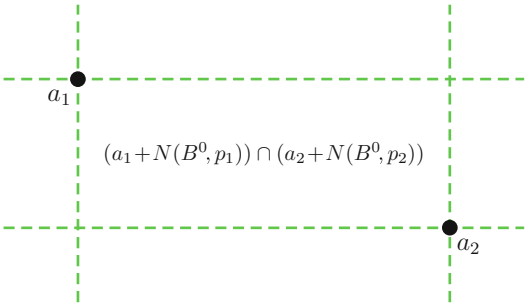


Fig. 9.7 Illustration of an elementary convex set for the ℓ_1-norm

(1985), a nonempty convex set C is called an elementary convex set if there exists a family π such that $C_\pi = C$. If the unit balls are polytopes, then we can obtain the elementary convex sets as intersections of cones generated by fundamental directions of these balls pointed at each demand point (for details, see Durier and Michelot 1985). The two-dimensional elementary convex sets are called cells. Let $\mathscr{C}$ denote to the set of these cells. Therefore each cell is a polyhedron whose vertices are the intersection points, which we denote by $\mathscr{I}\mathscr{P}$. Finally, in the case of $\mathbb{R}^2$ there exists an upper bound on the number of cells which is $O((MG_{\max})^2)$ (see Durier and Michelot 1985).

In Fig. 9.7 we show an elementary convex set for the ℓ_1-norm for two points a_1, a_2. In this example the dual norm is the ℓ_∞-norm where its unit ball B^0 has the extreme points $\{(1,1),(-1,1),(-1-,1),(1,-1)\}$. The normal cones to B^0 at $p_1 = (1,-1)$ and $p_2 = (-1,1)$ are given by $N(B^0, p_1) = cone((1,0),(0,-1))$ and $N(B^0, p_2) = cone((-1,0),(0,1))$, respectively, where *cone* stands for the conical hull of its argument. Thus, the elementary convex set C_π with $\pi = (p_1, p_2)$ is the rectangle defined by a_1 and a_2 with sides parallel to the coordinates axes.

9.2.1.1 Bicriteria Case

In this section we restrict ourselves to the bicriteria case, which, as will be seen later, is the basis for solving the Q-criteria case. To this end, we are looking for the Pareto solutions of the vector optimization problem in $\mathbb{R}^2$,

$$\min_{x\in\mathbb{R}^2}\left(f^1(x) := \sum_{i=1}^{M} w_i^1\gamma_i(x-a_i),\ f^2(x) := \sum_{i=1}^{M} w_i^2\gamma_i(x-a_i)\right),$$

where the weights w_i^q are non negative ($i = 1,\ldots,M; q = 1,2$). The following theorem provides a geometric characterization of the set $\mathscr{X}_{\text{Par}}^*$.

Theorem 9.4 *$\mathscr{X}_{\text{Par}}^*\left(f^1, f^2\right)$ is a connected chain from $\mathscr{X}^*(f^1)$ to $\mathscr{X}^*(f^2)$ consisting of faces or vertices of cells, or complete cells.*

Proof First, we note that $\mathscr{X}^*(f^q) \neq \emptyset$ for $q = 1,2$ (see Puerto and Fernández 2000). Moreover, $\mathscr{X}_{\text{Par}}^* \cap \mathscr{X}^*(f^q) \neq \emptyset$ for $q = 1,2$. Therefore, we know that $\mathscr{X}_{\text{Par}}^* \neq \emptyset$, so we can choose $x \in \mathscr{X}_{\text{Par}}^*$. There exists at least one cell $\mathrm{C} \in \mathscr{C}$ with $x \in \mathrm{C}$. We can assume without loss of generality that C is bounded. We also note that the functions f^1 and f^2 are linear within each cell (see Rodríguez-Chía et al. 2000). Given a set A, in what follows, conv(A), bd(A) and int(A) will denote the convex hull, the boundary and the interior of the set A, respectively. Hence three cases may occur:

Case 1: $x \in \text{int(C)}$. Since $x \in \mathscr{X}_{\text{Par}}^*$ we obtain

$$\bigcap_{q=1}^{2} L_{\leq}(f^q, f^q(x)) = \bigcap_{q=1}^{2} L_{=}(f^q, f^q(x))$$

and by linearity of the median problem in each cell we have

$$\bigcap_{q=1}^{2} L_{\leq}(f^q, f^q(y)) = \bigcap_{q=1}^{2} L_{=}(f^q, f^q(y)) \quad \forall\, y \in \mathrm{C}$$

which means $y \in \mathscr{X}_{\text{Par}}^*\ \forall\, y \in \mathrm{C}$, hence $\mathrm{C} \subseteq \mathscr{X}_{\text{Par}}^*$.

Case 2: $x \in \overline{ab} := \text{conv}(\{a,b\}) \subset \text{bd(C)}$ and $a, b \in \text{Ext(C)}$. We can choose $y \in \text{int(C)}$ and two cases can occur:

Case 2.1: $y \in \mathscr{X}_{\text{Par}}^*$. Hence we can continue as in Case 1.
Case 2.2: $y \notin \mathscr{X}_{\text{Par}}^*$. Therefore using the linearity we first obtain

$$\bigcap_{q=1}^{2} L_{\leq}(f^q, f^q(z)) \neq \bigcap_{q=1}^{2} L_{=}(f^q, f^q(z)) \quad \forall\, z \in \text{int(C)}.$$

Second, since $x \in \mathscr{X}^*_{\text{Par}}$, we have

$$\bigcap_{q=1}^{2} L_{\leq}(f^q, f^q(z)) = \bigcap_{q=1}^{2} L_{=}(f^q, f^q(z)) \quad \forall\, z \in \overline{ab}.$$

Hence, we have that $\mathrm{C} \not\subseteq \mathscr{X}^*_{\text{Par}}$ and $\overline{ab} \subseteq \mathscr{X}^*_{\text{Par}}$.

Case 3: $x \in \text{Ext(C)}$. We can choose $y \in \text{int(C)}$ and two cases can occur

Case 3.1: If $y \in \mathscr{X}^*_{\text{Par}}$, we can continue as in Case 1.
Case 3.2: If $y \notin \mathscr{X}^*_{\text{Par}}$, we choose $z_1, z_2 \in \text{Ext(C)}$ such that $\overline{xz_1}, \overline{xz_2}$ are faces of C,

- If z_1 or z_2 are in $\mathscr{X}^*_{\text{Par}}$, we can continue as in Case 2.
- If z_1 and z_2 are not in $\mathscr{X}^*_{\text{Par}}$, then using the linearity in the same way as before we obtain that $(\mathrm{C} \setminus \{x\}) \cap \mathscr{X}^*_{\text{Par}} = \emptyset$.

Hence, we conclude that the set of Pareto solutions consists of complete cells, complete faces, and vertices of these cells. Since we know that the set $\mathscr{X}^*_{\text{Par}}$ is connected, the proof is completed. □

In the following we develop an algorithm to solve the bicriteria planar minisum location problem. The idea of this algorithm is to start in a vertex x of the cell structure which belongs to $\mathscr{X}^*_{\text{Par}}$, say $x \in \mathscr{X}^*_{1,2} := \arg\min_{x \in \mathscr{X}^*(f^1)} f^2(x)$ (set of optimal lexicographical locations, see Nickel 1995). Then, using the connectivity of $\mathscr{X}^*_{\text{Par}}$, the algorithm proceeds by moving from vertex x to another Pareto-optimal vertex y of the cell structure which is connected with the previous one by an elementary convex set. This procedure is repeated until the end of the chain reaches $\mathscr{X}^*_{2,1} := \arg\min_{x \in \mathscr{X}^*(f^2)} f^1(x)$.

Let C be a cell and y, x and z three vertices of C enumerated counterclockwise (see Fig. 9.8). By the linearity of the level sets in each cell we can distinguish the following disjoint situations, if $x \in \mathscr{X}^*_{\text{Par}}$:

(S1) $\mathrm{C} \subseteq \mathscr{X}^*_{\text{Par}}$, i.e., C is contained in the chain.
(S2) $\overline{xy}$ and $\overline{xz}$ are candidates for $\mathscr{X}^*_{\text{Par}}$ and $\text{int}(C) \not\subset \mathscr{X}^*_{\text{Par}}$.
(S3) $\overline{xy}$ is candidate for $\mathscr{X}^*_{\text{Par}}$ and $\overline{xz}$ is not contained in $\mathscr{X}^*_{\text{Par}}$.
(S4) $\overline{xz}$ is candidate for $\mathscr{X}^*_{\text{Par}}$ and $\overline{xy}$ is not contained in $\mathscr{X}^*_{\text{Par}}$.
(S5) Neither $\overline{xy}$ nor $\overline{xz}$ are contained in $\mathscr{X}^*_{\text{Par}}$.

We denote by $sit(\mathrm{C}, x)$ the situations (S1–S4 or S5) in which the cell C is classified according to the extreme point x of C. The following lemma, whose proof is based on an exhaustive case analysis of the different relative positions of x within C, can be found in Weissler (1999). It states when a given segment belongs to the Pareto-set in terms of the $sit(\cdot,\cdot)$ function.

Lemma 9.1 *Let* $\mathrm{C}_1, \ldots, \mathrm{C}_{P_x}$ *be the cells containing the intersection point* x*, considered in counterclockwise order, and* $y_1, \ldots, y_{P_x}$ *the intersection points adjacent to* x*, considered in counterclockwise order (see Fig. 9.9). If* $x \in \mathscr{X}^*_{\text{Par}}$

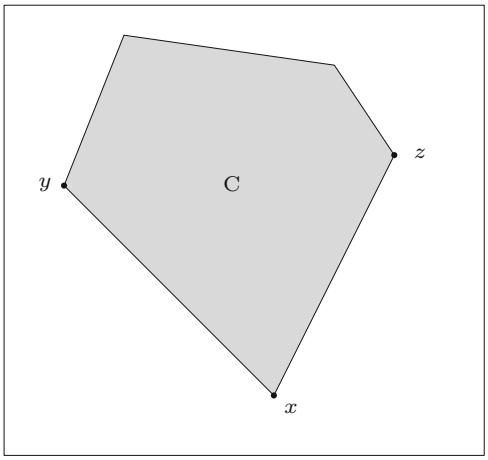


Fig. 9.8 Illustration to $y, x, z \in Ext(\mathrm{C})$ in counterclockwise order

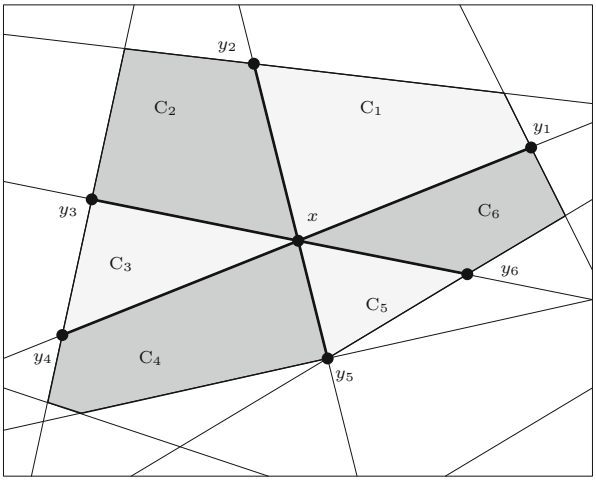


Fig. 9.9 Illustration to Lemma 9.1 with $P_x = 6$

and $i \in \{1, \ldots, P_x\}$*, then the following holds (assume that* $i + 1 = 1$ *whenever* $i = P_x$*):*

$$\overline{xy_{i+1}} \subseteq \mathscr{X}_{\mathrm{Par}}^{*} \iff \left\{ \begin{array}{ll} & sit(\mathrm{C}_i, x) = S1 \\ or & sit(\mathrm{C}_{i+1}, x) = S1 \\ or & \left\{ \begin{array}{l} sit(\mathrm{C}_i, x) \in \{S2, S3\} \\ sit(\mathrm{C}_{i+1}, x) \in \{S2, S4\} \end{array} \right. \end{array} \right\}$$

These results validate the following algorithm for finding $\mathscr{X}_{\mathrm{Par}}^{*}\left(f^1, f^2\right)$.

Algorithm 9.1 *Step 1. Compute the planar graph generated by the cells and the two sets of lexicographical locations* $\mathscr{X}^*_{1,2}$, $\mathscr{X}^*_{2,1}$.

Step 2. If $\mathscr{X}^*_{1,2} \cap \mathscr{X}^*_{2,1} \neq \emptyset$ *then set* $\mathscr{X}^*_{\text{Par}} := \text{conv}(\mathscr{X}^*_{1,2})$ *(trivial case* $\mathscr{X}^*(f^1) \cap \mathscr{X}^*(f^2) \neq \emptyset$*). Otherwise set* $\mathscr{X}^*_{\text{Par}} := \mathscr{X}^*_{1,2} \cup \mathscr{X}^*_{2,1}$ *(non trivial case* $\mathscr{X}^*(f^1) \cap \mathscr{X}^*(f^2) = \emptyset$*)*

Step 3. Choose $x \in \mathscr{X}^*_{1,2} \cap \mathscr{I}\mathscr{P}$.

Step 4. Scan the list of cells adjacent to x *until we get situation* $S1$ *for a cell* C *or two consecutive cells,* C, $\overline{C}$, *in situations* $C \in \{S2, S3\}$ *and* $\overline{C} \in \{S2, S4\}$, *respectively.*

Step 5. If situation A *occurs then* $\mathscr{X}^*_{\text{Par}} := \mathscr{X}^*_{\text{Par}} \cup C$ *(we have found a bounded cell.) Otherwise* $\mathscr{X}^*_{\text{Par}} := \mathscr{X}^*_{\text{Par}} \cup \overline{xy}$ *where* y *is a vertex of* C *defined in situations* $S2$ *and* $S4$ *(we have found a bounded face.)*

Step 6. Let C *be the last scanned cell. Choose* $y \in \mathscr{I}\mathscr{P} \cap C$ *and, such that,* y *is connected to* x. *If* $y \in \mathscr{X}^*_{2,1}$ *stop. Otherwise, set* $x := y$ *and go to Step 4.*

Output: $\mathscr{X}^*_{\text{Par}}(f^1, f^2)$. □

Edelsbrunner (1987) proved that the computation of a planar graph induced by n lines in the plane can be done in $O(n^2)$ time. This implies that in the case of the minisum location problem the computation of the planar graph generated by the fundamental direction lines is doable in $O(M^2 G^2_{\max})$ time.

The evaluation of the minisum location function needs $O(M \log(G_{\max}))$ for one point, therefore we obtain $O(M^3 G^2_{\max} \log(G_{\max}))$ time for the computation of lexicographic solutions. At the end, the complexity for computing the chain is $O(M^3 G^2_{\max} \log(G_{\max}))$, since we have to consider at most $O(M^2 G^2_{max})$ cells and the determination of $sit(.,.)$ can be done in $O(M \log(G_{max}))$ time. Hence, the overall complexity is $O(M^3 G^2_{\max} \log(G_{\max}))$. Notice that the polynomial complexity of this algorithm allows an efficient computation of the solution set.

Example 9.6 Consider a three-criteria median problem with nine existing facilities $A = \{a_1, \ldots, a_9\}$ (see Fig. 9.10). The coordinates $a_i = (x_i, y_i)$ of the existing facilities are given by the set: $\{(-3,0), (3,0), (0,-4), (11,-6), (17,-6), (14,-2), (11,2), (17,2), (14,6)\}$, and the weights $w^q, q = 1,2,3$ are given by $w^1 = (2,2,1,0,0,0,0,0,0)$, $w^2 = (0,0,0,2,2,1,0,0,0)$ and $w^3 = (0,0,0,0,0,0,2,2,1)$.

The optimal solutions of the location problems associated with the median functions f^1, f^2 and f^3 with $f^q = \sum_{i=1}^{M} w_i^q \parallel x - a_i \parallel_1, q = 1,2,3$, are unique and given by $\mathscr{X}_1^* = \{(0,0)\}$, $\mathscr{X}_2^* = \{(14,-6)\}$ and $\mathscr{X}_3^* = \{(14,2)\}$, respectively, all of them with the (optimal) objective value 16. The bicriteria chains (consisting of cells and edges with respect to the fundamental directions drawn in Fig. 9.10) are given by

$$\mathscr{X}^*_{\text{Par}}(f^1, f^3) = \overline{(0,0)(3,0)} \cup \text{conv}(\{(3,0),(3,2),(11,2),(11,0)\}) \cup \overline{(11,2)(14,2)},$$

$$\mathscr{X}^*_{\text{Par}}(f^2, f^3) = \overline{(14,2)(14,-6)},$$

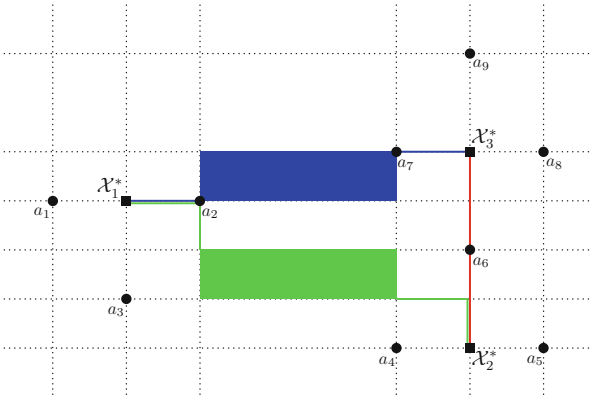


Fig. 9.10 Illustration to Example 9.6

$$\begin{aligned}\mathcal{X}_{\text{Par}}^*\left(f^1, f^2\right) = {} & \overline{(0,0)(3,0)} \cup \overline{(3,0)(3,-2)} \cup \\ & \operatorname{conv}(\{(3,-2),(3,-4),(11,-4),(11,-2)\}) \cup \\ & \overline{(11,-4)(14,-4)} \cup \overline{(14,-4)(14,-6)}.\end{aligned}$$

9.2.1.2 Three-Criteria Case

In this section we consider the three-criteria case and develop an efficient algorithm for computing $\mathcal{X}_{\text{Par}}^*\left(f^1, f^2, f^3\right)$ using the results for the bicriteria case. In particular, we obtain a characterization of the Pareto solution set for the three criteria case using the region surrounded by the chains of bicriteria Pareto solutions. We denote the union of the bicriteria chains including the one-criterion solutions by

$$\mathcal{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right) := \bigcup_{q=1}^{3} \mathcal{X}^*(f^q) \cup \bigcup_{q=1}^{2} \bigcup_{p=q+1}^{3} \mathcal{X}_{\text{Par}}^*\left(f^p, f^q\right).$$

We use "gen" since this set will *generate* the set $\mathcal{X}_{\text{Par}}^*\left(f^1, f^2, f^3\right)$ (see Fig. 9.11).

The next lemma provides useful geometric information to build $\mathcal{X}_{\text{Par}}^*\left(f^1, f^2, f^3\right)$. For a set A, let $\text{cl}(A)$ denote the topological closure of A.

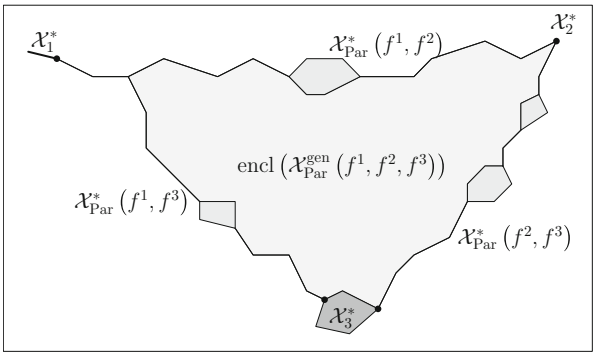


Fig. 9.11 The enclosure of $\mathcal{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right)$

Lemma 9.2 *The following inclusion of sets holds:*

$$\text{cl}\big(\text{encl}\left(\mathcal{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right)\right)\big) \subseteq \mathcal{X}_{\text{s-Par}}^*\left(f^1, f^2, f^3\right).$$

The interested reader is referred to Nickel et al. (2005b) for a detailed proof of this result.

Remark 9.4 Since $\mathcal{X}_{\text{Par}}^*\left(f^i, f^j\right) = \mathcal{X}_{\text{w-Par}}^*\left(f^i, f^j\right)$ for any $i, j \in \{1, 2, 3\}$, we have that:

$$\text{encl}\left(\mathcal{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right)\right) = \text{encl}\left(\mathcal{X}_{\text{w-Par}}^{\text{gen}}\left(f^1, f^2, f^3\right)\right).$$

Finally we obtain the following theorem which provides a subset as well as a superset of $\mathcal{X}_{\text{Par}}^*\left(f^1, f^2, f^3\right)$.

Theorem 9.5 *The following inclusions of sets hold:*

$$\begin{aligned}\text{encl}\left(\mathcal{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right)\right) &\subseteq \mathcal{X}_{\text{Par}}^*\left(f^1, f^2, f^3\right)\\ &\subseteq \mathcal{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right) \cup \text{encl}\left(\mathcal{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right)\right)\\ &= \mathcal{X}_{\text{w-Par}}^*\left(f^1, f^2, f^3\right).\end{aligned}$$

Proof Using Lemma 9.2 and Theorem 9.2 we have the following chain of inclusions that proves the thesis of the theorem.

$$\begin{aligned}\text{encl}\left(\mathcal{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right)\right) &\subseteq \mathcal{X}_{\text{s-Par}}^*\left(f^1, f^2, f^3\right)\\ &\subseteq \mathcal{X}_{\text{Par}}^*\left(f^1, f^2, f^3\right) \subseteq \mathcal{X}_{\text{w-Par}}^*\left(f^1, f^2, f^3\right)\\ &\subseteq \mathcal{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right) \cup \text{encl}\left(\mathcal{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right)\right). \quad \square\end{aligned}$$

Now it remains to consider the Pareto-optimality of the set $\mathscr{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right)$ with respect to the three objective functions f^1, f^2, f^3. For a cell $C \in \mathscr{C}$ we define the collapsing and the remaining part of C with respect to Q-criteria optimality by

$$\text{col}_Q(C) := \left\{x \in C \,:\, x \notin \mathscr{X}_{\text{Par}}^{*}\left(f^1, \ldots, f^Q\right)\right\}$$

$$\text{rem}_Q(C) := \left\{x \in C \,:\, x \in \mathscr{X}_{\text{Par}}^{*}\left(f^1, \ldots, f^Q\right)\right\}.$$

Summing up the preceding results we get a complete geometric characterization of the set of Pareto solutions for the three criteria case. For each cell C, $\text{col}_Q(C) \,\dot\cup\, \text{rem}_Q(C) = C$ and, as shown by Nickel et al. (2005b), determining both sets can be done with the gradients of the objective functions with a complexity of $O(Q \log Q)$.

Theorem 9.6 *The set of Pareto solutions satisfies:*

$$\mathscr{X}_{\text{Par}}^{*}\left(f^1, f^2, f^3\right) = \left(\mathscr{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right) \cup \text{encl}\left(\mathscr{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right)\right)\right) \setminus \{x \in \mathbb{R}^2 \,:\, \exists\, C \in \mathscr{C}, C \subseteq \mathscr{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right)\ x \in \text{col}_3(C)\}.$$

Proof Let $y \in \mathscr{X}_{\text{Par}}^{*}\left(f^1, f^2, f^3\right)$. Then we have, by Theorem 9.5, that $y \in \mathscr{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right) \cup \text{encl}\left(\mathscr{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right)\right)$. Moreover for $C \in \mathscr{C}$ with $y \in C$ we have $y \in \text{rem}_3(C)$, i. e., $y \notin \text{col}_3(C)$. This implies

$$y \in \left(\mathscr{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right) \cup \text{encl}\left(\mathscr{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right)\right)\right) \setminus \{x \in \mathbb{R}^2 \,:\, \exists\, C \in \mathscr{C}, C \subseteq \mathscr{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right)\ x \in \text{col}_3(C)\}.$$

We distinguish the following cases:

Case 1: $y \in \text{encl}\left(\mathscr{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right)\right)$. Then $y \in \mathscr{X}_{\text{Par}}^{*}\left(f^1, f^2, f^3\right)$ by Theorem 9.5.

Case 2: $y \in \mathscr{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right)$.

Case 2.1: $\exists\, C \in \mathscr{C},\ C \subseteq \mathscr{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right)$ with $y \in C$

$\Rightarrow\ y \notin \text{col}_3(C)\ \Rightarrow\ y \in \text{rem}_3(C)\ \Rightarrow\ y \in \mathscr{X}_{\text{Par}}^{*}\left(f^1, f^2, f^3\right)$.

Case 2.2: $\nexists C \in \mathscr{C},\ C \subseteq \mathscr{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right)$ with $y \in C$

$\Rightarrow\ L_{\leq}(f^p, f^p(y)) \cap L_{\leq}(f^q, f^q(y)) = \{y\}$ for some $p, q \in \{1, 2, 3\}, p < q$

$\Rightarrow\ \bigcap_{q=1}^{3} L_{\leq}(f^q, f^q(y)) = \{y\}\ \Rightarrow\ y \in \mathscr{X}_{\text{s-Par}}^{*}\left(f^1, f^2, f^3\right) \subseteq \mathscr{X}_{\text{Par}}^{*}\left(f^1, f^2, f^3\right)$. □

In the case of median functions the gradients $\nabla f^q(x)$, $q \in \{1, 2, 3\}$, (in those points where they are well-defined) can be computed in $O(M \log(G_{\max}))$ time (analogous to the evaluation of the function). Therefore, we can test in $O(M \log(G_{\max}))$ time if a cell $C \in \mathscr{C}$, $C \subseteq \mathscr{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right)$ collapses. We

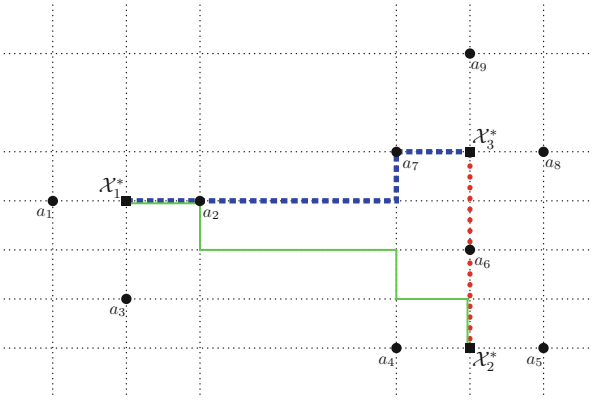


Fig. 9.12 Illustration of $\mathscr{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right)$ and $\mathscr{X}_{\text{Par}}^{*}\left(f^1, f^2, f^3\right)$ for the problem introduced in Example 9.6

obtain the following algorithm for the three-criteria median problem with time complexity $O(M^3 G_{\max}^2 \log(G_{\max}))$ (see Nickel et al. 2005b for more details).

Algorithm 9.2 *Step 1. Compute the subdivision of the plane generated $\mathscr{C}$, the family of elementary convex sets. Compute $\mathscr{X}_{\text{w-Par}}^{*}\left(f^1, f^2\right)$, $\mathscr{X}_{\text{w-Par}}^{*}\left(f^1, f^3\right)$, $\mathscr{X}_{\text{w-Par}}^{*}\left(f^2, f^3\right)$ using Algorithm 9.1.*

Step 2. Set $\mathscr{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right) := \mathscr{X}_{\text{w-Par}}^{*}\left(f^1, f^2\right) \cup \mathscr{X}_{\text{w-Par}}^{*}\left(f^1, f^3\right) \cup \mathscr{X}_{\text{w-Par}}^{*}\left(f^2, f^3\right)$ and $\mathscr{X}_{\text{Par}}^{*}\left(f^1, f^2, f^3\right) := \mathscr{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right) \cup \text{encl}\left(\mathscr{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right)\right)$.

Step 3. For any $C \in \mathscr{C}$ with $C \subseteq \mathscr{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right)$ compute $\text{col}_3(C)$ and set $\mathscr{X}_{\text{Par}}^{*}\left(f^1, f^2, f^3\right) := \mathscr{X}_{\text{Par}}^{*}\left(f^1, f^2, f^3\right) \setminus \text{col}_3(C)$.

Output: $\mathscr{X}_{\text{Par}}^{*}\left(f^1, f^2, f^3\right)$. □

Figure 9.12 illustrates the preceding results using the data introduced in Example 9.6. The dashed path joining $\mathscr{X}_1^*$ and $\mathscr{X}_3^*$ in the picture represents the set $\mathscr{X}_{\text{w-Par}}^{*}\left(f^1, f^3\right)$ after removing the $col_3(C)$. In the same way, the path joining $\mathscr{X}_1^*$ and $\mathscr{X}_2^*$ represents the set $\mathscr{X}_{\text{w-Par}}^{*}\left(f^1, f^2\right)$ after removing the $col_3(C)$. Finally, the dotted segment joining $\mathscr{X}_2^*$ and $\mathscr{X}_3^*$ is $\mathscr{X}_{\text{w-Par}}^{*}\left(f^2, f^3\right)$ (in this case there are not cells to be collapsed).

9.2.1.3 Case Where $Q > 3$

In this section we consider the general Q-Criteria case ($Q > 3$). We prove that the Pareto solution set can be obtained from the Pareto solution sets of all the three criteria problems. This construction requires the removal of the dominated points

from the union of all the three criteria Pareto solution sets. The reader may notice that all this process reduces to obtaining the bicriteria Pareto chains as proved in Theorem 9.6.

Theorem 9.7 *The following inclusions hold:*

I. $$\bigcup_{\substack{p,q,r\in\mathcal{Q}\\ p<q<r}} \mathrm{cl}\big(\mathrm{encl}\left(\mathscr{X}^{\mathrm{gen}}_{\mathrm{Par}}(f^p,f^q,f^r)\right)\big) \subseteq \mathscr{X}^{*}_{\mathrm{Par}}(f^1,\ldots,f^Q).$$

II. $$\begin{aligned}&\mathscr{X}^{*}_{\mathrm{Par}}\left(f^1,\ldots,f^Q\right)\\ &\subseteq \bigcup_{\substack{p,q,r\in\mathcal{Q}\\ p<q<r}} \mathscr{X}^{\mathrm{gen}}_{\mathrm{Par}}(f^p,f^q,f^r) \cup \bigcup_{\substack{p,q,r\in\mathcal{Q}\\ p<q<r}} \mathrm{encl}\left(\mathscr{X}^{\mathrm{gen}}_{\mathrm{Par}}(f^p,f^q,f^r)\right)\\ &= \mathscr{X}^{*}_{\mathrm{w-Par}}\left(f^1,\ldots,f^Q\right).\end{aligned}$$

Proof (I) Let $x \in \bigcup_{\substack{p,q,r\in\mathcal{Q}\\ p<q<r}} \mathrm{cl}\big(\mathrm{encl}\left(\mathscr{X}^{\mathrm{gen}}_{\mathrm{Par}}(f^p,f^q,f^r)\right)\big)$. This is equivalent to

$$x \in \mathrm{cl}\big(\mathrm{encl}\left(\mathscr{X}^{\mathrm{gen}}_{\mathrm{Par}}(f^p,f^q,f^r)\right)\big) \quad \text{for some} \quad p,q,r \in \mathcal{Q}, p<q<r.$$

Then, by Lemma 9.2, $x \in \mathscr{X}^{*}_{\mathrm{s-Par}}(f^p,f^q,f^r)$ for some $p,q,r \in \mathcal{Q}, p<q<r$. Applying characterization (9.4), this is equivalent to $L_{\preceq}(f^p,f^p(x)) \cap L_{\preceq}(f^q,f^q(x)) \cap L_{\preceq}(f^r,f^r(x)) = \{x\}$ for some $p,q,r \in \mathcal{Q}, p<q<r$ and since $x \in L_{\preceq}(f^q,f^q(x))$ for all $q \in \mathcal{Q}$ it follows that $\bigcap_{q=1}^{Q} L_{\preceq}(f^q,f^q(x)) = \{x\}$. Finally, again by (9.4), $x \in \mathscr{X}^{*}_{\mathrm{s-Par}}\left(f^1,\ldots,f^Q\right)$, which implies that $x \in \mathscr{X}^{*}_{\mathrm{Par}}\left(f^1,\ldots,f^Q\right)$.

(II) Let $x \in \mathscr{X}^{*}_{\mathrm{Par}}\left(f^1,\ldots,f^Q\right)$ then $x \in \mathscr{X}^{*}_{\mathrm{w-Par}}\left(f^1,\ldots,f^Q\right)$ and, by (9.2), this is equivalent to $\bigcap_{q=1}^{Q} L_{\prec}(f^q,f^q(x)) = \emptyset$. By Helly's theorem, there exists $p,q,r \in \mathcal{Q}, p<q<r$, such that, $L_{\prec}(f^p,f^p(x)) \cap L_{\prec}(f^q,f^q(x)) \cap L_{\prec}(f^r,f^r(x)) = \emptyset$. By characterization (9.2), this is equivalent to $x \in \mathscr{X}^{*}_{\mathrm{w-Par}}(f^p,f^q,f^r)$ for some $p,q,r \in \mathcal{Q}, p<q<r$ and, by Theorem 3.2 in Rodríguez-Chía and Puerto (2002), this implies that $x \in \mathscr{X}^{\mathrm{gen}}_{\mathrm{Par}}(f^p,f^q,f^r) \cup \mathrm{encl}\left(\mathscr{X}^{\mathrm{gen}}_{\mathrm{Par}}(f^p,f^q,f^r)\right)$ for some $p,q,r \in \mathcal{Q}, p<q<r$. Finally, this can be equivalently written as

$$x \in \bigcup_{\substack{p,q,r\in\mathcal{Q}\\ p<q<r}} \mathscr{X}^{\mathrm{gen}}_{\mathrm{Par}}(f^p,f^q,f^r) \cup \bigcup_{\substack{p,q,r\in\mathcal{Q}\\ p<q<r}} \mathrm{encl}\left(\mathscr{X}^{\mathrm{gen}}_{\mathrm{Par}}(f^p,f^q,f^r)\right).$$

□

In the Q-criteria case the crucial region is now given by the cells $C \in \mathscr{C}$ with

$$\begin{aligned}C &\subseteq \bigcup_{\substack{p,q,r\in\mathcal{Q}\\ p<q<r}} \mathscr{X}^{\mathrm{gen}}_{\mathrm{Par}}(f^p,f^q,f^r) \setminus \bigcup_{\substack{p,q,r\in\mathcal{Q}\\ p<q<r}} \mathrm{encl}\left(\mathscr{X}^{\mathrm{gen}}_{\mathrm{Par}}(f^p,f^q,f^r)\right)\\ &= \bigcup_{\substack{p,q\in\mathcal{Q}\\ p<q}} \mathscr{X}^{*}_{\mathrm{w-Par}}(f^p,f^q) \setminus \bigcup_{\substack{p,q,r\in\mathcal{Q}\\ p<q<r}} \mathrm{encl}\left(\mathscr{X}^{\mathrm{gen}}_{\mathrm{Par}}(f^p,f^q,f^r)\right).\end{aligned}$$

Similar to the situation in the previous section one can test whether the cell $C \in \mathscr{C}$ collapses with respect to $f^1, \ldots, f^Q$ by comparing the gradients of the objective functions in int(C). Finally we obtain the following theorem, which can be proven using the same reasoning as in the three-criteria case (see proof of Theorem 9.6).

Theorem 9.8

$$\mathscr{X}^*_{\text{Par}}\left(f^1, \ldots, f^Q\right) = \left(\bigcup_{\substack{p,q,r \in \mathscr{Q} \\ p<q<r}} \mathscr{X}^{\text{gen}}_{\text{Par}}\left(f^p, f^q, f^r\right) \cup \bigcup_{\substack{p,q,r \in \mathscr{Q} \\ p<q<r}} \text{encl}\left(\mathscr{X}^{\text{gen}}_{\text{Par}}\left(f^p, f^q, f^r\right)\right) \right)$$
$$\setminus \left\{ x \in \mathbb{R}^2 : \exists C \in \mathscr{C}, C \subseteq \bigcup_{\substack{p,q \in \mathscr{Q} \\ p<q}} \mathscr{X}^*_{\text{w-Par}}\left(f^p, f^q\right) \setminus \bigcup_{\substack{p,q,r \in \mathscr{Q} \\ p<q<r}} \text{encl}\left(\mathscr{X}^{\text{gen}}_{\text{Par}}\left(f^p, f^q, f^r\right)\right) \; x \in \text{col}_Q(C) \right\}$$

For the Q-criteria median problem we obtain the following algorithm.

Algorithm 9.3 *Step 1. Compute the subdivision of the plane generated $\mathscr{C}$, the family of elementary convex sets. Compute $\mathscr{X}^*_{\text{w-Par}}\left(f^p, f^q\right)$, $p, q \in \mathscr{Q}$, $p < q$, using Algorithm 9.1.*

Step 2. Set for any p, q and r with $p < q < r$

$$\mathscr{X}^{\text{gen}}_{\text{Par}}\left(f^p, f^q, f^r\right) := \mathscr{X}^*_{\text{w-Par}}\left(f^p, f^q\right) \cup \mathscr{X}^*_{\text{w-Par}}\left(f^p, f^r\right) \cup \mathscr{X}^*_{\text{w-Par}}\left(f^q, f^r\right),$$

and

$$\mathscr{X}^*_{\text{Par}}\left(f^1, \ldots, f^Q\right) := \bigcup_{\substack{p,q,r \in \mathscr{Q} \\ p<q<r}} \mathscr{X}^{\text{gen}}_{\text{Par}}\left(f^p, f^q, f^r\right) \cup \bigcup_{\substack{p,q,r \in \mathscr{Q} \\ p<q<r}} \text{encl}\left(\mathscr{X}^{\text{gen}}_{\text{Par}}\left(f^p, f^q, f^r\right)\right).$$

Step 3. For every cell $C \subseteq \bigcup_{\substack{p,q \in \mathscr{Q} \\ p<q}} \mathscr{X}^*_{\text{w-Par}}\left(f^p, f^q\right) \setminus \bigcup_{\substack{p,q,r \in \mathscr{Q} \\ p<q<r}} \text{encl}\left(\mathscr{X}^{\text{gen}}_{\text{Par}}\left(f^p, f^q, f^r\right)\right)$ compute $\text{col}_Q(C)$ and set $\mathscr{X}^*_{\text{Par}}\left(f^1, \ldots, f^Q\right) := \mathscr{X}^*_{\text{Par}}\left(f^1, \ldots, f^Q\right) \setminus \text{col}_Q(C)$.

Output: $\mathscr{X}^*_{\text{Par}}\left(f^1, \ldots, f^Q\right)$. □

The complexity of Algorithm 9.3 can be determined as follows. For each cell C, $col_Q(C)$ can be computed in $O(Q \log(Q))$ time. Algorithm 9.3 needs to solve $O(Q^3)$ three-criteria problems which dominates all other elementary operations of the algorithm. Each one of them has the same complexity as the two-criteria problem. Thus, the overall complexity is $O(M^3 G^2_{\max} Q^3 (\log G_{\max}) + M^2 G^2_{\max} Q \log Q) = O(M^3 G^2_{\max} Q^3 (\log G_{\max})$.

We would like to conclude this section pointing that the multi-facility versions of the problems analyzed in this section have been hardly studied in the literature, although an exception is the paper by Nickel (1997).

9.3 Network Location Problems

9.3.1 1-Facility Median Problems

9.3.1.1 Pareto Locations in General Networks

Let $G = (V, E)$ be a connected graph with node set $V = \{v_1, \ldots, v_n\}$ and edge set $E = \{e_1, \ldots, e_m\}$. Each edge $e \in E$ has a positive length $\ell(e)$, and is assumed to be rectifiable. Let $P(G)$ denote the continuum set of points on edges of G. We denote a point $x \in e = \{u, v\}$ as a pair $x = (e, t)$, where t $(0 \leq t \leq 1)$ gives the relative distance of x from node u along edge e. For the sake of readability, we identify $P(G)$ with G and $P(e)$ with e for $e \in E$. We also define $(e, (t_1, t_2)) := \{x = (e, t) : t \in (t_1, t_2)\}$; $(e, [t_1, t_2])$, $(e, (t_1, t_2])$, and $(e, [t_1, t_2))$ are used in an analogous way.

We denote by $d(x, y)$ the length of the shortest path connecting two points $x, y \in G$. Let $v_i \in V$ and $x = (\{v_r, v_s\}, t) \in G$. The distance from v_i to x entering the edge $\{v_r, v_s\}$ through v_r (v_s) is given as $D_i^+(x) = d(v_r, x) + d(v_r, v_i)$ $(D_i^-(x) = d(v_s, x) + d(v_s, v_i))$. Hence, the length of a shortest path from v_i to x is given by $D_i(x) = \min\{D_i^+(x), D_i^-(x)\}$. As $d(v_r, x) = t \cdot \ell(e)$ and $d(v_s, x) = (1-t) \cdot \ell(e)$, the functions $D_i^+(x)$ and $D_i^-(x)$ are linear in x and $D_i(x)$ is piecewise linear and concave in x (cf. Drezner 1995). The distance from v_i to a facility located at x is finally defined as $d(v_i, x) = D_i(x) = \min\{D_i^+(x), D_i^-(x)\}$.

We consider the objective function $f(x) = (f^1(x), \ldots, f^Q(x))$, where each $f^q(x)$, $q \in \mathcal{Q}$, is a median function defined as:

$$f^q(x) = \sum_{v_i \in V} w_i^q \, d(v_i, x) \, .$$

More formally, we assign a vector of weights

$$w_i = \begin{pmatrix} w_i^1 \\ \vdots \\ w_i^Q \end{pmatrix} \neq 0 \text{ to every vertex } v_i \in V, \text{ with } w_i^q \geq 0, \; q \in \mathcal{Q} := \{1, \ldots, Q\}.$$

The quality of a point $x \in P(G)$ in this multicriteria setting is defined by

$$f(x) := \begin{pmatrix} f^1(x) \\ \vdots \\ f^Q(x) \end{pmatrix} := \begin{pmatrix} \sum_{v_i \in V} w_i^1 d(x, v_i) \\ \vdots \\ \sum_{v_i \in V} w_i^Q d(x, v_i) \end{pmatrix}$$

in the undirected case and

$$f(x) := \begin{pmatrix} f^1(x) \\ \vdots \\ f^Q(x) \end{pmatrix} := \begin{pmatrix} \sum_{v_i \in V} w_i^1 \left(d(x, v_i) + d(v_i, x)\right) \\ \vdots \\ \sum_{v_i \in V} w_i^Q \left(d(x, v_i) + d(v_i, x)\right) \end{pmatrix}$$

in the directed case.

Let $S \subseteq P(G)$ and $W \subseteq \mathbb{R}^Q$. We define $W_{par} = \{f(x) \in W : \nexists f(y) \in W$ such that $f(y)$ dominates $f(x)$ in the objective space$\}$ and $\mathscr{X}_{par}^* := \{x \in S : f(x) \in W_{par}\}$. If $S = P(G)$ we simply write $\mathscr{X}_{par}^*$. A point $x \in \mathscr{X}_{par}^*(S)$ is called a Pareto location with respect to S, and the elements of $\mathscr{X}_{par}^*(V)$ are called Pareto nodes or Pareto vertices.

Computing $\mathscr{X}_{par}^*(V)$ can simply be done by pairwise comparison of the nodes. For $\mathscr{X}_{par}^*$ we first have to check if a multicriteria version of Hakimi's node dominance result holds (Hakimi 1964). For the directed case we even have $\mathscr{X}_{par}^*(V) = \mathscr{X}_{par}^*$. The proof relies on the concavity of the distance functions among the edges and also on the fact that in the directed case we have no choice on which side to exit or enter an edge. This implies that the objective function is strictly concave and therefore the nodes always dominate the edges. For the technical details and the proofs the reader is referred to Hamacher et al. (1999). In the case of undirected networks, this aspect is slightly more complicated as shown in the next example (Fig. 9.13).

Example 9.7 Consider the following network $N = (G, \ell)$ with $n = 6$ nodes and a distance matrix $D = (d_{ij})_{i,j=1,\ldots,6}$ given by

$$D = \begin{pmatrix} 0 & 1 & 1 & 4 & 3 & 2 \\ 1 & 0 & 2 & 3 & 4 & 1 \\ 1 & 2 & 0 & 3 & 2 & 3 \\ 4 & 3 & 3 & 0 & 5 & 2 \\ 3 & 4 & 2 & 5 & 0 & 3 \\ 2 & 1 & 3 & 2 & 3 & 0 \end{pmatrix}.$$

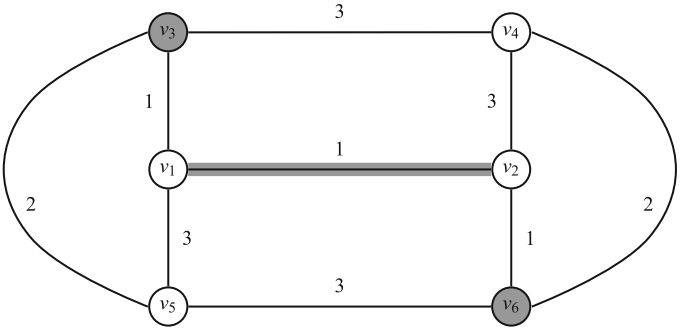


Fig. 9.13 Network of Example 9.7

Assume that the weight vectors are

$$w_1 = \begin{pmatrix}1\\2\end{pmatrix},\ w_2 = \begin{pmatrix}2\\1\end{pmatrix},\ w_3 = \begin{pmatrix}1\\2\end{pmatrix},\ w_4 = \begin{pmatrix}2\\2\end{pmatrix},\ w_5 = \begin{pmatrix}2\\2\end{pmatrix},\ w_6 = \begin{pmatrix}2\\1\end{pmatrix}.$$

Using this information we get

	v_1	v_2	v_3	v_4	v_5	v_6
$f(\cdot)$	$\binom{21}{19}$	$\binom{19}{21}$	$\binom{21}{17}$	$\binom{27}{29}$	$\binom{29}{27}$	$\binom{17}{21}$

By pairwise comparison we get

$$\mathscr{X}^*_{par}(V) = \{v_3\} \cup \{v_6\} = \mathscr{X}^*\left(f^1(V)\right) \cup \mathscr{X}^*\left(f^2(V)\right).$$

Now we look at the points on the edges and get (by using concavity in the objective functions):

- v_3 dominates all points on the edges $\{v_3, v_5\}$, $\{v_3, v_4\}$, $\{v_3, v_1\}$
- v_6 dominates all points on the edges $\{v_6, v_2\}$, $\{v_6, v_5\}$, $\{v_6, v_4\}$
- v_2 dominates all points on the edge $\{v_2, v_4\}$
- v_1 dominates all points on the edge $\{v_1, v_5\}$

We also observe that no vertex can dominate a point with both objective functions smaller than 21. The only edge left is now $\{v_1, v_2\}$ (Fig. 9.14).
We see that

I. For all points $x \in P(\{v_1, v_2\})$ with $x \neq v_1$, $x \neq v_2$ we have $f^1(x) < 21$, $f^2(x) < 21$.
II. No point on $\{v_1, v_2\}$ dominates another point on $\{v_1, v_2\}$

$$\Rightarrow \mathscr{X}^*_{par} = \{v_3\} \cup \{v_6\} \cup (\{v_1, v_2\},\ (0, 1)).$$

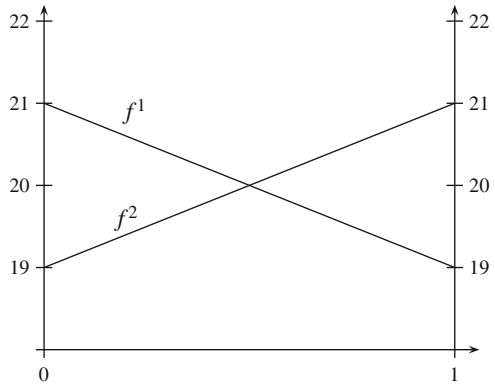


Fig. 9.14 Objective functions on the edge $\{v_1, v_2\}$ in Example 9.7

We conclude that we have no node dominance and that even on edges with endnodes not in $\mathscr{X}^*_{par}(V)$ we can find elements of $\mathscr{X}^*_{par}$.

Since we do not have node dominance in the undirected case, we have to explicitly solve a multicriteria global optimization problem. First we will identify local Pareto locations with respect to an edge $e = \{v_i, v_j\}$ for all edges of the network. In a second step we will compare all local Pareto locations to get $\mathscr{X}^*_{par}$. Due to the limited space and a possible overload of technicalities, we will describe the main ideas which allow the reader to understand the final algorithm. For the technical details and the proofs the reader is referred to Hamacher et al. (1999).

9.3.1.2 Bi-criteria Case

We will first deal with the bi-criteria case, since here we can derive a geometrical solution method. The main property of the objective functions we are using is the concavity on an edge $e = \{v_i, v_j\}$. In addition we have also piecewise linearity but this is not really needed. Suppose that $f(v_i) > f(v_j)$ or $f(v_j) > f(v_i)$. In the first situation we say that v_j dominates v_i and in the latter v_i dominates v_j. Both situations do not allow any location on the edge, which is not dominated by an endnode due to concavity.

Now assume that for an edge $e = \{v_i, v_j\}$ with v_i and v_j not dominating each other one of the functions f^1 or f^2 is constant. It is easy to see that this is only the case if $f(v_i) = f(v_j)$. If for an edge e only one of the objective functions is constant then $\mathscr{X}^*_{par}(e) = \{v_i\} \cup \{v_j\}$. If both objective functions are constant then $\mathscr{X}^*_{par}(e) = \big(\{v_i, v_j\}, [0, 1]\big)$. Again this is due to the concavity of the objective functions and can be seen in Fig. 9.15.

Now we have only one situation left (the most typical one), where the endnodes do not dominate each other and none of the two objective functions is constant. Without loss of generality we can assume $f^1(v_i) > f^1(v_j)$ and $f^2(v_i) < f^2(v_j)$ (otherwise exchange the roles of v_i and v_j). The behaviour of the objective functions can be seen in Fig. 9.16. First, both objectives functions are increasing (maybe for a small or zero interval only) and all points are dominated by the left endnode.

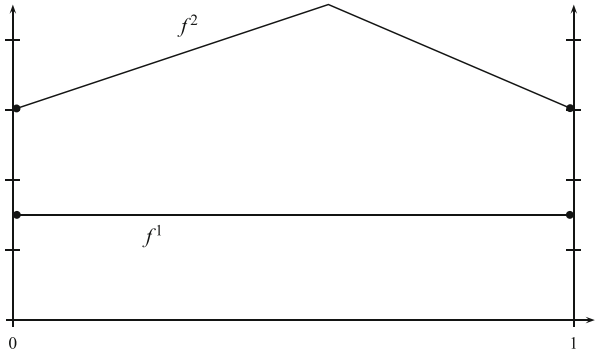


Fig. 9.15 Concavity on an edge with one objective function constant

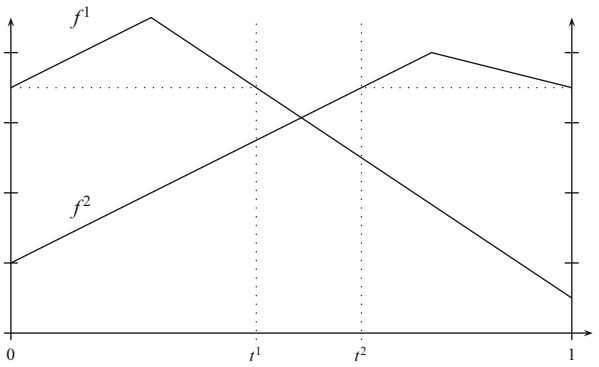


Fig. 9.16 Derivation of t^1 and t^2

Only after the first objective function is already decreasing and smaller than the left endnode value, the endnode cannot dominate the points of the edge. The same argument can be applied by starting from the right endnode. More formally we can define

$$t^1 := \max\{t \in [0,1] : f^1(v_i) = f^1\left(\left(\{v_i, v_j\}, t\right)\right)\}$$

and

$$t^2 := \min\{t \in [0,1] : f^2(v_j) = f^2\left(\left(\{v_i, v_j\}, t\right)\right)\}$$

Then

$$\mathscr{X}^*_{par}(e) = \{v_i\} \cup \{v_j\} \cup \left(\{v_i, v_j\},\ \left(t^1, t^2\right)\right).$$

Overall we have that for each $e \in E$ in (G, ℓ), $\mathscr{X}^*_{par}(e)$ is a (possibly empty) single subedge of e plus one or both endnodes. Now we can combine these results to get an efficient algorithm for determining $\mathscr{X}^*_{par}(e)$.

Algorithm 9.4 (Computation of $\mathscr{X}^*_{par}(e)$)

Input: *edge $e = \{v_i, v_j\} \in E$, undirected network (G, l), distance matrix D*

Step 1. *IF v_i dominates v_j then $\mathscr{X}^*_{par}(e) := \{v_i\}$, go to Step 7*
Step 2. *IF v_j dominates v_i then $\mathscr{X}^*_{par}(e) := \{v_j\}$, go to Step 7*
Step 3. *IF $f(v_i) = f(v_j)$ then*

A. IF $f\left(\left(\{v_i, v_j\}, \frac{1}{2}\right)\right) = f(v_i)$ then $\mathscr{X}^*_{par}(e) := P(\{v_i, v_j\})$, go to Step 7
B. IF $f\left(\left(\{v_i, v_j\}, \frac{1}{2}\right)\right) \neq f(v_i)$ then $\mathscr{X}^*_{par}(e) := \{v_i\} \cup \{v_j\}$, go to Step 7

Step 4. *IF $f^1(v_i) < f^1(v_j)$ and $f^2(v_i) > f^2(v_j)$ then exchange v_i and v_j*
Step 5. *Compute t^1 and t^2 as defined above*

Step 6. *IF* $t^1 < t^2$
THEN $\mathscr{X}^*_{par}(e) := \{v_i\} \cup \{v_j\} \cup \big(\{v_i, v_j\}, (t^1, t^2)\big)$
ELSE $\mathscr{X}^*_{par}(e) := \{v_i\} \cup \{v_j\}$

Output: $\mathscr{X}^*_{par}(e)$

To analyze the complexity of this algorithm, we need the following definition: A point $x = \big(\{v_i, v_j\}, t\big)$, $t \in [0, 1]$ on one edge $e = \{v_i, v_j\}$ is called a bottleneck point for f^q if there exists a vertex v_k with $w^q_k > 0$, such that

$$d(v_k, x) = d(v_k, v_i) + d(v_i, x) = d(v_k, v_j) + d(v_j, x).$$

Let B_{ij} denote the set of bottleneck points on the edge $\{v_i, v_j\}$. Note that $|B_{ij}| \leq |V|$.

If D is given, the only non constant operation in Algorithm 9.4 is the computation of t^1 and t^2. To plot f^q we have to determine the breakpoints of f^q which is piecewise linear on an edge. Since these breakpoints correspond to the bottleneck points on this edge we have to compute B_{ij} for $e = \{v_i, v_j\}$. This can be done in $O(|V| \log |V|)$ (see Hansen et al. 1991). Then t^1 and t^2 can be determined by exploring the sorted list of bottleneck points two times. The total complexity for finding $\mathscr{X}^*_{par}(e)$ is $O(|V| \log |V|)$ and the total complexity for finding $\bigcup_{e \in E} \mathscr{X}^*_{par}(e)$ is $O(|E||V| \log |V|)$ (Fig. 9.17).

Example 9.8 Consider the network in Fig. 9.17 with distance matrix

$$D = \begin{pmatrix} 0 & 1 & 2 & 2 \\ 1 & 0 & 2 & 1 \\ 2 & 2 & 0 & 1 \\ 2 & 1 & 1 & 0 \end{pmatrix}.$$

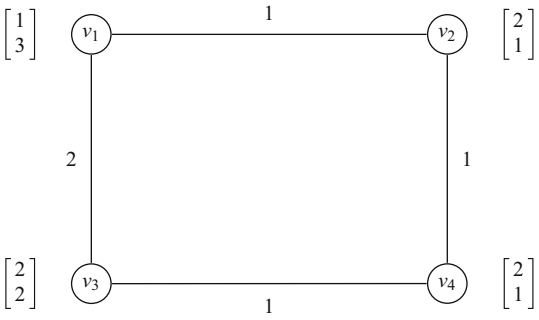


Fig. 9.17 Network of Example 9.8

We first compute

	v_1	v_2	v_3	v_4
f^1	10	7	8	6
f^2	7	8	9	9

and obtain $\mathscr{X}^*_{par}(V) = \{v_1, v_2, v_4\}$. Now we have to determine the set $\mathscr{X}^*_{par}(e)$ for every $e \in E$:

- $e = \{v_1, v_2\}$. v_1 and v_2 do not dominate each other and $f^1,\ f^2$ are not constant, i.e., we need to plot $f^1,\ f^2$ and therefore we have to find B_{12}

$$B_{12} = \left\{ b^1_{12} = \left(\{v_1, v_2\}, \frac{1}{2} \right) \right\}$$

$$f^1\left(b^1_{12}\right) = 9.5 \quad \text{and} \quad f^2\left(b^1_{12}\right) = 8.5$$

So the objective function can be drawn as shown in Figs. 9.18 and 9.19.

$$t^1 = \max\left\{ t \in [0,1] :\ f^1(v_1) = f^1\left(\{v_1, v_2\}, t\right) \right\} = 0$$

$$t^2 = \min\left\{ t \in [0,1] :\ f^2(v_2) = f^2\left(\{v_1, v_2\}, t\right) \right\} = \frac{1}{3}$$

$$\left(\text{in } [0, \frac{1}{2}], \quad f^2(x) \equiv 7 + 3t,\ 7 + 3t = 8 \Leftrightarrow t = \frac{1}{3}\right)$$

$$\mathscr{X}^*_{par}(e) = \{v_1\} \cup \{v_2\} \cup \left(\{v_1, v_2\}, \left(0, \frac{1}{3}\right) \right)$$

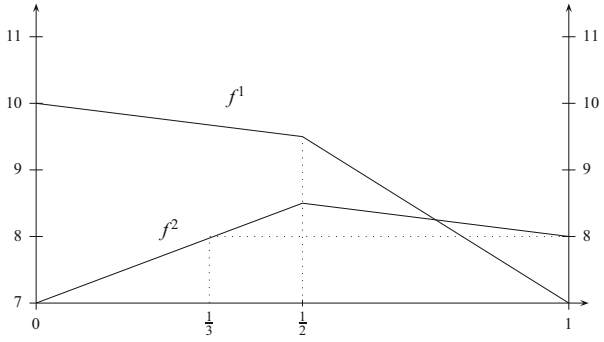


Fig. 9.18 Computing $\mathscr{X}^*_{par}(\{v_1, v_2\})$

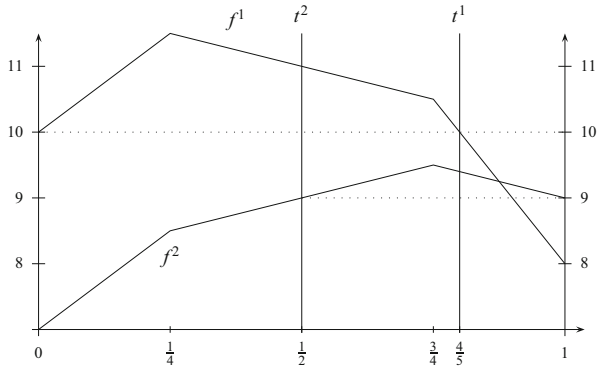


Fig. 9.19 Computing $\mathscr{X}^*_{par}(\{v_1, v_3\})$

- $e = \{v_2, v_4\}$. $f^1(v_2) = 7 > f^1(v_4) = 6$ and $f^2(v_2) = 8 < f^2(v_4) = 9$ and $B_{24} = \emptyset \Rightarrow t_1 = 0$, $t_2 = 1 \Rightarrow \mathscr{X}^*_{par}(e) = P(e)$.
- $e = \{v_3, v_4\}$. v_4 dominates $v_3 \Rightarrow \mathscr{X}^*_{par}(e) = \{v_4\}$.
- $e = \{v_1, v_3\}$. $B_{13} = \left\{ \left(\underbrace{\{v_1, v_3\}}_{b^1_{13}}, \frac{1}{4} \right), \left(\underbrace{\{v_1, v_3\}}_{b^2_{13}}, \frac{3}{4} \right) \right\}$

$$f\left(b^1_{13}\right) = \begin{pmatrix} 11.5 \\ 8.5 \end{pmatrix}, \quad f\left(b^2_{13}\right) = \begin{pmatrix} 10.5 \\ 9.5 \end{pmatrix}$$

$$t_1 = \frac{4}{5}, \quad t_2 = \frac{1}{2}$$

$$\mathscr{X}^*_{par}(e) = \{v_1\} \cup \{v_3\}$$

In a second step we have to compare all local Pareto locations $\mathscr{X}^*_{par}(e)$, $e \in E$ to get $\mathscr{X}^*_{par}$. With two objective functions we can map everything to the objective space where dominance can easily be computed. In the case of median objective functions on a network, we know that f^1 and f^2 are piecewise linear with the same potential breakpoints. This leads to the following mapping in the (z^1, z^2)-space (or objective space) as shown in Fig. 9.20. Essentially, this plot shows all pairs (z_1, z_2) of the objective function values $f_1(x)$ and $f_2(x)$ for all points x on the edge. Again we would like to skip the technical details and proofs and refer the reader to Hamacher et al. (1999).

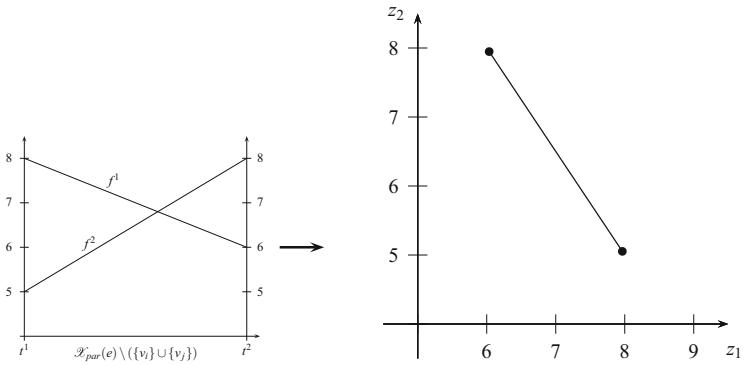


Fig. 9.20 Mapping $\mathscr{X}^*_{par}(e)$ to the objective space

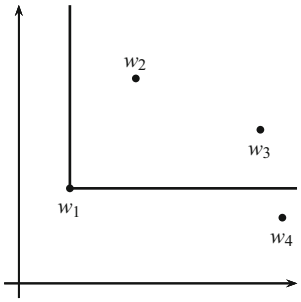


Fig. 9.21 w_1 is dominating w_2 and w_3

In the objective space, a point w dominates all other points in $w + \mathbb{R}^2_+\setminus\{0\} := \{w + y : y \in \mathbb{R}^2_+\setminus\{0\}\}$ (see Fig. 9.21).

In order to obtain $\mathscr{X}^*_{par}$ we draw $\mathrm{IM}(f)$ which is defined as the set of all images of $\mathscr{X}^*_{par}(e)$ for $e \in E$ in the objective space. The lower envelope for a set P of points in $\mathbb{R}^2$ is defined as

$$\bigcup \{(x, y) \in P : y \leq y' \text{ for all } (x, y') \in P\}.$$

Algorithm 9.5 (Combining the Local Pareto Locations)

Input: *$\mathscr{X}^*_{par}(e)$ for all $e \in E$*

Step 1. *Let* $z^1_{\max} := \max\left\{f^1(x) : x \in \bigcup_{e\in E} \mathscr{X}^*_{par}(e)\right\}$

Step 2. *Build* $\mathrm{IM}(f) = \bigcup_{e\in E} f\left(\mathscr{X}^*_{par}(e)\right)$

Step 3. *For each connected component l in $\mathrm{IM}(f)$, let (z^1_l, z^2_l) be the right-most point (largest z^1 value) and add to $\mathrm{IM}(f)$ the horizontal segment going from (z^1_l, z^2_l) to $(z^1_{\max}, z^2_l)$.*

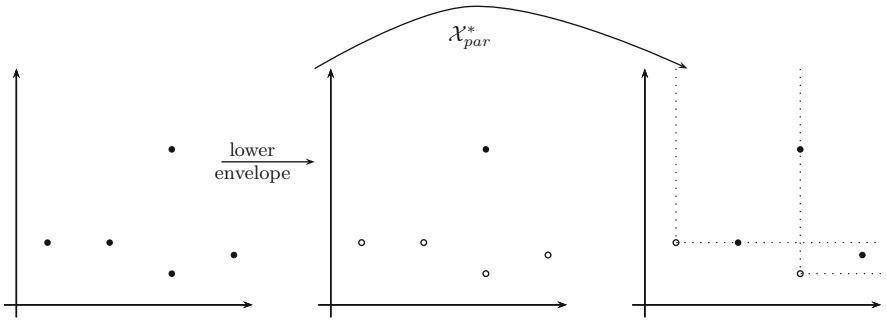


Fig. 9.22 Using the lower envelope to delete dominated solutions

Step 4. *Compute the lower envelope L of* IM(f)*, which is the lower envelope of $O(|E||V|)$ line segments.*
Step 5. *Eliminate every horizontal line segment of L, except its left-most point.*
Step 6. *Set* $\mathcal{X}^*_{par} := f^{-1}(L)$.

Output: $\mathcal{X}^*_{par}$

In order to get the same result from the dominance relation we have to add an artificial line segment and delete it from the solution (see Fig. 9.22).

Steps 1 and 3 are necessary to modify IM(f) such that we can get $\mathcal{X}^*_{par}$ form the lower envelope. These steps as well as Step 2 can be done in linear time. Step 4 can be done in a naive way in $O\left(|E|^2|V|^2\right)$ or in optimal time of $O\left(|E||V|\log\left(\max\left(|E||V|\right)\right)\right)$ by an algorithm of Hershberger (1989). Since Step 5 can be done in linear time the complexity of Step 4 determines the overall complexity. For easier handling of the segments, note that we may use instead of an open subedge $\left(\{v_i, v_j\}, (t_1, t_2)\right)$ the closed subedge $\left(\{v_i, v_j\}, [t_1, t_2]\right)$. After applying the algorithm we then have to test if we deleted a point directly above the left-most point.

Example 9.9 (Example 9.8 cont.) We first draw IM(f) and add the horizontal line segments. Finally, we get $\mathcal{X}^*_{par} = P\left(\{v_2, v_4\}\right) \cup \left(\{v_1, v_2\}, \left[0, \frac{1}{3}\right)\right)$ (Fig. 9.23).

9.3.1.3 Q-Criteria Case

We will now briefly explain how this approach generalizes to the Q-criteria case. Also in this situation we easily see that if for an edge $e = \{v_i, v_j\}$ one endnode dominates the other one, there are no Pareto locations in the interior of e. From now on assume that neither v_i dominates v_j nor v_j dominates v_i. Let $\mathcal{Q}_1$ and $\mathcal{Q}_2$ be a *partition* of $\mathcal{Q}$, such that $f^q(v_i) \geq f^q(v_j)$ for all $q \in \mathcal{Q}_1$ and $f^q(v_i) < f^q(v_j)$ for all $q \in \mathcal{Q}_2$. Of course, $\mathcal{Q}_1 \neq \emptyset$, $\mathcal{Q}_1 \cap \mathcal{Q}_2 = \emptyset$ and $\mathcal{Q}_1 \cup \mathcal{Q}_2 = \mathcal{Q}$. Also in case

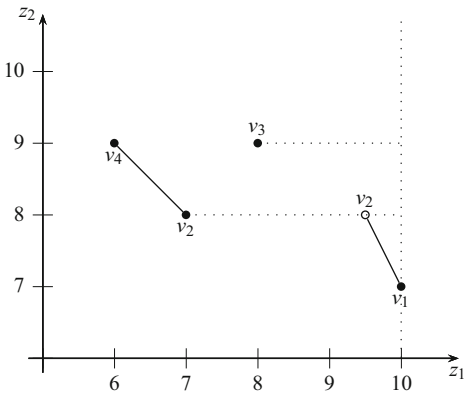


Fig. 9.23 Computing $\mathscr{X}^*_{par}$ for Example 9.8

of constant functions we get a similar result as in the bi-criteria case. Accordingly, assume that $f(v_i) \neq f(v_j)$ for an edge $e = \{v_i, v_j\}$ and let

$$t^1(f^q) := \max\left\{t \in [0,1] : \ f^q(v_i) = f^q\left((\{v_i, v_j\}, t)\right)\right\} \text{ for } q \in \mathscr{Q}_1$$

and

$$t^2(f^q) := \min\left\{t \in [0,1] : \ f^q(v_j) = f^q\left((\{v_i, v_j\}, t)\right)\right\} \text{ for } q \in \mathscr{Q}_2.$$

Then (see Hamacher et al. 1999 for the details)

$$\mathscr{X}^*_{par}(e) = \{v_i\} \cup \{v_j\} \cup \left(\{v_i, v_j\}, \left(\min_{q\in\mathscr{Q}_1}\left\{t^1(f^q)\right\}, \max_{q\in\mathscr{Q}_2}\left\{t^2(f^q)\right\}\right)\right).$$

For comparing the local Pareto locations, the mapping to the objective space becomes rather involved especially when we have to compute lower envelopes.

In order to compare $\mathscr{X}^*_{par}(e)$ for all $e \in E$ pairwise, we use the following iterative procedure: Let $\left(\{v_j, v_l\}, [t_r, t_{r+1}]\right)$ be a subedge of $\mathscr{X}^*_{par}(e_l)$, $e_l = \{v_j, v_l\}$ (to have closed subedges we neglect the vertices and handle first only the Pareto parts in the interior) where (t_r, t_{r+1}) are assumed to not include any further bottleneck points of e_l (if this is not true we subdivide the subedge further). This leads to

$$f^q\left((\{v_j, v_l\}, t)\right) = b^q_r + m^q_r t \quad \text{for all} \quad q \in \mathscr{Q}, \ t \in [t_r, t_{r+1}],$$

i.e., all f^q are affine linear on $\left(\{v_j, v_l\}, [t_r, t_{r+1}]\right)$. Take now a closed linear subedge from another edge $e_k = \{v_k, v_m\}$, then we get $\left(\{v_k, v_m\}, [s_p, s_{p+1}]\right) \subseteq \mathscr{X}^*_{par}(e_k)$. This leads to

$$f^q\left((\{v_k, v_m\}, s)\right) = b^q_p + m^q_p s \quad \text{for all} \quad q \in \mathscr{Q}, \ s \in [s_p, s_{p+1}],$$

If we apply the definition of a Pareto location to these two subedges, we get that a point $\big(\{v_j, v_l\},\ t\big),\ t \in [t_r,\ t_{r+1}]$ is dominated by some point $(\{v_k, v_m\},\ s),\ s \in [s_p,\ s_{p+1}]$

$$\Leftrightarrow \quad b_p^q + m_p^q s \le b_r^q + m_r^q t \quad \text{for all} \quad q \in \mathcal{Q},$$

where at least one inequality is strict. Now we define the polyhedron

$$\mathcal{F} := \Big\{(s,t) :\ m_r^q t - m_p^q s \ge b_p^q - b_r^q,\ \forall q \in \mathcal{Q}\Big\} \cap \big([s_p,\ s_{p+1}] \times [t_r,\ t_{r+1}]\big).$$

We have two cases: If $\mathcal{F} = \emptyset$, then $\big(\{v_j, v_l\},\ [t_r,\ t_{r+1}]\big)$ contains no point which is dominated by a point from $\big(\{v_k, v_m\},\ [s_p,\ s_{p+1}]\big)$. Otherwise, $\mathcal{F} \neq \emptyset$ is taken as a feasible solution of the two 2-variable linear programs

$$\mathrm{LB} = \min\{t :\ (s,t) \in \mathcal{F}\}, \quad \mathrm{UB} = \max\{t :\ (s,t) \in \mathcal{F}\}.$$

Let s_{LB} and s_{UB} be the optimal values for s corresponding to LB and UB, respectively. Now we still have to check if one inequality is strict: If $b_p^q + m_p^q s_{\mathrm{LB}} = b_r^q + m_r^q \mathrm{LB}$ and $b_p^q + m_p^q s_{\mathrm{UB}} = b_r^q + m_r^q \mathrm{UB}$ for all $q \in \mathcal{Q}$, then there is no dominance. Otherwise $\mathcal{X}_{par}^*(e_l) := \mathcal{X}_{par}^*(e_l) \setminus \big(\{v_j, v_l\},\ [\mathrm{LB},\ \mathrm{UB}]\big)$. Note that this procedure works also if $t_r = t_{r+1}$ or $s_p = s_{p+1}$ (in this case, we are testing a single point).

Algorithm 9.6 (Combining Local Pareto Location in the Q-Criteria Case)

Input: *Network as in Algorithm 9.4*

Step 1. *Determine $\mathcal{X}_{par}^*(e)$ for all $e \in E$ and set $\mathcal{X}_{par}^* := \bigcup_{e \in E} \mathcal{X}_{par}^*(e)$*

Step 2. *Compare all v_i and all edges, where all f^q, $q \in \mathcal{Q}$ are constant*

Step 3. *For all Pareto linear subedges do a pairwise comparison as described above and reduce $\mathcal{X}_{par}^*$ accordingly.*

Output: $\mathcal{X}_{par}^*$

The complexity of this algorithm is $O(|E|^2|V|^2Q)$.

9.3.1.4 Multicriteria Median Problems on a Tree

Many difficult problems on general networks become easier to solve if the underlying graph has a tree structure. We will show that this is also true for multicriteria problems. We relate our results with the research that has previously been done on trees and end up with a generalization of Goldman's algorithm (see Goldman 1971). The major concept which makes the analysis easier on trees is convexity. We first introduce this concept based on Dearing et al. (1976).

Let $N = (T, \ell)$ be a tree network, with $T = (V, E)$. For two points $a, b \in P(T)$ we define the line segment $L[a, b]$ between a and b as

$$L[a, b] := \{x \in P(T) : \ d(a, x) + d(x, b) = d(a, b)\},$$

which contains all points on the unique path between a and b. A subset $C \subseteq P(T)$ is called convex, if and only if for all $a, b \in C$, $L[a, b] \subseteq C$.

Now let $C \subseteq P(T)$ be convex and let $h : P(T) \to \mathbb{R}$ be a real valued function. This function h is called convex on C, if and only if for all $a, b \in C$,

$$h(x_\lambda) \leq \lambda h(a) + (1 - \lambda)h(b) \ , \ \forall \lambda \in [0, 1] \ ,$$

where x_λ is uniquely defined by

$$d(x_\lambda, b) = \lambda d(a, b) \text{ and } d(x_\lambda, a) = (1 - \lambda)d(a, b) \ . \tag{9.5}$$

A function is called convex on T if it is convex on $C = P(T)$. Note that it is possible to define convexity also on general networks. Then one can show that $d(x, c)$ for $c \in P(T)$ fixed is convex if and only if the underlying graph is a tree. Median and Center objective functions are convex functions on a tree (see Dearing et al. 1976).

Now let $L(a, b) := L[a, b]\backslash\{a, b\}$, $L(a, b] := L[a, b]\backslash\{a\}$ and $L[a, b) := L[a, b]\backslash\{b\}$. We have now the following important property (a proof can be found in Hamacher et al. 1999).

Theorem 9.9 *Let $a, b \in P(T)$ and $h := (h^1, \ldots, h^Q)$ be a vector of Q objective functions, with h^q convex on T, for all $q \in \mathcal{Q} = \{1, \ldots, Q\}$. Then the following holds:*

$$\{a, b\} \subseteq \mathcal{X}^*_{par} \ \textit{if and only if} \ L[a, b] \subseteq \mathcal{X}^*_{par} \ .$$

For $T = (V, E)$ and $V' \subseteq V$ let

$$W(V') := \begin{pmatrix} w^1(V') \\ w^2(V') \\ \vdots \\ w^Q(V') \end{pmatrix} ,$$

where $w^q(V') := \sum_{v_i \in V'} w_i^q$, $\forall q \in \mathcal{Q}$.

Proposition 9.1 *Let T be partitioned in such a way that $T = T_1 \cup T_2 \cup \{e\}$ (and $T_1 \cap T_2 = \emptyset$). Then $W(V(T_1))$ dominates $W(V(T_2))$ if and only if for all $x \in P(T_1)$ there exists some $y \in P(T_2)$ which dominates x.*

Now we can state a multicriteria version of Goldman's dominance algorithm (see Goldman 1971). We start with a subtree containing only one leaf of the tree

(check for dominance) and enlarge this subtree until we get a Pareto location using the criterion established in Proposition 9.1. This procedure is then repeated for all leaves and we end up with a subtree of all Pareto locations by using Theorem 9.9.

Algorithm 9.7 (Solving Q-Criteria Median Problems on a Tree)

Input: *$T = (V, E)$, with length function ℓ and node weight vectors w^q, $q \in \mathcal{Q}$.*

Step 0. *Set $W := W(V)$*
Step 1. *Choose a leaf v_k of T, which was not yet considered and give it the status "considered".*
Step 2. *IF $V = \{v_k\}$*
Set $\mathcal{X}^*_{par}(f(V)) := \mathcal{X}^*_{par}(f(T)) := \{v_k\}$ and go to Step 6
Step 3. *Let v_l be the only node adjacent to v_k*
IF $(w^1_k \dots w^Q_k)^T < \frac{1}{2} W$
THEN
- $w^q_l := w^q_l + w^q_k, \quad q = 1, \dots, Q$
- $T := T \setminus \{v_k\}$

Step 4. *IF there are any leaves left in T give them status "not considered" and go to Step 1*
Step 5. *Set $\mathcal{X}^*_{par}(f(V)) := V(T)$, $\mathcal{X}^*_{par}(f(T)) := T$*
Step 6. *STOP*

Output: *$\mathcal{X}^*_{par}(f(V))$ and $\mathcal{X}^*_{par}(f(T))$*

The complexity of this algorithm is $O(Q|V|)$. To illustrate the algorithm consider the following example:

Example 9.10 Consider the tree depicted in Fig. 9.24. We solve the following instance of a three-criteria median problem. Let $l(e) := 1$, $\forall e \in E$. The weights of the nodes are given in the following table:

	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}
w^1	14	6	8	4	1	2	1	3	2	2	7
w^2	11	3	3	24	5	2	2	3	2	2	5
w^3	16	2	1	1	2	3	3	1	6	4	21

Therefore $W = \begin{pmatrix} 50 \\ 62 \\ 60 \end{pmatrix}$ and $\frac{1}{2}W = \begin{pmatrix} 25 \\ 31 \\ 30 \end{pmatrix}$.

The adjacency structure of the tree is also given in Fig. 9.24. Now we check every leaf till there is none left with status *"not considered"*.

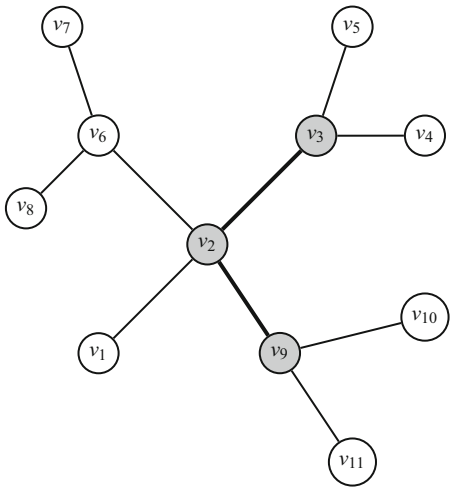


Fig. 9.24 Tree of Example 9.10. The *bold edges* and *nodes* indicate the set of Pareto locations

- Take v_1: $w_1 = \begin{pmatrix} 14 \\ 11 \\ 16 \end{pmatrix}$ dominates $\frac{W}{2} = \begin{pmatrix} 25 \\ 31 \\ 30 \end{pmatrix}$.
  Therefore $w_2 := \begin{pmatrix} 6+14 \\ 3+11 \\ 2+16 \end{pmatrix} = \begin{pmatrix} 20 \\ 14 \\ 18 \end{pmatrix}$.

By following the algorithm we delete v_8, v_7, v_6, v_5 and v_4. The actual value of w_3 is $\begin{pmatrix} 13 \\ 32 \\ 4 \end{pmatrix}$.

- Take v_3: $w_3 = \begin{pmatrix} 13 \\ 32 \\ 4 \end{pmatrix}$ does not dominate $\frac{W}{2}$.
- Take v_{11}: $w_{11} = \begin{pmatrix} 7 \\ 5 \\ 21 \end{pmatrix}$ dominates $\frac{W}{2}$. Therefore $w_9 := \begin{pmatrix} 9 \\ 7 \\ 27 \end{pmatrix}$.
- Take v_{10}: $w_{10} = \begin{pmatrix} 2 \\ 2 \\ 4 \end{pmatrix}$ dominates $\frac{W}{2}$. Therefore $w_9 := \begin{pmatrix} 11 \\ 9 \\ 31 \end{pmatrix}$.
- Take v_9: $w_9 = \begin{pmatrix} 11 \\ 9 \\ 31 \end{pmatrix}$ does not dominate $\frac{W}{2}$.

Since we delete after every domination step the corresponding node from the tree according to Algorithm 9.7 and no leaf with status not considered is left we end up with

$$\mathscr{X}_{par}^{*} = L[v_9, v_3] \ .$$

9.3.2 *Other Multicriteria Location Problems on Networks*

In the previous two subsections we presented optimal time algorithms for one facility median problems when looking for Pareto locations. We chose these two problems because the reader gets some insight into the needed properties. In addition, the simplification on trees caused by the uniqueness of paths can be seen. In the recent survey Nickel et al. (2005a) an overview on other location problems can be found. In Hamacher et al. (2002) an extension to 1-facility center problems as well as to positive and negative weight vectors on the nodes is developed. Those ideas have been further extended to problems with criteria dependent lengths in Skriver et al. (2004). A unified framework for multicriteria ordered median functions can be found in Nickel and Puerto (2005). In Colebrook and Sicilia (2007b) the location of undesirable facilities on multicriteria networks is looked at by using convex combinations of two objective functions. Some complexity analysis for the cent-dian location problem has been developed by Colebrook and Sicilia (2007a). Most approaches to the (in general NP-hard) multi-facility case are treated as discrete location problems (see Sect. 9.4). Only recently Kalcsics et al. (2014) started looking into polynomial cases of multi-facility multicriteria location problems on networks.

9.4 Discrete Location Problems

The previous sections show that planar and network multicriteria location problems have been widely developed from a methodological point of view so that important structural results and algorithms are known to determine solution sets. On the contrary, multicriteria analysis of discrete location problems has attracted less attention. In spite of that, several authors have dealt with problems and applications of multicriteria decision analysis in this field. An annotated bibliography with many references up to 2005 can be found in Nickel et al. (2005a). In general, very few papers focus in the complete determination of the whole set of Pareto-optimal solutions. Nevertheless, there are some exceptions, such as the paper by Ross and Soland (1980) that gives a theoretical characterization but does not exploit its algorithmic possibilities, as well as the work by Fernández and Puerto (2003) that addresses the computation of the entire set of Pareto-optimal solutions of the multiobjective uncapacitated plant location problem.

Nowadays, Multi-Objective Combinatorial Optimization (MOCO) (see Ehrgott and Gandibleux 2000; Ulungu and Teghem 1994) provides an adequate framework to tackle various types of discrete multicriteria problems as, for instance, the p-Median Problem (p-MP). Within this emergent research area, several methods are known to handle different problems. It is worth noting that most of MOCO problems are NP-hard and intractable (see Ehrgott and Gandibleux 2000, for further details). Even in most of the cases where the single objective problem is polynomially solvable the multiobjective version becomes NP-hard. This is the case of spanning tree problems and min-cost flow problems, among others. In the case of the p-MP, the single objective version is already NP-hard. This ensures that the multiobjective formulation is not solvable in polynomial time unless P=NP. In this context, when time and efficiency become a real issue, different alternatives can be used to approximate the Pareto-optimal set. One of them is the use of general-purpose MOCO heuristics (Gandibleux et al. 2000). Another possibility is the design of "ad hoc" methods based on one of the following strategies: (1) computing supported non-dominated solutions; and (2) performing partial enumerations of the solutions space. Obviously, the second strategy does not guarantee the non-dominated character of all the generated solutions although the reduction in computation time can be remarkable.

The aim of this section is to present methods to obtain the Pareto-optimal set for the multiobjective p-median problem (p-MP). In all cases, our approach to solve the multicriteria p-MP takes advantage of the problem's structure. The first method is exact and it determines the whole set of Pareto-optimal solutions based on new tools borrowed from the theory of short rational generating functions. The second method is an "ad hoc" approximate method that generates supported Pareto locations.

9.4.1 Model and Notation

Let $I = \{1, \ldots, M\}$ and $J = \{1, \ldots, N\}$ respectively denote the sets of indices for demand points and for plants, and $\mathscr{Q} = \{1, \ldots, Q\}$ denote the set of indices for the considered criteria. For each criterion $q \in \mathscr{Q}$, let $(c_{ij}^q)_{i \in I, j \in J} \in \mathbb{Q}^{M \times N}$ be the allocation costs of demand points to plants. The multicriteria p-median location problem is:

$$\text{v-Minimize} \left(\sum_{i=1}^{M} \sum_{j=1}^{N} c_{ij}^1 x_{ij}, \ldots, \sum_{i=1}^{M} \sum_{j=1}^{N} c_{ij}^q x_{ij} \right) \tag{9.6}$$

$$\text{subject to} \quad \sum_{j=1}^{N} x_{ij} = 1, \quad i \in I, \tag{9.7}$$

$$x_{ij} \leq y_j, \quad i \in I, \ j \in J, \tag{9.8}$$

$$\sum_{j=1}^{N} y_j = p, \tag{9.9}$$

$$x_{ij} \in \{0,1\},\ y_j \in \{0,1\}, \quad i \in I,\ j \in J. \tag{9.10}$$

As it is usual, v-min stands for vector minimum of the considered objective functions. Here variable y_j takes the value 1 if plant j is open and 0 otherwise. The binary variable x_{ij} is 1 if the demand point i is assigned to plant j and 0 otherwise. Constraints (9.7), together with integrality conditions on the x variables, ensure that each demand point is assigned to exactly one plant, while constraints (9.8) guarantee that no demand point is assigned to a non-open plant. Finally, constraint (9.9) ensures that exactly p plants are opened.

Recall that in the single criterion case the integrality conditions on the x variables need not be explicitly stated. The reason is that when the x_{ij} represent the proportion of demand of client i satisfied by plant j (i.e. $0 \leq x_{ij} \leq 1$), there exists an optimal solution with $x_{ij} = 0, 1, i \in I, j \in J$ This property is not necessarily true when multiple criteria are considered because, in general, there might be undominated solutions with non-integer values and even non-supported undominated integer solutions.

9.4.2 *Determining the Entire Set of Pareto-Optimal Solutions*

In order to characterize the set of Pareto locations of the p-MP we shall use rational generating functions. Short rational generating functions were used by Barvinok (1994) as a tool to develop an algorithm for counting the number of integer points inside convex polytopes, based on the previous geometrical paper by Brion (1988). The main idea is to encode those integer points in a rational function of as many variables as the dimension of the space where the polytope is defined. Let $P \subset \mathbb{R}^n_+$ be a given convex bounded polyhedron. Its integer points may be expressed in a formal sum $\mathrm{f}(P,z) = \sum_\alpha z^\alpha$ with $\alpha = (\alpha_1, \ldots, \alpha_n) \in P \cap \mathbb{Z}^n$, where $z^\alpha = z_1^{\alpha_1} \cdots z_n^{\alpha_n}$ Barvinok's goal was to represent that formal sum of monomials in the multivariate polynomial ring $\mathbb{Z}[z_1, \ldots, z_n]$, as a "short" sum of rational functions with the same variables. Actually, Barvinok (1994) developed a polynomial-time algorithm when the dimension, n, is fixed, to compute those functions. A clear example is the polytope $P = [0, T] \subset \mathbb{R}$ with $T \in \mathbb{N}$: the long expression of the generating function of the integer points inside P is $\mathrm{f}(P,z) = \sum_{i=0}^{T} z^i$, and it is easy to see that its representation as sum of rational functions is the well known formula $(1 - z^{T+1})/(1 - z)$.

The above approach, apart from counting lattice points, has been used to develop some algorithms to solve integer programming problems exactly. Specifically, De Loera et al. (2004, 2005), and Woods and Yoshida (2005) presented different

methods to solve this family of problems using Barvinok's rational function of the polytope defined by the feasible set of the given problem.

First of all, for the sake of readability, we recall some results on short rational functions for polytopes that shall be later used in our presentation. For further details the interested reader is referred to Barvinok (1994), Barvinok and Woods (2003).

Let $P = \{x \in \mathbb{R}^n : A\,x \leq b, x \geq 0\}$ be a rational polytope in $\mathbb{R}^n$. The main idea of Barvinok's Theory was to encode the integer points inside a rational polytope in a "long" sum of monomials:

$$\mathrm{f}(P, z) = \sum_{\alpha \in P \cap \mathbb{Z}^n} z^{\alpha},$$

where $z^{\alpha} = z_1^{\alpha_1} \cdots z_n^{\alpha_n}$, and then to re-encode, in polynomial-time for fixed dimension, these integer points in a "short" sum of rational functions in the form

$$\mathrm{f}(P; z) = \sum_{i \in I} \varepsilon_i \frac{z^{u_i}}{\prod_{j=1}^{n} (1 - z^{v_{ij}})},$$

where I is a polynomial-size indexing set, $\varepsilon_i \in \{1, -1\}$, and $u_i, v_{ij} \in \mathbb{Z}^n$ for all i and j (Theorem 5.4 in Barvinok and Woods 2003).

It is well-known that enumerating the entire set of Pareto-optimal solutions of general multiobjective integer linear problems is #P-hard even in fixed dimension (see, e.g., Ehrgott and Gandibleux 2002 and Chinchuluun and Pardalos 2007). Therefore listing these solutions, in general, is hopeless. Nevertheless, one can try to represent these sets in polynomial time using a different strategy by simply encoding their elements in an efficient way. This strategy has been recently applied by Blanco and Puerto (2012). In that paper, it is proved that using short generating functions of rational polytopes, one can encode the whole set of Pareto-optimal solutions of MOILP in polynomial time, fixing only the dimension of the space of variables. As an application of this result we can state the following theorem.

Theorem 9.10 *Assume that the number of facilities M and plants N is fixed. Then, in polynomial time, we can encode the entire set of Pareto-optimal solutions for* (9.6)–(9.10) *in a short sum of rational functions.*

Proof Apply Theorem 1 in Blanco and Puerto (2012) to the polytope of Problem (9.6)–(9.10). □

The combination of Theorem 9.10 and Theorem 7 in De Loera et al. (2009) results in the following theorem.

Theorem 9.11 *Assume M and N are constant. There exists a polynomial-delay polynomial-space procedure to enumerate the entire set of Pareto-optimal solutions of* (9.6)–(9.10).

This construction can be implemented for problems of small to medium size dimension using the open source software `barvinok`, see Verdoolaege (2008).

9.4.3 Determining Supported Pareto-Optimal Solutions

In some situations it suffices to generate the set of supported Pareto-optimal points. It is well-known that the set of supported Pareto-optimal solutions to a problem can be obtained by solving the scalarized problem for all possible values of the scalar weights in the standard Q-dimensional simplex $\Lambda^Q = \{\lambda \in \mathbb{R}^Q : \sum_{q=1}^{Q} \lambda^q = 1,\ \lambda^q \geq 0,\ \forall q = 1, \ldots, Q\}$.

In order to describe how to obtain these solutions in Problem (9.6)–(9.10) we need to introduce some additional notation. We denote by B any feasible basis of the linear relaxation of Problem (9.6)–(9.10); and by $\overline{N}$ all the columns that are not in B. Also, abusing notation, as usual in linear programming, we shall refer to the indices determining the basis B ($\overline{N}$) in the variables and the objective function by $(x, y)_B$ ($(x, y)_{\overline{N}}$) and c_B ($c_{\overline{N}}$), respectively.

For any $\lambda \in \Lambda^Q$, we shall denote by $c(\lambda) = (c_{ij}(\lambda))_{ij}$, where $c_{ij}(\lambda) = \sum_{q=1}^{Q} \lambda^q c_{ij}^q$.

For each feasible basis B, consider the subdivision of the space Λ^Q induced by the hyperplanes:

$$\lambda^q c_B^q B^{-1}\overline{N} - \lambda^q c_{\overline{N}}^q = 0, \quad q \in \mathcal{Q}.$$

Next, let $\lambda_B^Q \in \Lambda^Q$ be a parameter such that it belongs to the relative interior of one of the elements in the above subdivision and satisfies $c_B(\lambda^Q)B^{-1}\overline{N} - c_{\overline{N}}(\lambda^Q) \leq 0$. This choice of λ^Q ensures that the problem:

$$\text{Minimize} \sum_{i=1}^{M}\sum_{j=1}^{N} c_{ij}(\lambda_B^Q)x_{ij} \tag{9.11}$$

$$\text{subject to} \sum_{j=1}^{N} x_{ij} = 1, \quad i \in I, \tag{9.12}$$

$$x_{ij} \leq y_j, \quad i \in I,\ j \in J, \tag{9.13}$$

$$\sum_{j=1}^{N} y_j = p, \tag{9.14}$$

$$x_{ij} \geq 0,\ y_j \geq 0, \quad i \in I,\ j \in J; \tag{9.15}$$

will identify supported Pareto-optimal solutions of the linear relaxation of Problem (9.6)–(9.10). However, these Pareto-optimal solutions may result in fractional location variables since Problem (9.11)–(9.14) is a scalarization of the continuous version of our original multiobjective location problem. To avoid this inconvenience we shall solve the binary version of (9.11)–(9.14), namely

$$\text{Minimize} \sum_{i=1}^{M} \sum_{j=1}^{N} c_{ij}(\lambda_B) x_{ij} \tag{9.16}$$

$$\text{subject to} \sum_{j=1}^{N} x_{ij} = 1, \quad i \in I, \tag{9.17}$$

$$x_{ij} \leq y_j, \quad i \in I, \ j \in J, \tag{9.18}$$

$$\sum_{j=1}^{N} y_j = p, \tag{9.19}$$

$$x_{ij} \in \{0,1\}, \ y_j \in \{0,1\}, \quad i \in I, \ j \in J. \tag{9.20}$$

Any optimal binary solution of (9.16)–(9.20) gives a supported Pareto-optimal solution of our original multiobjective location problem. Repeating the above process for all feasible basis of Problem (9.6)–(9.10) will result in a set of supported Pareto-optimal solutions for the problem.

9.5 Conclusions

In this chapter we have presented and analyzed some of the most important models of multicriteria location problems considering three different decision spaces: continuous, networks and discrete. This material provides a general overview of the state-of-the-art of the field as well as a number of references that can be used by the interested readers to go for a further analysis of the topic. Emphasis was put on an efficient (if possible) description of the whole set of Pareto locations.

Acknowledgements The authors were partially supported by projects FQM-5849 (Junta de Andalucía\FEDER), Fundación Séneca, grant number 08716/PI/08, the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office and MTM2010-19576-C02-01/02 (Ministry of Economy and Competitiveness\FEDER, Spain).

References

Barvinok A (1994) A polynomial time algorithm for counting integral points in polyhedra when the dimension is fixed. Math Oper Res 19:769–779

Barvinok A, Woods K (2003) Short rational generating functions for lattice point problems. J Am Math Soc 16:957–979

Blanco V, Puerto J (2012) A new complexity result on multiobjective linear integer programming using short rational generating functions. Optim Lett 6:537–543

Brion M (1988) Points entiers dans les polyèdres convexes. Ann Sci Ecole Norm S Sér 4 21:653–663

Carrizosa E, Conde E, Fernández FR, Puerto J (1993) Efficiency in Euclidean constrained location problems. Oper Res Lett 14:291–295

Chinchuluun A, Pardalos PM (2007) A survey of recent developments in multiobjective optimization. Ann Oper Res 154:29–50

Colebrook M, Sicilia J (2007a) A polynomial algorithm for the multicriteria cent-dian location problem. Eur J Oper Res 179:1008–1024

Colebrook M, Sicilia J (2007b) Undesirable facility location problems on multicriteria networks. Comput Oper Res 34:1491–1514

De Loera JA, Haws D, Hemmecke R, Huggins P, Sturmfels B, Yoshida R (2004) Short rational functions for toric algebra and applications. J Symb Comput 38:959–973

De Loera JA, Haws D, Hemmecke R, Huggins P, Yoshida R (2005) A computational study of integer programming algorithms based on Barvinok's rational functions. Discrete Optim 2:135–144

De Loera JA, Hemmecke R, Köppe M (2009) Pareto optima of multicriteria integer linear programs. INFORMS J Comput 21:39–48

Dearing P, Francis R, Lowe T (1976) Convex location problems on tree networks. Oper Res 24:628–642

Drezner Z (1995) Facility location. A survey of applications and methods. Springer, New York

Durier R (1990) On Pareto optima, the Fermat–Weber problem, and polyhedral gauges. Math Program 47:65–79

Durier R, Michelot C (1985) Geometrical properties of the Fermat-Weber problem. Eur J Oper Res 20:332–343

Durier R, Michelot C (1986) Sets of efficient points in a normed space. J Math Anal Appl 117:506–528

Edelsbrunner H (1987) Algorithms in combinatorial geometry. Springer, New York

Ehrgott M (2005) Multicriteria optimization. Springer, Heidelberg

Ehrgott M, Gandibleux X (2000) A survey and annotated bibliography of multiobjective combinatorial optimization. OR Spectrum 22:425–460

Ehrgott M, Gandibleux X (2002) Multiple criteria optimization. State of the art annotated bibliographic surveys. Kluwer, Boston

Fernández E, Puerto J (2003) Multiobjective solution of the uncapacitated plant location problem. Eur J Oper Res 145:509–529

Gandibleux X, Jaszkiewicz A, Freville A, Slowinski RE (2000) Special issue 'multiple objective metaheuristics'. J Heuristics 6:291–431

Goldman A (1971) Optimal center location in simple networks. Transp Sci 5:212–221

Hakimi S (1964) Optimum location of switching centers and the absolute centers and medians of a graph. Oper Res 12:450–459

Hamacher H, Nickel S (1996) Multicriteria planar location problems. Eur J Oper Res 94:66–86

Hamacher HW, Labbé M, Nickel S (1999) Multicriteria network location problems with sum objectives. Networks 33:79–92

Hamacher HW, Labbé M, Nickel S, Skriver AJ (2002) Multicriteria semi-obnoxious network location problems (MSNLP) with sum and center objectives. Ann Oper Res 110:33–53

Hansen P, Perreur J, Thisse J (1980) Location theory, dominance and convexity: some further results. Oper Res 28:1241–1250
Hansen P, Labbé M, Thisse JF (1991) From the median to the generalized center. RAIRO 25:73–86
Hershberger J (1989) Finding the upper envelope of n line segments in o($n \log n$) time. Inf Process Lett 33:169–174
Kalcsics, J., Nickel, S., Puerto, J. and Rodríguez-Chía, A. M. (2014), Several 2-facility location problems on networks with equity objectives. NETWORKS. doi:10.1002/net.21568
Nickel S (1995) Discretization of planar location problems. Ph.D. dissertation, Fachbereich Mathematik, University of Kaiserslautern
Nickel S (1997) Bicriteria and restricted 2-facility weber problems. Math Method Oper Res 45:167–195
Nickel S, Puerto J (2005) Location theory: a unified approach. Springer, Berlin/Heildelberg
Nickel S, Puerto J, Rodríguez-Chía AM (2005a) MCDM location problems. In: Figueira JA, Greco S, Ehrogott M (eds) Multiple criteria decision analysis: state of the art surveys. International series in operations research & management science, vol 78. Springer, New York, pp 761–787
Nickel S, Puerto J, Rodríguez-Chía AM, Weissler A (2005b) Multicriteria planar ordered median problems. J Optim Theory Appl 126:657–683
Puerto J, Fernández F (1999) Multicriteria minisum facility location problem. J Multi-Criteria Decis Anal 8:268–280
Puerto J, Fernández F (2000) Geometrical properties of the symmetrical single facility location problem. J Nonlinear Convex A 1:321–342
Rockafellar R (1970) Convex analysis. Princeton University Press, Princeton
Rodríguez-Chía A, Puerto J (2002) Geometrical description of the weakly efficient solution set for multicriteria location problems. Ann Oper Res 111:179–194
Rodríguez-Chía A, Nickel S, Puerto J, Fernández F (2000) A flexible approach to location problems. Math Method Oper Res 51:69–89
Ross GT, Soland RM (1980) A multicriteria approach to the location of public facilities. Eur J Oper Res 4:307–321
Skriver AJ, Andersen KA, Holmberg K (2004) Bicriteria network location (BNL) problems with criteria dependent lengths and minisum objectives. Eur J Oper Res 156:541–549
Ulungu E, Teghem J (1994) Multi-objective combinatorial optimization problems: a survey. J Multi-Criteria Decis Anal 3:83–104
Verdoolaege S (2008) Software barvinok. http://freecode.com/projects/barvinok
Warburton A (1983) Quasiconcave vector maximization: connectedness of the sets of pareto-optimal and weak pareto-optimal alternatives. J Optim Theory Appl 40:537–557
Weissler A (1999) General bisectors and their application in planar location theory. Shaker, Aachen
Wendell R, Hurter AJ (1973) Location theory, dominance and convexity. Oper Res 21:314–320
Wendell R, Hurter A, Lowe T (1977) Efficient points in location problems. AIIE Trans 9:238–246
Woods K, Yoshida R (2005) Short rational generating functions and their applications to integer programming. SIAG/OPT Views News 16:15–19

Chapter 10
Ordered Median Location Problems

Justo Puerto and Antonio M. Rodríguez-Chía

Abstract This chapter analyzes the ordered median location problem in three different frameworks: continuous, discrete and networks; where some classical but also new results have been collected. For each solution space we study general properties that lead to resolution algorithms. In the continuous case, we present two solution approaches for the planar case with polyhedral norms (the most intuitive case) and a novel approach applicable for the general case based on a hierarchy of semidefinite programs that can approximate up to any degree of accuracy the solution of any ordered median problem in finite dimension spaces with polyhedral or ℓ_p-norms. We also cover the problems on networks deriving finite dominating sets for some particular classes of λ parameters and showing the impossibility of finding a FDS with polynomial cardinality for general lambdas in the multifacility case. Finally, we present a covering based formulation for the capacitated discrete ordered median problem with binary assignment which is rather promising in terms of gap and CPU time for solving this family of problems.

Keywords Finite dominating set • Mixed integer linear programming • Ordered median function

10.1 Introduction

The Ordered Median location problem, see Nickel and Puerto (2005), has been recognized as a powerful tool from a modeling point of view within the field of Location Analysis. Actually, this problem provides a common framework for most of the classical location problems (median, center, k-centrum, centdian, trimmed-mean, among others) as well as for others which have not been studied before. As an illustrative example, in the well-known case of logistics supply chain networks,

J. Puerto (✉)
IMUS, Universidad de Sevilla, Sevilla, Spain
e-mail: puerto@us.es

A.M. Rodríguez-Chía
Dpto. Estadística e Investigación Operativa, Universidad de Cádiz, Cádiz, Spain
e-mail: antonio.rodriguezchia@uca.es

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_10

this modeling tool allows to distinguish the roles played by the different parties in the network inducing new type of distribution patterns, see Kalcsics et al. (2010a,b). This type of formulation incorporates flexibility through rank dependent compensation factors, and it allows one to model that the driving force in a distribution problem is shared by its different parties.

The goal of the ordered median location problem is to minimize the ordered weighted average of the distances or transportation costs, between the clients/demand points and the server, once we have applied rank dependent compensation factors on them. These rank dependent weights allow, for instance, to compensate unfair situations. Indeed, if a solution places a set of facilities so that the accessibility cost of a demand point at j is in the s-th position in the ordered sequence of cost between each client and its corresponding server and the cost of a demand point at j' is in the t-th position with $s < t$, the model tries to favor j with respect to j' by assigning weights $\lambda_s \leq \lambda_t$. (Note that these weights do not penalize site j' but instead they compensate site j because these lambdas reduce the dispersion of the costs.) In order to incorporate this ordinal information in the overall transportation cost, the objective function applies a correction factor to the transportation cost for each demand point (to reach the facility) which is dependent on the position of that cost relative to similar costs from other demand points. For example, a different penalty might be applied if the transportation cost of a demand point at j was the 5th-most expensive cost rather than the 2nd-most expensive, see Boland et al. (2006), Marín et al. (2009), Nickel and Puerto (2005), Puerto and Fernández (2000), Rodríguez-Chía et al. (2000). It is even possible to neglect some costs by assigning a zero penalty. This adds a "sorting"-problem to the underlying location problem, making formulation and solution more challenging.

This type of objective function has been extensively studied and successfully applied in a variety of problems within the literature of Location Analysis. Puerto and Fernández (2000) and Papini and Puerto (2004) characterize the structure of optimal solutions sets. Rodríguez-Chía et al. (2000, 2010), Blanco et al. (2013, 2014a), Espejo et al. (2009), Nickel et al. (2005), Drezner (2007) and Drezner and Nickel (2009a,b), among others, develop algorithms for different continuous ordered median location problems. In addition, there are nowadays some successful approaches available when the framework space is either discrete (see Boland et al. 2006; Domínguez-Marín et al. 2005; Espejo et al. 2009; Marín et al. 2009, 2010; Puerto et al. 2011, 2013, 2014) or a network (see Berman et al. 2009; Kalcsics et al. 2003, 2002; Nickel and Puerto 1999; Puerto and Tamir 2005; Puerto and Rodríguez-Chía 2005).

The aim of this chapter is to introduce the reader into the field of ordered median location providing some modeling tools and properties. These elements will allow to formulate and solve location problems in different solution spaces (continuous, networks and discrete settings) using this unifying tool. To achieve this goal, in the next section we formally introduce the family of ordered median functions (OMf). Sections 10.3.2, 10.4 and 10.5 are devoted to analyze the ordered median location problem in three different frameworks: continuous, networks and discrete, respectively. The chapter ends with some concluding remarks.

10.2 The Ordered Median Function

As mentioned above, the structure of Ordered Median Functions involves a nonlinearity in the form of an ordering operation that introduces a degree of complication but at the same time gives an extra freedom which allows a lot of flexibility in modeling. In this section, we will review interesting properties of these functions in a first step to understand their behavior and then, we shall give a characterization of this objective function.

We start defining the ordered median function. This function is a weighted average of ordered elements. For any $x \in \mathbb{R}^n$ denote $x_{ord} = (x_{(1)}, \ldots, x_{(n)})$ where $x_{(1)} \leq x_{(2)} \leq \ldots x_{(n)}$. We consider the function:

$$\begin{aligned} sort_n : \mathbb{R}^n &\longrightarrow \mathbb{R}^n \\ x &\longrightarrow x_{ord}. \end{aligned} \tag{10.1}$$

Definition 10.1 The function $f_\lambda : \mathbb{R}^n \longrightarrow \mathbb{R}$ is an ordered median function, for short $f_\lambda \in \mathsf{OMf}(n)$, if $f_\lambda(x) = \langle \lambda, sort_n(x) \rangle$ for some $\lambda = (\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n$, where $\langle , \rangle$ denotes the usual scalar product in $\mathbb{R}^n$.

It is clear that ordered median functions are nonlinear. Whereas the nonlinearity is induced by the sorting. One of the consequences of this sorting is that the pseudolinear representation given in Definition 10.1 is pointwise defined. Nevertheless, one can identify its linearity domains. (See Puerto and Fernández 2000; Nickel and Puerto 2005; Rodríguez-Chía et al. 2000.) The identification of these regions provides us with a subdivision of the framework space where in each of its cells the function is linear. Obviously, the topology of these regions depends on the space and on the lambda vector. A detailed discussion can be found in Puerto and Fernández (2000). As mentioned in Sect. 10.1, different choices of lambda lead also to different functions within the same family: $\lambda = (1/n, \ldots, 1/n)$ is the mean average, $\lambda = (0, \ldots, 0, 1)$ is the center, $\lambda = (\alpha, \ldots, \alpha, \alpha, 1)$ is the α-centdian, $\alpha \in [0, 1]$, $\lambda = (0, \ldots, 0, 1, \overset{k}{\ldots}, 1)$ is the k-centrum or $\lambda = (\alpha, 0, \ldots, 0, 1 - \alpha)$ is Hurwicz's criterion, see Chaps. 1, 2 and 4 for further details.

These functions are not new and some operators related to them have been developed by other authors independently. This is the case of the ordered weighted operators (OWA) studied by Yager (1988) to aggregate semantic preferences in the context of artificial intelligence; as well as SAND functions (isotone and sublinear functions) introduced by Francis et al. (2000) to study aggregation errors in multifacility location models.

First, we recall some simple properties and remarks concerning ordered median functions. Most of them are natural questions that appear when a family of functions is considered. Partial answers are summarized in the following proposition.

Proposition 10.1 *Let $f_\lambda(x), f_\mu(x) \in \mathit{OMf}(n)$.*

I. $f_\lambda(x)$ *is a continuous function.*
II. $f_\lambda(x)$ *is a symmetric function, i.e. for any* $x \in \mathbb{R}^n$ $f_\lambda(x) = f_\lambda(sort_n(x))$.
III. $f_\lambda(x)$ *is a convex function iff* $\lambda_1 \leq \ldots \leq \lambda_n$.
IV. If c_1 *and* c_2 *are constants, then the function* $c_1 f_\lambda(x) + c_2 f_\mu(x) \in \mathit{OMf}(n)$.
V. If $\{f_{\lambda^r}(x)\}$ *is a set of ordered median functions that pointwise converges to a function* f, *then* $f \in \mathit{OMf}(n)$.
VI. If $\{f_{\lambda^r}(x)\}$ *is a set of ordered median functions, all bounded above in each point* x *of* $\mathbb{R}^n$, *then the pointwise maximum (or sup) function defined at each point* x *is not in general an* OMf.
VII. Let $p < n-1$ *and* $x^p = (x_1, \ldots, x_p)$, $x^{\setminus p} = (x_{p+1}, \ldots, x_r)$. *If* $f_\lambda(x) \in \mathit{OMf}(n)$ *then* $f_{\lambda^p}(x^p) + f_{\lambda^{\setminus p}}(x^{\setminus p}) \lesseqgtr f_\lambda(x)$.
VIII. Every ordered median function $\mathit{OMf}(n)$ *is a difference of two positively homogeneous convex functions and has a representation*

$$f_\lambda(x) = \sum_{i=1}^{n} \lambda_1 \varphi_i(x),$$

where $\varphi_r(x) = \min\left\{\max\{x_{i_1}, x_{i_2}, \ldots, x_{i_r}\} | i_1 < i_2 < \ldots < i_r \text{ and } i_1, i_2, \ldots, i_r \in \{1, \ldots, n\}\right\}$.

Proof The proof of (1) can be found in Rosenbaum (1950). The proof of (3) and (8) are in Grzybowski et al. (2011). The proofs of items (2) and (4) are straightforward and therefore are omitted. A proof of (5) and counterexamples for (6) and (7) are given in Nickel and Puerto (2005, Examples 1.1 and 1.2). □

In order to continue the analysis of the ordered median function we need to introduce some notation that will be used in the following. Let $\mathscr{P}(1 \ldots n)$ be the set of all the permutations of the first n natural numbers,

$$\mathscr{P}(1 \ldots n) = \{\pi : \pi \text{ is a permutation of } 1, \ldots, n\}. \tag{10.2}$$

We write $\pi = (\pi(1), \ldots, \pi(n))$.

The next result, that we include for the sake of completeness, is well-known and its proof can be found in the book by Hardy et al. (1952).

Lemma 10.1 *Let* $x = (x_1, \ldots, x_n)$ *and* $y = (y_1, \ldots, y_n)$ *be two vectors in* $\mathbb{R}^n$. *Suppose that* $x \leq y$, *then* $x_{ord} = (x_{(1)}, \ldots, x_{(n)}) \leq y_{ord} = (y_{(1)}, \ldots, y_{(n)})$.

To understand the nature of the OMf we need a precise characterization. This will be done in the following two results using the concepts of symmetry and sublinearity.

Theorem 10.1 *A function* f *defined over* $\mathbb{R}^n_+$ *is continuous, symmetric and linear over* $\{x : 0 \leq x_1 \leq \ldots \leq x_n\}$ *if and only if* $f \in \mathit{OMf}(n)$.

Proof Since f is linear over $X^{\leq} := \{x \geq 0 : 0 \leq x_1 \leq \ldots \leq x_n\}$, there exists $\lambda = (\lambda_1, \ldots, \lambda_n)$ such that for any $x \in X^{\leq}$ $f(x) = \langle \lambda, x \rangle$. Now, let us consider any $y \notin X^{\leq}$. There exists a permutation $\pi \in \mathscr{P}(1 \ldots n)$ such that $y_\pi \in X^{\leq}$. By the symmetry property it holds $f(y) = f(y_\pi)$. Moreover, for y_π we have $f(y_\pi) = \langle \lambda, y_\pi \rangle$. Hence, we get that for any $x \in \mathbb{R}^n$

$$f(x) = \langle \lambda, x_{ord} \rangle.$$

Finally, the converse is trivially true. □

There are particular instances of the λ vector that make their analysis interesting. One of them is the convex case, i.e., $\lambda_1 \leq \ldots \leq \lambda_n$, where we can obtain a characterization without the explicit knowledge of a linearity region.

Theorem 10.2 *Given $\lambda = (\lambda_1, \ldots, \lambda_n)$ with $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$; and $\lambda_\pi = (\lambda_{\pi(1)}, \ldots, \lambda_{\pi(n)})$ with $\pi \in \mathscr{P}(1 \ldots n)$, a symmetric function f defined over $\mathbb{R}^n$ is the support function of the set $S_\lambda = conv\{\lambda_\pi : \pi \in \mathscr{P}(1 \ldots n)\}$ if and only if f is the convex ordered median function*

$$f_\lambda(x) = \sum_{i=1}^{n} \lambda_i x_{(i)}. \tag{10.3}$$

Proof Let us assume that f is symmetric and the support function of S_λ. Then,

$$f(x) = \sup_{s \in S_\lambda} \langle s, x \rangle = \sup_{\pi \in \mathscr{P}(1 \ldots n)} \langle \lambda_\pi, x \rangle = \sup_{\pi \in \mathscr{P}(1 \ldots n)} \langle \lambda, x_\pi \rangle = \sum_{i=1}^{n} \lambda_i x_{(i)}.$$

Conversely, it suffices to apply Theorem 368 in Hardy et al. (1952) to (10.3). □

Convexity is an important property within the scope of continuous optimization. Thus, it is crucial to know the conditions that ensure this property. Nevertheless, in the context of discrete optimization convexity cannot even be defined. Nevertheless, in this case submodularity plays a similar role. (The interested reader is referred to the chapter of the Handbook Discrete Optimization by McCormick (2005).) In the following, we also prove a submodularity property of the convex ordered median function, Puerto and Tamir (2005).

Let $x = (x_i)$, $y = (y_i)$, be vectors in $\mathbb{R}^n$. Define the *meet* of x, y to be the vector $x \bigwedge y = (\min\{x_i, y_i\})$, and the *join* of x, y by $x \bigvee y = (\max\{x_i, y_i\})$. The meet and join operations define a lattice on $\mathbb{R}^n$.

Theorem 10.3 (Submodularity Theorem) *Given $\lambda = (\lambda_1, \ldots, \lambda_n)$, satisfying $0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$, $f_\lambda(x)$ is submodular over the lattice defined by the above meet and join operations, i.e.,*

$$f_\lambda(x \bigvee y) + f_\lambda(x \bigwedge y) \leq f_\lambda(x) + f_\lambda(y), \quad \forall x, y \in \mathbb{R}^n.$$

10.3 The Continuous Ordered Median Problem

This section is devoted to the analysis of the Ordered Median Location Problem in a continuous framework. For the ease of understanding, we have divided this section in two main parts. In the first one, we restrict ourselves to the polyhedral gauges emphasizing the planar case. In this setting one can derive nice geometrical properties that help to capture the main elements of the problem, namely its linearity domains, ordered regions and intuitive algorithms for obtaining the optimal solutions. Second, we address the general case where we shall apply a new global optimization technique that allows us to handle and solve a wide range of ordered median location problems.

10.3.1 The Single Facility Polyhedral Ordered Median Location Problem

Consider a set of demand points $A = \{a_1, a_2, \ldots, a_n\} \subset \mathbb{R}^n$ (representing existing facilities or clients) and two sets of non negative scalars $w = (w_1, \ldots, w_n)$ and $\lambda = (\lambda_1, \ldots, \lambda_n)$. The element w_i is the weight assigned to the existing facility a_i and it represents the importance of this demand point. The elements of λ allow us to choose between different kinds of objective functions. We also consider a gauge $\gamma(\cdot) : \mathbb{R}^n \longrightarrow \mathbb{R}$ to measure distances. Recall that any gauge is defined by the Minkowski functional of a compact, convex set with the zero in its interior (see Nickel and Puerto 2005).

The ordered median problem is given by:

$$\min_{x \in \mathbb{R}^n} F(x) = \langle \lambda, sort_n((\gamma(x - a_1), \ldots, \gamma(x - a_n))) \rangle. \tag{10.4}$$

Note that the problem is well-defined even if ties occur. In that case any order of the tied positions gives the same value.

Example 10.1 Consider two demand points $a_1 = (0, 0)$ and $a_2 = (10, 5)$, $\lambda_1 = 100$ and $\lambda_2 = 1$ with ℓ_1-norm and $w_1 = w_2 = 1$. We obtain only two optimal solutions to problem (10.4), lying in each demand point. Observe that a linear representation of the objective function is regionwise defined and that the objective function is not convex since we have a nonconvex optimal solution set. See Fig. 10.1.

$$F(a_1) = 100 \times 0 + 1 \times 15 = 15$$
$$F(a_2) = 100 \times 0 + 1 \times 15 = 15$$
$$F(\frac{1}{2}(a_1 + a_2)) = 100 \times 7.5 + 1 \times 7.5 = 757.5$$

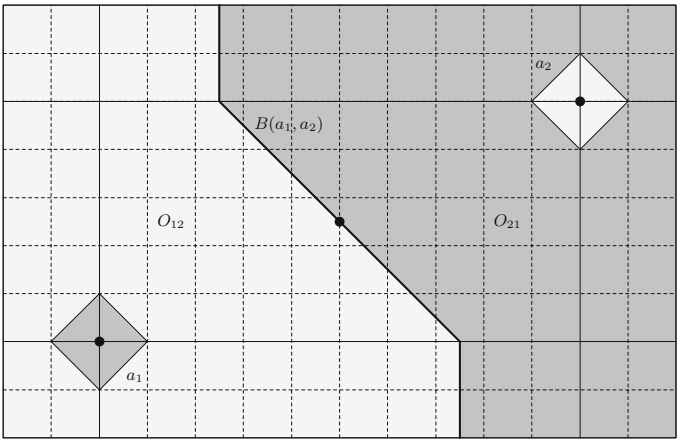


Fig. 10.1 Illustration to Example 10.1

In this section, for the sake of presentation, we restrict ourselves to study the particular case where the distances are measured with polyhedral gauges, i.e., the unit balls associated with these gauges are convex polytopes. For this reason we will assume in this subsection that $B \subseteq \mathbb{R}^n$ is a bounded polytope whose interior contains the zero and we denote the set of extreme points of B by $Ext(B) = \{e_g : g = 1, \ldots, G\}$. The polar set B^0 of B is given by $B^0 = \{x \in \mathbb{R}^n : \langle x, p\rangle \leq 1 \quad \forall p \in B\}$. In the polyhedral case, B^0 is also a polytope, see Ward and Wendell (1985) and Durier and Michelot (1985). The normal cone to B at x is given by $N(B, x) := \{p \in \mathbb{R}^n : \langle p, y - x\rangle \leq 0 \quad \forall y \in B\}$ and the boundary of B is denoted by $bd(B)$.

In what follows, we recall some geometrical properties of the planar formulation of problem (10.4) which give us specific insights into the considered model. In this case we define fundamental directions as the halflines defined by 0 and the extreme points of B. Let $\pi = (p_i)_{i=1,\ldots,n}$ be a family of elements of $\mathbb{R}^2$ such that $p_i \in B^0$ for each $i \in \{1, \ldots, n\}$ and let $C_\pi = \bigcap_{i=1}^n (a_i + N(B^o, p_i))$. A nonempty convex set C is called an elementary convex set (e.c.s.) if there exists a family π such that $C_\pi = C$.

It should be noted that if the unit balls are polytopes we can obtain the elementary convex sets as intersections of cones generated by fundamental directions of these balls pointed at each demand point. Therefore each elementary convex set is a polyhedron whose vertices are called intersection points (see Fig. 10.1). Finally, we recall that in the planar case an upper bound of the number of elementary convex sets is $O(n^2G^2)$ where G is the number of extreme points of B (see Durier and Michelot (1985) for further details).

Although the objective function of problem (10.4) may look like the one of the Weber problem we do not have a unified linear representation of such a function in the whole space. From the definition of the objective function, it is easy to see, that

the representation may change every time $\gamma(x-a_i)-\gamma(x-a_j)$ becomes 0 for some $i, j \in \{1,\ldots,n\}$ with $i \neq j$. Next, we analyze the sets where the representation of the objective function as a weighted sum stays unchanged.

Definition 10.2 The set $B_\gamma(a_i, a_j)$ consisting of points $\{x \ : \ w_i\gamma(x-a_i) = w_j\gamma(x-a_j), i \neq j\}$ is called bisector of a_i and a_j with respect to γ.

As an illustration of Definition 10.2 one can see in Fig. 10.1 the bisector line for the points a_1 and a_2 with the ℓ_1-norm. The set of bisectors builds a subdivision of the plane (very similar to the well-known order-k Voronoi diagrams, see the book Okabe et al. 1992). The cells of this subdivision will be called from now on ordered regions. We formally introduce this concept.

Definition 10.3 Given a permutation $\sigma \in \mathscr{P}(1,\ldots,n)$, the ordered region O_σ is the following set

$$O_\sigma = \{x \in \mathbb{R}^2 : w_{\sigma_1}\gamma(x-a_{\sigma_1}) \leq \ldots \leq w_{\sigma_n}\gamma(x-a_{\sigma_n})\}.$$

Observe that these regions need not be convex sets, see Fig. 10.1. The ordered regions play a very important role in the algorithmic approach developed for solving the problem. Moreover, under the above hypothesis the overall number of ordered regions in the planar case is $O(n^4G^2)$, see Rodríguez-Chía et al. (2000) for further details. The importance of these regions is that the ordered median function has a unique linear representation within the intersection of any ordered region with any elementary convex set. The sets resulting of these intersections are called generalized elementary convex sets and it is known that the entire set of optimal solutions of problem (10.4) always coincides with some generalized elementary convex sets, see Puerto and Fernández (2000) for further details.

Although the set of optimal solutions of problem (10.4) always coincides with a generalized elementary convex set, the large number of these regions and their intricate geometry requires some kind of good generation and enumeration schemes to derive an algorithm. This approach is possible in the plane for polyhedral gauges. One can easily derive an appealing geometrical algorithm to solve these problems in the plane. Compute the subdivision of the plane induced by the lines defining the fundamental directions of the gauges and the bisectors. Observe that this construction can be efficiently performed using any algorithm to generate subdivisions induced by arrangements of hyperplanes, see Edelsbrunner (1987). The complexity of computing the ordered regions and its number is $O(n^4G^2)$. Next, one needs to evaluate the objective function in each vertex of the subdivision. Each evaluation can be done in $O(nG\log nG)$. This results in an algorithm that solves the problem in the plane with a complexity of $O(n^5G^3\log nG)$.

In what follows we present an alternative, intuitive solution approach for the polyhedral version of the ordered median problem that consists in a enumerative algorithm that solves a linear program per visited ordered region. In order to do that, we first obtain some interesting properties of the following linear program where O_σ is an ordered region defined by the permutation σ:

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} \lambda_i z_{\sigma_i} \\ \text{subject to} & w_i \langle e_g^0, x - a_i \rangle \leq z_i, \; e_g^o \in B^o, i = 1, 2, \ldots, n, \\ & z_{\sigma_i} \leq z_{\sigma_i + 1} \qquad\qquad i = 1, 2, \ldots, n-1 \end{array} \qquad (P_\sigma)$$

where e_g^0 are the extreme points of B^0.

Lemma 10.2 *Let X^* be an optimal solution of P_σ.*

(i) If $X^* \in O_\sigma$ then X^* is also an optimal solution to the ordered median problem constrained to O_σ.

(ii) If $X^* \in O_{\sigma'} \neq O_\sigma$ then the optimal solution of the ordered median problem constrained to $O_{\sigma'}$ is better than the optimal solution of the ordered median problem constrained to O_σ.

Proof (i) At an optimal point X^* in O_σ we have

$$w_i \langle e_{g_i}^o, X^* - a_i \rangle = z_i \quad , i = 1, 2, \ldots, n \text{ , for some } g_i,$$

which means that $z_i = w_i \gamma(X^* - a_i)$ and the result follows.

(ii) At an optimal point X^* of P_σ in $O_{\sigma'}$ we have

$$\langle e_g^o, X^* - a_i \rangle < z_i \quad \text{for all } g$$

for at least one i. This means that we can decrease the objective function by moving from O_σ to $O_{\sigma'}$ and the result follows. □

Based on Lemma 10.2 we develop a another algorithm for this problem. For each ordered region we solve the problem as a linear program which geometrically means either finding the locally best solution in this ordered region or finding out that this region does not contain the global optimum by Lemma 10.2. In the former case two situations may occur. First, if the solution lies in the interior of the considered region (in $\mathbb{R}^n$) then we move to a different one not yet processed and secondly, if the solution is on the boundary we do a local search in the neighborhood regions where this point belongs to. It is worth noting that to accomplish this search a list $\mathscr{L}$ containing the already visited neighborhood regions is used in the algorithm. Besides, it is also important to realize that neither Step 2 nor Step 5 need to explicitly construct the corresponding ordered region. It suffices to evaluate and to sort the distances to the demand points. In addition, this algorithm can be improved in the interesting, important case where $\lambda_1 \leq \ldots \leq \lambda_n$. In this situation the objective function is globally convex and this fact can be exploited to reduce the enumeration of the entire list of ordered regions. Indeed, if one optimal solution of any Problem P_σ is interior to the ordered region O_σ or this solution cannot be improved in adjacent regions then by global convexity this implies that it is the global minimum. Otherwise, one can follow a descent iterative scheme moving from one region to another one not previously visited. The above arguments justify the validity of the following algorithm.

Algorithm 10.1 *Step 1. Choose x^o as an appropriate starting point. Initialize $\mathscr{L} := \emptyset$, $y^* = x^o$.*

Step 2. Consider O_{σ^o} which y^* belong to, where σ^o determines the order.

Step 3. Solve the linear program P_{σ^0}. Let $u^0 = (x_1^0, x_2^0, z_\sigma^0)$ be an optimal solution. If $x^0 = (x_1^0, x_2^0) \notin O_{\sigma^o}$ then let O_{σ^o} be such that $x^0 \in O_{\sigma^o}$ and go to Step 3.

Step 4. Let $y^o = (x_1^0, x_2^0)$.

Step 5. If y^o belongs to the interior of O_{σ^o} then set $y^* = y^0$ and go to Step 8.

Step 6. If $F(y^o) \neq F(y^*)$ then $\mathscr{L} := \{\sigma^0\}$

Step 7. If there exist i and j verifying $\gamma(y^o - a_{\sigma_i^o}) = \gamma(y^o - a_{\sigma_j^o})$ with $i < j$ such that $(\sigma_1^o, \ldots, \sigma_j^o, \ldots, \sigma_i^o, \ldots, \sigma_n^o) \notin \mathscr{L}$ then do

(a) $y^* := y^o$, $\sigma^o := (\sigma_1^o, \sigma_2^o, \ldots, \sigma_j^o, \ldots, \sigma_i^o, \ldots, \sigma_n^o)$
(b) $\mathscr{L} := \mathscr{L} \cup \{\sigma^o\}$
(c) go to Step 3

else go to Step 8 (Optimum found)

Step 8. Output y^*

The above algorithm is efficient in the sense that it is polynomially bounded in fixed dimension. Once the dimension of the problem is fixed, its complexity is dominated by the complexity of solving a linear program for each ordered region. Since the number of ordered regions is polynomially bounded, Algorithm 10.1 is polynomial.

The nice geometry of the problem in the plane allows us to derive the two above algorithms. Nevertheless, this geometry in higher dimension is rather intricate and the above approach, based on building ordered regions, is very difficult since no efficient algorithm for computing bisectors is available in dimension greater than 2.

In spite of that, we will present an alternative algorithm for solving the single facility ordered median problem in any dimension d. To this for, we shall introduce a valid MILP model that provides the optimal solution of the problem. Indeed, consider the following set of binary variables

$$z_{ij} := \begin{cases} 1 & \text{if the distance induced by facility } i \\ & \text{goes in sorted position } j \\ 0 & \text{otherwise.} \end{cases}$$

and the continuous variable

$\theta_j =$ distance between a facility and its server in the j-th position in the ordered sequence of distances between each facility and its corresponding server.

In order to minimize the ordered median function for a given set of nonnegative lambda parameters $\lambda_1, \ldots, \lambda_n$, we define the following problem.

$$\text{minimize} \sum_{j=1}^{n} \lambda_j \theta_j \tag{10.5}$$

$$\text{subject to } (1 - z_{ij})M + \theta_j \geq w_i \langle e_g^0, x - a_i \rangle, \text{ for } e_g^o \in B^o, \ i, j = 1, 2, \ldots, n \tag{10.6}$$

$$\sum_{i=1}^{n} z_{ij} = 1, \quad \text{for } j = 1, \ldots, n \tag{10.7}$$

$$\sum_{j=1}^{n} z_{ij} = 1, \quad \text{for } i = 1, \ldots, n \tag{10.8}$$

$$\theta_j \leq \theta_{j+1}, \quad \text{for } j = 1, \ldots, n-1 \tag{10.9}$$

$$\theta_j \geq 0, \quad \text{for } j = 1, \ldots, n \tag{10.10}$$

$$z_{ij} \in \{0, 1\}, \quad \text{for } i, j = 1, \ldots, n \tag{10.11}$$

$$x \in \mathbb{R}^d. \tag{10.12}$$

Constraints (10.7) and (10.8) define a permutation by placing at each position a single distance to a facility and each distance to a facility at a single sorted position. Constraints (10.6) relate distance values with the values placed in a sorted sequence. Constraint (10.9) imposes that the sorted values are ordered non-increasingly. Finally, (10.10)–(10.12) define the range of variables of the model.

The above approach solves efficiently the problem in any dimension provided that the gauges used to measure distances are polyhedral since problem (10.5)–(10.12) is a MILP that can be handled with any of the nowadays available MIP solvers.

We would like to conclude this section with some comments on several extensions of the considered problem. On the one hand, the multicriteria planar version of the above problem was analyzed in Nickel et al. (2005). On the other hand, the planar case of the ordered median problem using a ℓ_p-norm was also studied by Drezner and Nickel (2009a,b) where techniques of global optimization were used for solving it. In addition, Espejo et al. (2009), Rodríguez-Chía et al. (2010) proposed an adaptation of the Weiszfeld algorithm for the convex version of this problem, i.e., $0 \leq \lambda_1 \leq \ldots \leq \lambda_n$. Finally, we would like to mention some references that consider the multifacility version of particular classes of ordered median problems. These references can be seen as a starting point to dig into this challenging topic. The interested reader is referred to Blanco et al. (2014b), Ben-Israel and Iyigun (2010), Brimberg et al. (2000), Schöbel and Scholz (2010) for different approaches to the continuous multifacility location problem.

10.3.2 Generalized Continuous Ordered Median Location Problems

This section extends the analysis presented above, in Sect. 10.3.1, to the case of non-polyhedral norms and any dimension d. In doing that we shall cast that problem within the more general paradigm of polynomial programming. This approach allows us to apply powerful tools borrowed from the theory of global optimization to solve our original problem, see Blanco et al. (2013). This section contains advanced material which is self-contained. For this reason those non specialized readers not interested in global optimization techniques may decide to skip it without losing continuity with the remaining sections of this chapter.

We are given a set $A = \{a_1, \ldots, a_n\} \subset \mathbb{R}^d$ endowed with an ℓ_τ-norm (here ℓ_τ stands for the norm $\|x\|_\tau = \left(\sum_{i=1}^d |x_i|^\tau\right)^{1/\tau}$, for all $x \in \mathbb{R}^d$); and a feasible domain $\mathbf{K} := \{x \in \mathbb{R}^d : g_j(x) \geq 0, \quad j = 1, \ldots, \ell\} \subset \mathbb{R}^d$, assumed to be a closed semi-algebraic set, i.e. a set defined by a finite number of polynomial inequalities, where each $g_j(x) \in \mathbb{R}[x]$ is a polynomial, being $\mathbb{R}[x]$ the ring of real polynomials in $(x_1, \ldots, x_d)$. Since we are interested in solving location problems we shall assume without loss of generality that we wish to solve the problem in a bounded domain so that $\mathbf{K}$ is compact. The goal is to find a point $x^* \in \mathbf{K}$ minimizing some globalizing function of the distances to the set A. Here, we consider that the globalizing function is rather general and that it is given as an ordered weighted average of polynomials (the reader may observe that the same approach also extends to rational functions, Blanco et al. 2013).

Some well-known examples, that are formulated in the above terms, are the following (see, e.g., Blanquero and Carrizosa 2009; Drezner 2007; Espejo et al. 2009; López-de-los-Mozos et al. 2008 or Nickel and Puerto 2005): $f(u_1, \ldots, u_n) = \sum_{i<j}^n |u_i - u_j|$, is the absolute deviation or envy criterion, $f(u_1, \ldots, u_n) = \sum_{i=1}^n (u_i - 1/n \sum_{j=1}^n u_j)^2$, is the variance function, $f(u_1, \ldots, u_n) = \sum_{j=1}^n w_j / u_j^2$, where w_j are scalar weights, is the obnoxious facility criterion and $f(u_1, \ldots, u_n) = \sum_{j=1}^n b_j / (1 + h_j |u_j|^\lambda)$, with b_j and h_j appropriate weights, is the Huff competitive location objective function.

The main feature and what distinguishes location problems from other general purpose optimization problems, is that the dependence of the decision variables is given through the norms to the demand points in A, i.e. $\|x - a_i\|_\tau$. In this section, we consider a generalized version of the ordered continuous single facility location problems over closed semi-algebraic feasible sets, i.e., the Ordered Median of Polynomial Functions problem:

$$\rho_\lambda := \text{minimize} \{ \sum_{j=1}^m \lambda_j \tilde{f}_{(j)}(x) : x \in \mathbf{K} \}, \qquad \textbf{(OMPF)}$$

where:

- $\lambda_j \in \mathbb{R}$ $j = 1, \ldots, m$ are modeling weights.
- $f_j(u) : \mathbb{R}^n \mapsto \mathbb{R}$, with $f_j(u) \in \mathbb{R}[u_1, \ldots, u_n]$ (the ring of real polynomials in $(u_1, \ldots, u_n)$), $x \in \mathbb{K}$ for all $j = 1, \ldots, m$. We shall define the dependence of f_j to the decision variable $x \in \mathbb{R}^d$ via $u = (u_1, \ldots, u_n)$, where $u_i : \mathbb{R}^d \mapsto \mathbb{R}$, $u_i(x) := \|x - a_i\|_\tau$, $i = 1, \ldots, n$. Therefore, the j-th component of the ordered median objective function of our problems reads as:

$$\begin{aligned} \tilde{f}_j(x) : \mathbb{R}^d &\mapsto \mathbb{R} \\ x &\mapsto \tilde{f}_j(x) := f_j(\|x - a_1\|_\tau, \ldots, \|x - a_n\|_\tau). \end{aligned}$$

  In the classical ordered median problem these functions correspond with the distances from the demand points to the service facility, i.e. $f_j(\|x - a_1\|_\tau, \ldots, \|x - a_n\|_\tau) = \|x - a_j\|_\tau$; thus, in our application to the ordered median problem we will always assume to have $m = n$ and functions $\tilde{f}_j(x) := \|x - a_j\|_\tau$.
- $\mathbf{K} := \{x \in \mathbb{R}^d : g_j(x) \geq 0, \quad j = 1, \ldots, \ell\} \subseteq \mathbb{R}^d$ satisfies Archimedean property. (See Lasserre 2009 for a detail discussion on the Archimedean property and its implications in real algebraic geometry and global optimization. In our setting this property is essentially equivalent to assume compact feasible regions.)
- $\tau := r/s, r, s \in \mathbb{N}, r \geq s$ and $gcd(r, s) = 1$.

First of all, since $\mathbf{K}$ is compact there exist $M' > 0$ such that $\|x\|_2 \leq M'$ for all $x \in \mathbf{K}$. Then, we observe that any feasible solution of **OMPF** satisfies $\|x - a_i\|_2 \leq M' + \|a_i\|_2 \leq M' + \max_{1 \leq i \leq n} \|a_i\|_2 := M$. Then, since all norms are equivalent in $\mathbb{R}^d$, there exists $\gamma > 0$ such that $\|x\|_{2\tau}/\|x\|_2 \leq \gamma$, for all $x \in \mathbb{R}^d$. Hence, $\|x - a_i\|_{2\tau} \leq \gamma M =: \bar{M}$. This bound will allow us to derive the constraints (10.21) of our reformulation of Problem **OMPF**. These constraints ensure that the feasible region is bounded which in our framework is sufficient to imply compactness. For this reason, we will call them from now on *compactness* constraints.

Next, our goal is to cast the above problem within the framework of polynomial optimization. Associated with the above minimization problem we introduce an equivalent formulation that will be useful to apply the moment tools to solve the ordered median problem. For each $i = 1, \ldots, m$, $j = 1, \ldots, m$ consider the following family of decision variables for each $x \in \mathbf{K}$

$$w_{ij} = \begin{cases} 1 \text{ if } \tilde{f}_i(x) = \tilde{f}_{(j)}(x), \\ 0 \text{ otherwise.} \end{cases}.$$

However, we observe that ℓ_τ-norms are not, in general, polynomials. To avoid this inconvenience, we introduce the following auxiliary problem. Observe that this formulation embeds the original problem in a higher dimensional space to represent

the piecewise polynomials that appear in **OMPF** as polynomials in the new set of variables.

$$\overline{\rho}_\lambda = \text{minimize} \sum_{j=1}^{m} \lambda_j \sum_{i=1}^{m} f_i(u) w_{ij} := p_\lambda(x,u,v,w) \tag{10.13}$$

$$\text{subject to} \sum_{j=1}^{m} w_{ij} = 1, \text{ for } i = 1,\dots,m, \tag{10.14}$$

$$\sum_{i=1}^{m} w_{ij} = 1, \text{ for } j = 1,\dots,m, \tag{10.15}$$

$$\sum_{i=1}^{m} w_{ij} f_i(u) \le \sum_{i=1}^{m} w_{ij+1} f_i(u),\ j = 1,\dots,m-1, \tag{10.16}$$

$$w_{ij}^2 - w_{ij} = 0, \text{ for } i,j = 1,\dots,m, \tag{10.17}$$

$$v_{k\ell}^{2s} = (x_\ell - a_{k\ell})^{2r},\ k = 1,\dots,n,\ \ell = 1,\dots,d, \tag{10.18}$$

$$u_k^r = (\sum_{\ell=1}^{d} v_{k\ell})^s,\ k = 1,\dots,n, \tag{10.19}$$

$$\sum_{j=1}^{m} w_{ij}^2 \le 1,\ i = 1,\dots,m, \tag{10.20}$$

$$\sum_{j=1}^{d} v_{ij}^2 \le \bar{M}^{2\tau},\ i = 1,\dots,n, \tag{10.21}$$

$$w_{ij} \in \mathbb{R},\ \forall\, i,j = 1,\dots,m, \tag{10.22}$$

$$v_{k\ell} \ge 0, u_k \ge 0, k = 1,\dots,n, \ell = 1,\dots,d, \tag{10.23}$$

$$x \in \mathbf{K}. \tag{10.24}$$

By means of the w variables, the objective function (10.13) is the ordered weighted sum of the f_i polynomials which can be written as the polynomial p_λ. The first set of constraints (10.14) ensures that for each x, $\tilde{f}_i(x)$ is sorted in a unique position. The second set (10.15) ensures that the jth position is only assigned to one polynomial function. The next constraints (10.16) state that $f_{(1)}(u) \le \cdots \le f_{(m)}(u)$. Constraints (10.17) are added to assure that $w_{ij} \in \{0,1\}$. Next, the two families of constraints (10.18) and (10.19) set u_k^r as the correct value of $\|a_k - x\|_\tau$ (recall that $\tau = r/s$). The last set of constraints (10.20) and (10.21) ensure that Archimedean property holds for the new feasible region $\overline{\mathbf{K}}$ of the above auxiliary problem. (Note that this last set of constraints are redundant but it is convenient to add them for a better description of the feasible set.)

We also observe that the above problem simplifies for those cases where r is even. In these cases, we can replace the constraints (10.18) by the simplest constraints

$$v_{k\ell}^s = (x_k - a_{k\ell})^r, \quad \forall\, k, \ell.$$

This reformulation reduces the degree of the polynomials defining the feasible set.

We illustrate the above formulation with a standard model in location analysis: the k-centrum problem in the plane.

Example 10.2 Let us assume that we are given a set of demand points $A = \{a_1, \ldots, a_n\} \subset \mathbb{R}^2$, where $a_i = (a_{i1}, a_{i2})$, for $i = 1, \ldots, n$. We wish to model the k-centrum ($k < n$) with ℓ_3-distance, i.e. $r = 3$ and $s = 1$, with respect to the demand points in A and a feasible region defined by a set $\mathbf{K}$. It is clear that in this case $d = 2, m = n$ and each function $\tilde{f}_i(x) := \|x - a_i\|_3,\ i = 1, \ldots, n$.

According to the model above this problem can be formulated as follows:

$$\begin{aligned}
\text{minimize} \quad & \sum_{j=n-k+1}^{n} \sum_{i=1}^{n} u_i w_{ij} \\
\text{subject to} \quad & \sum_{i=1}^{n} w_{ij} = 1, && \text{for } j = 1, \ldots, n, \\
& \sum_{i=1}^{n} w_{ij} = 1, && \text{for } j = 1, \ldots, n, \\
& \sum_{i=1}^{n} w_{ij} u_i \le \sum_{i=1}^{n} w_{ij+1} u_i, && j = 1, \ldots, n-1 \\
& w_{ij}^2 - w_{ij} = 0, && \text{for } i, j = 1, \ldots, n, \\
& v_{k\ell}^2 = (x_\ell - a_{k\ell})^6, && k = 1, \ldots, n,\ \ell = 1, \ldots, 2, \\
& u_k^3 = (\sum_{\ell=1}^{d} v_{k\ell}), && k = 1, \ldots, n, \\
& \sum_{j=1}^{n} w_{ij}^2 \le 1, && i = 1, \ldots, n, \\
& \sum_{j=1}^{2} v_{ij}^2 \le \bar{M}^6, && i = 1, \ldots, n, \\
& w_{ij} \in \mathbb{R}, && \forall\, i, j = 1, \ldots, m, \\
& v_{k\ell} \ge 0, u_k \ge 0, && k = 1, \ldots, n, \ell = 1, \ldots, d, \\
& x \in \mathbf{K}.
\end{aligned}$$

Next, we get a result that shows the equivalence between the above polynomial optimization formulation and our location problem (**OMPF**).

Theorem 10.4 *Let x be a feasible solution of (**OMPF**) then there exists a solution (x, u, v, w) for (10.13)–(10.24) such that their objective values are equal. Conversely, if (x, u, v, w) is a feasible solution for (10.13)–(10.24) then there exists a solution (x) for (**OMPF**) having the same objective value. In particular $\rho_\lambda = \overline{\rho}_\lambda$. Moreover, if $\mathbf{K} \subset \mathbb{R}^d$ satisfies Archimedean property then $\overline{\mathbf{K}} \subset \mathbb{R}^{d+m^2+n(d+1)}$ also satisfies Archimedean property.*

The interested reader is referred to Blanco et al. (2013, Theorem 4) for a detailed proof.

Now, we can prove a convergence result that allows us to solve, up to any degree of accuracy, the above class of problems. In order to proceed further we need to introduce some additional material related to the Theory of Moments, Lasserre (2009).

Recall that by $\mathbb{R}[x]$ we denote the ring of real polynomials in the variables $x = (x_1, \ldots, x_d)$, for $d \in \mathbb{N}$ $(d \geq 1)$, and by $\mathbb{R}[x]_r \subset \mathbb{R}[x]$ the space of polynomials of degree at most $r \in \mathbb{N}$ (here $\mathbb{N}$ denotes the set of non-negative integers). We also denote by $\mathscr{B} = \{x^\alpha : \alpha \in \mathbb{N}^d\}$ a canonical basis of monomials for $\mathbb{R}[x]$, where $x^\alpha = x_1^{\alpha_1} \cdots x_d^{\alpha_d}$, for any $\alpha \in \mathbb{N}^d$. Note that $\mathscr{B}_r = \{x^\alpha \in \mathscr{B} : \sum_{i=1}^d \alpha_i \leq r\}$ is a basis for $\mathbb{R}[x]_r$.

For any sequence indexed in the canonical monomial basis $\mathscr{B}$, $\mathbf{y} = (y_\alpha)_{\alpha \in \mathbb{N}^d} \subset \mathbb{R}$, let $\mathrm{L}_\mathbf{y} : \mathbb{R}[x] \to \mathbb{R}$ be the linear functional defined, for any $f = \sum_{\alpha \in \mathbb{N}^d} f_\alpha\, x^\alpha \in \mathbb{R}[x]$, as $\mathrm{L}_\mathbf{y}(f) := \sum_{\alpha \in \mathbb{N}^d} f_\alpha\, y_\alpha$.

The *moment* matrix $\mathrm{M}_r(\mathbf{y})$ of order r associated with $\mathbf{y}$, has its rows and columns indexed by (x^α) and $\mathrm{M}_r(\mathbf{y})(\alpha, \beta) := \mathrm{L}_\mathbf{y}(x^{\alpha+\beta}) = y_{\alpha+\beta}$, for $|\alpha|, |\beta| \leq r$ (here $|a|$ stands for the sum of the coordinates of $a \in \mathbb{N}^d$).

For $g \in \mathbb{R}[x]$ $(= \sum_{\gamma \in \mathbb{N}^d} g_\gamma x^\gamma)$, the *localizing* matrix $\mathrm{M}_r(g\mathbf{y})$ of order r associated with $\mathbf{y}$ and g, has its rows and columns indexed by (x^α) and $\mathrm{M}_r(g\mathbf{y})(\alpha, \beta) := \mathrm{L}_\mathbf{y}(x^{\alpha+\beta} g(x)) = \sum_\gamma g_\gamma\, y_{\gamma+\alpha+\beta}$, for $|\alpha|, |\beta| \leq r$.

Let $\mathbf{y} = (y_\alpha)$ be a real sequence indexed in the monomial basis $(x^\beta u^\gamma v^\delta w^\zeta)$ of $\mathbb{R}[x, u, v, w]$ (with $\alpha = (\beta, \gamma, \delta, \zeta) \in \mathbb{N}^d \times \mathbb{N}^n \times \mathbb{N}^{nd} \times \mathbb{N}^{m^2}$).

Let $h_0(x, u, v, w) := p_\lambda(x, u, v, w)$, and denote $\xi_j := \lceil (\deg g_j)/2 \rceil$ and $v_j := \lceil (\deg h_j)/2 \rceil$, where $\{g_1, \ldots, g_\ell\}$, and $\{h_1, \ldots, h_{3m+m^2+n(d+3)}\}$ are, respectively, the polynomial constraints that define $\mathbf{K}$ and $\overline{\mathbf{K}} \setminus \mathbf{K}$ in (10.13)–(10.24). For $r \geq r_0 := \max\{\max_{k=1,\ldots,\ell} \xi_k, \max_{j=0,\ldots,3m+m^2+n(d+3)} v_j\}$, we introduce the hierarchy of semidefinite programs:

$$
\begin{array}{lll}
\text{minimize}_\mathbf{y} & \mathrm{L}_\mathbf{y}(p_\lambda) & \\
\text{subject to} & \mathrm{M}_r(\mathbf{y}) \succeq 0, & \\
& \mathrm{M}_{r-\xi_k}(g_k, \mathbf{y}) \succeq 0, & k = 1, \ldots, \ell, \\
& \mathrm{M}_{r-v_j}(h_j, \mathbf{y}) \succeq 0, & j = 1, \ldots, 3m + m^2 + n(d+3),
\end{array}
\qquad (\overline{\mathbf{Q}}_r)
$$

with optimal value denoted $\min \overline{\mathbf{Q}}_r$.

Theorem 10.5 *Let* $\overline{\mathbf{K}} \subset \mathbb{R}^{d+m^2+n(d+1)}$ *be the feasible domain of Problem* (10.13)–(10.24)*. Then, with the notation above:*

(a) $\min \overline{\mathbf{Q}}_r \uparrow \rho_\lambda$ *as* $r \to \infty$.
(b) Let $\mathbf{y}^r$ *be an optimal solution of the SDP relaxation* ($\overline{\mathbf{Q}}_r$)*. If*

$$\operatorname{rank} \mathrm{M}_r(\mathbf{y}^r) \;=\; \operatorname{rank} \mathrm{M}_{r-r_0}(\mathbf{y}^r) \;=\; t$$

then $\min \overline{\mathbf{Q}}_r \;=\; \rho_\lambda$ *and one may extract* t *points* $(x^*(k), u^*(k), v^*(k), w^*(k))_{k=1}^t \subset \overline{\mathbf{K}}$*, all global minimizers of problem* (**OMPF**)*.*

Proof The convergence of the semidefinite relaxation $\overline{\mathbf{Q}}_r$ follows from a result by Jibetean and de Klerk (2006, Theorem 9) that it is applied here to the polynomial function in (10.13) and the closed semi-algebraic set $\overline{\mathbf{K}}$. The second assertion on the rank condition, for extracting optimal solutions, follows from applying Lasserre (2009, Theorem 5.7) to the SDP relaxation $\overline{\mathbf{Q}}_r$. □

We also observe that one can exploit the block diagonal structure of the problem (10.13)–(10.21) since the only monomials that appear in that formulation are of the form $x^\alpha u_i^\beta \prod_{j=1}^m v_{ij}^{\gamma_j}$ for all $i = 1, \ldots, m$. Hence, a result similar to Theorem 12 in Blanco et al. (2013) about a sparse reformulation also holds for this problem.

Tables 10.1 and 10.2 present some computational results obtained applying the above technique for different planar ordered median problems. Programs were coded in MATLAB R2010b and executed in a PC with an Intel Core i7 processor at 2 × 2.93 GHz and 8 GB of RAM. The semidefinite programs were solved by calling SDPT3 4.0, Kim-Chuan et al. (2006). We report the CPU times for computing solutions as well as the gap, ϵ_{obj}, with respect to upper bounds obtained with the battery of functions in `optimset` of MATLAB, which only provide approximations on the exact solutions (optimality cannot be certified). In order to compute the accuracy of an obtained solution, we use the following measure for the error (see Blanco et al. 2013):

$$\epsilon_{\text{obj}} = \frac{|\text{the optimal value of the SDP} \;-\; \text{fopt}|}{\max\{1, \text{fopt}\}}, \tag{10.25}$$

where fopt is the approximated optimal value obtained with the functions in `optimset`. The interested reader is referred to Blanco et al. (2013, Section 5) for further details and computational results using the tools in this section applied to location problems.

Table 10.1 Computational results for different location problems in $\mathbb{R}^2$ with ℓ_2-norm

	Weber		Center		k-Centrum k = 0.1*n		k-Centrum k = 0.5*n		Range		Trimmed-mean	
	ℓ_2		ℓ_2		ℓ_2		ℓ_2		ℓ_2		ℓ_2	
n	CPU time	ϵ_{obj}	CPU time	ϵ_{obj}	CPU time	ϵ_{obj}	CPU time	ϵ_{obj}	CPU time	ϵ_{obj}	CPU time	ϵ_{obj}
10	0.31	0.00000127	1.33	0.00000978	1.34	0.00001760	1.34	0.00000455	1.26	−0.11849865	2.98	0.00018438
20	0.68	0.00000005	3.08	0.00001456	3.31	0.00000598	3.18	0.00000111	2.21	−0.06784203	6.34	0.00018729
30	1.00	0.00000003	5.35	0.00046734	6.34	0.00000465	5.50	0.00000123	3.10	−0.02626473	9.96	0.00013896
50	1.70	0.00000005	10.61	0.00001725	11.97	0.00000425	13.22	0.00000048	6.57	−0.07291619	20.89	0.00015183
100	3.55	0.00000004	30.83	0.00000542	38.59	0.00000292	37.58	0.00000020	14.58	−0.02572793	46.62	0.00015415
200	7.05	0.00000004	84.16	0.00001519	99.55	0.00000093	100.39	0.00000044	31.34	−0.03714671	118.09	0.00014847
300	10.66	0.00000003	139.36	0.00000386	164.28	0.00000055	159.49	0.00000005	74.49	−0.03314587	188.91	0.00014136
400	14.27	0.00000003	216.28	0.00000337	240.42	0.00000057	211.09	0.00000010	94.59	−0.04756016	304.58	0.00014574
500	17.74	0.00000003	305.36	0.00000336	328.64	0.00000028	285.02	0.00000012	172.06	−0.05599743	391.78	0.00014832

Table 10.2 Computational results for different location problems in $\mathbb{R}^2$ with ℓ_3-norm

	Weber		Center		k-Centrum k = 0.1*n		k-Centrum k = 0.5*n		Range		Trimmed-mean	
	ℓ_3		ℓ_3		ℓ_3		ℓ_3		ℓ_3		ℓ_3	
n	CPU time	ϵ_{obj}	CPU time	ϵ_{obj}	CPU time	ϵ_{obj}	CPU time	ϵ_{obj}	CPU time	ϵ_{obj}	CPU time	ϵ_{obj}
10	0.44	0.00000029	1.70	0.00000441	1.46	0.00000998	1.45	0.00000512	1.38	−0.10196862	2.87	0.00026887
20	1.01	0.00000007	3.59	0.00001389	3.71	0.00001100	4.15	0.00000065	2.70	−0.02628318	6.75	0.00017690
30	1.50	0.00000044	6.33	0.00001259	6.46	0.00000321	6.93	0.00000056	5.35	−0.09088091	11.19	0.00019343
50	2.50	0.00000018	12.91	0.00000947	13.92	0.00000554	16.20	0.00000048	10.51	−0.07220939	20.62	0.00021732
100	5.21	0.00000012	34.07	0.00000690	42.11	0.00000256	34.41	0.00000040	24.30	−0.03754705	52.83	0.00017720
200	10.73	0.00000010	87.18	0.00000663	111.38	0.00000043	98.39	0.00000028	55.67	−0.04069077	128.14	0.00018684
300	16.07	0.00000008	173.36	0.00001240	180.18	0.00000067	157.35	0.00000017	92.37	−0.07366743	191.46	0.00016696
400	21.30	0.00000015	240.12	0.00001163	262.77	0.00000053	233.61	0.00000010	154.74	−0.02080770	312.34	0.00020440
500	27.46	0.00000010	299.41	0.00000498	341.34	0.00000035	291.80	0.00000006	168.54	−0.01652014	391.24	0.00019197

10.4 The Ordered Median Problem on Networks

Let $N = (G, \ell)$ denote a network with underlying graph $G = (V, E)$, with node set $V = \{v_1, \ldots, v_n\}$ and edge set $E = \{e_1, \ldots, e_m\}$. We restrict ourselves to undirected graphs. Therefore, we write every edge $e \in E$ as $\{i, j\},\ v_i,\ v_j \in V$.

Each edge $e \in E$ is associated with a positive length by means of the function $\ell : E \to \mathbb{R}_+$. By $d(v_i, v_j)$, we denote the length of the shortest path between v_i and v_j measured by ℓ. Through $w : V \to \mathbb{R}_+ \cup \{0\}$, every vertex is assigned a non negative weight.

A point x on an edge $e = \{i, j\}$ is defined as a pair $x = (e, t),\ t \in [0, 1]$, with

$$d(v_k, x) := d(x, v_k) := \min\{d(v_k, v_i) + t\ell(e), d(v_k, v_j) + (1-t)\ell(e)\}. \quad (10.26)$$

The set of all the points of a network (G, ℓ) is denoted by $P(G)$. It should be noted that this set also contains the nodes V.

10.4.1 The Single Facility Ordered Median Problem

In this section we deal with the simplest version of the ordered median problem on networks where just a single location is placed. In order to do that, we consider the following notation. Let

$$d(x) := (w_1 d(v_1, x), \ldots, w_n d(v_n, x))$$

and

$$d_{\leq}(x) := (w_{(1)} d(v_{(1)}, x), \ldots, w_{(n)} d(v_{(n)}, x))$$

a permutation of the elements of $d(x)$, verifying

$$w_{(1)} d(v_{(1)}, x) \leq w_{(2)} d(v_{(2)}, x) \leq \ldots \leq w_{(n)} d(v_{(n)}, x).$$

For the sake of simplicity, let $d_{(i)}(x) := w_{(i)} d(v_{(i)}, x)$.

The ordered median problem on N is defined as

$$f_\lambda(d(x)) := \sum_{i=1}^{n} \lambda_i d_{(i)}(x) \quad \text{with} \quad \lambda = (\lambda_1, \ldots, \lambda_n) \geq 0\ , \quad (10.27)$$

and

$$M(\lambda) := \min_{x \in P(G)} f_\lambda(d(x)). \quad (10.28)$$

In this section we state the fundamental properties of Problem (10.28). We will present a localization result which generalizes the well-known results by Hakimi on finite dominating sets for the center and median problems on networks (Hakimi 1964) and gives some insight in the connection between median and center problems.

For all $v_i,\ v_j \in V,\ i \neq j$ define

$$EQ_{ij} := \{x \in P(G) : w_i d(v_i, x) = w_j d(v_j, x)\} \tag{10.29}$$

and let $EQ := \bigcup\{EQ_{ij} : i, j \text{ with } i \neq j\}$.

The points in EQ are called equilibria points of N. Two points $a, b \in EQ$ are called consecutive, if there is no other $c \in EQ$ on the shortest path between a and b. The points in EQ establish a partition on N with the property that for two consecutive elements $a, b \in EQ$ the permutation which gives the order of the vector $d_{\leq}(x)$ is the same for all $x \in [a, b]$.

Now we will give a finite dominating set (FDS) for the optimal locations of Problem (10.28), see Nickel and Puerto (1999) for further details.

Theorem 10.6 *An optimal solution for Problem* (10.28) *can always be found in the set* $Cand := EQ \cup V$.

Proof Starting from the original graph G, build a set of new graphs $G_1, \ldots, G_K$ by inserting all points of EQ as new nodes. Now every subgraph G_i is defined by either

I. Two consecutive elements of EQ on an edge or
II. An element $v_i \in V \backslash EQ$ and the adjacent elements of EQ

and the corresponding edges. In this situation for every subgraph G_i the permutation of $d_{\leq}(x)$ is constant (by definition of EQ). Therefore for all $x \in P(G_i)$ we have

$$\sum_{i=1}^{n} \lambda_i d_{(i)}(x) = \sum_{i=1}^{n} \lambda_i w_{\pi(i)} d(v_{\pi(i)}, x)\ ,$$

where $\pi \in P(1, \ldots, n)$, and $P(1, \ldots, n)$ is defined as the set of all permutations of $\{1, \ldots, n\}$. Therefore we can replace the objective by a classical median-objective. Now we can apply Hakimi's node dominance result in every G_i and the result follows. □

Theorem 10.6 also gives rise to some geometrical subdivision of the network N. Like indicated in the proof of Theorem 10.6 we can assign to every subgraph $G_i, i = 1, \ldots, k$ a n-tuple giving in the i-th position the i-th nearest vertex to all points in G_i. As an example we have in Fig. 10.2 a graph with three nodes and all weights w_i and all lengths are 1.

This partition can be seen as a kind of higher order Voronoi diagram of N quite related to the Voronoi partition of networks introduced in Hakimi et al. (1992).

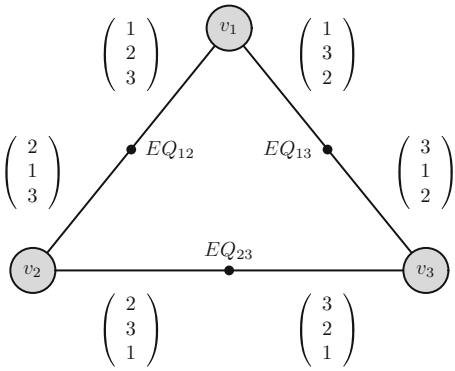


Fig. 10.2 A three-node network with $EQ = \{EQ_{12}, EQ_{13}, EQ_{23}, v_1, v_2, v_3\}$ and the geometrical subdivision

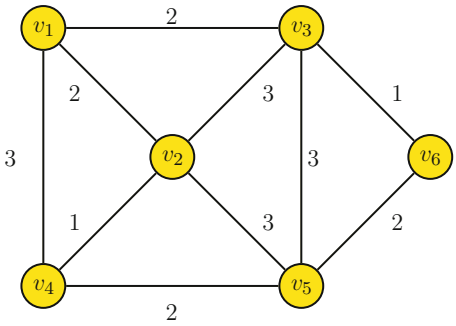


Fig. 10.3 The network used in Example 10.3

For algorithmic purposes one should note that the set EQ can be computed by intersection of all distance functions, see (10.26) on all edges. Since a distance function has maximally one breakpoint on every edge we can use a line sweep technique to determine EQ on one edge in $O((n+k)\log n)$, where $k \leq n^2$ is the number of intersection points. Therefore we can compute EQ for the whole network in $O(m(n+k)\log n)$ time. Of course, this is a worst-case bound and the set of candidates can be further reduced by some domination arguments: Take for two candidates x, y the corresponding weighted (and sorted) distance vectors $d_{\leq}(x)$, $d_{\leq}(y)$. If $d_{\leq}(x)$ is in every component strictly smaller than $d_{\leq}(y)$ then there is no positive λ with which $f_\lambda(d(y)) \leq f_\lambda(d(x))$. This domination argument can be integrated in any line sweep technique reducing, in most cases, the number of candidates.

Example 10.3 Consider the network given in Fig. 10.3 with $w_1 = w_2 = w_5 = 1$ and $w_3 = w_4 = w_6 = 2$. Table 10.3 lists the set EQ, where the labels of the rows EQ_{ij} indicate that i, j are the vertices under consideration and the columns indicate the edge $e = \{r, s\}$. The entry in the table gives for a point $x = (e, t)$ the value of t (if t is not unique an interval of values is shown).

Now we only have to evaluate the objective function with a given set of λ-values for EQ and determine the optima. Table 10.4 gives the solutions for some specific

Table 10.3 List of the set EQ for Example 10.3

	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{2,3\}$	$\{2,4\}$	$\{2,5\}$	$\{3,5\}$	$\{3,6\}$	$\{4,5\}$	$\{5,6\}$
EQ_{12}	$\frac{1}{2}$		$\frac{2}{3}$	$\frac{5}{6}$			$\frac{2}{3}$			$\frac{1}{2}$
EQ_{13}		$\frac{2}{3}$		$\frac{4}{9}$			$\frac{2}{3}$			$\frac{1}{2}$
EQ_{14}	1		$\frac{2}{3}$	0	0	$\frac{8}{9}$	$\frac{8}{9}$			$\frac{1}{6}$
EQ_{15}			$\frac{5}{6}$		$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{2}$		
EQ_{16}		1		1		$\frac{8}{9}$	$\frac{8}{9}$	0	$\frac{5}{6}$	
EQ_{23}		$\frac{1}{3}$		$\frac{2}{3}$			$\frac{2}{3}$			$\frac{1}{2}$
EQ_{24}			$\frac{2}{3}$		$\frac{2}{3}$				$\frac{1}{2}$	
EQ_{25}		$[\frac{3}{4}, 1]$		1		$\frac{1}{2}$	0	0	$\frac{1}{4}$	
EQ_{26}		$\frac{2}{3}$		$\frac{8}{9}$			$\frac{1}{3}$			$\frac{1}{6}$
EQ_{34}	$\frac{1}{4}$		$\frac{1}{6}$	$\frac{1}{3}$			$\frac{5}{6}$			$\frac{1}{4}$
EQ_{35}	$\frac{1}{6}$		$\frac{1}{9}$	$\frac{1}{3}$			$\frac{1}{3}$	1		1
EQ_{36}			$[\frac{5}{6}, 1]$		1	$\frac{1}{3}$	$\frac{5}{6}$	$\frac{1}{2}$	0	
EQ_{45}	$\frac{1}{2}$		$\frac{1}{3}$	$\frac{1}{3}$		$\frac{1}{9}$			$\frac{1}{3}$	
EQ_{46}	0	0	0	$\frac{1}{2}$		$[\frac{2}{3}, 1]$	$[\frac{2}{3}, 1]$		1	0
EQ_{56}		$\frac{1}{2}$		$\frac{2}{3}$			$\frac{1}{9}$			$\frac{2}{3}$

Table 10.4 Solutions for some specific choices for λ in Example 10.3

Obj. function	Corresponding λ	Set of optimal solutions	Obj. value
Center	$\lambda = (0,0,0,0,0,1)$	$EQ^{23}_{46}, EQ^{35}_{46}, EQ^{56}_{34}$	5
2-Centra	$\lambda = (0,0,0,0,\frac{1}{2},\frac{1}{2})$	$[EQ^{23}_{35}, EQ^{23}_{56}]$, $[EQ^{35}_{36}, EQ^{35}_{14}]$, $[EQ^{56}_{14}, EQ^{56}_{13}]$	5
3-Centra	$\lambda = (0,0,0,\frac{1}{3},\frac{1}{3},\frac{1}{3})$	EQ^{23}_{26}	$\frac{40}{9}$
Median	$\lambda = (1,1,1,1,1,1)$	$EQ^{23}_{16} = v_3$	18
Cent-dian	$\lambda = (\frac{\hat{\lambda}}{6}, \frac{\hat{\lambda}}{6}, \frac{\hat{\lambda}}{6}, \frac{\hat{\lambda}}{6}, \frac{\hat{\lambda}}{6}, \frac{6-5\hat{\lambda}}{6})$	EQ^{56}_{34}, $0 \le \hat{\lambda} \le \frac{36}{43}$, v_3 otherwise	$-\frac{17}{12}\hat{\lambda} + 5$, $-5\hat{\lambda} + 8$
Noname	$\lambda = (1,1,0,0,1,1)$	$EQ^{23}_{14}, EQ^{56}_{12}$	13

choices for λ. To describe the solution set we use the notation EQ^{ij}_{kl} to denote the part of EQ_{kl} which lies on the edge $\{i, j\}$.

Kalcsics et al. (2002) gives an FDS for the single facility ordered median problem with general node weights (the w-weights can be negative). Moreover, for the case of a directed network with non-negative w-weights, they prove that there is always an optimal solution in V.

10.4.2 The p-Facility Ordered Median Problem

In this section we deal with the multi-facility extension of the ordered median problem. The p-facility ordered median problem consists of finding a set

$X_p = \{x_1, \ldots, x_p\}$ that minimizes the following objective function

$$\text{minimize}_{X_p} \sum_{i=1}^{n} \lambda_i d_{(i)}(X_p), \tag{10.30}$$

where $d(v, X_p) := \min_{i=1,\ldots,p} d(v, x_i)$ with $v \in V$; $d(X_p) := (w_1 d(v_1, X_p), \ldots, w_n d(v_n, X_p))$ and $d_{\leq}(X_p) := (w_{(1)} d(v_{(1)}, X_p), \ldots, w_{(n)} d(v_{(n)}, X_p))$ a permutation of the elements of $d(X_p)$, verifying:

$$w_{(1)} d(v_{(1)}, X_p) \leq \ldots \leq w_{(n)} d(v_{(n)}, X_p).$$

The main result of this section establishes a generalization of the well-known theorem of Hakimi which states that always exists an optimal solution in V.

Theorem 10.7 *If $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ then Problem (10.30) has always an optimal solution X_p^* contained in V.*

Proof Since by hypothesis $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ we have that

$$f_\lambda(d(X_p)) = \sum_{i=1}^{n} \lambda_i d_{(i)}(X_p) = \text{minimize}\{\sum_{i=1}^{n} \lambda_i d_{\pi(i)}(X_p) : \pi \in \Pi(\{1, \ldots, n\})\}.$$

Assume that $X_p \not\subset V$.

Then there must exist $x_i \in X_p$ with $x_i \notin V$. Let $e = \{v, w\}$ be the edge containing x_i and $\ell(e)$ its length. Denote by $X_p(s) = X_p \setminus \{x_i\} \cup \{x(s)\}$ where $x(s)$ is the point on e with $d(v, x(s)) = s$, $s \in [0, l(e)]$.

The function g defined as $g(s) = \sum_{i=1}^{n} \lambda_i d_{(i)}(X_p(s))$ is concave for all $s \in [0, \ell(e)]$ because it is the composition of a concave and a linear function, i.e.

$$g(s) = \min_{\pi \in \Pi(\{1,\ldots,n\})} \left\{ \sum_{i=1}^{n} \lambda_i d_{\pi(i)}(X_p(s)) \right\}$$

and each

$$\begin{aligned} d_{\pi(j)}(X_p(s)) = \min\{&d(v_{\pi(j)}, x_1), \ldots, \min\{d(v_{\pi(j)}, a) + s, d(v_{\pi(j)}, b) \\ &+ \ell(e) - s\}, \ldots, d(v_{\pi(j)}, x_n)\} \end{aligned}$$

is concave.

Hence, $g(s) = F(X_p(s)) \geq \min\{F(X_p(0)), F(X_p(\ell(e))\}$ and the new solution set $X_p(s)$ contains instead of x_i one vertex of V.

Repeating this scheme a finite number of times the result follows. □

In the previous section we proved that the set $V \cup EQ$ always contains the set of optimal solutions of the problem (independent of the structure of λ). It might seem natural to expect that the same result holds for the p-facility case as it happens for the p-center problem. However, Example 10.4 shows that this property fails to be true.

This easy example shows the limit for the set $Cand = V \cup EQ$ to be a FDS (finite dominating set) for the multifacility extension of our model. In the literature we can find some characterizations of FDS for particular cases of the p-facility ordered median problem. For instance, Kalcsics et al. (2003) studies the multifacility ordered median problem where the λ-weights are defined as:

$$a = \lambda_1 = \ldots = \lambda_k \neq \lambda_{k+1} = \ldots = \lambda_n = b,$$

for a fixed k, such that, $1 \leq k < n$. They prove that the set Y, defined by (10.31), is a FDS for this problem.

However, none of these papers deals with the general case of the multifacility ordered median problem. In fact, these papers impose very restrictive hypotheses such that their respective results can not be extended further. In the following section we characterize a FDS for the general 2-facility ordered median problem.

10.4.2.1 A Finite Set of Candidates for the Two Facility Case

In this section we identify a finite set of candidates to be optimal solutions of the 2-facility ordered median problem. In order to consider the set of equilibrium points as a finite set we will assume that EQ only contains the equilibrium points that are isolated and the extreme points of the subedges in equilibrium, see Rodríguez-Chía et al. (2005) for further details.

Theorem 10.8 *Consider the following sets:*

$$\begin{aligned} R &= \{r \ : \ r = w_i d(v_i, y),\ v_i \in V,\ y \in V \cup EQ\}, \\ Y(r) &= \{y \in P(G) \ : \ w_i d(v_i, y) = r,\ v_i \in V\} \quad \textit{with } r \in R, \\ Y &= \bigcup_{r \in R} Y(r), \end{aligned} \tag{10.31}$$

$T = \{X_2 = (x_1, x_2) \in P(G) \times P(G) \ : \ \exists v_r, v_s$ *served by* x_1 *and* $v_{r'}, v_{s'}$ *served by* x_2, *such that* $w_r d(v_r, x_1) = w_{r'} d(v_{r'}, x_2)$ *and* $w_s d(v_s, x_1) = w_{s'} d(v_{s'}, x_2)$. *Moreover, if* $w_r = w_{r'}$ *and* $w_s = w_{s'}$, *then the slopes of the functions* $d(v_r, \cdot)$ *and* $d(v_s, \cdot)$, *in the edge that* x_1 *belongs to, must have the same signs at* x_1 *and the slopes of the functions* $d(v_{r'}, \cdot)$ *and* $d(v_{s'}, \cdot)$, *in the edge that* x_2 *belongs to, must have different signs at* x_2 $\}$.

$$F = ((EQ \cup V) \times Y) \cup T \subset P(G) \times P(G). \tag{10.32}$$

The set F *is a finite set of candidates to be optimal solutions of the* 2*-facility ordered median problem in the network* N.

Remark 10.1 The structure of the set F is different from previous FDS which appeared in the literature. Indeed, the set F is itself a set of candidates for optimal solutions because it is a set of pairs of points. That means that we do not have to choose the elements of this set by pairs to enumerate the whole set of candidates. The candidate solutions may be either a pair of points belonging to $(EQ \cup V) \times Y$ or a pair belonging to T, but they never can be one point of Y and another point of any pair in T.

The following examples show that the set F can not be shrunk because even in easy cases on the real line all the points are needed. The first example shows a graph where the optimal solution $X_2 = (x_1, x_2)$ verifies that x_1 is an equilibrium point and x_2 is not an equilibrium point which belongs to $Y(r) \setminus (EQ \cup V)$ for a given r. In the second example the optimal solution $X_2 = (x_1, x_2)$ belongs to the set T.

Example 10.4 Let $N = (G, \ell)$ be a network with underlying graph $G = (V, E)$ where $V = \{v_1, v_2, v_3, v_4\}$ and $E = \{\{1, 2\}, \{2, 3\}, \{3, 4\}\}$. The length function is given by $\ell(\{1, 2\}) = 3$, $\ell(\{2, 3\}) = 20$, $\ell(\{3, 4\}) = 6$. The w-weights are all equal to one and the λ-weights are $\lambda_1 = 0.1, \lambda_2 = 0.2, \lambda_3 = 0.4, \lambda_4 = 0.3$, see Fig. 10.4.

It should be noted that this example can not have optimal solutions on the edge $\{2, 3\}$ because any point of this edge is dominated by v_2 or v_3. In addition, using the symmetry of the problem we have omitted the evaluation of some of the elements of Y.

In Fig. 10.4 we represent the nodes (dots), the equilibrium points (ticks) and elements of Y (small ticks). Notice that in this case there are no pairs in T.

In this example the optimal solution is given by $x_1 = p(\{1, 2\}, 1.5)$ and $x_2 = p(\{3, 4\}, 1.5)$ (see Table 10.5). It is easy to check that x_1 is an equilibrium point between v_1 and v_2, and $x_2 \in Y(1.5)$. It is worth noting that the radius 1.5 is given by the distance from the equilibrium point, $p(\{1, 2\}, 1.5)$, generated by v_1 and v_2 to any of these nodes.

Example 10.5 Let $N = (G, \ell)$ be a network with underlying graph $G = (V, E)$ where $V = \{v_1, v_2, v_3, v_4, v_5\}$ and $E = \{\{1, 2\}, \{2, 3\}, \{3, 4\}, \{4, 5\}\}$. The length function is given by $\ell(\{1, 2\}) = 5, \ell(\{2, 3\}) = 20, \ell(\{3, 4\}) = 5.1, \ell(\{4, 5\}) = 1$.

Fig. 10.4 Illustration of Example 10.4

Table 10.5 Evaluation of the candidate pairs of Example 10.4

Candidate pair X_2	Value	Candidate pair X_2	Value
$p(\{1, 2\}, 0), p(\{3, 4\}, 0)$	3	$p(\{1, 2\}, 1.5), p(\{3, 4\}, 0)$	2.7
$p(\{1, 2\}, 0), p(\{3, 4\}, 1.5)$	2.85	$p(\{1, 2\}, 1.5), p(\{3, 4\}, 1.5)$	2.4
$p(\{1, 2\}, 0), p(\{3, 4\}, 3)$	2.7	$p(\{1, 2\}, 1.5), p(\{3, 4\}, 3)$	2.55

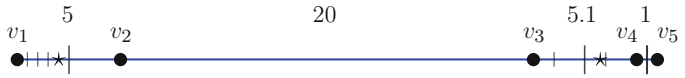


Fig. 10.5 Illustration of Example 10.5

Table 10.6 Evaluation of the candidate pairs of Example 10.5

Candidate pair X_2	Value	Candidate pair X_2	Value
$p(\{1,2\},0), p(\{3,4\},0)$	11.81	$p(\{1,2\},2.05), p(\{3,4\},3.05)$	8.455
$p(\{1,2\},0), p(\{3,4\},2.55)$	11.6	$p(\{1,2\},2.45), p(\{3,4\},2.55)$	9.005
$p(\{1,2\},0), p(\{3,4\},3.05)$	10.6	$p(\{1,2\},2.5), p(\{3,4\},0)$	14.31
$p(\{1,2\},0), p(\{4,5\},0)$	10.61	$p(\{1,2\},2.5), p(\{3,4\},2.5)$	9.06
$p(\{1,2\},0), p(\{4,5\},0.5)$	11.66	$p(\{1,2\},2.5), p(\{3,4\},2.55)$	8.955
$p(\{1,2\},0), p(\{4,5\},1)$	11.71	$p(\{1,2\},2.5), p(\{3,4\},2.6)$	8.95
$p(\{1,2\},0.5), p(\{4,5\},0.5)$	11.16	$p(\{1,2\},2.5), p(\{3,4\},3.05)$	8.905
$p(\{1,2\},1), p(\{4,5\},0)$	10.61	$p(\{1,2\},2.5), p(\{3,4\},3.6)$	8.96
$p(\{1,2\},1), p(\{4,5\},1)$	11.71	$p(\{1,2\},2.5), p(\{4,5\},0)$	9.11
$p(\{1,2\},1.45), p(\{3,4\},2.55)$	10.005	$p(\{1,2\},2.5), p(\{4,5\},0.5)$	9.16
$p(\{1,2\},1.95), p(\{3,4\},3.05)$	8.455	$p(\{1,2\},2.5), p(\{4,5\},1)$	10.21
$p(\{1,2\},2), p(\{3,4\},3.1)$	8.41		

The w-weights are all equal to one and the λ-weights are $\lambda_1 = 0, \lambda_2 = 1, \lambda_3 = 0, \lambda_4 = 1, \lambda_5 = 1.1$, see Fig. 10.5.

In Fig. 10.5, we use the same notation as in Fig. 10.4 and pairs of T are represented by $(\star)$. By domination and symmetry arguments not all the candidates are necessary and therefore, they are not depicted.

In this example the optimal solution is given by $x_1 = p(\{1,2\},2)$ and $x_2 = p(\{3,4\},3.1)$ (see Table 10.6). Therefore the optimal pair (x_1, x_2) belongs to the set T. Indeed, $d(v_1, x_1) = d(v_4, x_2)$ and $d(v_2, x_1) = d(v_5, x_2)$ and the slopes of $d(v_1, \cdot), d(v_2, \cdot)$ in the edge $\{1,2\}$ at x_1 are $1, -1$ respectively; and the slopes of $d(v_4, \cdot), d(v_5, \cdot)$ in the edge $\{3,4\}$ at x_2 are $-1, -1$ respectively.

Once we have proved that F is an essential set to describe the set of optimal solutions of the 2-facility ordered median problem we want to know its cardinality.

Proposition 10.2 *The cardinality of F is $O(m^3n^6)$.*

Proof In each edge there are at most two equilibrium points associated with each pair of nodes. Thus $|EQ| = O(mn^2)$ and $|R| = O(mn^3)$. The maximum degree of a node $v_i \in V$ is m (the star network) so $|Y(r)| = O(mn)$ with $r \in R$. Thus, $|Y| = O(m^2n^4)$. On the second hand, on each edge, each pair of nodes may determine an element of a pair in T. Therefore, the set T has a cardinality $O((n^2m)^2)$. In conclusion $|F| = O(m^3n^6 + m^2n^4) = O(m^3n^6)$. □

It is worth noting that F is an actual set of finite elements to be optimal solutions of problem (10.30). The difference with previous approaches is that this set is not a

set of candidates for each individual facility but it is the set of candidate pairs to be optimal solutions.

10.4.2.2 A Discouraging Result for the p-Facility Case

It is well-known that FDS of polynomial size exist for the classical p-median, p-center, p-centdian and p-k-centrum problems (see Hooker et al. 1991; Kalcsics et al. 2003). In addition, our previous section has shown a finite set of candidates to be optimal solutions of the 2-facility ordered median problem in a network. However, despite the similarity existing between those problems and the general p-facility ordered median problem, these results can not be extended to our model.

The reason for this is the following. For the 1-facility ordered median problem we have that the set of candidates to be optimal solutions is EQ, that means, the equilibrium points (see Nickel and Puerto 1999). For the 2-facility ordered median problem we have obtained that the set of candidates to be optimal solutions is $EQ \times Y \cup T$, that means, the points generated by the distances between each node and each equilibrium point and the set T. It should be noted that in this case we have added these points because there may exist ties which do not allow to move the service facility improving the objective function. In the 3-facility ordered median problem, the previous candidate set is not enough because if $x_1 \in EQ$ and $x_2 \in Y \setminus EQ$, the distances between each node and x_2 don't have to be included in the set of radius, R. Therefore, it may occur that there exists a tie between two nodes and the service facilities x_2 and x_3 respectively, so that there is no movement of the facilities at x_2 and x_3 which improves the objective function (see Example 10.6).

Example 10.6 Let $N = (G, \ell)$ be a network with underlying graph $G = (V, E)$ where $V = \{v_1, v_2, v_3, v_4, v_5, v_6\}$ and $E = \{\{1, 2\}, \{2, 3\}, \{3, 4\}, \{4, 5\}, \{5, 6\}\}$. The length function is given by $\ell(\{1, 2\}) = 3, \ell(\{2, 3\}) = 50, \ell(\{3, 4\}) = 6, \ell(\{4, 5\}) = 50, \ell(\{5, 6\}) = 10$. The w-weights are all equal to one and the λ-modeling weights are $\lambda_1 = 0.1, \lambda_2 = 0.2, \lambda_3 = 0.4, \lambda_4 = 0.3, \lambda_5 = 0.6, \lambda_6 = 0.55$, see Fig. 10.6 (in this figure we use the same notation used in Fig. 10.4).

In this example the optimal solution is given by $x_1 = p(\{1, 2\}, 1.5)$, $x_2 = p(\{3, 4\}, 1.5)$ and $x_3 = p(\{4, 5\}, 4.5)$ (see Table 10.7). It can be seen that x_1 is an equilibrium point, $x_2 \in Y(1.5)$ and x_3 neither belongs to Y nor is a component of a pair of T.

This example illustrates that in order to obtain the optimal solution for the 3-facility problem new points have to be added. Our conjecture is that these points

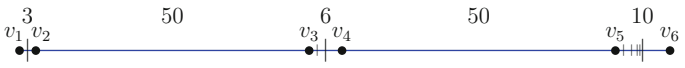


Fig. 10.6 Illustration of Example 10.6

Table 10.7 Evaluation of the candidate solutions of Example 10.6

Candidate pair X_3	Val.	Candidate pair X_3	Val.
$p(\{1,2\},0)$, $p(\{3,4\},0)$, $p(\{4,5\},0)$	10	$p(\{1,2\},1.5)$, $p(\{3,4\},0)$, $p(\{4,5\},0)$	10.1
$p(\{1,2\},0)$, $p(\{3,4\},0)$, $p(\{4,5\},1.5)$	9.77	$p(\{1,2\},1.5)$, $p(\{3,4\},0)$, $p(\{4,5\},1.5)$	9.62
$p(\{1,2\},0)$, $p(\{3,4\},0)$, $p(\{4,5\},3)$	9.55	$p(\{1,2\},1.5)$, $p(\{3,4\},0)$, $p(\{4,5\},3)$	9.25
$p(\{1,2\},0)$, $p(\{3,4\},0)$, $p(\{4,5\},4)$	9.3	$p(\{1,2\},1.5)$, $p(\{3,4\},0)$, $p(\{4,5\},4)$	9
$p(\{1,2\},0)$, $p(\{3,4\},0)$, $p(\{4,5\},4.5)$	9.15	$p(\{1,2\},1.5)$, $p(\{3,4\},0)$, $p(\{4,5\},4.5)$	8.85
$p(\{1,2\},0)$, $p(\{3,4\},0)$, $p(\{4,5\},5)$	9	$p(\{1,2\},1.5)$, $p(\{3,4\},0)$, $p(\{4,5\},5)$	8.75
$p(\{1,2\},0)$, $p(\{3,4\},1.5)$, $p(\{4,5\},0)$	9.7	$p(\{1,2\},1.5)$, $p(\{3,4\},1.5)$, $p(\{4,5\},0)$	9.55
$p(\{1,2\},0)$, $p(\{3,4\},1.5)$, $p(\{4,5\},1.5)$	9.17	$p(\{1,2\},1.5)$, $p(\{3,4\},1.5)$, $p(\{4,5\},1.5)$	8.87
$p(\{1,2\},0)$, $p(\{3,4\},1.5)$, $p(\{4,5\},3)$	8.95	$p(\{1,2\},1.5)$, $p(\{3,4\},1.5)$, $p(\{4,5\},3)$	8.5
$p(\{1,2\},0)$, $p(\{3,4\},1.5)$, $p(\{4,5\},4)$	8.7	$p(\{1,2\},1.5)$, $p(\{3,4\},1.5)$, $p(\{4,5\},4)$	8.25
$p(\{1,2\},0)$, $p(\{3,4\},1.5)$, $p(\{4,5\},4.5)$	8.57	$p(\{1,2\},1.5)$, $p(\{3,4\},1.5)$, $p(\{4,5\},4.5)$	8.12
$p(\{1,2\},0)$, $p(\{3,4\},1.5)$, $p(\{4,5\},5)$	8.6	$p(\{1,2\},1.5)$, $p(\{3,4\},1.5)$, $p(\{4,5\},5)$	8.15
$p(\{1,2\},0)$, $p(\{3,4\},3)$, $p(\{4,5\},0)$	11.2	$p(\{1,2\},1.5)$, $p(\{3,4\},3)$, $p(\{4,5\},0)$	9.1
$p(\{1,2\},0)$, $p(\{3,4\},3)$, $p(\{4,5\},1.5)$	8.87	$p(\{1,2\},1.5)$, $p(\{3,4\},3)$, $p(\{4,5\},1.5)$	8.42
$p(\{1,2\},0)$, $p(\{3,4\},3)$, $p(\{4,5\},3)$	8.35	$p(\{1,2\},1.5)$, $p(\{3,4\},3)$, $p(\{4,5\},3)$	8.2
$p(\{1,2\},0)$, $p(\{3,4\},3)$, $p(\{4,5\},4)$	8.4	$p(\{1,2\},1.5)$, $p(\{3,4\},3)$, $p(\{4,5\},4)$	8.25
$p(\{1,2\},0)$, $p(\{3,4\},3)$, $p(\{4,5\},4.5)$	8.42	$p(\{1,2\},1.5)$, $p(\{3,4\},3)$, $p(\{4,5\},4.5)$	8.27
$p(\{1,2\},0)$, $p(\{3,4\},3)$, $p(\{4,5\},5)$	8.45	$p(\{1,2\},1.5)$, $p(\{3,4\},3)$, $p(\{4,5\},5)$	8.3

can be generated using recursively the construction of the set of radii but now regarding the distances from the points in $\pi_2(F) := \{x_2 : (x_1, x_2) \in F\}$, that is, the points in $P(G)$ which correspond to the second candidate of any pair in F, and the node set:

$$R_1 = \{r : r = w_i d(v_i, y), v_i \in V, y \in \pi_2(F)\},$$
$$Y_1(r) = \{y : y \in P(G), w_i d(v_i, y) = r, v_i \in V\},$$
$$Y_1 = \bigcup_{r \in R_1} Y_1(r).$$

The same situation occurs in the p-facility case, so that in general this construction must be repeated p-times in order to obtain a finite candidate set to be optimal solutions for that problem. Therefore the structure of the candidate set defined in the previous section depends on the number of facilities to be located. Actually, Puerto and Rodríguez-Chía (2005) prove that there is no polynomial size FDS for the general ordered p-median problem even on path networks. The proof consists of building a family of $O(n^n)$ problems on the same graph with different solutions (each solution contains at least one point not included in the remaining), n being the number of nodes.

10.5 The Capacitated Discrete Ordered Median Problem

In this section our goal is to introduce the family of discrete ordered median location problems. As we have seen in previous sections, the main feature of these models is their flexibility to generalize the most popular objective functions studied in the location analysis literature and to allow modeling a wide variety of new problems appearing in logistics and manufacturing.

The uncapacitated version of the discrete ordered median location models has been analyzed in several papers, Boland et al. (2006), Nickel (2001), Nickel and Puerto (2005), Marín et al. (2009, 2010), Puerto et al. (2011, 2013), and different formulations and algorithms to solve medium sized problems have been developed. Recently, these models were extended to deal with capacities in Kalcsics et al. (2010a,b). However, although the approach in the initial papers leads to satisfactory results concerning motivations, applications and interpretations the solution times of larger problem instances need further improvements.

The goal of this section is to present, first, an intuitive formulation of the problem based on three-indexed variables, see Boland et al. (2006); and second, a formulation which makes use of the coverage ideas in Marín et al. (2009, 2010), applied to the capacitated version of the Discrete Ordered Median Problem, CDOMP, with binary assignment, see Puerto (2008), Puerto et al. (2011, 2013). To perform this task, first we introduce the Capacitated Discrete Ordered Median Problem formally and give these two mathematical programming formulations.

Then, the last part of this section is devoted to test the efficiency of the last approach by providing some preliminary numerical experiments.

10.5.1 A Three-Index Formulation

In order to introduce this formulation let A denote the given set of n sites and identify these with the integers $1, \ldots, n$, i.e., $A = \{1, \ldots, n\}$. We assume without loss of generality that the set of candidate sites for new facilities is identical to the set of clients. Let $C = (c_{ij})_{i,j=1,\ldots,n}$ be the given non-negative $n \times n$ cost matrix, where c_{ij} denotes the cost of satisfying the demand of client i from a facility located at site j. Let $p \leq n$ be the number of facilities to be located. Each client i has a demand a_i that must be served and each server j has an upper bound b_j on the capacity that it can fulfill. We assume further that assignment is binary, that is, the demand of each client must be served by a unique server.

A solution to the location problem is given by a set of p sites; we use $X \subseteq A$, with $|X| = p$, to denote a solution. Then, the problem consists of finding the set of sites X with $|X| = p$, which can supply the overall demand at a minimum cost with respect to the ordered median objective function.

A natural way to attack the formulation of the discrete ordered median problem is to use variables that keep track of the order of the transportation costs from each client and its server. This approach gives rise to a formulation with three-index variables, one for the order and the remaining two indices, for the client-server allocation. In order to formulate this model we consider a set of λ-weights, where λ_i can be seen as a correction factor to the *ith*-position with $i = 1, \ldots, n$. In addition, we define the following set of variables:

$$x_{ij}^k = \begin{cases} 1, & \text{if client } i \text{ is supplied by server } j \text{ and is the } k\text{-th} \\ & \text{cheapest cost allocation} \\ 0, & \text{otherwise,} \end{cases} \qquad \forall i, j, k = 1, \ldots, n,$$

$$y_j = \begin{cases} 1, & \text{if the server at } j \text{ is open} \\ 0, & \text{otherwise,} \end{cases} \qquad \forall j = 1, \ldots, n.$$

Hence, the formulation of the model is:

$$\text{minimize} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \lambda_k c_{ij} x_{ij}^k \tag{10.33}$$

$$\text{subject to} \sum_{j=1}^{n} \sum_{k=1}^{n} x_{ij}^k = 1, \quad \forall i = 1, \ldots, n \tag{10.34}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}^{k} = 1, \quad \forall k = 1,\dots,n \tag{10.35}$$

$$\sum_{i=1}^{n}\sum_{k=1}^{n} a_i x_{ij}^{k} \le b_j y_j, \quad \forall j = 1,\dots,n, \tag{10.36}$$

$$\sum_{j=1}^{n} y_j = p, \tag{10.37}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} x_{ij}^{k} \le \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} x_{ij}^{k+1}, \quad \forall k = 1,\dots,n-1, \tag{10.38}$$

$$x_{ij}^{k} \in \{0,1\}, \quad \forall i,j,k = 1,\dots,n, \tag{10.39}$$

$$y_j \in \{0,1\}, \quad \forall j = 1,\dots,n. \tag{10.40}$$

The objective function accounts for the weighted sum of the transportation cost using the lambda parameters. Constraints (10.34) ensure that each origin site i is allocated exactly to one server j. Constraints (10.35) guarantee that any position in the sorted vector of *client-server* costs is allocated to just one pair. Constraints (10.36) are the capacities constraint and also ensure that one origin may be allocated to a specific server only if it is open. Constraint (10.37) fixes the number of facilities to be located. Finally, constraints (10.38) ensure that the transportation cost assigned to the k-position is smaller than the one assigned to the $(k+1)$-position.

10.5.2 A Covering Formulation and Some Properties

In this subsection, we introduce a formulation for the binary assignment capacitated discrete ordered median problem based on covering variables. This formulation was first presented in Puerto (2008).

We first define G as the number of different non-zero elements of the cost matrix C. Hence, we can order the different values of C in non-decreasing sequence: $c_{(0)} := 0 < c_{(1)} < c_{(2)} < \cdots < c_{(G)} := \max_{1\le i,j\le n}\{c_{ij}\}$.

Given a feasible solution, we can use this ordering to perform the sorting process of the allocation costs. This can be done by the following variables ($j = 1,\dots,n$ and $k = 1,\dots,G$):

$$u_{jk} := \begin{cases} 1, \text{ if the } j\text{-th smallest allocation cost is at least } c_{(k)}, \\ 0, \text{ otherwise.} \end{cases} \tag{10.41}$$

With respect to this definition the j-th smallest cost element is equal to $c_{(k)}$ if and only if $u_{jk} = 1$ and $u_{j,k+1} = 0$. Therefore, we can reformulate the objective

function of the CDOMP (i.e. the capacitated ordered median problem), using the variables u_{jk}, as $\sum_{j=1}^{n}\sum_{k=1}^{G}\lambda_j\cdot(c_{(k)}-c_{(k-1)})\cdot u_{jk}$.

First of all, we need to impose the following group of sorting constraints on the u_{jk}-variables: $u_{j+1,k}\geq u_{jk}\qquad j=1,\ldots,n-1;\ k=1,\ldots,G$. To guarantee that exactly p servers will be opened among the n possibilities, we consider constraint (10.37) defined in the previous formulation.

Then, we need to ensure that demand is covered and capacity is satisfied. For these reasons we introduce the variables x_{ij}:

$$x_{ij}=\begin{cases}1, & \text{if the client } i \text{ is allocated to server } j\\ 0, & \text{otherwise}\end{cases} \tag{10.42}$$

(binary allocation) and constraints $\sum_{j=1}^{n}x_{ij}=1,\quad i=1,\ldots,n$ (each client is just assigned to one server) and $\sum_{i=1}^{n}a_i x_{ij}\leq b_j y_j,\quad j=1,\ldots,n$ (all the demand and capacity requirements must be satisfied and clients can only be assigned to servers which are open).

In addition, the relationship that links the variables u and x is: $\sum_{j=1}^{n}u_{jk}=\sum_{i=1}^{n}\sum_{j:c_{ij}\geq c_{(k)}}x_{ij}$. The meaning being clear. The number of allocations with a cost at least $c_{(k)}$ must be equal to the number of servers that support demand from facilities at a cost greater than or equal to $c_{(k)}$.

Summing up all these constraints and the objective function, the CDOMP can be formulated as

$$\text{minimize}\ \sum_{j=1}^{n}\sum_{k=1}^{G}\lambda_j(c_{(k)}-c_{(k-1)})u_{jk} \tag{10.43}$$

$$\text{subject to}\ \sum_{j=1}^{n}x_{ij}=1,\quad \forall i=1,\ldots,n \tag{10.44}$$

$$\sum_{i=1}^{n}a_i x_{ij}\leq b_j y_j,\quad \forall j=1,\ldots,n \tag{10.45}$$

$$x_{ij}\leq y_j\quad \forall i,j=1,\ldots,n \tag{10.46}$$

$$\sum_{j=1}^{n}y_j=p \tag{10.47}$$

$$\sum_{j=1}^{n}u_{jk}=\sum_{i=1}^{n}\sum_{\substack{j=1\ldots,n\\ c_{ij}\geq c_{(k)}}}x_{ij},\quad \forall k=1,\ldots,G \tag{10.48}$$

$$u_{j+1k}\geq u_{jk},\quad \forall j=1,\ldots,n-1;\ k=1,\ldots,G \tag{10.49}$$

$$u_{jk}\in\{0,1\},\quad \forall j=1,\ldots,n;\ k=1,\ldots,G \tag{10.50}$$

$$x_{ij},y_j\in\{0,1\},\quad \forall i,j=1,\ldots,n. \tag{10.51}$$

Since the proposed formulation contains $O(nG)$ binary variables and $O(nG)$ constraints, fast solution times for larger problem instances, using standard software-tools, are very unlikely.

First of all, the following proposition states that we can relax the y_j variables to be continuous and the solution will not change.

Proposition 10.3 *(CDOMP) admits a formulation with* $y_j \in [0,1]$ *and for each optimal solution of the relaxed problem one can obtain an optimal solution of the original problem.*

Proof Use (10.46) and (10.47) to ensure that any fractional y solution can be modified to be binary and feasible without increasing the objective value. □

The above formulation admits some valid inequalities that, at times, reinforce the linear relaxation improving the lower bound and reducing the computation time to solve the problem. In the following, we list three families of them.

The first one are the natural inequalities $u_{jk} \geq u_{jk+1}, \quad j = 1,\ldots,n, \quad k = 1,\ldots,G-1$. They come from the fact that the rows of the u-matrix are sorted. We have observed in our experiments that these constraints are not always satisfied by the optimal solution of the linear relaxation and thus they are useful in improving the formulation. This family of inequalities were introduced in Marín et al. (2009) for tightening the formulation of the Uncapacitated Discrete Ordered Median Problem.

Our next set of inequalities state that the number of assignments done by the x-variables at a cost at least $c_{(j)}$ for clients in S cannot exceed the number of ones in the last $|S| = r$ rows of the j-th column of the u-matrix. Then, if there are r allocations of demand points in S at a costs at least $c_{(j)}$, since the columns in the u-matrix are ordered in non-decreasing sequence, we get the following: $\sum_{i\in S}\sum_{k:c_{ik}\geq c_{(j)}} x_{ik} \leq \sum_{i=n-r+1}^{n} u_{ij}, \quad \forall\ S \subseteq \{1,\ldots,n\}, \quad |S| = r, \quad r = 1,\ldots,n, j = 1,\ldots,G$. Note that there are an exponential number of inequalities in this family.

Another set of valid inequalities are those stating that either client i is allocated at a cost at least $c_{(k)}$ or there must exist an open server j such that the allocation cost of client i is smaller than $c_{(k)}$. This results in: $\sum_{j:c_{ij}\geq c_{(k)}} x_{ij} + \sum_{j:c_{ij}<c_{(k)}} y_j \geq 1,\ i = 1,\ldots,n$.

Finally, the set of valid inequalities $x_{ij} \leq y_j \quad \forall i = 1,\ldots,n,\ j = 1,\ldots,n$, that reinforce the idea that the clients can only be assigned to servers which are open, also provide very good results from the computational point of view.

The rest of this section presents some computational results for this formulation of the capacitated discrete ordered problem. We restrict ourselves to consider just the second formulation, because although the first one is very intuitive and good to have a better understanding of the problem, its running times are much bigger than those obtained by the second one. (See e.g. Puerto (2008).) In order to test the performance of the considered formulation, we report on an experimental design that consists of the following factors: (1) *Size of the problem*: The number of sites, n, determines the dimensions of the cost matrix and the λ vectors. Moreover, it is an upper bound of the number of suppliers (p) to be located. We consider five different levels of

$n = 10, 20, 30, 40, 60$. (2) *Number of suppliers*: p is the second factor with three levels for each choice of n: $p = \lfloor n/5 \rfloor + 1$, $\lfloor n/2 \rfloor$, $4 \times \lfloor n/5 \rfloor$. (3) *Type of problem*: Each λ-vector is associated with a different objective function. Its levels are designed depending on the value of n as follows: (a) λ-vector corresponding to the p-median problem, i.e. $\lambda = (1, \ldots, 1) \in \mathbb{R}^n$; (b) λ-vector corresponding to the p-center problem, i.e. $\lambda = (0, \ldots, 0, 1) \in \mathbb{R}^n$; (c) λ-vector corresponding with the $\lfloor n/4 \rfloor$-centrum problems; and (d) λ-vector corresponding to the (k_1, k_2)-trimmed mean problem, i.e. $\lambda = (0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0) \in \mathbb{R}^n$ where $k_1 = \lfloor 0.2n \rfloor$, $k_2 = \lfloor 0.2n \rfloor$. (4) *Demand of facilities*: Each demand is considered integer and uniformly drawn from $[10, 20]$. (5) *Capacity of suppliers*: We consider that the capacities are uniformly discrete random variables in the interval $[1.1 \sum_{i=1}^{n} a_i / p, 1.4 \sum_{i=1}^{n} a_i / p]$. This choice ensures feasibility of the considered problems. (6) *Transportation cost*: We assume free self service and integer costs. The values c_{ij}, $i \neq j$, are drawn uniformly in $[0, 200]$.

We solve five instances for each possible combination of levels and we report the average and maximum: running time, gap at the root node and number of nodes in the branch-and-bound tree for this formulation. All computational studies were performed on a PC with a *Genuine Intel(R) CPU U4100* with two processors at 1.30 GHz and 4 GB of RAM. To solve the different instances of the problems we used XPRESS-IVE solver version 7.5, with a code implemented in XPRESS-MOSEL version 3.4.2.

The information of our computational test is reported in Table 10.8 that summarizes the results for the four considered problems types. The organization of the table is the following: columns show the results for the different sizes of n and p. A superindex in some values of p states the number of instances for the corresponding combination of n and p exceeding the CPU time limit (1 h). Each block of rows reports the results of the instances based on the formulation (10.43)–(10.51).

Within each block of rows we report on the *GAP* at the root node (average -Ag- and maximum -Mg-), *CPU*-time to solve the integer problems (average -At- and maximum -Mt-) and number of *NODES* in the branch-and-bound tree (average -An- and maximum -Mn-).

We observe, from the results in Table 10.8 that we could solve most of the instances, even medium sized $n = 60$, within 1 h of CPU time. This fact shows a good performance of the formulation. In addition, it is worth noting that the quality of the lower bounds provided by this formulation depends on the type of problem. In general, the lower bounds are rather poor for larger values of p relative to n. On the other hand, for small to medium values of p relative to n the performance of the lower bounds are good for median and trimmed mean problems, reasonable for k-centrum (less than 50 %) and poor for the center problem. These results show that there is room for further investigation on the polyhedral structure of this formulation in order to develop valid inequalities that could be integrated in a Branch & Cut algorithm to solve faster, larger problem sizes.

In conclusion, the formulation of the CDOMP based on covering, (10.43)–(10.51), is a promising approach. Moreover, it can be also strengthen with known

Table 10.8 Numerical results obtained with the covering formulation for the median, center, k-centrum and Trimmed-mean problems

Median															
n	10			20			30			40			60		
p	3	5	8	5	10	16	7	15	24	9	20	32	13	30^2	48^2
At	0.6	4.3	2.5	5.6	11.1	31.4	41.4	44.8	23	38.7	116.2	718.4	213.9	1,939.3	2,092.6
Mt	1.7	9.9	4.5	12.3	20.2	66.6	102	135.9	59.6	63.5	198.3	2,644.6	427.6	3,600.8	3,600.4
An	47	10.2	51.6	12.2	191	1,557.8	344	1,055.8	607.4	124.4	2,357	31,512.6	716.4	23,586.4	42,081.6
Mn	231	31	127	39	451	4,077	767	4,865	1647	485	4,221	129,917	1,848	46,523	91,324
Ag	2.3	21.2	77.3	6	15.3	83.6	6.7	23.3	71.8	5.2	25.3	76.4	7.7	27	76.9
Mg	4.1	41.2	89.8	8.7	25.3	92.1	10.1	43.3	77.7	7.6	35.8	91.9	10.8	35.3	88
Center															
n	10			20			30			40			60		
p	3	5	8	5	10	16	7	15^1	24^2	9	20	32^2	13^1	30^1	48^3
At	13.7	9.4	2.4	47.3	19.8	236.3	90.1	874.6	1,786.9	338.9	260.2	2,162.7	1,977.3	1,416.2	2,578.6
Mt	16.7	13.9	4.3	81.4	34.4	629.9	130.6	3,599.7	3,599.3	568	713	3,600	3,599.9	3,600	3,600.4
An	17.4	391	65	605.6	1,558	21,804	685.8	51,162.4	82,292.4	3,052.4	7,549	68,845.4	16,741.2	22,036.8	41,283.6
Mn	37	925	123	1,189	3,467	47,542	1,391	211405	167,386	4,465	25,227	105,734	43,411	40,891	64,023
Ag	74.2	78.8	94.7	69.2	80.9	96.9	70.1	80.9	97.5	70.6	81.6	97.6	71.5	85.4	97.6
Mg	77.6	83.2	97.6	74.5	83.3	99	76.3	85.9	99.2	72	82.9	98.3	72.9	99.4	98.6

k-Centrum															
n	10			20			30			40			60		
p	3	5	8	5	10	16	7	15	24	9	20	32[3]	13	30	48[1]
At	10.8	6.1	5.3	14.1	33.2	57.7	74.6	84.6	921.2	173	442.2	2,336.9	676.1	1,155.4	2,222
Mt	18.5	10.6	9.2	22.7	115.3	191.7	162	209.2	3,012.1	394.9	1,051.8	3,600	814.2	2,265	3,599.7
An	20.6	101.4	32.2	140.6	1,933.8	3,527.2	1,662.2	3,552.8	44,945	2,563.4	14,941.6	71,892.6	5,663.4	24,109.4	29,005.6
Mn	81	487	127	493	8,047	12,769	5,623	9,626	17,1469	7,745	40,269	125,596	11,227	52,339	46,295
Ag	28	37	60.2	26.6	46.2	74.5	29.3	42.1	86.2	30.3	42.6	87.8	30.3	39.7	79.2
Mg	37.9	70.3	84.2	28.9	58.8	92.3	33.3	48.3	91.9	37.8	52.1	92	32.2	48.2	88.2
Trimmean															
n	10			20			30			40			60		
p	3	5	8	5	10	16	7	15	24	9	20	32	13	30	48
At	6.8	0.9	1	5.1	7	17.6	22.9	31.5	4.8	24.5	84.3	36.8	79	179.9	339
Mt	12.6	2.4	2.9	11.9	11.9	34.2	51.6	46	7.5	48.3	296.9	97.3	195.4	252.8	520.5
An	2.6	1	3.4	38.2	147.6	890.6	108	561	6.2	9	4,487.6	998.2	225	2,540.6	6,100.2
Mn	8	1	13	187	335	2,235	406	1,191	27	29	19,803	3,223	753	3,806	11,259
Ag	26.8	25	0	25.2	29.3	0	26.6	36.1	0	26.1	31.7	0	25.5	37	0
Mg	33.2	25	0	26.2	42.3	0	28.2	48.3	0	29.7	39.1	0	26.5	43.5	0

valid inequalities, as for instance in Puerto et al. (2011), leading to solve larger problem sizes of capacitated discrete ordered median problems.

10.6 Conclusions

This chapter provides an overview of the ordered median function and its corresponding Ordered Median Location Problem as a powerful tool from a modeling point of view within the area of Location Analysis. We have included some of their most important insights considering three different framework spaces: continuous, networks and discrete. Our goal has been to structure this chapter as an useful tool for those readers that wish to start the study of the ordered functions and their related ordered median location problems. Moreover, the extensive list of references that have been included may result, for expert readers, an interesting source of literature to carry out a deeper study of this topic.

Acknowledgements The authors were partially supported by projects FQM-5849 (Junta de Andalucía\FEDER), the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office, MTM2010-19576-C02-01/02 and MTM2013-46962-C02-01/02 (Ministry of Economy and Competitiveness\FEDER, Spain).

References

Ben-Israel A, Iyigun C (2010) A generalized Weiszfeld method for the multi-facility location problem. Oper Res Lett 38:207–214

Berman O, Kalcsics J, Krass D, Nickel S (2009) The ordered gradual covering location problem on a network. Discrete Appl Math 157:3689–3707

Blanco V, Ben Ali SEH, Puerto J (2013) Minimizing ordered weighted averaging of rational functions with applications to continuous location. Comput Oper Res 40:1448–1460

Blanco V, Ben Ali SEH, Puerto J (2014a) Revisiting several problems and algorithms in continuous location with l_p norms. Comput Optim Appl 58:563–595

Blanco V, El-Haj Ben-Ali S, Puerto J (2014b) Continuous multifacility ordered median location problems. ArXiv:1401.0817v1, ArXiv.org

Blanquero R, Carrizosa E (2009) Continuous location problems and big triangle small triangle: constructing better bounds. J Global Optim 45:389–402

Boland N, Domínguez-Marín P, Nickel S, Puerto J (2006) Exact procedures for solving the discrete ordered median problem. Comput Oper Res 33:3270–3300

Brimberg J, Hansen P, Mladenovic N, Taillard ED (2000) Improvement and comparison of heuristics for solving the uncapacitated multisource weber problem. Oper Res 48:444–460

Domínguez-Marín P, Nickel S, Hansen P, Mladenović N (2005) Heuristic procedures for solving the discrete ordered median problem. Ann Oper Res 136:145–173

Drezner Z (2007) A general global optimization approach for solving location problems in the plane. J Global Optim 37:305–319

Drezner Z, Nickel S (2009a) Constructing a DC decomposition for ordered median problems. J Global Optim 45:187–201

Drezner Z, Nickel S (2009b) Solving the ordered one-median problem in the plane. Eur J Oper Res 195:46–61
Durier R, Michelot C (1985) Geometrical properties of the Fermat–Weber problem. Eur J Oper Res 20:332–343
Edelsbrunner H (1987) Algorithms in combinatorial geometry. Springer, New York
Espejo I, Marín A, Puerto J, Rodríguez-Chía AM (2009) A comparison of formulations and solution methods for the minimum-envy location problem. Comput Oper Res 36:1966–1981
Espejo I, Rodríguez-Chía AM, Valero C (2009) Convex ordered median problem with l_p-norms. Comput Oper Res 36:2250–2262
Francis R, Lowe T, Tamir A (2000) Aggregation error bounds for a class of location models. Oper Res 48:294–307
Grzybowski J, Nickel S, Pallaschke D, Urbański R (2011) Ordered median functions and symmetries. Optimization 60:801–811
Hakimi S (1964) Optimal location of switching centers and the absolute centers and medians of a graph. Oper Res 12:450–459
Hakimi S, Labbé M, Schmeichel E (1992) The Voronoi partition of a network and its applications in location theory. Orsa J Comput 4:412–417
Hardy GH, Littlewood JE, Pólya G (1952) Inequalities, 2nd edn. Cambridge University Press, Cambridge
Hooker J, Garfinkel R, Chen C (1991) Finite dominating sets for network location problems. Oper Res 39:100–118
Jibetean D, de Klerk E (2006) Global optimization of rational functions: a semidefinite programming approach. Math Program 106:93–109
Kalcsics J, Nickel S, Puerto J, Tamir A (2002) Algorithmic results for ordered median problems. Oper Res Lett 30:149–158
Kalcsics J, Nickel S, Puerto J (2003) Multifacility ordered median problems on networks: a further analysis. Networks 41:1–12
Kalcsics J, Nickel S, Puerto J, Rodríguez-Chía AM (2010a) Distribution systems design with role dependent objectives. Eur J Oper Res 202:491–501
Kalcsics J, Nickel S, Puerto J, Rodríguez-Chía AM (2010b) The ordered capacitated facility location problem. TOP 18:203–222
Kim-Chuan T, Todd MJ, Tutuncu RH (2006) On the implementation and usage of sdpt3—a matlab software package for semidefinite-quadratic-linear programming, version 4.0. Optimization software. http://www.math.nus.edu.sg/~mattohkc/sdpt3/guide4-0-draft.pdf
Lasserre J (2009) Moments, positive polynomials and their applications. Imperial College Press, London
López-de-los-Mozos M, Mesa JA, Puerto J (2008) A generalized model of equality measures in network location problems. Comput Oper Res 35:651–660
Marín A, Nickel S, Puerto J, Velten S (2009) A flexible model and efficient solution strategies for discrete location problems. Discrete Appl Math 157:1128–1145
Marín A, Nickel S, Velten S (2010) An extended covering model for flexible discrete and equity location problems. Math Method Oper Res 71:125–163
McCormick S (2005) Submodular function minimization. In: Discrete optimization. Elsevier, Amsterdam, pp 321–391
Nickel S (2001) Discrete ordered Weber problems. In: Operations research proceedings 2000. Selected papers of the symposium, OR 2000, Dresden, 9–12 September 2000. Springer, Berlin, pp 71–76
Nickel S, Puerto J (1999) A unified approach to network location problems. Networks 34:283–290
Nickel S, Puerto J (2005) Location theory. A unified approach. Springer, Berlin
Nickel S, Puerto J, Rodríguez-Chía AM, Weissler A (2005) Multicriteria planar ordered median problems. J Optim Theory Appl 126:657–683
Okabe A, Boots B, Sugihara K (1992) Spatial tessellations: concepts and applications of Voronoĭ diagrams. Wiley series in probability and mathematical statistics: applied probability and statistics. Wiley, Chichester, with a foreword by D.G. Kendall

Papini P, Puerto J (2004) Averaging the k largest distances among n: k-centra in Banach spaces. J Math Anal Appl 291:477–487

Puerto J (2008) A new formulation of the capacitated discrete ordered median problems with $\{0, 1\}$ assignment. In: Operations research proceedings 2007. Selected papers of the annual international conference of the German Operations Research Society (GOR), Saarbrücken, 5–7 September 2007. Springer, Berlin, pp 165–170

Puerto J, Fernández F (2000) Geometrical properties of the symmetric single facility location problem. J Nonlinear Convex Anal 1:321–342

Puerto J, Rodríguez-Chía AM (2005) On the exponential cardinality of FDS for the ordered p-median problem. Oper Res Lett 33:641–651

Puerto J, Tamir A (2005) Locating tree-shaped facilities using the ordered median objective. Math Program 102:313–338

Puerto J, Ramos AB, Rodríguez-Chía AM (2011) Single-allocation ordered median hub location problems. Comput Oper Res 38:559–570

Puerto J, Ramos AB, Rodríguez-Chía AM (2013) A specialized branch & bound & cut for single-allocation ordered median hub location problems. Discrete Appl Math 161:2624–2646

Puerto J, Pérez-Brito D, García-González C (2014) A modified variable neighborhood search for the discrete ordered median problem. Eur J Oper Res 234(1):61–76. doi:10.1016/j.ejor.2013.09.029

Rodríguez-Chía AM, Nickel S, Puerto J, Fernández FR (2000) A flexible approach to location problems. Math Method Oper Res 51:69–89

Rodríguez-Chía AM, Puerto J, Pérez-Brito D, Moreno JA (2005) The p-facility ordered median problem on networks. TOP 13:105–126

Rodríguez-Chía AM, Espejo I, Drezner Z (2010) On solving the planar k-centrum problem with Euclidean distances. Eur J Oper Res 207:1169–1186

Rosenbaum R (1950) Subadditive functions. Duke Math J 17:227–247

Schöbel A, Scholz D (2010) The big cube small cube solution method for multidimensional facility location problems. Comput Oper Res 37:115–122

Ward J, Wendell R (1985) Using block norms for location modeling. Oper Res 33:1074–1090

Yager R (1988) On ordered weighted averaging aggregation operators in multicriteria decision making. IEEE Trans Syst Man Cybern 18:183–190

Chapter 11
Multi-Period Facility Location

Stefan Nickel and Francisco Saldanha da Gama

Abstract In this chapter, we cover basic aspects related with facility location problems involving time dependent parameters. The emphasis is put on problems defined over a multi-period finite planning horizon. A brief overview of continuous and network problems is presented. Nevertheless, most of the chapter focus on a discrete setting. Basic modeling aspects and solution techniques are discussed. Additionally, some features of practical relevance are considered. The value of the multi-period solution is introduced as a measure for the relevance of considering a multi-period modeling framework instead of a static one. Current challenges and future trends on the topic are discussed.

Keywords Discrete models • Multi-period facility location • Value of the multi-period solution

11.1 Introduction

Facility location decisions are usually made taking into account the values of some parameters, such as the setup costs for the facilities and the demand levels. If variations are predictable for such values, it may be desirable to plan in advance for future adjustments in the location of facilities and in other related decisions (e.g., shipment decisions). In this case, locating a set of facilities becomes a question not only of "where" but also of "when". A new dimension is introduced in the decision space: the time. This is the topic of the current chapter.

S. Nickel (✉)
Institute for Operations Research, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Fraunhofer Institute for Industrial Mathematics (ITWM), Kaiserslautern, Germany
e-mail: stefan.nickel@kit.edu

F. Saldanha da Gama
Centro de Investigação Operacional/Departamento de Estatística e Investigação Operacional, Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal
e-mail: fsgama@fc.ul.pt

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_11

In order to capture predictable variations in the parameters of a facility location problem, we often have to consider a dynamic or time-dependent model. From a practical point of view, this type of model can be quite relevant because it allows for embedding other decisions, such as those related with (1) inventory management, (2) opening new facilities and removing existing ones, and (3) adjustment of the operating capacities (which, from a cost point of view is often better than opening new facilities). Even when the underlying parameters do not induce a dynamic model, some other conditions may do so. For instance, if a budget constraint exists say, per year, for installing new facilities, then locating the facilities over time may be unavoidable.

When facility location decisions are to be made over time, it is important to define the *planning horizon* beforehand. This is the time frame for which the decision maker wishes to plan. Only a few papers have investigated facility location problems over an infinite planning horizon. In this case, a static or a finite-horizon decision is usually sought that is "the best" for an infinitely long planning horizon. Some works in this direction include Chand (1988) and Daskin et al. (1992). Nevertheless, in most cases, decision makers assume a finite planning horizon (see the recent review paper by Arabani and Zanjirani Farahani 2012). This is the case we consider in this chapter.

When working with dynamic models, we can make a distinction between continuous and discrete-time models. In the first case, there are no specific moments for implementing the decisions; the best timing for performing changes in the system is itself a decision to make. Some works exploring this feature include Drezner and Wesolowsky (1991), Orda and Rom (1991), Puerto and Rodríguez-Chía (1999), and Zanjirani Farahani et al. (2009). In our opinion, continuous-time facility location problems are better addressed in the context of optimal control. Therefore, in this chapter we do not focus on this type of problems. Instead, we consider a discrete-time setting in which we have several moments in time for implementing the decisions. These moments induce a partition of the planning horizon into several time periods.

Facility location problems are often classified, according to the location space, as being continuous, on a network, or discrete (Hamacher and Nickel 1998). In recent years, due to successful applications of location theory to many areas, discrete models have increasingly played a major role. For this reason, in this chapter, special emphasis is given to this type of problems.

The remainder of the chapter is organized as follows: in Sects. 11.2 and 11.3 we present a brief overview of continuous and network multi-period facility location problems, respectively. In Sects. 11.4 and 11.5 we focus on discrete problems. Section 11.6 is used for introducing the value of the multi-period solution. Finally, in Sect. 11.7, we discuss some challenges and future trends on the topic.

11.2 Continuous Problems

One of the best-known facility location problems is the Weber problem: given a set of weighted nodes in the Euclidean plane, where to locate a single facility minimizing the weighted sum of the distances to the points? A multi-period extension of this problem was first proposed by Wesolowsky (1973). A finite planning horizon T, divided into several time periods, is assumed. In each period $t \in T$, a set of weighted nodes J_t is considered. The goal is to find the optimal location for the single facility in each period. When the facility changes from one location to another (in consecutive periods), a relocation cost is paid. The conceptual model proposed by Wesolowsky (1973) is the following:

$$\text{Minimize} \quad \sum_{t \in T} \sum_{j \in J_t} c_{tj}(x_t, y_t) + \sum_{t=2}^{|T|} f_t z_t \tag{11.1}$$

$$\text{subject to} \quad z_t = 0 \text{ if } d_{t-1,t} = 0, \quad t \in T \backslash \{1\} \tag{11.2}$$

$$z_t \in \{0, 1\}, \quad t \in T. \tag{11.3}$$

In this model, $c_{tj}(x_t, y_t)$ represents the present value of the cost for shipping from a facility located at (x_t, y_t) to demand point $j \in J_t$ in period $t \in T$; f_t denotes the cost for relocating the facility at the beginning of period $t \in T$; $d_{t-1,t}$ is the distance by which the facility is moved at the beginning of period $t \in T \backslash \{1\}$. All the costs are assumed to be forecasted in advance and therefore known to the model. For tackling this problem, Wesolowsky (1973) proposed an incomplete dynamic programming algorithm. The stages are associated with the time periods, the states correspond to a set of possible locations for the facility and the decisions correspond to the possible changes in the location of the facility. The relevance of this work arises from the fact that it represents the first attempt to extend the Weber problem to a multi-period setting. Nevertheless, the first work addressing the location and relocation of a single facility in the plane over a multi-period finite planning horizon is due to Ballou (1968). The goal is to maximize the total profit generated by a distribution system involving factories, markets and the single warehouse to be located and relocated. In that paper, a restricted set of potential locations for the warehouse was defined considering the optimal location for the facility in the different periods. This set then defined the states for all periods, and (incomplete) dynamic programming was then applied. The method was later converted into an exact one by Sweeney and Tatham (1976) who extended the restricted set just mentioned. In fact, a set of potential locations for the warehouse can be found in each time period, thus ensuring that the optimal solution of the problem is not lost when dynamic programming is applied. It is worth noting that the methodologies proposed by Ballou (1968) and Sweeney and Tatham (1976) can be applied to problems defined in a discrete setting.

Drezner and Wesolowsky (1991), investigated a different type of problem. Like in all of the above works, a single facility is considered, which can be relocated over time as a reaction to predictable changes in the demand. The set J of demand nodes is the same throughout the planning horizon. The demand of each node $j \in J$, is represented by a continuous function of time $w_j(.)$. A planning horizon T divided into several time periods is assumed. The following optimization model can be considered for each period $t \in T$:

$$C_t = \min_{x_t, y_t} \left\{ \sum_{j \in J} W_{jt} d_j(x_t, y_t) \right\}. \tag{11.4}$$

In this expression, (x_t, y_t) denotes the coordinates of the facility in period $t \in T$; $W_{jt} = \int_{a_{t-1}}^{a_t} w_j(\tau) d\tau$; a_{t-1} and a_t are the lower and upper time limits for period t, respectively; $d_j(x_t, y_t)$ denotes the distance between demand point $j \in J$ and point (x_t, y_t). The cost for the entire planning horizon is given by $\sum_{t \in T} C_t$. Drezner and Wesolowsky (1991) made use of the above model to solve a more general problem which consists of making a decision about the division of the planning horizon into time periods. In this case, the number of time periods and the "break points" are decisions to make. This work was later extended by Zanjirani Farahani et al. (2009) that included a cost for relocating the facility.

Scott (1971) studied a multi-facility, multi-period continuous location problem, assuming a finite planning horizon T divided into several time periods, and a set of demand nodes, J. In each time period, a single facility is to be located and must remain operating until the end of the planning horizon. A sequence of $|T|$ problems can be considered. In particular, the following mathematical model holds for period $t \in T$ (the coordinates (x_τ, y_τ), $\tau = 1, \ldots, t-1$, were already determined):

$$\text{Minimize} \quad \sum_{j \in J} \sum_{\tau=1}^{t-1} u_{j\tau} d_j(x_\tau, y_\tau) + \sum_{j \in J} u_{jt} d_j(x_t, y_t) \tag{11.5}$$

$$\text{subject to} \quad \sum_{\tau=1}^{t} u_{j\tau} = 1, \quad j \in J \tag{11.6}$$

$$u_{j\tau} \in \{0, 1\}, \quad \tau = 1, \ldots, t, \ j \in J. \tag{11.7}$$

In this model, (x_t, y_t) are the coordinates (to be determined) of the facility to install at the beginning of period $t \in T$; u_{jt} is a binary variable equal to 1 if demand point $j \in J$ is allocated to the facility installed in period $t \in T$ (such allocation can only occur in periods $t, \ldots, |T|$), and 0 otherwise; $d_j(x_t, y_t)$ is the Euclidean distance between demand node $j \in J$ and the facility to be installed in period $t \in T$. By solving the full sequence of problems (one for each $t \in T$), a solution is obtained for the multi-period problem. Nevertheless, using such a myopic procedure, optimality cannot be guaranteed for the whole planning horizon.

A multi-period extension of the planar p-median problem was proposed by Drezner (1995) who considered a finite planning horizon divided into $|T| = p$ time periods. The set of demand nodes is denoted by J and demand changes over time. The demand of node $j \in J$ is represented by a continuous function of time $w_j(.)$. At the beginning of each time period $t \in T$, exactly one facility is to be installed. The decision variables are the coordinates of the p locations for the facilities, (x_t, y_t), $t \in T$. The problem can be formulated as follows:

$$\text{Minimize} \quad \sum_{t \in T} \sum_{j \in J} W_{jt} \min_{\tau=1,\ldots,t} \left\{ d_j(x_\tau, y_\tau) \right\}, \tag{11.8}$$

where $d_j(x_t, y_t)$, $t \in T$, represents the distance between demand node $j \in J$ and the facility established at the beginning of period $t \in T$; $W_{jt} = \int_{a_{t-1}}^{a_t} w_j(\tau) d\tau$; a_{t-1} and a_t are, respectively, the lower and upper time limits for period t. The function to be minimized in (11.8) results from adding the costs for all periods. Drezner (1995) proposed a specially tailored algorithm for the 2-facility problem and suggested the use of a standard non-linear solver for the general case.

11.3 Network Problems

One of the earliest works on multi-period facility location problems on networks is due to Cavalier and Sherali (1985). The problems under consideration consist of progressively installing a set of facilities on a chain or on a tree considering a multi-period finite planning horizon. In each period, at most one facility can be installed. Demand occurs continuously on the edges, according to a uniform distribution. Different strategies were analyzed for obtaining solutions to the problems.

Considering general networks, Mesa (1991) addressed several multi-period facility location problems. Different concepts were introduced in that paper, such as the vertex $|T|$-period p-median, the vertex multi-period $(\alpha_1, \ldots, \alpha_{|T|})$-median and the absolute multi-period $(\alpha_1, \ldots, \alpha_{|T|})$-median. Among the different problems studied, the absolute multi-period $(\alpha_1, \ldots, \alpha_{|T|})$-median problem was, at the time, the one which was closer to what could be referred to as an extension of the p-median problem to a multi-period setting. In that problem, α_t points must be located in each period $t \in T$, satisfying $\sum_{t \in T} \alpha_t = p$. The author proved that the initial infinite set of possible choices for facilities can be reduced to a discrete set of nodes. This is due to the vertex-optimality property (Hakimi 1964, 1965), which holds for this multi-period problem.

The extension of the network p-median problem to a multi-period setting was proposed by Hakimi et al. (1999). Considering a time varying network, $N = (V, E, T)$, with T representing the planning horizon, it is assumed that the weight of each vertex $v_j \in V$ and the length of each edge $e \in E$ are functions of time and are invariant in each period. Assuming moving costs for the facilities, the multi-period,

1-median problem on network N can be formulated as follows:

$$\text{Minimize} \quad \sum_{t \in T} \left(\sum_{j \in V} w_{jt} d_t(v_j, x_t) + g(t) d_t(x_t, x_{t+1}) \right). \tag{11.9}$$

In this model, w_{jt} denotes the weight of vertex $v_j \in V$ in period $t \in T$; x_t represents the location of the median in period $t \in T$ (the exact location x_t, is defined with respect to the edge to which the median belongs and is given by its distance to the closest end node of the edge); $d_t(v_j, x_t)$ is the shortest path between v_j and x_t in period $t \in T$; $g(t)$ is a function representing the unitary cost for relocating the facility in the end of period t moving it from location x_t in period t to location x_{t+1} in period $t+1$ ($t \in T$, $x_{|T|+1} = x_{|T|}$). Hakimi et al. (1999) proved that the vertex-optimality property holds for this problem. The above model and this result can be easily extended to the p-facility case. The formulation is the following:

$$\text{Minimize} \quad \sum_{t \in T} \left(\sum_{j \in V} w_{jt} d_t(v_j, X_t) + g(t) d_t(X_t, X_{t+1}) \right). \tag{11.10}$$

In this case, $X_1, \ldots, X_{|T|}$ are the sets of locations for the p facilities during the planning horizon with $X_{|T|+1} = X_{|T|}$; $d_t(v_j, X_t) = \min\{d_t(v_j, x_k) \mid x_k \in X_t\}$; $d_t(X_t, X_{,t+1})$ is defined by the total weight of a minimum weight perfect matching in the complete bipartite graph $G_t(X_t, X_{t+1})$ defined as follows: X_t and X_{t+1} define the partition; for every point x' in X_t and for every point x'' in X_{t+1} the weight of the edge (x', x'') is given by $d_t(x', x'')$. In (11.10), $g(t)$ denotes the unitary cost for relocating a facility in (the end of) time period $t \in T$. This problem is NP-hard since it includes the static network p-median problem as a particular case. For this reason, the authors developed a heuristic procedure.

One important class of facility location problems on networks are center problems. The multi-period extension of the 1-center problem on a network was proposed also by Hakimi et al. (1999). The model is the following (the notation is the same presented above):

$$\underset{x_1, x_2, \ldots, x_{|T|+1}}{\text{Minimize}} \quad \sum_{t \in T} \max_{j \in V} \left\{ w_{jt} d_t(v_j, x_t) + g(t) d_t(x_t, x_{t+1}) \right\}. \tag{11.11}$$

Again, $X_{|T|+1} = X_{|T|}$. If the choice for x_t is restricted to a finite number of points in the network, the problem can be handled using a technique similar to the one presented in the same paper for the multi-period p-median problem.

The existing literature reveals that for most of the multi-period extensions proposed so far for well-known minsum facility location problems, the vertex-optimality property holds. This reduces the location space to a discrete set. Accordingly, models and techniques from integer programming and combinatorial

optimization emerge as a possibility for tackling these problems. Multi-period minmax facility location problems on networks have been scarcely investigated.

11.4 Discrete Problems

We start with one of the best-known discrete facility location problems, the p-median problem (see Chap.2), which can be easily extended to a multi-period setting. Consider a set J, of nodes, whose demand must be supplied during a finite multi-period planning horizon, T. Let $I \subseteq J$ be the set of nodes where the facilities can be located and assume that p facilities have to be operating in each period. The problem of deciding the best location for the facilities throughout the planning horizon, minimizing the total cost for satisfying the demand can be formulated as follows:

$$\text{Minimize} \quad \sum_{t\in T}\sum_{i\in I}\sum_{j\in J} c_{ijt}x_{ijt} \tag{11.12}$$

$$\text{subject to} \quad \sum_{i\in I} x_{ijt} = 1, \quad t \in T,\ j \in J \tag{11.13}$$

$$\sum_{j\in J} x_{ijt} \leq |J|x_{iit}, \quad t \in T,\ i \in I \tag{11.14}$$

$$\sum_{i\in I} x_{iit} = p, \quad t \in T \tag{11.15}$$

$$x_{ijt} \in \{0,1\}, \quad t \in T,\ i \in I,\ j \in J. \tag{11.16}$$

In this formulation, c_{ijt} represents the cost for allocating demand node $j \in J$ to facility $i \in I$ in period $t \in T$; x_{ijt} is a binary variable equal to 1 if demand node $j \in J$ is allocated to facility $i \in I$ in period $t \in T$ and 0 otherwise; $x_{iit} = 1$ indicates that a facility is operating at $i \in I$ in period $t \in T$ (i is allocated to itself). When $I = J$ we have a multi-period p-median problem.

In order to progressively build models that are more relevant from a practical point of view, we first note that the above problem still has little "multi-period flavor" because it can be decoupled, leading to $|T|$ single-period problems. Nevertheless, this model is an excellent basis for what we present next. In fact, a more interesting multi-period problem emerges if we include opening and closing costs for the facilities. This was first done by Wesolowsky and Truscott (1975). The extended problem can be formulated as follows:

$$\text{Minimize} \quad \sum_{t\in T}\sum_{i\in I}\sum_{j\in J} c_{ijt}x_{ijt} + \sum_{t\in T}\sum_{i\in I} g_{it}z'_{it} + \sum_{t\in T}\sum_{i\in I} h_{it}z''_{it} \tag{11.17}$$

$$\text{subject to} \quad (11.13)\text{–}(11.16)$$

$$\sum_{i \in I} z'_{it} \leq m_t, \qquad t \in T \tag{11.18}$$

$$x_{iit} - x_{ii,t-1} + z''_{i,t-1} - z'_{it} = 0, \qquad t \in T \setminus \{1\},\ i \in I \tag{11.19}$$

$$z'_{it}, z''_{it} \in \{0, 1\}, \qquad t \in T,\ i \in I. \tag{11.20}$$

In this model, facilities are assumed to be opened (closed) at the beginning (end) of time periods; m_t is the maximum number of facilities that can be opened in each period $t \in T$, whereas the binary variable z'_{it} (z''_{it}) is equal to 1 if a facility is opened (closed) at $i \in I$ in period $t \in T$ and 0 otherwise; g_{it} and h_{it} ($i \in I, t \in T$) denote the opening and closing costs, respectively. Wesolowsky and Truscott (1975) solve the above problem using dynamic programming. However, the method can only be used for instances with a small number of potential locations for the facilities because the dimension of the state space is exponential in this number.

Galvão and Santibañez-Gonzalez (1992) do not consider closing decisions and assume that the number of operating facilities does not have to be the same in all periods. Their formulation can be obtained from the above model by ignoring the variables and costs associated with closing the facilities and by replacing p with p_t in (11.15). For each period $t \in T$, p_t denotes the number of facilities to be operating in that period. Furthermore, in their model constraints (11.18) are redundant ($m_t = |I|, t \in T$) and constraints (11.14) are disaggregated, yielding

$$x_{ijt} \leq x_{iit}, \quad t \in T,\ i \in I,\ j \in J. \tag{11.21}$$

Without closing decisions, constraints (11.19) can be written as

$$z'_{it} \geq x_{iit} - x_{ii,t-1}, \quad t \in T \setminus \{1\},\ i \in I. \tag{11.22}$$

For this problem, Galvão and Santibañez-Gonzalez (1992) proposed two Lagrangean relaxation based procedures for computing lower and upper bounds: in the first one, constraints (11.13) and (11.22) are dualized; in the second, the choice involves constraints (11.21) and (11.22).

In all of the problems presented so far in this section, facilities can be opened and closed more than once during the planning horizon. However, in many applications this is not realistic. In order to illustrate how this aspect can be captured, we consider another well-known problem: the uncapacitated facility location problem (UFLP) described in Chap. 3. Like for the p-median problem, the extension of the UFLP to a multi-period setting is straightforward. Again we consider a finite multi-period planning horizon, T. The set of potential locations for the facilities is denoted by $I = \{1, \ldots, m\}$ and the set of demand nodes by $J = \{1, \ldots, n\}$. Additionally, let f_{it} be the cost for operating facility $i \in I$ in period $t \in T$, and c_{ijt} the cost for satisfying all the demand of customer $j \in J$ in period $t \in T$ from facility $i \in I$. A

multi-period uncapacitated facility location problem is the following:

$$\text{Minimize} \quad \sum_{t\in T}\sum_{i\in I} f_{it}y_{it} + \sum_{t\in T}\sum_{i\in I}\sum_{j\in J} c_{ijt}x_{ijt} \tag{11.23}$$

$$\text{subject to} \quad \sum_{i\in I} x_{ijt} = 1, \quad t \in T,\ j \in J \tag{11.24}$$

$$\sum_{j\in J} x_{ijt} \le n y_{it}, \quad t \in T,\ i \in I \tag{11.25}$$

$$x_{ijt} \ge 0, \quad t \in T,\ i \in I,\ j \in J \tag{11.26}$$

$$y_{it} \in \{0, 1\}, \quad t \in T,\ i \in I. \tag{11.27}$$

In this formulation, x_{ijt} represents the fraction of the demand of customer $j \in J$ in period $t \in T$ that is supplied by facility $i \in I$; y_{it} is a binary variable equal to 1 if a facility is operating at $i \in I$ in period $t \in T$ and 0 otherwise. Again, this problem can be decomposed into $|T|$ single-period problems. Nevertheless, it contains the basic ingredients for building more interesting models. In fact, one extension of this problem was proposed by Warszawski (1973), who included opening costs for the facilities. These costs are incurred whenever a facility is opened (even if the same facility has operated in some past period). Denoting by g_{it} the cost for opening a facility at $i \in I$ in the beginning of period $t \in T$, the model proposed by Warszawski (1973) differs from (11.23)–(11.27) by considering the following quadratic objective function:

$$\sum_{t\in T}\sum_{i\in I} g_{it}y_{it}\,(1 - y_{i,t-1}) + \sum_{t\in T}\sum_{i\in I} f_{it}y_{it} + \sum_{t\in T}\sum_{i\in I}\sum_{j\in J} c_{ijt}x_{ijt}, \tag{11.28}$$

with $y_{i0} = 0$, $i \in I$. Warszawski (1973) considered dynamic programming for solving instances with a small number of potential locations for the facilities, $|I|$, and a local search heuristic for larger instances. Chardaire et al. (1996) studied the same problem starting by disaggregating constraints (11.25). They developed a Langragean relaxation based algorithm for computing lower and upper bounds. A linearized model was also proposed and compared with the quadratic one in terms of the quality of the lower bounds produced.

Another extension of model (11.23)–(11.27) was proposed by Canel and Khumawala (1997) for locating facilities across different countries. They explicitly considered binary decision variables z_{it} indicating whether or not a new facility is opened at $i \in I$ in period $t \in T$. They proposed a profit maximization problem as follows:

$$\text{Maximize} \quad \sum_{t\in T}\sum_{i\in I}\sum_{j\in J} r_{ijt}x_{ijt} - \sum_{t\in T}\sum_{i\in I} f_{it}y_{it} - \sum_{t\in T}\sum_{i\in I} g_{it}z_{it} \tag{11.29}$$

$$\text{subject to} \quad (11.24), (11.26), (11.27)$$

$$\sum_{j \in P_{it}} x_{ijt} \leq n_{it} y_{it}, \quad t \in T,\ i \in I \tag{11.30}$$

$$z_{it} \geq y_{it} - y_{i,t-1}, \quad t \in T,\ i \in I \tag{11.31}$$

$$z_{it} \in \{0, 1\}, \quad t \in T,\ i \in I, \tag{11.32}$$

with $y_{i0} = 0$, $i \in I$. In this model, r_{ijt} represents the revenue obtained when supplying all the demand of customer $j \in J$ in period $t \in T$ from facility $i \in I$. For each facility $i \in I$ there is a maximum number of customers, n_{it}, it can supply in period $t \in T$. Furthermore, not all the facilities can supply all customers. In particular, P_{it} represents the set of customers that can be served from facility $i \in I$ in period $t \in T$. As we will see below, constraints (11.30) had been proposed before for another problem. Canel and Khumawala (1997) developed a branch-and-bound procedure for this problem adapting the algorithm proposed by Khumawala (1972), and Canel and Khumawala (2001) proposed a heuristic approach for the same problem.

In all of the above problems, facilities can be opened and closed more than once during the planning horizon. Dias et al. (2007) point out that these models ignore the fact that re-opening a facility has in general a smaller cost than opening it for the first time (for instance, land acquisition costs are incurred only once). They propose a model taking this aspect into account. Additional decision variables are required to distinguish whether a facility is being opened for the first time or is being re-opened. A primal-dual heuristic is proposed for obtaining lower and upper bounds for the problem. The gap is closed using a branch-and-bound procedure.

11.5 Modular Construction of Intrinsic Multi-Period Facility Location Models

In many practical situations it is not acceptable to install and remove a facility, say, in consecutive periods. This may make sense for seasonal facilities, such as warehouses if, for instance, they can be rented for short time intervals. Nevertheless, this cannot be assumed in general. Accordingly, the models presented in the previous section may be short for capturing some real-world problems. Early, researchers have noticed this fact and have considered models involving constraints that impose a limit on the number of changes performed in each location during the planning horizon. Often, such constraints state that once a facility is installed (removed), it must remain opened (closed) until the end of the planning horizon.

We consider again the multi-period p-median problem, i.e., we assume that a plan is to be made for locating exactly p facilities in a finite multi-period planning horizon T. Let us assume that removing facilities is not allowed. One additional feature that may be worth considering for this type of problem is the speed at which p changes. The adequate model is the following (the notation was introduced in

Sect. 11.4):

$$\text{Minimize} \quad \sum_{t\in T}\sum_{i\in I}\sum_{j\in J} c_{ijt}x_{ijt} \tag{11.33}$$

$$\text{subject to} \quad \sum_{i\in J} x_{ijt} = 1, \quad t \in T, j \in J \tag{11.34}$$

$$\sum_{j\in J} x_{ijt} \leq n x_{iit}, \quad t \in T, i \in J \tag{11.35}$$

$$\sum_{i\in J} x_{iit} = p_t, \quad t \in T \tag{11.36}$$

$$x_{iit} \geq x_{ii,t-1}, \quad t = 2, \ldots, |T|, i \in J \tag{11.37}$$

$$x_{ijt} \geq 0, \quad t \in T, i \in J, j \in J, \tag{11.38}$$

where $1 \leq p_1 \leq p_2 \leq \ldots \leq p_{|T|} = p$.

Constraints of type (11.37) were first proposed for a multi-period facility location problem by Roodman and Schwarz (1975, 1977). The latter paper was pioneering in the assumption that a set of facilities may be operating before the beginning of the planning horizon. These are the facilities that can be removed. Therefore, the possibility of adapting an existing system to predictable changes in some parameters, becomes explicitly considered in the models. The set of locations I can now be partitioned into two subsets: I^c and I^o. The former represents the facilities that are operating before the beginning of the planning horizon; the latter represents the set of locations for new facilities. A more comprehensive model for the multi-period facility location problem emerges:

$$\text{Minimize} \quad (11.23)$$

$$\text{subject to} \quad (11.24)\text{–}(11.27)$$

$$y_{it} \leq y_{i,t-1}, \quad t = 2, \ldots, |T|, i \in I^c \tag{11.39}$$

$$y_{it} \geq y_{i,t-1}, \quad t = 2, \ldots, |T|, i \in I^o. \tag{11.40}$$

Roodman and Schwarz (1977) were also pioneering by considering a maximum number of customers that can be served by each facility in each period and assumed that not all facilities can serve all customers. These aspects are easily accommodated in the above model if we replace (11.25) by (11.30). As mentioned before, the latter constraints would be later considered by Canel and Khumawala (1997). The research done by Roodman and Schwarz (1977) extends the work by the same authors published 2 years before (Roodman and Schwarz 1975) in which a pure phase-out problem had been considered.

The above models allow the removal of an existing facility before the beginning of period 1 with no costs imputed to the planning horizon. Imposing that the existing facilities must operate in at least one period, can be easily done by setting $y_{i1} = 1$, $i \in I^c$.

Van Roy and Erlenkotter (1982) proposed a reformulation of model (11.23)–(11.27), (11.39), and (11.40). Their idea, which can be extended to every multi-period facility location problem, consists of considering binary decision variables representing a change in a location instead of considering the traditional location variables. In particular, for an existing facility $i \in I^c$, a new binary variable z_{it}, can be defined that is equal to 1 if the facility is removed at the end of period t (i.e., it operates in periods $1, \ldots, t$) and 0 otherwise. For facility $i \in I^c$, $z_{i|T|} = 1$, indicates that the facility is operating during the entire planning horizon. For a potential new facility $i \in I^o$, the binary variable, z_{it}, is equal to 1 if it is installed at the beginning of period t (i.e., it operates in periods $t, \ldots, |T|$) and 0 otherwise. Using the new set of variables, we obtain the following model:

$$\text{Minimize} \quad \sum_{t \in T} \sum_{i \in I} F_{it} z_{it} + \sum_{t \in T} \sum_{i \in I} \sum_{j \in J} c_{ijt} x_{ijt} \tag{11.41}$$

$$\text{subject to} \quad \sum_{i \in I} x_{ijt} = 1, \quad t \in T,\ j \in J \tag{11.42}$$

$$x_{ijt} \leq \sum_{\tau \in \overline{T}_{it}} z_{i\tau}, \quad t \in T,\ i \in I,\ j \in J \tag{11.43}$$

$$x_{ijt} \geq 0, \quad t \in T,\ i \in I,\ j \in J \tag{11.44}$$

$$z_{it} \in \{0, 1\}, \quad t \in T,\ i \in I. \tag{11.45}$$

In this model, F_{it} ($i \in I, t \in T$) represents the total operation cost for facility i if $z_{it} = 1$, i.e., $F_{it} = f_{i1} + \ldots + f_{it}$ for $i \in I^c$, $t \in T$ and $F_{it} = f_{it} + \ldots + f_{i|T|}$ for $i \in I^o$, $t \in T$. The set $\overline{T}_{it}$ contains the periods in which it is possible to remove (install) a facility at $i \in I^c$ ($i \in I^o$) if we want to have it operating in period $t \in T$. More formally, $\overline{T}_{it} = \{t, \ldots, |T|\}$ if $i \in I^c$ and $\overline{T}_{it} = \{1, \ldots, t\}$ if $i \in I^o$. It is important to note that the aggregated costs F_{it} can be easily extended to more general situations, such as the one in which we have fixed setup and removal costs for the facilities. In fact, suppose that a fixed cost g_{it} is incurred when removing (installing) a facility $i \in I^c$ ($i \in I^o$) in period t. We can simply set $F_{it} = g_{it} + f_{i1} + \ldots + f_{it}$ for $i \in I^c$, $t \in T$ and $F_{it} = g_{it} + f_{it} + \ldots + f_{i|T|}$ for $i \in I^o$, $t \in T$.

The relation between the previous y-variables and the new z-variables is straightforward:

$$\begin{aligned}
z_{i|T|} &= y_{i|T|}, && i \in I^c \\
z_{it} &= y_{it} - y_{i,t+1}, && t \in \{1, \ldots, |T| - 1\},\ i \in I^c \\
z_{i1} &= y_{i1}, && i \in I^o \\
z_{it} &= y_{it} - y_{i,t-1}, && t \in \{2, \ldots, |T|\},\ i \in I^o
\end{aligned}$$

Using these relations, it is straightforward to prove that model (11.23)–(11.27), (11.39), and (11.40) is equivalent to model (11.41)–(11.45). The relevance of the

latter arises from the fact that it is particularly suited for the application of a dual-based heuristic, which is a popular method for obtaining sharp lower and upper bounds for discrete facility location problems. This fact was explored by Van Roy and Erlenkotter (1982). Multiplying constraints (11.43) by -1 the dual of the linear relaxation of model (11.41)–(11.45) becomes:

$$\text{Maximize} \quad \sum_{t\in T}\sum_{j\in J} v_{jt} \tag{11.46}$$

$$\text{subject to} \quad v_{jt} - w_{ijt} \le c_{ijt}, \quad t \in T,\ i \in I,\ j \in J \tag{11.47}$$

$$\sum_{j\in J}\sum_{\tau\in T_{it}} w_{ij\tau} \le F_{it}, \quad t \in T,\ i \in I \tag{11.48}$$

$$w_{ijt} \ge 0, \quad t \in T,\ i \in I,\ j \in J. \tag{11.49}$$

In this model, v_{jt} and w_{ijt} ($t \in T, i \in I, j \in J$) are the dual variables associated with constraints (11.42) and (11.43), respectively (with the latter previously multiplied by -1). The set T_{it} ($i \in I, t \in T$) contains the operating periods for facility i if a change (installation or removal) occurs in this location in period t. In particular, $T_{it} = \{1, \ldots, t\}$ if $i \in I^c$ and $T_{it} = \{t, \ldots, |T|\}$ if $i \in I^o$.

From (11.47) and (11.49) we may set

$$w_{ijt} = \max\{0, v_{jt} - c_{ijt}\}, \quad t \in T,\ i \in I,\ j \in J,$$

which yields the following condensed dual:

$$\text{Maximize} \quad (11.46)$$

$$\text{subject to} \quad \sum_{j\in J}\sum_{\tau\in T_{it}} \max\{0, v_{jt} - c_{ijt}\} \le F_{it}, \quad t \in T,\ i \in I. \tag{11.50}$$

The complementary slackness conditions for the linear relaxation of model (11.41)–(11.45) are the following:

$$v_{jt}\left(\sum_{i\in I} x_{ijt} - 1\right) = 0 \qquad t \in T,\ j \in J$$

$$w_{ijt}\left(\sum_{\tau\in \overline{T}_{it}} z_{i\tau} - x_{ijt}\right) = 0, \quad t \in T,\ i \in I,\ j \in J$$

$$x_{ijt}\left(v_{jt} - c_{ijt} - w_{ijt}\right) = 0, \quad t \in T,\ i \in I,\ j \in J$$

$$z_{it} S_{it} = 0, \qquad t \in T,\ i \in I,$$

where S_{it} represent the slack variables for constraints (11.50).

Van Roy and Erlenkotter (1982) proposed a heuristic for the condensed dual just presented. Starting from the trivial dual feasible solution defined by $v_{jt} = \min_{i \in I}\{c_{ijt}\}$ $(t \in T,\ j \in J)$ an ascent procedure is performed for increasing the values of the dual variables v_{jt}, thus increasing the value of the dual objective function. When this procedure does not lead to further improvements, a primal solution is constructed using the slackness conditions. Finally, a primal-dual adjustment phase is performed in order to reduce the gap between the values of the primal and dual objective functions. When no further gap reduction is achieved, a branch-and-bound procedure is applied to complete the search for an optimal solution for the problem. The reader should refer to Van Roy and Erlenkotter (1982) for further details.

The procedure developed by Van Roy and Erlenkotter (1982) is quite efficient to solve instances of moderate size. Nevertheless, this multi-period facility location problem includes the UFLP as a special case and thus, it is NP-hard. For this reason, Saldanha-da-Gama and Captivo (1998) proposed a two-phase heuristic procedure for the problem. The first phase is a drop procedure which starts with all facilities operating in all periods, and progressively removes operating periods to the facilities. This is done while a reduction in the total cost is observed. Losing feasibility is never allowed during the process. The second phase consists of a local search procedure.

Although representing an important basis for describing real problems, the above models still miss one important feature found in many applications: capacity constraints. Denote by Q_i the capacity of a facility located at $i \in I$, and by d_{jt} the demand of customer $j \in J$ in period $t \in T$. A capacitated multi-period facility location problem consists of minimizing (11.41) subject to (11.42), (11.44), (11.45), and

$$\sum_{j \in J} d_{jt} x_{ijt} \leq Q_i \sum_{\tau \in \overline{T}_{it}} z_{i\tau}, \quad t \in T,\ i \in I. \tag{11.51}$$

This model was addressed by Saldanha da Gama (2002) who developed a dual-based procedure for obtaining lower and upper bounds. The model was previously enhanced with (11.43) and

$$\sum_{t \in T} \sum_{i \in I} R_{kit} z_{it} \leq r_k, \quad k \in K. \tag{11.52}$$

By choosing appropriate values for R_{kit} and r_k, these generic constraints can accommodate every inequality involving the binary variables. This is important because the linear relaxation of capacitated facility location problems can often be strengthened through the inclusion of valid inequalities involving the location variables. For instance, a set of constraints often used in (static) capacitated facility location problems, state that the operational capacity must be at least equal to the

total demand. In the multi-period case, these constraints are written as

$$\sum_{i \in I} \left(Q_i \sum_{\tau \in \overline{T}_{it}} z_{i\tau} \right) \geq \sum_{j \in J} d_{jt}, \quad t \in T, \tag{11.53}$$

which can be easily accommodated in (11.52).

For the linear relaxation of model (11.41)–(11.45), (11.51), and (11.52), Saldanha da Gama (2002) extended the dual-based procedure proposed by Van Roy and Erlenkotter (1982), thus obtaining sharp lower and upper bounds for the problem.

The inclusion of capacity constraints is an important step towards building more comprehensive multi-period facility location models. Nevertheless, the capacity constraints (11.51) are rather restrictive when it comes to real applications, namely those arising in logistics (see Chap. 16). By considering a fixed capacity in each location, these constraints neglect the possibility of making future adjustments in the capacity of the facilities, which is a feature quite relevant in practice. In fact, it is often the case that adjusting the capacity of an existing facility is more advantageous from a cost point of view than installing a new facility in some other location. One attempt to overcome such restrictive representation for the capacities was made by Van Roy and Erlenkotter (1982) who considered exogenous time-dependent capacities Q_{it} ($i \in I$, $t \in T$). Nevertheless, this is still unsatisfactory from a practical point of view because no connection is established between the capacities in different periods.

The problem of planning for the capacity expansion of existing facilities was very much in focus in the 1970s and in the 1980s (see, for instance, Erlenkotter 1981, and Lee and Luss 1987). However, at that time, the focus was put mainly on the expansion of existing facilities. In many cases, the location of facilities was not even a decision to make. Furthermore, many of these works considered continuous adjustments in the capacities, which is often not adequate from a practical point of view. In fact, if we think of production or sorting lines, we immediately realize that changes in the capacities should be modular, or at least discrete.

One paper that clearly interconnects multi-period facility location decisions with discrete capacity expansion is due to Shulman (1991). A set of facility types P is considered, and in each location, facilities of different types can be progressively established during the planning horizon, as a way of adjusting the operating capacity of the system. In each period, at most one facility of each type can be installed in each location but several facilities can be installed if they are of different types. For each location $i \in I$, a set $P_i \subseteq P$ is assumed, corresponding to the set of facility types that can be located at i. Denote by c_{ijpt} the cost for supplying all the demand of customer $j \in J$ in period $t \in T$ from a facility operating at $i \in I$ that is of type $p \in P_i$. Let f_{ipt} be the cost for installing a facility of type $p \in P_i$ at $i \in I$ in period $t \in T$. Additionally, let Q_p be the capacity of a facility of type $p \in P$. Finally, let n_{ip0} denote the number of facilities of type $p \in P_i$ operating at location $i \in I$ before the beginning of the planning horizon (i.e., the problem captures the situation in which the system is not built from scratch but is to be adapted to future

changes in demands). The demand of customer $j \in J$ in period $t \in T$ is again denoted by d_{jt}. Two sets of decision variables were proposed by Shulman (1991): x_{ijpt}, representing the fraction of the demand of customer $j \in J$ in period $t \in T$ that is satisfied from a facility operating at $i \in I$ that is of type $p \in P_i$, and y_{ipt} denoting a binary variable that is equal to 1 if in period $t \in T$ a facility of type $p \in P_i$ is installed at $i \in I$ and 0 otherwise. Assuming that the capacity expansions occur at the beginning of the time periods, the problem can be formulated as follows:

$$\text{Minimize} \quad \sum_{t\in T}\sum_{i\in I}\sum_{p\in P_i} f_{ipt}y_{ipt} + \sum_{t\in T}\sum_{i\in I}\sum_{j\in J}\sum_{p\in P_i} c_{ijpt}x_{ijpt} \tag{11.54}$$

$$\text{subject to} \quad \sum_{i\in I}\sum_{p\in P_i} x_{ijpt} = 1, \quad t \in T,\ j \in J \tag{11.55}$$

$$\sum_{j\in J} d_{jt}x_{ijpt} \le n_{ip0}Q_p + \sum_{\tau=1}^{t} Q_p y_{ip\tau}, \quad t \in T,\ i \in I,\ p \in P_i \tag{11.56}$$

$$x_{ijpt} \ge 0 \quad t \in T,\ i \in I,\ j \in J,\ p \in P_i \tag{11.57}$$

$$y_{ipt} \in \{0,1\}, \quad t \in T,\ i \in I,\ p \in P_i. \tag{11.58}$$

The values c_{ijpt} may include the transportation costs between facilities and customers as well as handling costs at the facilities. Shulman (1991) proposed a Lagrangean relaxation based procedure for obtaining lower and upper bounds for the problem. Constraints (11.55) are dualized. The relaxed problem can be decomposed into $|I|$ problems, each of which to be solved exactly by dynamic programming. However, the complexity of this algorithm is exponential in the number of facilities. Therefore, it can only be used when $|I|$ is small. Nevertheless, for the particular case in which it is not possible to mix different facility types in the same location (i.e., $|P_i| = 1, i \in I$), a polynomial algorithm for the relaxed problem was proposed in the same paper.

The need for more comprehensive multi-period facility location models suited for being applied to real-world problems has led to further important developments. Hinojosa et al. (2000) proposed the first multi-period, multi-echelon, multi-product capacitated discrete facility location problem, setting one important foundation for the strong link that we observe nowadays between multi-period facility location and logistics network design (see Chap. 16). Two facility echelons are considered in that work: plants and warehouses. Location decisions are to be made for both. This paper extends the models proposed by Roodman and Schwarz (1977) by considering more than one facility echelon and multiple commodities. Existing facilities are assumed to be operating before period 1 and can be removed during the planning horizon. Additionally, a set of potential locations for establishing new facilities during the planning horizon is considered. Once removed, a facility cannot be opened again, and once installed, a facility must remain opened until the end of the

planning horizon. Hinojosa et al. (2000) proposed a Lagrangean relaxation based procedure in order to compute lower and upper bounds. The problem would be later extended by Hinojosa et al. (2008) to include inventory decisions. The new model proposed extends the reformulation proposed by Van Roy and Erlenkotter (1982) (i.e., the decision variables represent the changes in the locations—installation of new facilities and removal of existing ones—in the different periods of the planning horizon). A Lagrangean relaxation based procedure was also proposed.

Canel et al. (2001) also investigated a system with two echelons: factories and facilities (e.g., distribution centers). Unlike the problems investigated by Hinojosa et al. (2000, 2008), location decisions are to be made only for the lower echelon. Furthermore, facilities can be opened and closed more than once during the planning horizon. Multiple commodities are considered as well as an important feature much relevant is real logistic systems: the possibility of making direct shipments from the upper echelon to the customers. The authors proposed an exact approach for the problem based on branch-and-bound and dynamic programming.

Jena et al. (2012) investigated a multi-period capacitated facility location problem that in addition to the decisions about where to locate new facilities, consider the possibility of relocating existing facilities or expanding the capacity of existing ones. The authors also consider the possibility of temporarily closing facility parts. The problem arises within the context of logging companies that wish to plan for locating accommodation camps for their workers over some finite planning horizon. The authors proposed several mixed integer linear programming formulations for the problem that they compared in terms of the bounds provided by linear relaxation and tested in instances that use data provided by a real company. They also observed that the problem calls for a very specific cost structure associated with capacity changes. This motivated a more recent work (Jena et al. 2014) in which a general cost structure is associated with capacity changes. A mixed integer linear programming modeling framework was then proposed and shown to generalize two important special cases: facility closing and reopening and capacity expansion and reduction. Alternative formulations were also proposed for these special cases which were compared with the above general modeling framework in terms of the linear relaxation bounds. A combination of the above mentioned cases can also be accommodated in the general modeling framework proposed. In that work, the general model was solved using an off-the-shelf solver. Computational tests were performed using a large set of generated instances.

Albareda-Sambola et al. (2009) extended the model proposed by Roodman and Schwarz (1977) for handling the so-called multi-period incremental service facility location problem. In each time period, a minimum number of facilities is to be established that should be kept operating until the end of the planning horizon. All the customers must start being served in some period and remain served until the end of the planning horizon. The problem is motivated by some practical problems requiring a multi-period plan for progressively extending some service to the population in some region. Accordingly, the service level is progressively increased over time until all customers are being served. A Lagrangean relaxation based procedure was proposed in that paper for obtaining lower and upper bounds. A particular

case of this problem was addressed by Albareda-Sambola et al. (2010), assuming that each customer requires service only in a subset of periods. Additionally, it is possible not to fulfil the request in one or several of those periods but in this case, a penalty cost is paid. Several mathematical programming formulations were proposed for the problem, which were compared computationally.

A multi-period discrete facility location problem was also investigated by Gourdin and Klopfenstein (2008). The problem is motivated within the context of telecommunications network design and consists of planning for the location of modular equipment over a finite planning horizon. Operating capacity constraints are considered for the nodes and for the links. The goal is to progressively expand the capacity of the equipment as well as the capacity of its links to the demand nodes. In that paper, the mathematical programming model initially proposed for the problem was enhanced via polyhedral analysis.

11.6 The Value of the Multi-Period Solution

Multi-period modeling frameworks like those proposed in the previous sections, involve one extra dimension in the decision space: the time. Models tend to be large and thus more difficult to tackle, even for instances of moderate size. Accordingly, one may ask whether it is worth considering this extra dimension. In other words, let us consider a situation in which it is possible to make a static decision even with costs, demands (and possibly other parameters) varying over time. Is it still worth considering a multi-period modeling framework? An answer to this question can be given by the *value of the multi-period solution*, which is a concept first introduced by Alumur et al. (2012) in the context of a multi-period reverse logistics network design problem.

The value of the multi-period solution compares the optimal value of the multi-period problem and the value of a solution found by solving a static counterpart. A static counterpart is a problem that takes into account the information available for the planning horizon and looks for a static (time invariant) solution. Given the optimal solution to a static counterpart, one can consider again the original multi-period problem and set such solution for all periods of the planning horizon. If, by doing so, we obtain a feasible solution to the multi-period problem, the difference between its value and the optimal value of the multi-period problem gives the value of the multi-period solution. In general, several static counterparts can be associated with a multi-period problem. Depending on the one that is considered, a different static solution may be obtained. Accordingly, the value of the multi-period solution may not be unique.

In a multi-period facility location problem, costs, demands, and possibly other parameters are assumed to change over the planning horizon. A static counterpart is a problem that looks for a static location for the facilities, i.e., that can be implemented at the beginning of period 1 and remain unchanged until the end of the planning horizon. One possibility for building a static counterpart is to somehow

aggregate the information available for all periods. For instance, consider time varying demands. If facilities are uncapacitated, then several possibilities emerge for aggregating this information: (1) the demands can be averaged over the planning horizon, or (2) a reference value can be determined (e.g., the maximum value observed throughout the planning horizon). If additional constraints exist (e.g., capacity constraints) then, choosing a reference value may render the resulting static solution infeasible in some periods. In this case, one possibility for building a static counterpart is to define the (time invariant) demand of each customer according to the maximum value observed across all periods. In any case, the adequate aggregation of multi-period data is very much problem-dependent.

In order to clarify the above explanation, we consider problem (11.23)–(11.27), (11.39), and (11.40). A static counterpart can be obtained simply by considering the UFLP with operation costs f_i, $i \in I$, equal to the average of the values f_{it}, $t \in T$ and distribution costs c_{ij}, $i \in I$, $j \in J$, given by the average of the values c_{ijt}, $t \in T$.

When the value of the multi-period solution is obtained by aggregating the data for all periods we refer to it as a *weak* value of the multi-period solution. On the other hand, we obtain a *strong* value of the multi-period solution when no aggregation is performed in the data. This is a possibility in some cases, namely when we can add a set of constraints to the problem stating that some or all decisions are to be the same in all periods of the planning horizon. In the case of a multi-period facility location problem, a static counterpart must define a static location, i.e., a solution in which the location of the facilities is the same for all periods of the planning horizon. Consider, for instance, problem (11.41), (11.42), (11.44), (11.45), and (11.51). A static counterpart yielding a strong value of the multi-period solution is obtained by setting

$$z_{it} = 0 \quad t = 1, \ldots, |T| - 1, \; i \in I^c,$$
$$z_{it} = 0 \quad t = 2, \ldots, |T|, \; i \in I^o.$$

These conditions simply impose that the status of each location does not change during the planning horizon. Therefore, the set of operating facilities will be the same across all periods.

To the best of our knowledge, the only paper within the context of facility location, in which the relevance of using a multi-period modeling framework is measured is the one by Alumur et al. (2012).

11.7 Conclusions

In this chapter, we have presented and discussed several essential aspects related with multi-period facility location problems. The existing literature reveals that the topic has achieved a significant level of maturity. From a modeling point of view, it is now clear how to capture several features of practical relevance and how to tackle

the resulting models. We discussed the weak and strong values of the multi-period solution as measures for the relevance of using a multi-period modeling framework.

In recent years, much work has been developed on facility location problems arising in the context of logistics systems (see, e.g., Melo et al. 2009). As it will be discussed in Chap. 16, an adequate modeling framework can hardly neglect the multi-period nature of such problems. Some papers within this context that somehow extend some multi-period models discussed in the previous sections are those by Melo et al. (2006) and Manzini and Gebennini (2008).

Another aspect of relevance in many applications regards the uncertain nature of the data underlying the problems. Aghezzaf (2005) addressed a multi-period facility location problem under uncertainty. A robust optimization modeling framework was proposed. Recently, multi-period stochastic facility location problems were addressed by Nickel et al. (2012) and Albareda-Sambola et al. (2013). These works show that capturing uncertainty in multi-period facility location problems is still a challenge.

Another challenging area in multi-period facility location concerns the location of public facilities. One first work in this direction is due to Antunes and Peeters (2001). Although static models for public facilities location have attracted much attention in the past, the same does not happen with multi-period problems.

One class of problems which is still much unexplored, regards multi-criteria, multi-period facility location problems. To the best of our knowledge only a few papers exist within this context. Dias et al. (2008) proposed a memetic algorithm for multi-period problems when it is possible to install and remove a facility more than once during the planning horizon. Hugo and Pistikopoulos (2005) and Melachrinoudis and Min (2007) study multi-criteria, multi-period facility location problems in the context of logistics network design.

Most of the contents in this chapter are a basis for addressing more complex real-world problems. In fact, several models presented in the previous sections have already been extended to problems arising in other areas (see, for instance, Chaps. 12, 15 and 16). Nevertheless, some challenges still exist. The research done so far is scarce when it comes to some classes of multi-period facility location problems, such as those just mentioned above. These are existing research directions worth exploring in order to broaden the scope and knowledge on multi-period facility location, making the topic an even stronger basis for being applied to real-world systems.

References

Aghezzaf E (2005) Capacity planning and warehouse location in supply chains with uncertain demands. J Oper Res Soc 56:453–462

Albareda-Sambola M, Fernández E, Hinojosa Y, Puerto J (2009) The multi-period incremental service facility location problem. Comput Oper Res 36:1356–1375

Albareda-Sambola M, Alonso-Ayuso A, Escudero LF, Fernández E, Hinojosa Y, Pizarro-Romero C (2010) A computational comparison of several formulations for the multi-period incremental service facility location problem. TOP 18:62–80

Albareda-Sambola M, Alonso-Ayuso A, Escudero LF, Fernández E, Pizarro C (2013) Fix-and-relax coordination for a multi-period location-allocation problem under uncertainty. Comput Oper Res 40:2878–2892

Alumur SA, Nickel S, Saldanha da Gama F, Verter V (2012) Multi-period reverse logistics network design. Eur J Oper Res 220:67–78

Antunes A, Peeters D (2001) On solving complex multi-period location models using simulated annealing. Eur J Oper Res 130:190–201

Arabani AB, Zanjirani Farahani R (2012) Facility location dynamics: an overview of classifications and applications. Comput Ind Eng 62:408–420

Ballou RH (1968) Dynamic warehouse location analysis. J Mark Res 5:271–276

Canel C, Khumawala BM (1997) Multi-period international facilities location: an algorithm and application. Int J Prod Res 35:1891–1910

Canel C, Khumawala BM (2001) International facilities location: a heuristic procedure for the dynamic uncapacitated problem. Int J Prod Res 39:3975–4000

Canel C, Khumawala BM, Law J, Loh A (2001) An algorithm for the capacitated, multi-commodity multi-period facility location problem. Comput Oper Res 28:411–427

Cavalier TM, Sherali HD (1985) Sequential location-allocation problems on chains and trees with probabilistic link demands. Math Program 32:249–277

Chand S (1988) Decision/forecast horizon results for a single facility dynamic location/relocation problem. Oper Res Lett 7:247–251

Chardaire P, Sutter A, Costa M-C (1996) Solving the dynamic facility location problem. Networks 28:117–124

Daskin MS, Hopp WJ, Medina B (1992) Forecast horizons and dynamic facility location planning. Ann Oper Res 40:125–151

Dias J, Captivo ME, Clímaco J (2007) Efficient primal-dual heuristic for a dynamic location problem. Comput Oper Res 34:1800–1823

Dias J, Captivo ME, Clímaco J (2008) A memetic algorithm for multi-objective dynamic location problem. J Global Optim 42:221–253

Drezner Z (1995) Dynamic facility location: the progressive p-median problem. Locat Sci 3:1–7

Drezner Z, Wesolowsky GO (1991) Facility location when demand is time dependent. Nav Res Logist 38:763–777

Erlenkotter D (1981) A comparative study of approaches to dynamic location problems. Eur J Oper Res 6:133–143

Galvão RD, Santibañez-Gonzalez ER (1992) A Lagrangean heuristic for the p_k-median dynamic location problem. Eur J Oper Res 58:250–262

Gourdin É, Klopfenstein O (2008) Multi-period capacitated location with modular equipments. Comput Oper Res 35:661–682

Hakimi SL (1964) Optimum location of switching centers and the absolute centers and medians of a graph. Oper Res 12:450–459

Hakimi SL (1965) Optimum distribution of switching centers in a communication network and some related graph theoretic problems. Oper Res 13:462–475

Hakimi SL, Labbé M, Schmeichel EF (1999) Locations on time-varying networks. Networks 34:250–257

Hamacher HW, Nickel S (1998) Classification of location problems. Locat Sci 6:229–242

Hinojosa Y, Puerto J, Fernández FR (2000) A multiperiod two-echelon multicommodity capacitated plant location problem. Eur J Oper Res 123:271–291

Hinojosa Y, Kalcsics J, Nickel S, Puerto J, Velten S (2008) Dynamic supply chain design with inventory. Comput Oper Res 35:373–391

Hugo A, Pistikopoulos EN (2005) Environmentally conscious long-range planning and design of supply chain networks. J Clean Prod 13:1471–1491

Jena S, Cordeau J-F, Gendron B (2012) Modeling and solving a logging camp location problem. Ann Oper Res. doi: 10.1007/s10479-012-1278-z

Jena S, Cordeau J-F, Gendron B (2014) Dynamic facility location with generalized modular capacity. Transp Sci. doi: 10.1287/trsc.2014.0575

Khumawala BM (1972) An efficient branch and bound algorithm for the warehouse location problem. Manag Sci 18:718–731
Lee S-B, Luss H (1987) Multifacility-type capacity expansion planning: algorithms and complexities. Oper Res 35:249–253
Manzini R, Gebennini E (2008) Optimization models for the dynamic facility location and allocation problem. Int J Prod Res 46:2061–2086
Melachrinoudis E, Min H (2007) Redesigning a warehouse network. Eur J Oper Res 176:210–229
Melo MT, Nickel S, Saldanha da Gama F (2006) Dynamic multi-commodity capacitated facility location: a mathematical modeling framework for strategic supply chain planning. Comput Oper Res 33:181–208
Melo MT, Nickel S, Saldanha da Gama F (2009) Facility location and supply chain management. Eur J Oper Res 196:401–412
Mesa J (1991) Multiperiod medians on networks. RAIRO - Rech Oper 25:87–95
Nickel S, Saldanha da Gama F, Ziegler H-P (2012) A multi-stage stochastic supply network design problem with financial decisions and risk management. Omega 40:511–524
Orda A, Rom R (1991) Location of central nodes in time varying computer networks. Oper Res Lett 10:143–152
Puerto J, Rodríguez-Chía A (1999) Location of a moving service facility. Math Method Oper Res 49:373–393
Roodman GM, Schwarz LB (1975) Optimal and heuristic facility phase-out strategies. AIIE Trans 7:177–184
Roodman GM, Schwarz LB (1977) Extensions of the multi-period facility phase-out model: new procedures and application to a phase-in/phase-out problem. AIIE Trans 9:103–107
Saldanha da Gama F (2002) Modelos e algoritmos para o problema de localização dinâmina (in Portuguese). Ph.D. thesis, Faculty of science, University of Lisbon, Portugal
Saldanha da Gama F, Captivo ME (1998) A heuristic approach for the discrete dynamic location problem. Locat Sci 6:211–223
Scott AJ (1971) Dynamic location-allocation systems: some basic planning strategies. Environ Plan 3:73–82
Shulman A (1991) An algorithm for solving dynamic capacitated plant location problems with discrete expansion sizes. Oper Res 39:423–436
Sweeney D, Tatham RL (1976) An improved long-run model for multiple warehouse location. Manag Sci 22:748–758
Van Roy T, Erlenkotter D (1982) A dual-based procedure for dynamic facility location. Manag Sci 28:1091–1105
Warszawski A (1973) Multi-dimensional location problems. Oper Res Q 24:165–179
Wesolowsky GO (1973) Dynamic facility location. Manag Sci 19:1241–1248
Wesolowsky GO, Truscott WG (1975) The multi-period location-allocation problem with relocation of facilities. Manag Sci 22:57–65
Zanjirani Farahani R, Drezner Z, Asgari N (2009) Single facility location and relocation problem with time dependent weights and discrete planning horizon. Ann Oper Res 167:353–368

Chapter 12
Hub Location Problems

Ivan Contreras

Abstract *Hub Location Problems* (HLPs) lie at the heart of network design planning in transportation and telecommunication systems. They are a challenging class of optimization problems that focus on the location of hub facilities and on the design of hub networks. This chapter overviews the key distinguishing features, assumptions and properties commonly considered in HLPs. We highlight the role location and network design decisions play in the formulation and solution of HLPs. We also provide a concise overview of the main developments and most recent trends in hub location research. We cover various topics such as hub network topologies, flow dependent discounted costs, capacitated models, uncertainty, dynamic and multi-modal models, and competition and collaboration. We also include a summary of the most successful integer programming formulations and efficient algorithms that have been recently developed for the solution of HLPs.

Keywords Hub location • Hub networks • Integer programming

12.1 Introduction

Transportation, telecommunications and computer networks frequently employ hub-and-spoke architectures to efficiently route flows between many origins and destinations. Their key feature lies in the use of transshipment, consolidation, or sorting points, called *hub facilities*, to connect a large number of origin/destination (O/D) pairs by using a small number of links. Flows having the same origin but different destinations are consolidated when routed to the hubs and then, combined with other flows having different origins but the same destination. This helps reduce setup costs, centralize commodity handling and sorting operations, and achieve economies of scale on routing costs through the consolidation of flows. Broadly speaking, *Hub Location Problems* (HLPs) consist of locating hub facilities and of designing hub networks so as to optimize a cost-based (or service-based) objective.

I. Contreras (✉)
Concordia University and Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT), Montreal, QC, Canada H3G 1M8
e-mail: icontrer@encs.concordia.ca

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_12

HLPs constitute a challenging class of NP-hard problems involving joint location and network design decisions. Their main difficulty stems from the inherent interrelation between two levels of the decision process. The first level considers the selection of a set of nodes to locate hub facilities, whereas the second level deals with the design of the hub network, by selecting the links to connect origins, destinations and hubs, as well as the routing of flows through the network.

HLPs lie at the heart of network design planning in transportation and telecommunication systems. Application areas of HLPs in transportation are abundant. These include express package delivery, air freight and passenger travel, postal delivery, trucking, and rapid transit systems. Demand corresponds to commodities (i.e. express packages, passengers, mail, goods) carried by vehicles (i.e. trucks, trains, airplanes, vessels) moved on physical networks such as roads and railways or through the air of water. Hub facilities correspond to sorting centers or transportation terminals in which one or more transportation modes interact. Consolidation of flows at hubs enable economies of scale on the transportation costs, not only on the routing of flows between hubs, but also between O/D nodes and hubs.

Applications of HLPS in telecommunications arise in the design of various distributed data networks, where demand corresponds to electronic data that are routed over a variety of physical links (i.e. fiber optic links and co-axial cables) or through the air (i.e. satellite channels and microwave links). Hub facilities are hardware such as switches, concentrators, multiplexors, and routers. Economies of scale in data transmission and network utilization, in combination with large set-up costs for hub facilities and communication links, motivate the use of hub-and-spoke architectures.

The study of HLPs began with the pioneering work of O'Kelly (1986a), for continuous models, and O'Kelly (1986b, 1987), for discrete models, and has since evolved into a rich research area. Over the last three decades hub location has been studied by researchers around the globe from different disciplines such as location science, geography, regional science, network optimization, transportation, telecommunications, and computer science. There exist several reviews and surveys on HLPs, each one of them focusing on different aspects of these problems. The early reviews dealing with HLPs, by O'Kelly and Miller (1994) and Campbell (1994a), contain classification schemes for fundamental models and for the topological structures applicable to hub networks. Klincewicz (1998) concentrate on the design of hub networks in the context of telecommunication networks, and Bryan and O'Kelly (1999) present a survey focused on air transportation networks. Campbell et al. (2001) wrote a comprehensive survey of HLPs in which the location of hubs is the key decision. Alumur and Kara (2008) provide a classification and review of the growing literature on network hub location models before 2008. Campbell and O'Kelly (2012) provide an insight into early motivations for analyzing HLPs and highlight recent research directions. Zanjirani Farahani et al. (2013) review solution methods and applications for several classes of HLPs.

This chapter focuses on the role location and network design decisions play in the formulation and solution of HLPs. It overviews features and assumptions commonly considered in discrete HLPs, providing insights on their modeling implications

and consequences. We point out how these assumptions simplify network design decisions, creating a first generation of HLPs that focuses mostly on the location and allocation decisions. We also show how network decisions become more involved when relaxing some of these assumptions.

We start with an introduction to the fundamentals of HLPs, including their distinguishing features, assumptions, properties, as well as commonly used objectives. A review of the most interesting and useful *Mixed Integer Programming* (MIP) formulations for fundamental HLPs considering cost-based objectives is then presented. We also highlight some of the main developments and most recent trends in hub location. We would like to clarify that, due to space limitations, this is not intended to be a comprehensive survey of all diverse topics associated with hub location research, but rather our personal view-point on some of the most interesting research on this filed. In particular, we include hub network topologies, flow dependent discounted cost models, capacitated models, models dealing with uncertainty, dynamic and multi-modal models, and competition and collaboration. A summary of successful integer programming methods that have given rise to efficient approximate and exact solution algorithms for solving HLPs is also presented.

This chapter does not cover continuous HLPs or models in which locational decisions are not present. The reader is referred to O'Kelly (1986a), O'Kelly and Miller (1991), Aykin (1988), Campbell (1990, 2013), Saberi and Mahmassani (2013), and references therein for continuous variants of HLPs, and to Klincewicz (1998), Gendron et al. (1999), and Wieberneit (2008) for hub-and-spoke network design models in which the set of hub facilities is given a priori. The reader is also referred to Contreras and Fernández (2012) for a survey of other general network design problems that also combine location and network design decisions.

12.2 Fundamentals

HLPs are closely related to classical *Facility Location Problems* (FLPs). As a result, for several classical facility location problems such as p-median, uncapacitated facility location, p-center, and covering problems, analogous HLPs have been studied: p-hub median, uncapacitated hub location, p-hub center, and hub covering problems. Due to their multiple applications, inside these classes of HLPs there exist several variants that differ with respect to a number of assumptions like the topological structure, the allocation pattern of O/D nodes to hubs, and capacity constraints on the hub network, among others.

The key difference between FLPs and HLPs relies on the type of service demand required by the users and on the function the facilities provide. In the case of FLPs, service is given at (from) the facilities and flows thus originate at demand nodes (facilities) and their destination are the facilities (demand nodes). Network design and routing decisions are usually determined by the assignment pattern of demand nodes to their allocated facilities. In HLPs, service demand is between

O/D nodes and hub facilities are intermediate nodes in the O/D paths which act as transshipment and consolidation points. When a hub serves as transshipment (switching or sorting) point, it allows flows to be processed and redirected to other hubs or O/D nodes with many fewer links than would be needed with direct connections. As a consolidation (concentration or breakbulk) point, a hub allows flows to be aggregated and disaggregated, creating economies of scale in the transportation (or communication) cost between hubs and between O/D nodes and hubs. The interaction of hub facilities and O/D nodes increases the complexity of network design and routing decisions, since these are not necessarily determined by the assignment pattern of O/D nodes to hubs.

Another difference between FLPs and HLPs is that, when dealing with uncapacitated hub location models, a single assignment pattern of non-hub nodes to hubs is not necessarily an optimal allocation strategy. In most uncapacitated FLPs, once the facility locations are known the flow (or allocation) cost is minimized by assigning each demand node to its closest (or least costly) open facility. In the case of HLPs, once the hub locations are known, the flow cost is minimized by finding the shortest path on the network induced by the selected hubs for each O/D pair, resulting in a multiple allocation pattern of O/D nodes to hubs. For this reason, both single and multiple assignments versions of HLPs exist. In a hub location problem with single assignments, O/D nodes must be assigned to exactly one hub facility. All demand flows with the same origin (or destination) are thus routed via the same hub. In a hub location problem with multiple assignments, each O/D node can be allocated to more than one hub facility. Multiple assignment patterns simplify the routing decisions and provide greater flexibility on hub networks, allowing lower flow cost solutions. However, they can considerably increase the network design cost as a larger number of links must be activated on the hub network.

12.2.1 Features, Assumptions and Properties

The key distinguishing features of HLPs can be summarized as follows: *(i)* service demand is associated with flows between O/D pairs, *(ii)* hub facilities are intermediate nodes in the O/D paths which act as transshipment or consolidation points, *(iii)* there is a benefit (or requirement) of routing flows via hubs, *(iv)* there is a cost-based (or service-based) objective that depends on the design of the hub network (location of hubs and selection of links) and the routing of flows.

We can provide a description of a generic hub location problem as follows. Consider a complete graph $G = (N, E)$, where N is the set of nodes representing the origins and destinations of flows, and E is the set of edges. Let N be the set of potential hub locations as well. For each node pair (i, j), let $W_{ij} \geq 0$ and $d_{ij} \geq 0$ denote the amount of flow to be routed and the distance, respectively, from the origin $i \in N$ to the destination $j \in N$. For each node $i \in N$, f_i is the fixed set-up cost for locating a hub, whereas for each $e \in E$, g_e denotes the fixed set-up cost for locating a hub arc. A hub arc $e = (i, j) \in E$ connects two different hub nodes

i and j and has a per unit flow cost of αd_{ij}. The parameter α ($0 \leq \alpha \leq 1$) is used as a discount factor to provide reduced unit flow costs on hub arcs to reflect economies of scale resulting from consolidation of flows between hubs. The per unit flow cost between O/D pairs is given by the length of the path between the origin and destination nodes in the solution network. Each O/D path has a *collection* leg from the origin node to the first hub, possibly a *transfer* leg between the first and the last hubs, and a *distribution* leg from the last hub to the destination node. A generic hub location problem consists of locating a set of hub facilities and a set of hub arcs, and of determining the routing of flows through the hub network, with the objective of minimizing the total set-up and flow cost.

Most of the hub location literature has focused on *Hub Node Location Problems* (HNLPs), which consider the location of a set of hub facilities and the assignment of O/D nodes to these facilities. Arc selection and routing decisions are usually determined by the assumptions made on the cost structure and the assignment pattern. The network induced by the solution of a HNLP consists of three types of arcs: *(i) hub arcs* connecting two hubs, *(ii) access arcs* connecting non-hub nodes and hubs, and *(iii) direct arcs* connecting two non-hub nodes. A more general class of hub location models, known as *Hub Arc Location Problems* (HALPs), have received less attention in the literature. HALPs consider the location of a set of hub arcs, that induce a set of hub nodes, and the assignment of O/D nodes to these hub arcs. In HALPs, the possibility of connecting two hub nodes with a fourth type of arc arises. A *bridge arc* is an arc that connects two different hub nodes, without benefiting from the reduced unit flow cost of a hub arc. HNLPs can be seen as particular cases of HALPs in which additional conditions are imposed.

There are four common assumptions underlying most HLPs:

1. Flows have to be routed via a set of hubs.
2. Access arcs and bridge arcs have no set-up cost.
3. The discount factor α is the same for all hub arcs and does not depend on the amount of flow that is actually routed on each hub arc.
4. Distances d_{ij} satisfy the triangle inequality.

A consequence of Assumption 1 is that direct connections between O/D nodes which are not hubs are not allowed and thus, O/D paths must include at least one hub node. In most HNLPs an additional fifth assumption stating that the set-up cost of hub arcs is equal to zero (i.e., $g_e = 0$ for each $e \in E$) is also considered. This allows hubs to be interconnected at no extra cost and, together with Assumptions 3 and 4, an important resulting property in solution networks of HNLPs is that the set of hub arcs define a complete subgraph on the set of hub nodes (i.e. hubs are fully interconnected). As a consequence, hub arc selection decisions become trivial once the location of hub nodes is known. Another important property, obtained when combining all assumptions, is that paths between O/D pairs will contain at least one and at most two hubs. However, it is important to note that whenever Assumption 4 is not satisfied, paths may contain more than two hubs and more than one hub arc.

The above properties simplify the network design decisions and characterize the structure of O/D paths. In HNLPs, all O/D paths include either a single hub

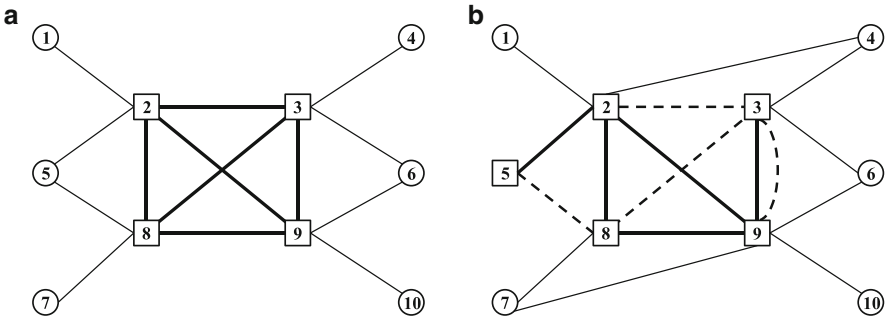


Fig. 12.1 Solution network of a hub node location problem (**a**) and a hub arc location problem (**b**)

node and no hub arc, or two hub nodes and a single hub arc. Moreover, because of Assumptions 2 and 4, each collection and distribution leg, if present, contains only one access arc. O/D paths are thus of the form (i, k, m, j), where $(k, m) \in N \times N$ is the ordered pair of hubs to which i and j are allocated, respectively. Note that these paths contain one, two or at most three arcs, depending on the number of visited hubs and on the function of origins and destinations (i.e. hub or non-hub nodes). For each O/D pair, the flow cost of routing W_{ij} along the path (i, k, m, j) is then given by $F_{ijkm} = W_{ij}\left(\chi d_{ik} + \alpha d_{km} + \delta d_{mj}\right)$, where χ, α, and δ represent the collection, transfer and distribution costs along the path. To reflect economies of scale between hubs, we assume that $\tau < \chi$ and $\tau < \delta$.

Figure 12.1a shows an example of a solution network of a HNLP in which different structures on O/D paths arise (squares represent hub nodes and circles represent non-hub nodes). The path $(1, 2, 9, 10)$ is a two-hub path formed by the access arcs $(1, 2)$, $(9, 10)$ and the hub arc $(2, 9)$. The path $(2, 2, 9, 6)$ is also a two-hub path but containing only the access arc $(9, 6)$ and the hub arc $(2, 9)$. The path $(3, 3, 9, 9)$ is yet another two-hub path formed only by the hub arc $(3, 9)$. The path $(1, 2, 2, 8)$ is a one-hub path containing only the access arcs $(1, 2)$ and $(2, 8)$. The path $(7, 8, 8, 8)$ is also a one-hub path containing the single access arc $(7, 8)$.

In HALPs, hubs are not necessarily fully interconnected due to the set up cost on the hub arcs or because additional conditions on the network topology are imposed. This causes O/D paths to become more involved, since they may use more than three arcs and visit more than two hub nodes. Similar to HNLPs, because of Assumptions 2 and 4, each collection and distribution leg, if present, employs either one access arc or one bridge arc. However, the transfer leg can now use several bridge and hub arcs, depending on the particular assumptions considered on the structure of O/D paths.

To simplify the routing decisions in HALPs, an additional assumption stating that O/D paths contain at most one hub arc can be considered. This limits paths to have at most three arcs, being the first and last ones either access or bridge arcs and the intermediate arc, if it exists, a hub arc. As mentioned in Campbell et al. (2005a), this assumption is used to increase service level in classical HLPs

and is also consistent with practice. In air transportation, for example, it ensures that a passenger will never have to change flights more than twice. In ground transportation, it is convenient to restrict the number of hub facilities that each route has to pass through so as to reduce handling and congestion at hubs and to provide a form of performance guarantee. O/D paths are once more of the form (i, k, m, j), and thus, defining their flow cost as F_{ijkm}.

Figure 12.1b shows an example of a solution network of a HALP in which different structures on O/D paths arise (dashed lines represent bridge arcs). The path $(5, 8, 2, 3)$ is a four-hub path formed by the bridge arcs $(5, 8)$, $(2, 3)$ and the hub arc $(8, 2)$. The path $(5, 8, 9, 10)$ is a three-hub path containing the bridge arc $(5, 8)$, the hub arc $(8, 9)$ and the access arc $(9, 10)$.

12.2.2 Supermodular Properties

We next show how a general class of HLPs can be stated as the minimization of a real-valued supermodular set function. This fundamental property, which is also known for other types of classical facility location problems (p-median, uncapacitated and capacitated facility location), can be exploited to develop mathematical formulations and solution algorithms with worst case bounds.

This class of HLPs, referred to as *Supermodular Hub Location Problems* (SHLPs), considers Assumptions 1–4 and the additional assumption that limits O/D paths to contain at most one hub arc. SHLPs consist of locating a set of at most q hub arcs ($q \geq 1$), that induce a set of at most p hub nodes ($p \geq 2$), and of determining the routing of commodity flows through the hub network, with the objective of minimizing the total set-up and flow cost. We can state SHLPs as the following combinatorial problem. Let $U = N \cup E$ be a finite set containing both the set of nodes N and the set of edges E of G. For each non-empty subset $(S, R) \subseteq U$, where $S \subseteq E$ and $R \subseteq N$, define

$$c(S,R) = \sum_{i \in R} c_i; \quad g(S,R) = \sum_{e \in S} g_e; \quad h(S,R) = \sum_{i,j \in N} h^{ij}(S) = \sum_{i,j \in N} \min_{(k,m) \in S} F_{ijkm},$$

and

$$f(S,R) = c(S,R) + g(S,R) + h(S,R) = \sum_{i \in R} c_i + \sum_{e \in S} g_e + \sum_{i,j \in N} \min_{(k,m) \in S} F_{ijkm}, \tag{12.1}$$

and $f(\emptyset) = 0$. For nonempty sets of hub nodes $R \subseteq N$ and hub arcs $S \subseteq E$, $c(S, R)$ is the total set-up costs for setting hub nodes, $g(S, R)$ is the total set-up cost of the hub arcs, and $h(S, R)$ is the total cost for routing the flows when the set of hub arcs S is chosen. Thus, $f(S, R)$ is the objective function value associated with the set of hub nodes R and the set of hub arcs S. Therefore, SHLPs can be

stated as to find a set of arcs $S \subseteq E$ of cardinality at most q $(q \leq |E|)$ and R of cardinality at most p $(p \leq |N|)$ such that $f(S, R)$ is minimum, i.e.,

$$\min_{(S,R)\subseteq U} \{f(S, R) : |S| \leq q,\ |R| \leq p,\ N(S) = R\}, \tag{12.2}$$

where $N(S) = \{i \in N : (i, j) \in S \text{ or } (j, i) \in S\}$ is the set of nodes incident with some edge in S. In order to deal only with feasible problems, we assume that $p \geq \lceil \frac{q}{2} \rceil$. When $p \geq \min\{|N|, 2q\}$ the maximum cardinality constraint on the number of hub nodes becomes redundant. Similarly, if $q \geq \min\{|E|, \binom{p}{2}\}$ the maximum cardinality constraint on the number of hub arcs becomes redundant. A fundamental property of f is that, for $(S, R) \subset (T, Q)$ and $e \in E \setminus T$, adding e to T will decrease f by no more than by adding e to S. A real-valued set function with such property is called *supermodular set function*.

Proposition 12.1

a. $h(S, R) = \sum_{i,j \in N} h^{ij}(S, R)$ *is supermodular and nonincreasing.*

b. $f(S, R) = c(S, R) + g(S, R) + h(S, R)$ *is supermodular.*

Problem (12.2) can thus be stated as the minimization of a supermodular set function, which is known to be in the class of NP-hard problems. We use SHLP to describe any problem that can be formulated as (12.2). SHLPs are a quite general class of HLPs and include several special cases which are of particular interest such as p-hub median, uncapacitated hub location, and q-hub arc location. Other classical facility location problems, such as the p-median or the uncapacitated facility location problem, are also relevant special cases of SHLPs. However, we note that not every HLP can be stated as problem (12.2). For instance, when a single assignment pattern is imposed the flow cost associated with a given set of hub arcs S is no longer $h(S, R)$, since all flow with the same origin (destination) must be routed through the same collection (transfer) leg. That is, HLPs with single assignments cannot be formulated as SHLPs. Moreover, even if multiple allocation is allowed, the addition of capacity constraints also preclude the supermodularity property when commodities cannot be splitted.

12.2.3 Objectives

Most of the hub location research has focused on HLPs that consider either a cost-based or a service-based objective. Transportation applications tend to focus on the flow transportation costs and travel times, whereas telecommunication applications focus more on the set-up costs of the hub network. Analogously to facility location, HLPs can be classified based on the type of objective they use.

- *p-Hub Median Problems* assume that the number of hubs to locate is given as an input of the problem. They consist of locating a set of p hub facilities with the objective of minimizing the total flow cost for routing the flows through the hub network.
- *Hub Location Problems* consider that the number of hubs to locate is not known a priori, but a fixed set-up cost for each hub is considered. The objective is to minimize the sum of hub fixed costs and of demand flow costs over the hub network.
- *p-Hub Center Problems* are minmax problems that focus on the minimization of a maximum service or cost measure between O/D pairs. Some of these measures are: *(i)* the maximum flow cost (or travel time) of all O/D pairs, *(ii)* the maximum flow cost (or travel time) of all arcs of the hub network, and *(iii)* the maximum flow cost (or travel time) associated with an access arc.
- *Hub Covering Problems* impose a maximum threshold value on the service level (travel time) and focus on the minimization of the set-up cost of the hub network. They assume demand is covered if both origin and destination nodes are within a specified distance of a hub node. They differ on their considered coverage criteria. An O/D pair (i, j) is covered by hubs k and m if: *(i)* the length of the path (i, k, m, j) is within a specified value, *(ii)* the length of each arc in the path (i, k, m, j) does not exceed a specified value, or *(iii)* each of the access arcs meet different specified values.

Both single and multiple assignment models, as well as uncapacitated and capacitated models have been considered in the literature for most of these classical objectives. We refer to Campbell (1994a), Campbell et al. (2001), and Alumur and Kara (2008) for a detailed overview of these models.

HLPs considering more complex classes of objective functions have also been studied. Costa et al. (2008) and Köksalan and Soylu (2010) consider HLPS with multiple objectives. Puerto et al. (2011) introduce a general class of HLPs that consider an ordered median function (see Chap. 10) for which the above mentioned objectives (and others) are particular cases. O'Kelly (2012) considers objectives related to the fuel burn and environmental impact in airline hub networks. Campbell and O'Kelly (2012) review some recent HLPs that integrate both cost and service objectives.

12.3 Formulating Hub Location Problems

One of the major modeling challenges in HLPs is that knowing the hub network structure is not necessarily sufficient to evaluate the objective function. Formulations must be able to model the path used for routing each flow to determine the flow cost. Significant progress has been made toward the development of *Mixed Integer Programming* (MIP) formulations for fundamental HLPs. These exploit the structure of the solution network obtained when considering the modeling

assumptions presented in Sect. 12.2.1. We next introduce the most important families of MIP formulations for both single and multiple assignment variants of p-hub median and hub location problems. These have been successfully used in combination with sophisticated solution algorithms to obtain optimal solutions for large-scale instances. They have also been extended to model more complex variants of HLPs including additional features of real applications. We refer to Campbell et al. (2007), Alumur and Kara (2009), Wagner (2008a), Ernst et al. (2009), Yaman and Elloumi (2012), Hwang and Lee (2013), and Lowe and Sim (2013) for formulations of p-hub center and hub covering problems.

12.3.1 Single Assignments

A *natural* way to formulate HLPs with single assignments is to consider them as facility location problems with additional quadratic costs associated with the interaction of hub facilities. For each pair $i, k \in N$, we define location/allocation variables z_{ik}, equal to one if node i is assigned to hub k and zero otherwise. When $i = k$, variable z_{kk} represents the establishment or not of a hub at node k. The *Uncapacitated Hub Location Problem with Single Assignments* (UHLPSA) can be stated as the following quadratic mixed integer program (O'Kelly 1987):

$$\text{minimize} \quad \sum_{k \in N} f_k z_{kk} + \sum_{i,k \in N} (\chi O_i + \delta D_i)\, d_{ik} z_{ik} + \sum_{i,j,k,m \in N} \alpha W_{ij} d_{km} z_{ik} z_{jm} \tag{12.3}$$

$$\text{subject to} \quad \sum_{k \in N} z_{ik} = 1 \qquad i \in N \tag{12.4}$$

$$z_{ik} \leq z_{kk} \qquad i, k \in N \tag{12.5}$$

$$z_{ik} \in \{0, 1\} \qquad i, k \in N, \tag{12.6}$$

where $O_i = \sum_{j \in N} W_{ij}$ and $O_i = \sum_{j \in N} W_{ji}$. The first term of the objective function represents the total set-up cost of the hub facilities, whereas the second and third term are the flow cost on the access and hub arcs, respectively. Constraints (12.4) guarantee that every O/D node is assigned to exactly one hub, whereas constraints (12.5) impose that they can only be assigned to open hubs. Note that constraints (12.4)–(12.6) define the set of feasible solutions of the *Uncapacitated Facility Location Problem* (see Chap. 3). However, objective (12.3) contains an additional quadratic term associated with the inter-hub flow cost. Several linearized formulations have been proposed to overcome this added difficulty of UHLPSAs.

An important family of formulations, referred to as *path-based formulations*, use decision variables to characterize O/D paths visiting either one or two hub nodes. We introduce binary routing variables x_{ijkm}, $i, j, k, m \in N$, equal to 1 if and only if

the flow originated at i and destination j transits via a first hub node k and a second hub node m. The UHLPSA can be stated as follows (Skorin-Kapov et al. 1997):

$$\begin{aligned}
\text{minimize} \quad & \sum_{k\in N} f_k z_{kk} + \sum_{i,j,k,m\in N} F_{ijkm} x_{ijkm} \\
\text{subject to} \quad & (12.4)\text{–}(12.6) \\
& \sum_{m\in N} x_{ijkm} = z_{ik} \qquad i,j,k \in N \qquad (12.7) \\
& \sum_{k\in N} x_{ijkm} = z_{jm} \qquad i,j,m \in N \qquad (12.8) \\
& x_{ijkm} \geq 0 \qquad i,j,k,m \in N. \qquad (12.9)
\end{aligned}$$

Constraints (12.7) state that if node i is assigned to hub k then all the flow from node i to any other node j must go through some other hub m. Constraints (12.8) have a similar interpretation relative to the flow arriving to a node j assigned to hub m from some node i. There is no need to explicitly state the integrality on the x_{ijkm} variables because there always exists an optimal solution of (12.4)–(12.8) in which all x_{ijkm} variables are integer. One of the attractive features of this formulation is that it usually provides tight *Linear Programming* (LP) relaxation bounds, at the expense of requiring $O(n^4)$ variables and $O(n^3)$ constraints. Saito et al. (2009) study the polyhedral structure of the quadratic semi-assignment polytope, a relaxation of this formulation, and provides strong valid inequalities to further improve its LP bound.

It is possible to project out the path-based variables x_{ijkm} to obtain a formulation with fewer variables (see Labbé and Yaman 2004; Labbé et al. 2005). We define continuous variables $y_{km}, k,m \in N$, equal to the amount of flow routed on hub arc (k,m). The UHLPSA can be formulated as:

$$\begin{aligned}
\text{minimize} \quad & \sum_{k\in N} f_k Z_k + \sum_{i,k\in N} (\chi O_i + \delta D_i) d_{ik} z_{ik} + \sum_{k,m\in N} \alpha d_{km} y_{km} \\
\text{subject to} \quad & (12.4)\text{–}(12.6) \\
& y_{km} \geq \sum_{(i,j)\in K} W_{ij} \left(z_{ik} + z_{jm} - 1\right) \qquad k,m \in N, K \subseteq N \times N \qquad (12.10) \\
& y_{km} \geq 0 \qquad k,m \in N. \qquad (12.11)
\end{aligned}$$

For each arc (k,m), constraints (12.10) and (12.11) imply

$$y_{km} = \max_{K\subseteq N\times N} \sum_{(i,j)\in K} W_{ij} \left(z_{ik} + z_{jm} - 1\right) = \sum_{(i,j)\in K_{km}} W_{ij} \left(z_{ik} + z_{jm} - 1\right),$$

where K_{km} is the set of all demands which are routed on hub arc (k,m). This formulation contains only $O(n^2)$ variables but an exponential number of constraints. Labbé and Yaman (2004) show that constraints (12.10) are a particular case of

a more general class of facet defining inequalities which can be separated in polynomial time.

Another important family of formulations, referred to as *flow-based formulations*, use continuous variables to compute the amount of flow routed on a particular arc originated at a given node. In the case of single assignments, we only need to use one set of flow variables associated with the hub arcs. We thus define continuous variables Y_{ikm}, $i, j, k \in N$, equal to the amount of flow originated at node i and passing through hub arc (k, m). The UHLPSA can be formulated as follows (Ernst and Krishnamoorthy 1996):

$$\text{minimize} \quad \sum_{k \in N} f_k z_{kk} + \sum_{i,k \in N} (\chi O_i + \delta D_i) d_{ik} z_{ik} + \sum_{i,k,m \in N} \alpha d_{km} Y_{ikm}$$

$$\text{subject to} \quad (12.4)\text{–}(12.6)$$

$$\sum_{j \in N} W_{ij} z_{jk} + \sum_{m \in N} Y_{ikm} = \sum_{m \in N} Y_{imk} + O_i z_{ik} \qquad i, k \in N \tag{12.12}$$

$$Y_{ikm} \geq 0 \qquad i, k, m \in N. \tag{12.13}$$

Constraints (12.12) are the well-known flow conservation constraints for each O/D node i at each (potential) hub node k, where the supply and demand at each node is determined by the allocation pattern. The above formulation contains $O(n^3)$ variables and $O(n^2)$ constraints and thus, fewer variables and constraints as compared with the path-base formulation. However, it usually produces weaker LP bounds. Contreras et al. (2010, 2013) present some families of extended cut-set inequalities that can help improve the LP bounds.

12.3.2 Multiple Assignments

Given that in HLPs with multiple assignments O/D nodes can be connected to more than one hub facility, we can exploit the properties on the structure of O/D paths to obtain path-based formulations with less variables than the ones required for single assignment models. In particular, it is known that every flow uses at most one direction of a hub arc, the one with lower flow cost (Hamacher et al. 2004). We thus define an *undirected* flow cost F_{ije} for each $e = (k, m) \in E$ and $i, j \in N$ as $F_{ije} = \min\{F_{ijkm}, F_{ijmk}\}$. We also define binary location variables Z_i, $i \in N$, equal to 1 if and only if a hub is located at node i. The *Uncapacitated Hub Location Problem with Multiple Assignments* (UHLPMA) can be stated as follows (Hamacher et al. 2004; Marín 2005a):

$$\text{minimize} \quad \sum_{k \in N} f_k Z_k + \sum_{i,j \in N} \sum_{e \in E} F_{ije} x_{ije}$$

$$\text{subject to} \quad \sum_{e \in E} x_{ije} = 1 \qquad i, j \in N \tag{12.14}$$

$$\sum_{e \in E: k \in e} x_{ije} \leq z_k \qquad i, j, k \in N \tag{12.15}$$

$$x_{ije} \geq 0 \qquad i, j, k \in N \tag{12.16}$$

$$Z_i \in \{0, 1\} \qquad i \in N. \tag{12.17}$$

Constraints (12.14) guarantee that there is a single path connecting the origin and destination nodes of every commodity. Constraints (12.15) prohibit commodities from being routed via a non-hub node. As in UHLPSA, there is no need to explicitly state the integrality on the x_{ije} variables because there always exists an optimal solution of (12.14)–(12.17) in which all x_{ije} variables are integer. This formulation has $O(n^4)$ variables and $O(n^3)$ constraints and usually provides tight LP bounds. Hamacher et al. (2004) and Marín (2005a) independently prove that constraints (12.15) are indeed facet-defining inequalities. Marín (2005a) provide other classes of inequalities associated with the set-packing polytope which also define facets.

The number of routing variables x_{ije} can be further reduced by defining a set of candidate hub arcs for each O/D pair (see Contreras et al. 2011b). This is done by using the property that no flow will be routed through a hub arc containing two different hubs whenever it is cheaper to route it through only one of them (Boland et al. 2004; Marín 2005a).

In HLPs with multiple assignments it is also possible to completely eliminate the undirected routing variables x_{ije} by exploiting the supermodular properties presented in Sect. 12.2.2. We define binary hub arc location variables y_e, $e \in E$, equal to 1 if and only if a hub arc is located at e. For each $i, j \in N$, we order the elements of E by non-decreasing values of their coefficients F_{ije}, and we denote e_{rk} to the r-th element according to that ordering. That is, $F_{ije_1} \leq F_{ije_2} \leq \cdots \leq F_{ije_{|E|}k} \leq F_{e_{ij|E|+1}}$, where $F_{e_{ij|E|+1}} = F_{ije^*}$ is the cost for the fictitious edge e^* such that (i) $F_{e^*k} > \max_{e \in E} F_{ek}$, for all $k \in K$; and (ii) $\sum_{k \in K} F_{e^*k} > \max_{e \in E}(f_e + \sum_{k \in K} F_{ek})$. This assumption guarantees that at least one hub variable y_e is at value one in any optimal solution. The UHLPMA can be stated as the following MIP (see Contreras and Fernández 2014):

$$\text{minimize} \quad \sum_{k \in N} f_k Z_k + \sum_{i,j \in N} \eta_{ij}$$

$$\text{subject to} \quad \eta_{ij} \geq F_{ije_r} + \sum_{e \in E} (F_{ije} - F_{ije_r})^{-} y_e \qquad r = 1, \ldots, |E| + 1, \; i, j \in N \tag{12.18}$$

$$y_e \leq z_k \qquad e = (k, m) \in E \tag{12.19}$$

$$y_e \leq z_m \qquad e = (k, m) \in E \tag{12.20}$$

$$y_e, z_i \in \{0, 1\} \qquad e \in E, i \in N, \tag{12.21}$$

where η_{ij} are continuous decision variables used to evaluate the flow cost of O/D pair (i, j). This new formulation has $O(n^2)$ variables and $O(n^4)$ constraints. It is interesting to note that, for the particular case of the p-hub median problem, the above supermodular formulation coincides with the *radius-based formulation* of García et al. (2012).

As in the case of single assignments, we can also use flow-based formulations to model the UHLPMA. However, we now need additional flow variables for the collection and distribution legs. We define continuous variables $X_{ijm}, i, j, m \in N$, equal to the amount of flow from hub m to destination j that originates at node i. We also define continuous variables $Z_{ik}, i, k \in N$ equal to the amount of flow from origin node i to hub k. Using these sets of decision variables, we can formulate the UHLPMA as follows (Ernst and Krishnamoorthy 1998b):

$$\text{minimize} \quad \sum_{k \in N} f_k Z_k + \sum_{i,k \in N} \chi d_{ik} Z_{ik} + \sum_{i,k,m \in N} \alpha d_{km} Y_{ikm} + \sum_{ijm} \delta d_{jm} X_{ijm}$$

subject to (12.17)–(12.13)

$$\sum_{k \in N} Z_{ik} = O_i \qquad i \in N \tag{12.22}$$

$$\sum_{m} X_{ijm} = W_{ij} \qquad i, j \in N \tag{12.23}$$

$$Z_{ik} + \sum_{m \in N} Y_{ikm} = \sum_{m \in N} Y_{imk} + \sum_{j} X_{ijm} \qquad i, k \in N \tag{12.24}$$

$$Z_{ik}, X_{ijm} \geq 0 \qquad i, j, m \in N. \tag{12.25}$$

Constraints (12.22) ensure that all flow from each origin is sent to a subset of hubs. Constraints (12.23) forces the flow of each O/D pair to arrive at its destination. Constraints (12.24) are the flow conservation constraints at hub facilities. The above formulation contains $O(n^3)$ variables and $O(n^2)$ constraints. Boland et al. (2004) presents some preprocessing procedures that can be used to reduce the number of variables and constraints, and some valid inequalities to improve the LP bounds of capacitated variants.

12.4 Main Developments and Recent Trends

Early hub location research focused mostly on a first generation of HLPs which consider the assumptions introduced in Sect. 12.2.1. In this section we present some research areas that have attracted most attention in the literature over the last decade, leading to more realistic models that relax some of these assumptions and incorporate additional features of real applications. We focus on six particular areas: hub network topologies, flow dependent discounted costs, capacitated models,

models dealing with uncertainty, dynamic and multi-modal models, and competition and collaboration.

12.4.1 Hub Network Topologies

Full interconnection between hub nodes may be prohibitive in applications where there is a considerable setup cost associated with the hub arcs (see O'Kelly and Miller 1994; Klincewicz 1998). To overcome this difficulty, several models considering incomplete hub networks have been studied. HALPs, originally introduced in Campbell et al. (2005a,b), relax the assumption of full interconnection between hubs and consider the location of a set of hub arcs that may (or may not) require a particular topological structure of their induced network. Some of these models do not even require the hub arcs to define a single connected component. Alumur et al. (2009) and Calık et al. (2009) study the design of incomplete hub networks with single assignments in which no network structure other than connectivity is imposed on the backbone network. Other works study models that do not consider a complete backbone network but rather, a particular topological structure. Kim and Tcha (1992), Contreras et al. (2009b, 2010) and Martins de Sá et al. (2013), study the design of tree-star hub networks in which the hubs are connected by means of a tree and the O/D nodes are assigned to exactly one hub. Labbé and Yaman (2008) and Yaman (2008) consider the design of star-star networks in which hub nodes are directly connected to a central node (i.e. star backbone network) and the O/D nodes are assigned to exactly one hub node. Martins de Sá et al. (2015) study the problem of designing a hub-line network in which hubs are connected by means of a line and the aim is to minimize the total service time between pairs of nodes. Martins de Sá et al. (2014) present and extension of this problem to the case in which multiple hub-lines are to be located. Lee et al. (1993) and Contreras et al. (2013) focus on the design of cycle-star networks in which the hubs are connected by means of a cycle. Figure 12.2 shows some examples of different hub network structures.

Yaman (2009) studies the problem of designing a three-layer hub-and-spoke network, where the top layer consists of a complete network connecting the central hubs, and the second and third layers are unions of start networks connecting the remaining hubs to central hubs and the O/D nodes to hubs, respectively. Yaman and Elloumi (2012) consider the design of two-level start networks, while taking into consideration the service quality in terms of the length of paths between pair of O/D nodes. Adler and Smilowitz (2007) focus on the design of global three-layer hub networks in which two types of hub facilities are considered, international gateways and regional hubs. The backbone network associated with each hub-layer is assumed to be complete.

Some papers focus on the design of more complex access networks that are not longer determined by a single or multiple assignment pattern of O/D nodes to hubs. Aykin (1994, 1995) and Sung and Jin (2001) present models that explicitly consider direct connections between non-hub nodes (i.e. they relax Assumption 1).

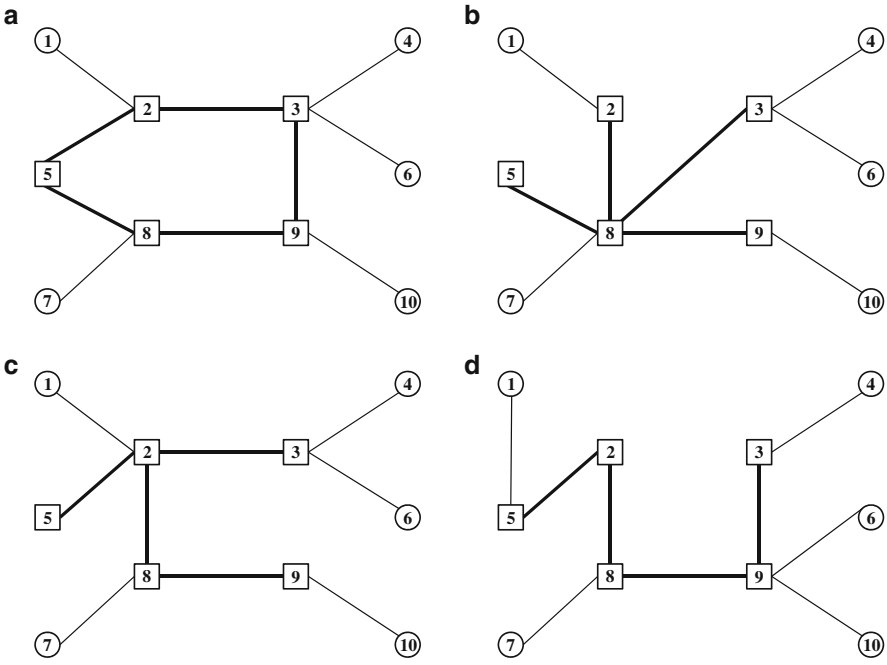


Fig. 12.2 Structure of a cycle-star (**a**), star-star (**b**), tree-star (**c**), and line-star (**d**) hub network

Klincewicz (1998) and Yaman et al. (2007) consider multi-stop access paths that may visit more than one O/D nodes on the way to a hub node. Nagi and Salhi (1998), Camargo et al. (2013), Rodríguez-Martín et al. (2014), and Rieck et al. (2014) study problems in which collection and distribution tours have to be designed. Thomadsen and Larsen (2007) and Saboury et al. (2013) describe HLPs in which both the backbone and access networks are fully interconnected. Figure 12.3 shows some examples of various access network structures.

12.4.2 Modeling Flow Costs

The assumption of flow-independent discounted costs (Assumption 3) is most appropriate in applications where hub arcs are associated with faster transportation modes. However, this can be an oversimplification in applications where the costs represent the economies of scale due to the bundling of flows on the hub arcs. For instance, this assumption could lead to solution networks where hub arcs send considerable less flow than access arcs, yet the flow cost is only discounted on the hub arcs. It may also happen that the amount of flow that is actually routed on each hub arc is quite variable, yet the same discount factor is always applied. For these

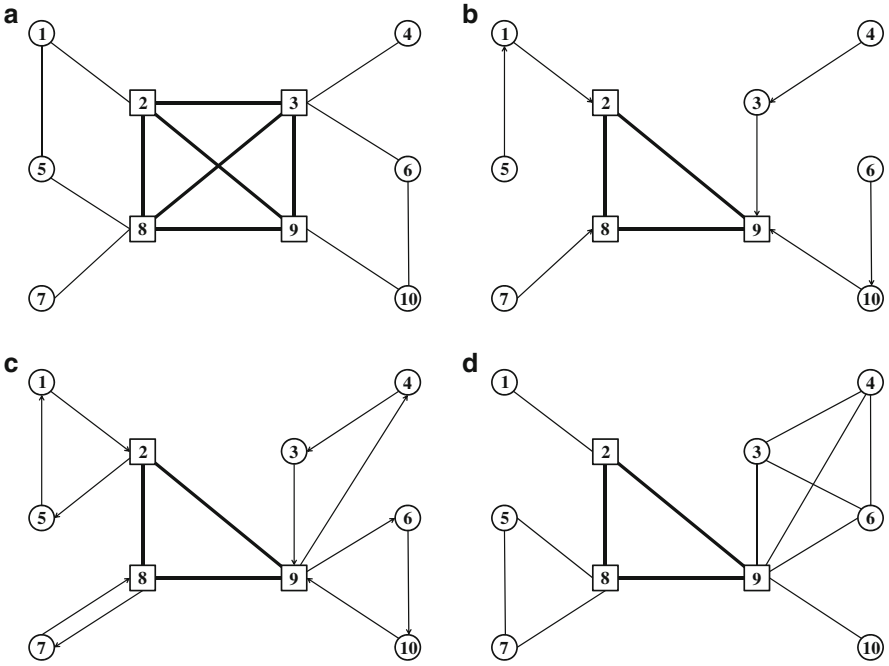


Fig. 12.3 Access network with direct connections (**a**), multi-stops (**b**), tours (**c**), and complete subgraphs (**d**)

reasons, the use of flow-independent costs may not only miscalculate the overall flow cost of the hub network, but could also erroneously select the optimal set of hub nodes and the assignment pattern of O/D nodes to hubs.

Several authors have pointed out these anomalies and different hub location models able to capture the flow-dependency of discounted costs have been proposed. The first hub location model that explicitly accounts for scale economies by allowing discount factors on hub arcs to be a function of flows was introduced in O'Kelly and Bryan (1998). This model, referred to as FLOWLOC, uses a non-linear cost function, in which costs increase at a decreasing rate as flows increase, to compute the flow cost in each hub arc. For any amount of flow, the cost is assumed to be always less than the linear cost associated with a constant discount factor. This function is approximated with a piecewise linear function to obtain a linear integer programming formulation for the problem. Bryan (1998) provides some extensions of the FLOWLOC model that relax the assumption of full interconnection between hubs, by using a minimum threshold value to activate a hub arc, and that incorporate a flow-dependent cost function for both the hub and access arcs. Klincewicz (2002) shows that, once the location of the hubs is known, the FLOWLOC model can be reduced to a classical UFLP. Horner and O'Kelly (2001) present a different non-linear flow cost function based on link performance functions commonly used in

urban transportation planning. This function is used to model flow-dependent costs in both hub and access arcs.

Racunica and Wynter (2005) study an extension of HLPs arising in the design of intermodal transportation networks for freight rail. Their model uses another type of non-linear concave function to model flow-dependent discounted costs only on the transfer and distribution legs. Contrary to the FLOWLOC model, this function is based on an efficiency threshold that considers that discounted flow costs should be higher than the linear cost up to a threshold, and less costly thereafter.

Kimms (2006) introduces a different approach for modeling flow-dependent discounted costs in all the arcs of the network, which is based on fixed-charge cost functions commonly used in other network design problems. This function consist of a fixed flow-independent set-up cost and of a variable flow-dependent (or marginal) cost. This paper presents three different models: an uncapacitated model, a capacitated model, and a multimodal model with different capacities for each mode of transportation. Mirzaghafour (2013) consider a stepwise function to model flow-dependent costs on both hub and access arcs. This type of functions are commonly used to model the transportation cost in most vehicle routing problems (see Laporte 2009).

12.4.3 Capacitated Models

Similar to FLPs, an important extension to HLPs is the incorporation of capacity considerations when designing hub networks. However, in the case of HLPs the capacity constraints may arise not only at the hub facilities but also at the arcs of the network. Moreover, when considering capacitated models with multiple assignment patterns, commodities may be split over several paths and thus, splittable and non-splittable commodity variants arise. In the former case, commodities are allowed to be split over several paths between their origins and destinations. However, in the latter case the commodities cannot be split, meaning that each commodity will be routed through the network from its origin to its destination through a unique path. Note that a multiple assignment pattern that allows splitting is highly desirable when minimizing the total flow cost. However, splitting commodities may not be feasible in some applications.

Capacitated versions of HLPs with multiple assignments are studied by Campbell (1994b), Ebery et al. (2000), and Boland et al. (2004), with capacity constraints on the incoming or outgoing flow at the hubs. Bryan (1998) introduces a model in which capacities are associated with the hub arcs rather than with the hub nodes. Marín (2005b) studies a capacitated model in which commodities are splittable. Rodríguez-Martín and Salazar-González (2008) study another model where commodities can be split into several routes. Capacity constraints are imposed on the incoming flow of each hub, whether it originated from non-hub nodes or from hub nodes. In addition, an upper limit is imposed on the flow traversing any link of the network.

Capacitated versions of HLPs with single assignment have also been studied by Campbell (1994b), Ernst and Krishnamoorthy (1999), Labbé et al. (2005), Correia et al. (2010), Contreras et al. (2009a), and Contreras et al. (2011d). All these models only consider capacity constraints on the incoming or outgoing flow at the hub nodes. Aykin (1994, 1995) have considered HLPs with capacity constraints on the incoming flow at the hubs as well as on direct O/D links. Carello et al. (2004), Yaman and Carello (2005) and Yaman (2008) have studied capacitated HLPs with modular link capacities. They considered capacity constraints on the incoming and outgoing flow at hubs.

All of the above mentioned capacitated models consider that both hub and arc capacities are exogenous, i.e. capacity levels for potential hub nodes and hub arcs are determined a priori. Given that capacities can have a determining impact on locational and routing decisions, some researchers have started studying more realistic capacitated models in which the amount of installed capacity is part of the decision process. Correia et al. (2010) studied an extension of capacitated HLPs with single assignment in which the hub capacity is a decision variable. Elhedhli and Wu (2010) introduced a capacitated model in which hub capacity is also a decision variable. Contreras et al. (2012) presented models with multiple assignments in which the amount of capacity installed at the hubs is part of the decision process, for both splittable and non-splittable commodity cases.

12.4.4 Uncertainty in Hub Location

The design of hub networks corresponds to long-term strategic decisions which are typically made within an uncertain environment. That is, costs, demands, distances, and other parameters may change after location and network design decisions have been made. Nevertheless, most HLPs treat data as known and deterministic. This can result in highly sub-optimal solutions given the inherent uncertainty surrounding future conditions. Some researchers have thus started to study how different uncertainty aspects can be taken into account when designing hub networks.

Marianov and Serra (2003) is probably the first paper dealing with uncertainty, focusing on stochasticity at the hub nodes by representing hub airports as $M/D/c$ queues and limiting through chance constraints the number of airplanes that can queue at an airport. Sim et al. (2009) introduce the stochastic p-hub center problem and employ a chance-constrained formulation to model the minimum service-level requirement. This model takes into account the variability in travel times when designing the hub network so that the maximum travel time through the network is minimized.

Contreras et al. (2011a) study how the classical UHLPMA can be modeled as a two-stage integer stochastic program with recourse in the presence of uncertainty on demands and flow costs. In particular, three different stochastic versions are introduced. The first considers the flow between O/D nodes to be stochastic. The

second assumes that uncertainty is given by a single parameter equally influencing the flow cost for all links of the network. The third considers the more general case in which the uncertainty of transportation costs is independent for each link of the network. the authors show that the first to variants are equivalent to their associated expected value problem in which uncertain amount of flows and flow costs are replaced with their expected value. However, this equivalence does not hold for the third case. Alumur et al. (2012b) consider HLPs under uncertainty in the set-up cost for the location of hubs and in the demand flows for both single and multiple assignments models. The first class of models deals with uncertainty on the set-up costs in the absence of a known probability distribution for these random parameters. The authors propose the use of a minimax regret model in which the objective is to minimize the worst-case regret over a finite set of scenarios. The second class considers uncertainty on the demand flows and uses a two-stage stochastic program with recourse. However, as shown in Contreras et al. (2011a) these problems are equivalent to their associated expected value problem. The third class considers uncertainty in both set-up costs and demand flows and are modeled as two-stage minmax regret programs with recourse.

Demand uncertainty has also been studied in hub location from a congestion perspective. When demand flows increase unexpectedly within a short time, they are likely to congest the hub network. This causes an increase in the operational cost of the network due to delays at hub facilities. Elhedhli and Hu (2005) present a single allocation hub location model that considers hub congestion-related costs as an exponential function of the hub flow. Camargo et al. (2009) propose the multiple allocation analogue of the previous model. Elhedhli and Wu (2010) study a different approach in which the hub network is modeled as a network of $M/M/1$ queues where each hub behaves as a single server with a given exponential service rate determined by its capacity. The congestion cost is modeled using a Kleinrock average delay function. Camargo and Miranda (2012) provide extensions to the previous single allocation models by considering two different perspectives: a network owner perspective in which the goal is to design a hub network with the least congestion cost, and a user perspective in which the goal is to minimize the maximum congestion effect.

An important uncertainty aspect neglected until very recently is the reliability of hub networks. Kim and O'Kelly (2009) presents a reliable p-hub location problem arising in the design of telecommunication networks. This problem considers the reliability of O/D paths by taking into account the probability of successful communication to deliver traffic without congestion or loss between O/D pairs. It focuses on maximizing the total network flow that can be routed when incorporating the reliability of O/D paths. An et al. (2011) and Aziz et al. (2014) study models in which disruptions at hub nodes are taken into account when designing the hub network. The proposed models mitigate the resulting hub unavailability by using backup hubs and alternative routes for demand flows. The objective of this model is to minimize the total expected flow cost considering both the regular and the disruptive situation.

12.4.5 Dynamic and Multi-modal Models

One common feature of real applications is the dynamic nature of the problem. Parameters such as costs, demand, and resources often vary over the planning horizon. From the location point of view this gives rise to different types of multi-period, or dynamic problems. In this type of problems, not only a routing plan has to be made, but the times at which facilities are opened or closed must be determined.

Campbell (1990) develops a continuous approximation model to locate transportation terminals (hubs) for a general freight carrier serving an increasing demand in a fixed region. It can be seen as a continuous dynamic hub location model in which it is assumed that the O/D points are scattered randomly over the service region. Contreras et al. (2011c) studies a dynamic model with multiple assignments which includes strategic decisions related to the location, operation and closing of hub facilities over time. It is assumed that the forecast demand between O/D pairs is known with certainty but varies over the time horizon. Moreover, the proposed model allows hubs to be opened and closed at different time periods to provide a flexible hub network. Gelareh (2008) presents another multi-period hub location model arising in the design of public transportation networks in which it is relaxed the full interconnection assumption and thus, additional hub arc selection are considered.

Another important feature in some applications is the presence of strategic decisions related to the choice for mode of transportation. Most HLPs consider only one mode of transportation is available and thus, only one type of hub facility. However, global hub networks usually employ a mixture of air, ground and water transportation modes. In a multi-modal hub network, each mode can be characterized by its flow cost structure, modal connectivity, availability of transfer points, and service time performance.

Racunica and Wynter (2005) address the design of hub networks for inter-modal freight transport on dedicated or semi-dedicated freight rail lines which could make use of shuttle trains on the hub arcs. Groothedde et al. (2005) develop a multi-modal hub location model that focus on the design of a collaborative hub network for the distribution of fast moving consumer goods using a combination of trucking and inland barges. Ishfaq and Sox (2011) present a multiple allocation model to design a rail-road inter-modal network. It considers the location of two different types of hubs with different modal connectivity costs and the incorporation of service time requirements. Meng and Wang (2011) study the design of an inter-modal hub network for multi-type container transportation with multiple stakeholders: the network planner, carriers, hub operations and inter-modal operators. The proposed model incorporates the user equilibrium behavior of inter-modal operators in route choice. Alumur et al. (2012a) introduce a more general hub network design problem in which the full interconnection of hubs assumption is relaxed and hub arc location decisions, that include the selection of the type of transportation mode, are considered. This model incorporates set-up costs, transportation costs and service levels when designing the multi-modal hub network. Alumur et al. (2012c) study

a related hub covering problem to locate two types of hub nodes and hub arcs associated with ground and air transportation. The model uses a cost-oriented objective while ensuring time-definite deliveries.

12.4.6 Competition and Collaboration

Most HLPs studies assume that the decision maker is a monopolist firm in a market and thus can capture all demand flow in the market, regardless of the design of the hub network. As a result, location and network design decisions are usually determined by the firm's cost-based objective without taking into account customer preferences. However, in practice many telecommunication and transportation networks operate in a competitive environment where several firms exist in a market and compete to provide service to customers. Customers must determine which competing firm to use based on several criteria such as the travel time and the costs charged. Competitive hub location models focus on the design of hub networks so as to maximize the market share of competing firms. In these models, customers (or demand flow) are captured from competitor's hub networks whenever the new hub network offers a reduction of the travel time or distance needed by the customers to go from their origins to their destinations.

Most competitive hub location models use a sequential location approach, in which an existing company (the leader) serves the demand flow in a region, and a new company (the follower) wants to enter the market and will attempt to capture the maximum possible demand and thus, maximize its market share. Marianov et al. (1999) introduce competitive hub location models in which the follower wants to locate a set of hub nodes so as to maximize the captured demand flow. In the first proposed model it is assumed that demand is fully captured when the flow cost does not exceed the current competitor's cost. The second model considers a more realistic version in which a stepwise linear function is used to model the proportion of demand captured depending on the new flow cost as compared to the competitor's cost. In both models, at most one path is used to route flow between each O/D pair. Wagner (2008b) points out that if the new company is assumed to capture demand flow when its flow cost is equal to the current competitor's cost, then the optimal solution is always to locate a hub node in each location where the leader has one, making the new company to capture all demand. Therefore, the author suggests modifying the definition of the problem so that demand is captured by the follower if and only if the new cost is strictly smaller than the competitor's cost. Eiselt and Marianov (2009) provide an extension to the models presented in Marianov et al. (1999), in which each competitor can have more than one path between O/D pairs. The proportion of flow that is captured on a particular path is modeled with a gravity-like attraction function that does not only depend on the flow cost but also on the travel time. Gelareh et al. (2010) present a competitive model arising in liner shipping networks, where a new liner service provider wants to design a hub network to maximize its market share, using an stepwise attraction function

which depends on the service time and flow cost. This model allows O/D paths to contain more than one hub arc or to have direct connections between origins and destinations. Luer-Villagra and Marianov (2013) study a competitive model in which an existing firm uses a hub network and charges its flow costs plus a fixed additional percentage to their customers. A new company wants to enter into the same market using an incomplete hub network and to determine prices so as to maximize its profit, rather than its market share. The profit comes from the revenues from captured flows, minus the a fixed and variable costs. Customer preferences on selected firm and route are modeled using a logit model.

Using a game theoretic framework, Sasaki and Fukushima (2001) introduce a continuous Stackelberg hub location model where a large company competes with several medium-size companies to maximize its profit. The large company first locates a new hub on a plane as a leader, and the other companies then locate their new hubs. The authors use a nonlinear logit function to model the level of captured customers and formulate the leader's problem as a bilevel program and the follower's problems as lower level programs. Sasaki (2005) provides an extension to the discrete case assuming there is a leader and only one follower. The proposed model considers that companies cannot provide any service whose captured market share does not reach to a threshold lower limit value. Sasaki et al. (2009) study a more general model in which the full interconnection assumption is relaxed and a set of hub arcs must be located. As in Sasaki (2005), two firms compete for customers in a Stackelberg framework, where the leader firm locates hub arcs to maximize its market share, knowing that the follower will later locate its own hub arcs to maximize its market share.

Instead of considering a pure competitive environment, some studies have looked at hub network alliances and mergers, as well as user cooperation employing a game theoretic approach. In Skorin-Kapov (1998) a cooperative game theory is used to analyze several cost allocation problems referred to as hub network games. In particular, the flow routing cost is distributed among the hub network users with possibly conflicting interests, but their cooperation is essential for the exploitation of economies of scale on the routing of flows. Lin and Lee (2010) propose a non-cooperative game theoretic model to study the competition hub network design in an oligopolistic market with few dominant firms. In this model, each firm will first observe the hub network and demand flows of other firms and will then simultaneously determine its hub network, demand, and routing plan in order to maximize its profits. The firms' decisions jointly determine the market prices, which include the reassessment and redesign of hub networks of all other firms. The process of observation, design and reassessment will continue until a long-term Cournot-Nash equilibrium is established.

Adler and Smilowitz (2007) present hub location models to analyze global alliances and mergers in the airline industry under competition. In particular, the authors develop a game theoretic approach in which merger and hub location decisions are considered to evaluate hub networks under competition. The proposed problems are modeled as games played among multiple airlines, consisting of selecting the optimal hubs to develop, expand or remove in the newly merged hub

network. Asgari et al. (2013) study a game theoretic hub network design model that investigates the competition and cooperation amongst two major hub ports and the shipping companies, with the objective of minimizing the shipping companies' cost and maximizing the hub ports' revenue.

12.5 Solving Hub Location Problems

The interrelation of location and network design decisions make HLPs particularly difficult to solve. A considerable effort has thus been made over the past two decades to develop algorithms capable of obtaining high quality solutions of various classes of HLPs, particularly when considering more realistic, large-scale instances. Some of these algorithms are able to provide an estimation of the quality of the obtained solutions and some them are able to prove that the obtained solution is optimal. In this section, we point out recent papers describing the most effective solution algorithms for various classes of HLPs. The interested reader is referred to Alumur and Kara (2008) and Zanjirani Farahani et al. (2013) for a detailed survey of approximate and exact algorithms for HLPs.

12.5.1 Complexity Results

Most HLPs are known to be NP-hard. However, very little research has been done to analyze the complexity and polynomial-time approximability of particular classes of HLPs. In the case of fundamental HLPs with single assignments, in which the full interconnection assumption is used, even if the location of the hub nodes is given the remaining subproblem is still NP-hard. This problem is known as the *quadratic semi-assignment problem* or the *single allocation problem* (see Saito et al. 2009; Sohn and Park 2000, and references therein). Sohn and Park (1997) show that for the particular case of the *uncapacitated p-hub median problem with single assignments* (UpHLPSA), when $p = 2$ the problem can be polynomial solved by reducing it to $n(n-1)/2$ independent minimum cut problems. Sohn and Park (2000) prove that the single allocation problem becomes NP-hard as soon as the number of hubs is three and thus, the UpHLPSA is NP-hard for $p \geq 3$. Iwasa et al. (2009) describe a deterministic 3-approximation algorithm and a randomized 2-approximation algorithm for the single allocation problem. Moreover, they provide a $(5/4)$-approximation algorithm for the particular case in which the number of hubs is three.

When considering HLPs with incomplete hub networks, even if the location of hubs and the assignment of O/D nodes to hubs is given, the subproblem associated with the location of hub arcs remains challenging. For instance, when considering tree-star topologies the design of a tree spanning the set of hub nodes is equivalent to the so-called *optimum communication spanning tree problem*, known to be NP-hard

(Contreras et al. 2010). In the case of cycle-star topologies, connecting the hub nodes by means of a cycle is equivalent to the *minimum flow cost Hamiltonian cycle problem*, known to be NP-hard (Contreras et al. 2013).

In the case of uncapacitated HLPs with multiple assignments, in which the full interconnection assumption is used, once the location of the hubs is known the allocation subproblem is equivalent to an *all pairs shortest path problem* and thus, can be solved in polynomial time (Ernst and Krishnamoorthy 1998a). When considering capacities on the hub nodes and commodities can be split, Contreras et al. (2012) show that the allocation subproblem remains polynomially solvable as it is equivalent to a classical *transportation problem*. However, when commodities cannot be split the subproblem is equivalent to a *generalized assignment problem* and thus becomes NP-hard.

Contreras and Fernández (2014) show that a general class of HLPs with multiple assignments, known as *supermodular hub location problems* (Sect. 12.2.2), is NP-hard. We recall that SHLPs include several special cases such as p-hub median, uncapacitated hub location, and q-hub arc location. The authors also present worst-case performance results for simple greedy and local improvement heuristics for particular classes of SHLPs in which the objective functions are also non-increasing, as in p-hub median and q-hub arc location problems.

Kara and Tansel (2003) show that *hub set-covering problems with single assignments* are NP-hard. Kara and Tansel (2000) prove that the *uncapacitated p-hub center problem with single assignments* is also NP-hard for $p < n - 1$. Ernst et al. (2009) show that the multiple assignments version of this problem is also NP-hard. They also prove that the single allocation subproblem with respect to a given set of hubs is already NP-hard, whereas for the multiple assignment case is not. Liang (2013) considers the *star p-hub center problem* and shows that is strongly NP-hard and that there is no $(5/4 - \epsilon)$-approximation algorithm for it for any $\epsilon > 0$, unless P = NP. This paper also provides a $7/2$-approximation algorithm for this problem.

12.5.2 Heuristic Algorithms

A considerable amount of hub location research on heuristic algorithms has focused on fundamental HLPs. To the best of our knowledge, the best heuristic for the *uncapacitated p-hub location problem with single assignments* is the variable neighborhood search algorithm of Ilić et al. (2010). It outperforms all previous heuristics and it yields solutions for very large-scale instances with up to 1,000 nodes and $p = 20$ within reasonable CPU times. The best results for the UHLPSA seem to be obtained using the memetic algorithm recently designed by Marić et al. (2013). This heuristic has the best performance, especially on large instances with up to 900 nodes. Contreras et al. (2011d) provide GRASP heuristics for capacitated versions of this problem. Contreras et al. (2011b) design a GRASP heuristic for the UHLPMA capable of obtaining high quality solutions for instances with up to

500 nodes within reasonable CPU times. Meyer et al. (2009) present an ant colony optimization algorithm for the *p-hub center problem with single assignments* which is able to obtain high quality solutions for large-scale instances with up to 400 nodes.

Some researchers have recently focused on the development of efficient heuristic algorithms for more realistic extensions of HLPs. Calık et al. (2009) describe a tabu search to solve hub covering problems over incomplete hub networks. Köksalan and Soylu (2010) study evolutionary algorithms for two bicriteria uncapacitated p-hub location problems considering congestion-related costs. Contreras et al. (2013) describe a GRASP algorithm for the design of incomplete hub networks with a cycle-star topology. Saboury et al. (2013) present two hybrid heuristics to design of hub networks with fully interconnected backbone and access networks. Martins de Sá et al. (2014) propose an adaptive large neighborhood search and GRASP algorithms to design hub networks with multiple hub lines.

12.5.3 Lower Bounding Procedures and Exact Algorithms

Dual ascent and dual adjustments techniques have been used to efficiently obtain the LP bound of MIP formulations for various HLPs. Yoon and Current (2008) use dual based heuristics to solve HLPs with additional arc selection decisions. Cánovas et al. (2007) present a *Branch-and-Bound* (BB) algorithm based on dual techniques to obtain optimal solutions to uncapacitated HLPs with multiple assignments. Meyer et al. (2009) develop a two-phase exact algorithm for the p-hub center problem with single assignments. In this algorithm the BB method presented in Ernst and Krishnamoorthy (1998a) is used during the first phase to obtain a set of potential optimal hub locations. This algorithm seems to be the best exact algorithm for hub center problems, being able to solve to optimality large-scale instances with up to 400 nodes.

Lagrangean relaxation (LR) has been successfully used to obtain tight lower and upper bounds on the value of the optimal solution of several classes of HLPs. Pirkul and Schilling (1998) present efficient LR heuristics to approximately solve uncapacitated HLPs with single assignments, whereas Yaman (2008), Contreras et al. (2009a,b), and Elhedhli and Wu (2010) propose LR heuristics to solve various capacitated HLPs. Exact BB methods based on LR have also been developed to optimally solve HLPs. Marín (2005a) propose a relax-and-cut algorithm for the UHLPMA, which adds violated facet-defining inequalities to a LR of the path-based formulation presented in Sect. 12.3.2, to optimally solve instances with up to 50 nodes. Contreras et al. (2011c) present an exact BB method, that uses a LR of an extension of the path-based formulation presented in Sect. 12.3.2, to obtain optimal solutions for uncapacitated dynamic hub location problems with up to 100 nodes and ten time periods.

Benders decomposition (BD) is another successful method used to optimally solve several classes of HLPs. Camargo et al. (2009) use a BD algorithm to solve large-scale instances of the challenging flow-dependent cost (FLOWLOC) model.

Contreras et al. (2011b) describe an exact algorithm for the UHLPMA which applies an enhanced BD to the path-based formulation presented in Sect. 12.3.2, to obtain optimal solutions for large-scale instances with up to 500 nodes. Contreras et al. (2012) provide an extension of the previous BD to solve multi-capacity HLPs with multiple assignments, with splittable and non-splittable commodities, for instances with up to 300 nodes. Contreras et al. (2011a) develops a Monte–Carlo simulation-based algorithm that integrates a BD to solve uncapacitated HLPs having stochastic flow costs. Camargo et al. (2013) describe a BD algorithm to solve hub location-routing problems, in which additional routing decisions to serve O/D nodes are considered. This algorithm can solve instances with up to 100 nodes. Several BD algorithms have also been implemented for HLPs with congestion costs for both multiple (Camargo et al. 2009) and single (Camargo et al. 2011; Camargo and Miranda 2012) assignments versions, HALPs with particular topological structures such as tree-start networks (Martins de Sá et al. 2013) and hub-line networks (Martins de Sá et al. 2015, 2014), HLPs arising in public transportation networks (Gelareh and Nickel 2011), and liner shipping applications (Gelareh and Nickel 2011; Gelareh and Pisinger 2011).

Branch-and-cut (BC) methods have also been developed to optimally solve various HLPs. Labbé et al. (2005) develop a BC algorithm based on the two-index formulation presented in Sect. 12.3.1 for various classes of capacitated HLPs with single assignments. This method is able to solve to optimality instances with up to 50 nodes. García et al. (2012) presents a BC algorithm for the uncapacitated p-hub median problem with multiple assignments. This algorithm uses an extension of the two-index formulation presented in Sect. 12.3.2 and is able to optimally solve large-scale instances with up to 200 nodes with very large values of p. Contreras and Fernández (2014) also introduce a BC algorithm based on the two-index formulation for the general class of *supermodular hub location problems* presented in Sect. 12.2.2. This method is able to solve q-hub arc location problems with up to 125 nodes. Contreras et al. (2010) and Contreras et al. (2013) use an adaptation of the flow-based formulation introduced in Sect. 12.3.1 to develop BC algorithms to solve HLPs with tree-star and cycle-star topologies, respectively. Contreras et al. (2013) is able to solve to optimality instances with up to 100 nodes. Catanzaro et al. (2011) study a incomplete hub network design problem with additional graph partitioning and routing decisions. Rodríguez-Martín et al. (2014) introduce a BC algorithm for a hub location-routing problem, which is able to solve instances with up to 50 nodes.

Column generation (CG) is the method that has received the least attention in the hub location literature. Thomadsen and Larsen (2007) presents a branch-and-price method for solving a HLP with fully interconnected access networks. Contreras et al. (2011d) presents and exact algorithm, that combines LR and CG methods as a bounding procedure, to obtain optimal solutions of large-scale capacitated HLPs with single assignments with up to 200 nodes.

12.6 Conclusions

In this chapter we have provided an overview of hub location problems in which both the location of hubs and the design of the hub network are key decisions. We have highlighted how the commonly used assumptions presented in Sect. 12.2.1 simplify network design decisions, which have created a first generation of *idealized* hub location models focusing mostly on location and allocation decisions. Several researchers have exploited the rich structure of these models and as a consequence, a significant progress has been made on the development of strong MIP formulations and efficient algorithms for their solution.

Strong path-based formulations, used in combination with sophisticated decomposition methods, have proven to be amongst the most effective formulations to solve to optimality large-scale instances (with hundreds of nodes) for several classes of hub location problems. Flow-based formulations, having fewer variables and constraints, have been particularly useful when used with general purpose MIP solvers to solve small to medium-size instances (containing usually no more than 50 nodes) for a wide range of problems without having to develop ad-hoc solution algorithms. These formulations have also been strengthened with the addition of valid inequalities and used within a cutting plane framework to solve challenging hub location variants. Over the past few years, promising two-index formulations have started to arise. However, a substantial amount of work still needs to be done to analyze how these can be used as a basis for sophisticated algorithms.

We have also pointed out how location and network design decisions become more involved when relaxing some of the *simplifying* assumptions presented in Sect. 12.2.1. In particular, Sect. 12.4.1 described several classes of hub network topologies, arising from different areas of application, that have started to be studied. The resulting hub location problems contain additional hub arc and access arc selection decisions, making them substantially more difficult to model and solve than first generation problems considering fully interconnection between hubs and access networks characterized by single or multiple assignment patterns. Section 12.4.2 focused on more realistic models with discounting levels that depend on the amount of flow passing through each arc to better model the flow cost. Although some flow-dependent models have already been presented in the literature, alternative modeling approaches need to be studied to more accurately represent flow costs, specially on transportation applications. Section 12.4.3 reviewed several capacitated hub location models, most of which focus on capacity restrictions on the hub nodes and only a few of them on the links. More complex problems combining both types of capacities need to be studied. Section 12.4.4 described some models in which specific sources of uncertainty were considered, mostly from a stochastic programming perspective. However, additional aspects such as congestion on hubs and arcs, reliability, and disruptions, among other things, need to be further studied. Very few models considering dynamic and multi-modal features have been proposed (Sect. 12.4.5). Additional models need to be developed to better model the optimal evolution of hub networks and the choice for mode of transportation. Given that

most companies using hub networks are not monopolists in a market and are also not redesigning their network from scratch, competition and collaboration are very important aspects in most hub location applications (Sect. 12.4.6). For this reason, additional models that consider a competitive environment, collaborations, mergers, acquisitions, and divestments of companies, need to be further studied.

References

Adler N, Smilowitz K (2007) Hub-and-spoke network alliances and mergers: Price-location competition in the airline industry. Transp Res B Methodol 41:394–409

Alumur S, Kara BY (2008) Network hub location problems: The state of the art. Eur J Oper Res 190:1–21

Alumur S, Kara BY (2009) A hub covering network design problem for cargo applications in Turkey. J Oper Res Soc 60:1349–1359

Alumur S, Kara BY, Karasan OE (2009) The design of incomplete single allocation hub networks. Transp Res B Methodol 43:936–951

Alumur S, Kara BY, Karasan OE (2012a) Multimodal hub location and hub network design. Omega 40:927–939

Alumur S, Nickel S, Saldanha da Gama F (2012b) Hub location under uncertainty. Transp Res B Methodol 46:529–543

Alumur S, Yaman H, Kara BY (2012c) Hierarchical multimodal hub location problem with time-definite deliveries. Transp Res E Logist 48:1107–1120

An Y, Zhang Y, Zeng B (2011) The reliable hub-and-spoke design problem: Models and algorithms. Optimization Online

Asgari N, Zanjirani Farahani R, Goh M (2013) Network design approach for hub ports-shipping companies competition and cooperation. Transp Res A Pol 48:1–18

Aykin T (1988) On the location of hub facilities. Transp Sci 22:155–157

Aykin T (1994) Lagrangian relaxation based approaches to capacitated hub-and-spoke network design problem. Eur J Oper Res 79:501–523

Aykin T (1995) Networking policies for hub-and-spoke systems with applications to the air transportation system. Transp Sci 3:201–221

Aziz N, Chauhan S, Vidyarthi N (2014) The impact of hub failure in hub-and-spoke networks: mathematical formulations and solution techniques. Comput Oper Res DOI: 10.1016/j.cor.2014.05.012

Boland N, Krishnamoorthy M, Ernst AT, Ebery J (2004) Preprocessing and cutting for multiple allocation hub location problems. Eur J Oper Res 155:638–653

Bryan DL (1998) Extensions to the hub location problem: Formulations and numerical examples. Geogr Anal 30:315–330

Bryan DL, O'Kelly ME (1999) Hub-and-spoke networks in air transportation: An analytical review. J Reg Sci 39:275–295

Çalık H, Alumur, SA, Kara BY, Karasan OE (2009) A tabu-search based heuristic for the hub covering problem over incomplete hub networks. Comput Oper Res 36:3088–3096

Camargo RS, Miranda Jr G (2012) Single allocation hub location problem under congestion: Network owner and user perspectives. Expert Syst Appl 39:3385–3391

Camargo RS, Miranda Jr G, Ferreira RPM (2011) A hybrid outer-approximation / Benders decomposition algorithm for the single allocation hub location problem under congestion. Oper Res Lett 39:329–337

Camargo RS, Miranda Jr G, Ferreira RPM, Luna HP (2009) Multiple allocation hub-and-spoke network design under hub congestion. Comput Oper Res 36:3097–3106

Camargo RS, Miranda Jr G, Lokketagen A (2013) A new formulation and an exact approach for the many-to-many hub location-routing problem. Appl Math Model 37:12–13
Camargo RS, Miranda Jr G, Luna HP (2009) Benders decomposition for hub location problems with economies of scale. Transp Sci 43:86–97
Campbell JF (1990) Locating transportation terminals to serve an expanding demand. Transp Res B Methodol 3:173–192
Campbell JF (1994a) A survey of network hub location. Stud Locational Anal 6:31–43
Campbell JF (1994b) Integer programming formulations of discrete hub location problems. Eur J Oper Res 72:387–405
Campbell JF (2013) A continuous approximation model for time definite many-to-many transportation. Transp Res B Methodol 54:100–112
Campbell JF, O'Kelly ME (2012) Twenty-five years of hub location research. Transp Sci 46:153–169
Campbell JF, Ernst AT, Krishnamoorthy M (2001) Hub location problems. In: Drezner Z, Hamacher HW (eds) Facility location. Applications and theory, Springer, Heidelberg, New York, Berlin, pp 373–408
Campbell JF, Ernst AT, Krishnamoorthy M (2005a) Hub arc location problems: part I Introduction and results. Manage Sci 51:1540–55
Campbell JF, Ernst AT, Krishnamoorthy M (2005b) Hub arc location problems: part II formulations and optimal algorithms. Manage Sci 51:1556–71
Campbell AM, Lowe TJ, Zhang L (2007) The p-hub center allocation problem. Eur J Oper Res 176:819–835
Cánovas L, Garcia S, Marín A (2007) Solving the uncapacitated multiple allocation hub location problem by means of a dual-ascent technique. Eur J Oper Res 179:990–1007
Carello G, Della Croce F, Ghirardi M, Tadel R (2004) Solving the hub location problem in telecommunications network design: A local search approach. Networks 44:94–105
Catanzaro D, Gourdin É, Labbé M, Ozsoy FA (2011) A branch-and-cut algorithm for the partitioning-hub location-routing problem. Comput Oper Res 38:539–549
Contreras I, Cordeau J-F, Laporte G (2011a) Stochastic uncapacitated hub location. Eur J Oper Res 212:518–528
Contreras I, Cordeau J-F, Laporte G (2011b) Benders decomposition for large-scale uncapacitated hub location. Oper Res 9:1477–1490
Contreras I, Cordeau J-F, Laporte G (2011c) The dynamic uncapacitated hub location problem. Transp Sci 45:18–32
Contreras I, Cordeau J-F, Laporte G (2012) Exact solution of large-scale hub location problems with multiple capacity levels. Transp Sci 46:439–459
Contreras I, Díaz JA, Fernández E (2009a) Lagrangean relaxation for the capacitated hub location problem with single assignment. OR Spectr 31:483–505
Contreras I, Díaz JA, Fernández E (2011d) Branch and price for large-scale capacitated hub location problems with single assignment. INFORMS J Comput 23:41–55
Contreras I, Fernández E (2012) General network design: A unified view of combined location and network design problems. Eur J Oper Res 219:680–697
Contreras I, Fernández E (2014) Hub location as the minimization of a supermodular set function. Oper Res 62, 557–570
Contreras I, Fernández E, Marín A (2009) Tight bounds from a path based formulation for the tree of hubs location problem. Comput Oper Res 36:3117–3127
Contreras I, Fernández E, Marín A (2010) The tree of hubs location problem. Eur J Oper Res 202:390–400
Contreras I, Tanash M, Vidyarthi N (2013) The cycle hub location problem. Technical Report CIRRELT-2013-59
Correia I, Nickel S, Saldanha da Gama F (2010a) Single-assignment hub location problems with multiple capacity levels. Transp Res B Methodol 44:1047–1066
Correia I, Nickel S, Saldanha da Gama F (2010b) The capacitated single-allocation hub location problem revisited: A note on a classical formulation. Eur J Oper Res 207:92–96

Costa MG, Captivo ME, Climaco J (2008) Capacitated single allocation hub location problem - a bi-criteria approach. Comput Oper Res 35:3671–3695
Ebery J, Krishnamoorthy M, Ernst AT, Boland N (2000) The capacitated multiple allocation hub location problem: Formulations and algorithms. Eur J Oper Res 120:614–631
Elhedhli S, Hu FX (2005) Hub-and-spoke network design with congestion. Comput Oper Res 32:1615–1632
Elhedhli S, Wu H (2010) A Lagrangean heuristic for hub-and-spoke system design with capacity selection and congestion. INFORMS J Comput 22:282–296
Eiselt HA, Marianov V (2009) A conditional p-hub location problem with attraction functions. Comput Oper Res 36:3128–3135
Ernst AT, Hamacher HW, Jiang H, Krishnamoorthy M, Woenginger G (2009) Uncapacitated single and multiple allocation p-hub center problems. Comput Oper Res 36:2230–2241
Ernst AT, Krishnamoorthy M (1996) Efficient algorithms for the uncapacitated single allocation p-hub median problem. Locat Sci 4:139–154
Ernst AT, Krishnamoorthy M (1998a) An exact solution approach based on shortest-paths for p-hub median problems. INFORMS J Comput 10:149–162
Ernst AT, Krishnamoorthy M (1998b) Exact and heuristic algorithms for the uncapacitated multiple allocation p-hub median problems. Eur J Oper Res 104:100–112
Ernst AT, Krishnamoorthy M (1999) Solution algorithms for the capacitated single allocation hub location problem. Ann Oper Res 86:141–159
García S, Landete M, Marín A. (2012) New formulation and a branch-and-cut algorithm for the multiple allocation p-hub median problem. Eur J Oper Res 220:48–57
Gelareh S (2008) Hub location models in public transportation planning. PhD thesis.
Gelareh S, Nickel S (2011) Hub location in transportation networks. Transp Res E-Logist 47:1092–1111
Gelareh S, Nickel S, Pisinger D (2010) Liner shipping hub network design in a competitive environment. Transp Res E Logist 46:991–1004
Gelareh S, Pisinger D (2011) Fleet deployment, network design and hub location of liner shipping companies. Transp Res E Logist 47:947–964
Gendron B, Crainic TG, Frangioni A (1999) Multicommodity capacitated network design. In: Sansó B and Soriano P (eds) Telecommunications Network Planning, Kluwer, Norwell, MA, pp 1–19
Groothedde B, Ruijgrok C, Tavasszy L (2005) Towards collaborative, intermodal hub networks: a case study in the fast moving consumer good market. Transp Res E Logist 41:567–583
Hamacher HW, Labbé M, Nickel S, Sonneborn T (2004) Adapting polyhedral properties from facility to hub location problems. Discrete Appl Math 145:104–116
Horner MW, O'Kelly ME (2001) Embedding economies of scale concepts for hub network design. J Transp Geogr 9:255–265
Hwang YH, Lee YH (2013) Uncapacitated single allocation p-hub maximal covering problem. Comput Ind Eng 63:382–389
Ilić A, Urošević D, Brimberg J, Mladenović N (2010) A general variable neighborhood search for solving the uncapacitated single allocation p-hub median problem. Eur J Oper Res 206:289–300
Ishfaq R, Sox CR (2011) Hub location-allocation in intermodal logistic networks. Eur J Oper Res 210:213–230
Iwasa M, Saito H, Matsui T (2009) Approximation algorithms for the single allocation problem in hub-and-spoke networks and related metric labeling problems. Discrete Appl Math 157:2078–2088
Kara BY, Tansel BÇ (2000) On the single-assignment p-hub center problem. Eur J Oper Res 125:648–655
Kara BY, Tansel BÇ (2003) The single-assignment hub covering problem: Models and linearizations. J Oper Res Soc 54:59–64
Kim H, O'Kelly ME (2009) Reliable p-hub location problem in telecommunication networks. Geogr Anal 41:283–306

Kim J-G, Tcha D-W (1992) Optimal design of a two-level hierarchical network with tree-star configuration. Comput Ind Eng 22:273–281
Kimms A (2006) Economies of scale in hub and spoke network design: we have it all wrong. In: Morlock M, Schwindt C, Trautmann N, Zimmermann J (eds) Perspectives on operations research, Weisbaden, Germany, pp 293–317
Klincewicz JG (1998) Hub location in backbone/tributary network design: A review. Loc Sci 6:307–335
Klincewicz JG (2002) Enumeration and search procedures for a hub location problem with economies of scale. Ann Oper Res 110:107–122
Köksalan M, Soylu B (2010) Bicriteria p-hub location problems and evolutionary algorithms. INFORMS J Comput 22:528–542
Labbé M, Yaman H (2004) Projecting the flow variables for hub location problems. Networks 44:84–93
Labbé M, Yaman H (2008) Solving the hub location problem in a start-start network. Networks 51:19–33
Labbé M, Yaman H, Gourdin É (2005) A branch and cut algorithm for hub location problems with single assignment. Math Program 102:371–405
Laporte G (2009) Fifty years of vehicle routing. Trans Sci 43:408–416
Lee C-H, Ro H-B, Tcha D-W (1993) Topological design of a two-level network with ring-star configuration. Comput Oper Res 20:625–637
Liang H (2013) The hardness and approximation of the star p-hub center problem. Oper Res Lett 41:138–141
Lin C-C, Lee S-C (2010) The competition game on hub network design. Transp Res B Methodol 44:618–629
Lowe TJ, Sim T (2013) The hub covering flow problem. J Oper Res Soc 64:973–981
Luer-Villagra A, Marianov V (2013) A competitive hub location and pricing problem. Eur J Oper Res 231:734–744
Marianov V, Serra D, ReVelle, CS (1999) Location of hubs in a competitive environment. Eur J Oper Res 114:363–371
Marianov V, Serra D (2003) Location models for airline hubs behaving as M/D/c queues. Comput Oper Res 30:983–1003
Marić M, Stanimirović Z, Stanojević P (2013) An efficient memetic algorithm for the uncapacitated single allocation hub location problem. Soft Comput 17:445–466
Marín A (2005a) Uncapacitated Euclidean hub location: Strengthened formulation, new facets and a relax-and-cut algorithm. J Glob Optim 33:393–422
Marín A (2005b) Formulating and solving splittable capacitated multiple allocation hub location problems. Comput Oper Res 32:3093–3109
Martins de Sá E, Contreras I, Cordeau J-F (2014) Exact and heuristic algorithms for the design of hub networks with multiple lines. Submitted to Eur J Oper Res
Martins de Sá E, Contreras I, Cordeau J-F, de Camargo RS, de Miranda R (2015) The hub line location problem. Transp Sci, forthcoming
Martins de Sá E, de Camargo RS, de Miranda R (2013) An improved Benders decomposition algorithm for the tree of hubs location problem. Eur J Oper Res 226:185–202
Meng Q, Wang X (2011) Intermodal hub-and-spoke network design: Incorporating multiple stakeholders and multi-type containers. Transp Res B Methodol 45:724–742
Meyer T, Ernst AT, Krishnamoorthy M (2009) A 2-phase algorithm for solving the single allocation p-hub center problem. Comput Oper Res 36:3143–3151
Mirzaghafour F (2013) Modular hub location problems. Msc thesis, Concordia University, Montreal, Canada
Nagi G, Salhi S (1998) The many-to-many location-routing problem. TOP 6:261–275
O'Kelly ME (1986a) The location of interacting hub facilities. Transp Sci 20:92–106
O'Kelly ME (1986b) Activity levels at hub facilities in interacting networks. Geogr Anal 18:343–356

O'Kelly ME (1987) A quadratic integer program for the location of interacting hub facilities. Eur J Oper Res 32:393–404
O'Kelly ME (1992) Hub facility location with fixed costs. Pap Reg Sci 20:293–306
O'Kelly ME (2012) Fuel burn and environmental implications of airline hub networks. Transp Res D-Tr E 17:555–567
O'Kelly ME, Bryan DL (1998) Hub location with flow economies of scale. Transp Res B Methodol 32:605–616
O'Kelly ME, Miller HJ (1991) Solution strategies for the single facility minimax hub location problem. Pap Reg Sci 70:367–380
O'Kelly ME, Miller HJ (1994) The hub network design problem: A review and synthesis. J Transp Geogr 2:31–40
Pirkul H, Schilling DA (1998) An efficient procedure for designing single allocation hub and spoke systems. Manage Sci 44:235–242
Puerto J, Ramos AB, Rodriguez-Chia AM (2011) Single-allocation ordered median hub location problems Comput Oper Res 38:559–570
Racunica I, Wynter L (2005) Optimal location of intermodal freight hubs. Transp Res B Methodol 39:453–477
Rieck J, Ehrenberg C, Zimmermann J (2014) Many-to-many location-routing with inter-hub transport and multi-commodity pickup-and-delivery. Eur J Oper Res 236:863–878
Rodríguez-Martín I, Salazar-González JJ (2008) Solving a capacitated hub location problem. Eur J Oper Res 184:468–479
Rodríguez-Martín I, Salazar-González J-J, Yaman H (2014) A branch-and-cut algorithm for the hub location and routing problem. Comput Oper Res 50:161–174.
Saberi M, Mahmassani HS (2013) Modeling the airline hub location and optimal market problems with continuous approximation techniques. J Transp Geogr 30:68–76
Saboury A, Ghaffari-Nasab N, Barzinpour F, Jabalameli MS (2013) Applying two efficient hybrid heuristics for hub location problem with fully interconnected backbone and access networks. Comput Oper Res 40:2493–2507
Saito H, Fujie T, Matsui T, Matuura S (2009) A study of the quadratic semi-assignment polytope. Discret Optim 6:37–50
Sasaki M (2005) Hub network design model in a competitive environment with flow threshold. J Oper Res Soc Jpn 48:158–171
Sasaki M, Campbell JF, Ernst AT, Krishnamoorthy M (2009) Hub arc location with competition. Technical Report NANZAN-TR-2009-02
Sasaki M, Fukushima M (2001) Stackelberg hub location problem. J Oper Res Soc Jpn 44:390–405
Sim T, Lowe TJ, Thomas BW (2009) The stochastic p-hub center problem with service-level constraints. Comput Oper Res 36:3166–3177
Skorin-Kapov D (1998) Hub network games. Networks 31:293–302
Skorin-Kapov D, Skorin-Kapov J, O'Kelly ME (1997) Tight linear programming relaxations of uncapacitated p-hub median problems. Eur J Oper Res 94:582–593
Sohn J, Park S (1997) A linear program for the two-hub location problem. Eur J Oper Res 100:617–622
Sohn J, Park S (2000) The single allocation problem in the interacting three-hub network. Networks 35:17–25
Sung CS, Jin HW (2001) Dual-based approach for a hub network design problem under non-restrictive policy. Eur J Oper Res 132:88–105
Thomadsen T, Larsen J (2007) A hub location problem with fully interconnected backbone and access networks. Comput Oper Res 34:2520–2531
Wagner B (2008a) Model formulations for hub covering problems. J Oper Res Soc 59:932–938
Wagner B (2008b) A note on location of hubs in a competitive environment. Eur J Oper Res 184:57–62
Wieberneit N (2008) Service network design for freight transportation: A review. OR Spectr 30:77–112

Yaman H (2008) Star p-hub median problem with modular arc capacities. Comput Oper Res 35:3009–3019

Yaman H (2009) The hierarchical hub median problem with single assignment. Transp Res B Methodol 43:643–658

Yaman H, Carello G (2005) Solving the hub location problem with modular link capacities. Comput Oper Res 32:3227–3245

Yaman H, Elloumi S (2012) Star p-hub center problem and star p-hub median problem with bounded path lengths. Comput Oper Res 39:2725–2732

Yaman H, Kara BY, Tansel BÇ (2007) The latest arrival hub location problem for cargo delivery systems with stopovers. Transp Res B Methodol 41:906–919

Yoon MG, Current JR (2008) The hub location and network design problem with fixed and variable arc costs: formulation and dual-based solution heuristic. J Oper Res Soc 59:80–89

Zanjirani Farahani R, Hekmatfar M, Arabani AB, Nikbakhsh E (2013) Hub location problems: A review of models, classification, solution techniques, and applications. Comput Ind Eng 64:1096–1109

Chapter 13
The Quadratic Assignment Problem

Zvi Drezner

Abstract The quadratic assignment problem is reviewed in this chapter. Weights between pairs of facilities and distances between the same number of locations are given. The problem is to find the assignment of facilities to locations that minimizes the weighted sum of distances. This problem is considered to be one of the most difficult combinatorial optimization problems. The construction of efficient solution algorithms (exact or heuristic) is challenging and has been extensively investigated by the communities working in Operations Research/Management Science, Industrial Engineering, or Computer Science. Examples of applications are given, the related layout problem is briefly described, exact and heuristic solution algorithms are reviewed, and a list of test problem instances and results are reported.

Keywords Exact methods • Metaheuristics • Quadratic assignment

13.1 Introduction

The quadratic assignment problem (QAP) is considered one of the most difficult optimization problems to solve optimally. The QAP is a combinatorial optimization problem stated for the first time by Koopmans and Beckmann (1957). Early papers on the subject include Gilmore (1962), Pierce and Crowston (1971), Lawler (1973), and Love and Wong (1976). The problem is defined as follows. A set of n possible sites are given and n facilities are to be located on these sites, one facility at a site. Let c_{ij} be the cost per unit distance between facilities i and j and d_{ij} be the distance between sites i and j. The cost f to be minimized over all possible permutations, calculated for an assignment of facility i to site $p(i)$ for $i = 1, \ldots, n$, is:

$$f = \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} d_{p(i)p(j)} \tag{13.1}$$

Z. Drezner (✉)
California State University, Fullerton, CA 92834, USA
e-mail: zdrezner@fullerton.edu

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_13

There are $n!$ possible permutations. The optimal solution is the best such permutation. Note that adding a constant to all c_{ij} does not change the solution. The objective function is increased by a constant (the common weight increase multiplied by the sum of the distances). Therefore, if there are negative weights, a constant can be added to all weights so that all weights are positive if this is required for a solution procedure.

The original formulation of the QAP by Koopmans and Beckmann (1957) is based on defining n^2 binary variables so that $x_{ij} = 1$ if facility i is located at site j and $x_{ij} = 0$ otherwise. The problem is then:

$$\text{minimize} \quad \left\{ f = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{r=1}^{n} \sum_{s=1}^{n} c_{ij} d_{rs} x_{ir} x_{js} \right\} \tag{13.2}$$

subject to

$$\sum_{i=1}^{n} x_{ij} = 1, \qquad j = 1, \ldots, n,$$

$$\sum_{j=1}^{n} x_{ij} = 1, \qquad i = 1, \ldots, n,$$

$$x_{ij} \in \{0, 1\}, \qquad i, j = 1, \ldots, n,$$

hence the name "Quadratic Assignment Problem". The constraints are identical to those of the linear assignment problem (Burkard and Cela 1999) but the objective function is quadratic rather than linear.

The QAP was proven to be NP-hard by Sahni and Gonzalez (1976). Even obtaining an ε-approximation for a given $\varepsilon > 0$ cannot be done in polynomial time unless P=NP.

Reviews of the quadratic assignment problem include Burkard (1990, 2013), Cela (1998), Rendl (2002), Taillard (1995), Drezner et al. (2005), Drezner (2008a), Drezner and Misevičius (2013), and Loiola et al. (2007).

The web site QAPLIB (http://www.seas.upenn.edu/qaplib) includes comprehensive and up to date information on the quadratic assignment problem such as research papers and solution results of test instances. There are sets of problems for which the optimal solution is known by design. Two of them are reported in Li and Pardalos (1992) and Drezner et al. (2005).

13.2 Applications

There are many applications that can be formulated and solved as quadratic assignment problems. Examples of such applications are listed here.

1. Office assignment. There are n planned offices, and n employees or special equipment to be assigned to them. There are several possible interpretations for the interaction weights c_{ij} (Drezner 1975, 1980).
   (a) The interaction between any two employees (c_{ij}) and the physical distance between any two offices (d_{ij}) are known. The problem is to assign employees to offices such that those who interact extensively are as close as possible to one another. The objective is to find the best assignment of employees to offices so that the sum of the products of the interactions and the distances is minimized.
   (b) The weight c_{ij} is the probability that a customer of this complex needs to visit both offices i and j in the same visit to complete the service. The objective in this case is to minimize the total distance an average customer needs to walk between offices.
   (c) In a hospital setting (e.g., Elshafei 1977 and Hahn and Krarup 2001) the offices represent different types of specific purpose rooms and the interaction is the probability that a patient needs the service of two different rooms.
2. Planning a complex of buildings. For example, Dickey and Hopkins (1972) explore campus building arrangement; Drezner (1980) explores the building arrangement of a military base. Most pairs of buildings have a positive interaction. Some pairs of buildings may have zero interaction whereas others may have a negative interaction (such as in planning a military base, top secret intelligence offices should be as far as possible from the cafeteria or other frequently visited offices).
3. The wiring problem of an electronic board or the construction of a computer chip was suggested by Steinberg (1961). The total wiring distance between components that send signals to one another has to be minimized.
4. Planning a keyboard of 26 letters was suggested by Burkard and Offermann (1977). The interaction c_{ij} is the probability of typing letter i following or preceding letter j. The distances between the letter-keys on the keyboard are to be considered. Different languages may suggest different key configurations even for the same letters.
5. The problem of finding the tightest cluster was suggested by Drezner (2006). Consider n objects (such as points in the plane or nodes of a network) with a given distance between every pair of points. We wish to find a cluster of m points which minimizes the total distance between all pairs of points in the cluster. This cluster can be interpreted as the "tightest" cluster of m points. We define the interaction matrix $\{c_{ij}\}$ as $c_{ij} = 1$ for $i, j \leq m$ and $c_{ij} = 0$ otherwise. Every permutation of the points defines the selected group as the first m points of the permutation and the value of the objective function is the sum of all the distances among the selected group members. For this problem there are $m!$ equivalent optimal permutations.

6. The grey pattern problem instances were suggested by Taillard (1995). It is based on a rectangle of dimensions n_1 by n_2. A grey pattern of m black points is selected from the $n = n_1 \times n_2$ points in the rectangle while the rest of the points remain white. This forms a "grey pattern" of density m/n. The objective is to have a grey pattern where the black points are distributed as uniformly as possible. This objective is achieved by defining a distance between pairs of points according to some rule. The interaction matrix $\{c_{ij}\}$ is the same as that of the tightest cluster formulation because only distances between the m "black" points are counted in the objective function. For more details see Taillard (1995) and Drezner (2006).
7. The turbine balancing problem was suggested by Laporte and Mercure (1988). Consider the manufacturing of a turbine engine, such as a hydro turbine or a jet engine, with n blades. The blades are inserted into equally spaced slots. To properly function, the turbine must be balanced. If all blades are identical, the turbine engine is balanced. In reality, there are slight variations in the weights of different blades, therefore the turbine is not perfectly balanced. Suppose that the weights are designed to be 5 kg each and the variations across blades are in the order of magnitude of milligrams. The problem is to find the "correct" assignment of blades into slots so that the turbine will be as balanced as possible. Let δ_i be the deviation of blade i from the target weight. The objective function is to find the permutation $p(i)$ $i = 1, \ldots, n$ that minimizes

$$\left\{\sum_{i=1}^{n} \delta_i \cos\left(\frac{2\pi p(i)}{n}\right)\right\}^2 + \left\{\sum_{i=1}^{n} \delta_i \sin\left(\frac{2\pi p(i)}{n}\right)\right\}^2.$$

   Following some algebraic manipulations the objective is equivalent to minimizing

$$\sum_{i \neq j=1}^{n} -\delta_i \delta_j \sin^2 \frac{\pi(p(i) - p(j))}{n}.$$

   The weights are $c_{ij} = -\delta_i \delta_j$ and the distances are $d_{ij} = \sin^2 \frac{\pi(i-j)}{n}$. The weights can be either positive or negative.
8. The arrangement of Microarray layouts was suggested by de Carvalho Jr and Rahmann (2006). The engineering component of this problem is quite complicated. The production of commercial DNA microarrays is based on a light-directed chemical synthesis driven by a set of masks or micromirror arrays. Because of the natural properties of light and the ever shrinking feature sizes, the arrangement of the probes on the chip and the order in which their nucleotides are synthesized play an important role on the quality of the final product. The reader is referred to de Carvalho Jr and Rahmann (2006) for complete information.

9. Configuration of a large airport (Drezner et al. 2005). Many airports have several terminals arranged in a partial star shape. Travel from one gate to another in a different terminal requires passengers to go first to the center and then to the other terminal. The weight c_{ij} is the probability that a customer has a connecting flight involving gates i and j.
10. Zoning a forest for different uses was proposed by Bos (1993). Land of a particular suitability and location has to be assigned land use objectives in such a way that the highest value is derived from zoning.
11. Scheduling parallel production lines proposed by Geoffrion and Graves (1976). Production orders for a number of products must be scheduled on a number of similar production lines so as to minimize the sum of product-dependent changeover costs, production costs, and time-constraint penalties.
12. Assigning runners in a relay team. Heffley (1977) observed that, for example, in a four person relay swimming competition each swimmer swims in a different style. For each swimmer the time for each style is known. The coach needs to select four swimmers, one for each style, such that the total time of all four swimmers is minimized. This leads to a linear assignment problem. However, in a runners relay when a baton needs to be transferred from one runner to the next, the transfer time depends on the runner handing the baton and the one receiving it. Suppose that the relay spans n runners and the problem is which runner to assign to the first position, which one to the second, and so on. Let us assume that run times following the transfer of the baton do not depend on the position. Therefore, total run time depends on the total baton transfer times. The cost matrix c_{ij} is the baton transfer time from runner i to runner j. The distances are all zeroes except that $d_{i,i+1} = 1$ for $i = 1, \ldots, n-1$. The objective function of the resulting quadratic assignment problem is the sum of all times of baton transfers.

13.3 The Layout Problem

The quadratic assignment problem seeks a permutation of equally sized facilities to a given equally numbered set of locations. A similar situation is formulated as the layout problem (Francis et al. 1992) where the locations of different sized facilities are sought. Applications of such a problem are a floor layout of a plant, the dashboard of an airplane instruments, layout of facilities, such as planed buildings, in an area where the locations of the various facilities are flexible but an interaction matrix representing the desirability of one facility to be close to another one is given. The QAP can be viewed as a discrete location problem because the potential locations for the facilities are given. In the layout problem the locations for the facilities are not restricted to a given set of locations. Drezner (1980) suggested that the facilities are circles of a given radius that cannot intersect but can be freely located on the plane to minimize an expression similar to (13.1) where $d_{p(i)p(j)}$ is

replaced by the distance between the centers of circles i and j. The solution method is termed DISCON (dispersion concentration). In the dispersion phase all facilities are put very close to one another and an explosion like the "big bang" disperses them while they are attracted to one another by their weights as springs. At the end of the process the circles are far from one another and in the concentration phase they are moved back still attracted to one another by the weights. The solution for this layout formulation of non-intersecting circles can be bounded or unbounded (when some the weights are negative). Drezner (1975, 2010) analyzed the issue of when the solution is bounded or unbounded, and formulated necessary conditions and sufficient conditions on the set of weights to determine whether the solution to this formulation is bounded or unbounded. A different approach to heuristically solve this problem by replacing the dispersion phase with the eigenvectors associated with the second and third smallest eigenvalues of a certain matrix based on the weights, was suggested by Drezner (1987). Armour and Buffa (1963) defined basic square shaped "building blocks" and each facility consists of a given number of building blocks and the shape of the facilities can be formed by having control over the configuration of the set of building blocks associated with each facility.

13.4 Extensions

The three-index assignment problem (three-dimensional AP or 3AP), first suggested by Pierskalla (1967, 1968), is based on weights and distances defined by three indices and the minimization requires two permutations rather than one.

The Generalized Quadratic Assignment problem (GQAP) was introduced by Lee and Ma (2005). In this formulation the number of facilities is not necessarily equal to the number of sites. Each site has a limited capacity to accommodate facilities. The GQAP reduces to the standard QAP when the number of facilities is equal to the number of sites, and the capacity of each site is one. Solution algorithms for the GQAP can be found in Cordeau et al. (2006) and Hahn et al. (2008).

13.5 Exact Solution Algorithms

Designing an exact algorithm for solving the QAP is very difficult. Recently, Fischetti et al. (2012) and Nyberg and Westerlund (2012) reformulated the problem in ways that non-linear programming software can be applied to solve such problems with limited success.

Even designing an effective lower bound to be used in a branch-and-bound algorithm is not easy. The first lower bound was developed by Gilmore (1962) and Lawler (1973). A linear programming based lower bound was suggested by

Resende et al. (1995). A quadratic programming based lower bound was proposed by Anstreicher and Brixius (2001) who reported in a follow-up paper (Anstreicher et al. 2002) the optimal solution of the Nug30 (Nugent et al. 1968) instance. The process was run in parallel on hundreds of computers that would take about 7 years on a single computer. The "Reformulation-Linearization Technique" (RLT) was developed by Sherali and Adams (1990, 1998) and utilized by Hahn and Grant (1998) as a Level-1 LRT. Later it was extended to Level-2 LRT by Adams et al. (2007), and Level-3 by Hahn et al. (2012). Other lower bounds include the Level-2 RLT interior point bound by Ramakrishnan et al. (2002), the SDP bound by Roupin (2004), the lift-and-project SDP bound by Burer and Vandenbussche (2006), and the bundle method bound by Rendl and Sotirov (2007).

Most of the problems solved optimally are based on no more than 30 facilities. Nyström (1999) reported the optimal solution of the $n = 36$ (Steinberg 1961) problem. Drezner et al. (2005) proposed problems whose optimal solution is known and problems with up to 72 facilities are optimally solved. The special structure of grey pattern problems enables more efficient solution algorithms. The $n = 64$ grey pattern problem of uniformly placing $m = 13$ black points in a 64 points square (termed Tai64c) was optimally solved by Drezner (2006) in about 2 h of computer time. Drezner (2006) also optimally solved $n = 256$ problems with $3 \leq m \leq 8$ black points. The Tai64c problem was also solved later by Fischetti et al. (2012) in about 5 h, and by Nyberg and Westerlund (2012) in about 50 h. Drezner et al. (2014) developed a more efficient branch-and-bound approach which optimally solved the Tai64c problem in about 15 s of computer time.

13.6 Heuristic Solution Algorithms

Optimal algorithms can solve relatively small problems. Consequently, considerable effort has been devoted to constructing heuristic algorithms.

The first heuristic approaches based on a descent type heuristic of checking some or all exchanges between facilities were proposed by Gilmore (1962), CRAFT (Buffa et al. 1962) and Hillier and Connors (1966). Nugent et al. (1968) suggested a biased sampling of exchanges rather than checking all of them. All exchanges between pairs of facilities define a "neighborhood" of solutions which serves as a basis for more recent metaheuristic algorithms.

In order to calculate all the values of the objective function in the neighborhood, $n(n-1)/2$ possible pair exchanges need to be evaluated. Evaluating each value directly by (13.1) requires $O(n^2)$ time leading to a total of $O(n^4)$ time. Burkard and Rendl (1984) suggested a short cut that we present for symmetric problems with zero diagonal (i.e., the cost between a facility and itself, and the distance between the same two locations is zero). Note, however, that this can be easily generalized to non symmetric problems. Let Δf_{rs} be the change in the cost f, calculated by

Eq. (13.1), by exchanging the sites of facilities r and s. There are $n(n-1)/2$ such values. It can be easily verified that:

$$\begin{aligned}\Delta f_{rs} &= 2\sum_{i=1}^{n}\left\{c_{ir}\left[d_{p(i)p(s)}-d_{(p(i)p(r)}\right]+c_{is}\left[d_{p(i)p(r)}-d_{(p(i)p(s)}\right]\right\}\\ &= 2\sum_{i=1}^{n}\left\{\left[c_{ir}-c_{is}\right]\left[d_{p(i)p(r)}-d_{p(i)p(s)}\right]\right\}. \end{aligned} \tag{13.3}$$

Calculating Δf_{rs} by (13.3) requires only $O(n)$ time rather than $O(n^2)$ time.

Taillard (1991) points to yet a faster formula for calculating Δf_{rs}. Let $\Delta_{uv} f_{rs}$ be the variation in the value of the objective function corresponding to exchanging u and v given that the previous exchange involved r and s, and assuming u and v different from r and s. This change in the value of the objective function can be calculated in $O(1)$ time starting from the second iteration. The formula is based on Δf_{uv} (the change in the value of the objective function from the previous permutation by exchanging the pair uv). Therefore, one needs to keep all the values of Δf_{ij} for all pairs i, j. Saving these values requires $O(n^2)$ time for each evaluation of all pair exchanges. It can be easily verified by (13.3) that:

$$\Delta_{uv} f_{rs} = \Delta f_{uv} + 2\left[c_{su}+c_{rv}-c_{sv}-c_{ru}\right]\left[d_{p(s)p(u)}+d_{p(r)p(v)}-d_{p(s)p(v)}-d_{p(r)p(u)}\right],$$

which is calculated in $O(1)$ time. Note that only $2n-3$ pairs are not mutually exclusive and formula (13.3) can be used in these cases to evaluate Δf_{rs}. Therefore, evaluating the change in the value of the objective function for all $n(n-1)/2$ possible pair exchanges requires $O(n^2)$ time rather than $O(n^4)$ time.

Many metaheuristic approaches have been suggested for solving the QAP. For example, simulated annealing was proposed by Wilhelm and Ward (1987), Connoly (1990), Misevičius (2003), ant colonies was investigated by Gambardella et al. (1999), Taillard (1998, 2000), Talbi et al. (2001), migrating birds optimization was suggested by Duman et al. (2012), scatter search was implemented by Cung et al. (1997), simulated jumping was proposed by Amin (1999) and a greedy randomized adaptive search procedure was designed by Li et al. (1994) and Oliveira et al. (2004).

Various versions of the metaheuristic tabu search (Glover 1977, 1986; Glover and Laguna 1997) were suggested for the solution of the QAP. Skorin-Kapov (1990) proposed the first application of tabu search followed by Taillard (1991) who proposed the Robust Tabu. The latter remained as the most powerful heuristic approach for many years. The tabu list is set to contain pairs of facility-site (i.e., there are n^2 possible entries in the tabu list). There is a short term and long term tabu memory.

Short Term Memory: When a facility is removed from a site, the iteration number is recorded. An exchange between two facilities is not allowed (unless the objective function is better than the best one found so far) if both facilities move

back to a site they were removed from in the last t iterations. The best solutions found by Taillard (1991) were obtained considering the tabu tenure t randomly generated in $[0.9n, 1.1n]$ in every iteration.

Long Term Memory: Every iteration after x iterations (for example, $x = 3n^2$): if there is an exchange between two facilities such that one facility moves to a site it was never there in the last x iterations, such an exchange preempts any other exchange and is executed. The long term memory serves as a diversification of the tabu search.

The robust tabu search approach proposed by Taillard (1991) was improved by Drezner (2008a) who suggested a small change of the tabu tenure by selecting it randomly in the range $[0.2n, 1.8n]$ (rather than $[0.9n, 1.1n]$) in each iteration.

The "reactive tabu search" was proposed by Battiti and Tecchiolli (1994). The "concentric tabu search" was proposed in Drezner (2002) and extended in Drezner (2005b). Various tabu searches approaches were proposed and computationally tested in Misevičius and Blonskis (2005) and Misevičius et al. (2006). The iterated tabu search was proposed by Misevičius (2012).

Many versions of genetic algorithms which are inspired by biological evolution and survival of the fittest (Holland 1975; Drezner and Drezner 2005) have also been proposed for the QAP. The first two papers are due to Fleurent and Ferland (1994) and Tate and Smith (1995). Other works investigating this type of algorithm include Misevičius (2008), Wu and Ji (2008), Ahuja et al. (2000).

The most successful heuristic algorithms seem to be the hybrid genetic algorithms (Drezner 2003, 2008a; Misevičius 2004, 2005; Misevičius and Rubliauskas 2009; Misevičius et al. 2009; Misevičius and Guogis 2012). For a review of the application of such heuristic algorithms for the solution of the QAP see Drezner and Misevičius (2013). Hybrid genetic algorithms apply a local search on the generated offspring before considering its inclusion into the population. Two parameters are given: the population size P and the number of generations G. A specific local search, such as tabu search, is selected. The general framework of a simple hybrid genetic algorithms is the following:

1. A starting population of size P is randomly selected, and the local search heuristic is applied on each starting population member. The current generation number is set to $g = 1$.
2. Two population members are randomly selected and merged by a crossover operator to produce an offspring.
3. The local search heuristic is applied to the merged solution, possibly improving it.
4. If the value of the offspring's objective function is not better than the worst population member's objective function, the offspring is ignored and go to Step 5. Else,

   (a) If the offspring is identical to an existing population member, it is ignored. Go to Step 5.

(b) If the offspring is different from all population members, the offspring replaces the worst population member.

5. Set $g = g + 1$. If $g \leq G$ go to Step 2.
6. Otherwise ($g = G + 1$), stop with the best population member as the final solution.

Many modifications for such a simple framework have been proposed. For example, Schaffer et al. (1989), Drezner and Marcoulides (2003), Fox and McMahon (1991), Wu and Ji (2008), Drezner and Drezner (2006), Misevičius (2008), Drezner (2005a), and Cantú-Paz (2001).

An important component for the success of genetic algorithms is the crossover operator. The following crossover operator, suggested by Drezner (2003), exploits the structure of the problem and works well when distances are life-like distances such as Manhattan or Euclidean distances. When problems are randomly generated like the Taia instances (Taillard 1991) it may not work well. Two parents are to be merged to produce an offspring. In Drezner (2003) the following crossover operator is repeated for all n facilities considered separately as pivot sites. The best merged offspring is selected for the local search heuristic. The merge of the two selected parents for one pivot site is as follows:

1. The median distance from the pivot site to all sites is calculated.
2. A site that is closer than the median to the pivot site is assigned the facility located there in the first parent.
3. All other sites are assigned a facility from the second parent.
4. It is possible that some facilities are assigned twice. The same number of facilities are not assigned at all. Therefore,

   (a) Go over all the facilities from left to right and create a list of unassigned facilities.
   (b) Find all facilities that are assigned twice, and replace the site that is farther than the median (i.e., from the second parent) with a facility that is not assigned at all.

Drezner (2003) applied the concentric tabu search (Drezner 2002) as the local search heuristic. Drezner (2008a) found that the modified robust tabu (applying only the short term memory) performed better as a local search heuristic.

13.7 Test Problem Instances

There are many test problems instances listed in the web-site QAPLIB. Commonly used sets of test problem instances for evaluating the effectiveness of algorithms are listed in Table 13.1.

Note that if at least one of the sets of weights or distances is symmetric, the problem can be formulated as a symmetric problem. Suppose that $d_{ij} = d_{ji}$. Define

Table 13.1 Problem instances

Name	Range	Reference	Comments
Nug	12–30	Nugent et al. (1968)	All optimal solutions found
Taia	12–100	Taillard (1991)	Random weights & distances
Taib	12–150	Taillard (1991)	Real-life like
Taic	64–256	Taillard (1995)	Grey pattern problems
Taie	27–343	Drezner et al. (2005)	Large airport configuration
Dre	30–90	Drezner et al. (2005)	Known optimum
Tho	30–150	Thonemann and Bölte (1994)	
Sko	42–100	Skorin-Kapov (1990)	
BL,CI	36–144	de Carvalho Jr and Rahmann (2006)	Non-symmetric weights

the weights as $c'_{ij} = c_{ij} + c_{ji}$ and the problem becomes symmetric with double the value of the objective function. Symmetric problems can be solved in about half the time because most of the calculations are not replicated twice unnecessarily.

The best known solutions to some of the bigger problems are listed in Table 13.2. The results for the Sko and Tho problems are taken from Drezner (2008a) and the results for the other problems are taken from Drezner and Misevičius (2013).

The best known solution values for the grey pattern problems for $n = 256$ and $3 \leq m \leq 128$ are available in Drezner (2006, 2008b), Misevičius (2011) and reported in Table 13.3. For $m > 128$ the solution is obtained by exchanging between the locations of black and white points. The results for $3 \leq m \leq 8$ are proven optimal in Drezner (2006). The original Tai256c (Taillard 1995) is defined for $m = 92$. Misevičius (2011) also reports the best known solutions for various value of m for the $n = 64$ grey pattern problems and Misevičius et al. (2013) define the largest QAP test problems using a grey pattern with $n = 1024$ points and report the best known solutions for these problems up to $m = 512$.

Two pictorial solutions to the grey pattern problems reported in Drezner et al. (2014) illustrate the grey pattern results. First we present in Fig. 13.1 the optimal configuration of locating 20 black points in a square of dimensions 8×8 replicated 9 times. This problem was optimally solved in Drezner et al. (2014) in less than six and a half hours. The configuration shows groups of 5 points in a "V" shape alternating up and down. The other (heuristic, but probably optimal) solution found in a few papers is for locating 64 black points in a 16×16 square. The pattern is depicted in Fig. 13.2. It is interesting that this pattern is very close to an hexagonal pattern that is known to be the densest packing, see Coxeter (1973) and Hilbert and Cohn-Vossen (1956). The distance to the four points in a diagonal direction is $\sqrt{5} = 2.236$ while the distance to the two points on the left and on the right is 2. In a hexagonal pattern these six distances are the same and therefore this pattern can be viewed as "hexagonal-like". The hexagonal pattern is also preferred to a square pattern for a large number of points in many location problems (Drezner and Suzuki 2010; Drezner and Zemel 1992; Okabe and Suzuki 1987; Suzuki and Drezner 1996; Suzuki and Okabe 1995; Szabo et al. 2007).

Table 13.2 Best known results

Instance	Result	Instance	Result	Instance	Result	Instance	Result
Sko49	23,386	Tho40	240,516	Tai25b	344,355,646	BL81	7,532
Sko56	34,458	Tho150	8,133,398	Tai30b	637,117,113	BL100	9,264
Sko64	48,498	Tai20a	703,482	Tai35b	283,315,445	BL121	11,400
Sko72	66,256	Tai25a	1,167,256	Tai40b	637,250,948	BL144	13,460
Sko81	90,998	Tai30a	1,818,146	Tai50b	458,821,517	CI36	168,611,971
Sko90	115,534	Tai35a	2,422,002	Tai60b	608,215,054	CI49	236,355,034
Sko100a	152,002	Tai40a	3,139,370	Tai80b	818,415,043	CI64	325,671,035
Sko100b	153,890	Tai50a	4,938,796	Tai100b	1,185,996,137	CI81	427,447,820
Sko100c	147,862	Tai60a	7,205,962	Tai150b	498,896,643	CI100	523,146,366
Sko100d	149,576	Tai80a	13,499,184	BL36	3,296	CI121	653,409,588
Sko100e	149,150	Tai100a	21,052,466	BL49	4,548	CI144	794,811,636
Sko100f	149,036	Tai20b	122,455,319	BL64	5,988		

Table 13.3 Results for grey pattern problems ($n = 256$)

m	Result	m	Result	m	Result	m	Result	m	Result	m	Result
3	7,810	24	2,010,846	45	8,674,910	66	20,648,754	87	39,389,054	108	63,582,416
4	15,620	25	2,215,714	46	9,129,192	67	21,439,396	88	40,416,536	109	64,851,966
5	38,072	26	2,426,298	47	9,575,736	68	22,234,020	89	41,512,742	110	66,120,434
6	63,508	27	2,645,436	48	10,016,256	69	23,049,732	90	42,597,626	111	67,392,724
7	97,178	28	2,871,704	49	10,518,838	70	23,852,796	91	43,676,474	112	68,666,416
8	131,240	29	3,122,510	50	11,017,342	71	24,693,608	92	44,759,294	113	69,984,758
9	183,744	30	3,373,854	51	11,516,840	72	25,529,984	93	45,870,244	114	71,304,194
10	242,266	31	3,646,344	52	12,018,388	73	26,375,828	94	46,975,856	115	72,630,764
11	304,722	32	3,899,744	53	12,558,226	74	27,235,240	95	48,081,112	116	73,962,220
12	368,952	33	4,230,950	54	13,096,646	75	28,114,952	96	49,182,368	117	75,307,424
13	457,504	34	4,560,162	55	13,661,614	76	29,000,908	97	50,344,050	118	76,657,014
14	547,522	35	4,890,132	56	14,229,492	77	29,894,452	98	51,486,642	119	78,015,914
15	644,036	36	5,222,296	57	14,793,682	78	30,797,954	99	52,660,116	120	79,375,832
16	742,480	37	5,565,236	58	15,363,628	79	31,702,182	100	53,838,088	121	80,756,852
17	878,888	38	5,909,202	59	15,981,086	80	32,593,088	101	55,014,262	122	82,138,768
18	1,012,990	39	6,262,248	60	16,575,644	81	33,544,628	102	56,202,826	123	83,528,554
19	1,157,992	40	6,613,472	61	17,194,812	82	34,492,592	103	57,417,112	124	84,920,540
20	1,305,744	41	7,002,794	62	17,822,806	83	35,443,938	104	58,625,240	125	86,327,812
21	1,466,210	42	7,390,586	63	18,435,790	84	36,395,172	105	59,854,744	126	87,736,646
22	1,637,794	43	7,794,422	64	19,050,432	85	37,378,800	106	61,084,902	127	89,150,166
23	1,820,052	44	8,217,264	65	19,848,790	86	38,376,438	107	62,324,634	128	90,565,248

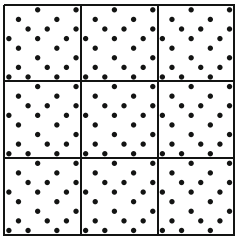


Fig. 13.1 Optimal configuration of 20 *black points* in an 8×8 square replicated 9 times

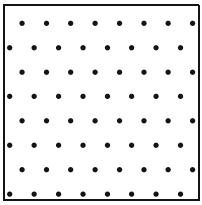


Fig. 13.2 *Grey patterns* of 64 points in a 16×16 square

13.8 Conclusions

The Quadratic Assignment Problem (QAP) is considered to be one of the most difficult combinatorial optimization problems. The problem was presented and applications were described. The related layout problem and two extensions were briefly presented. Exact and heuristic solution methods were listed and best known results for some test problems reported. We concluded with a depiction of two grey pattern problems which are a special case of the QAP.

Current exact algorithms can solve mostly problems of up to 30–40 facilities while heuristic algorithms require long run times to obtain reasonably good solutions. Research in developing more effective exact and heuristic algorithms will be very helpful and should be pursued.

References

Adams W, Guignard M, Hahn P, Hightower W (2007) A level-2 reformulation-linearization technique bound for the quadratic assignment problem. Eur J Oper Res 180:983–996

Ahuja R, Orlin J, Tiwari A (2000) A descent genetic algorithm for the quadratic assignment problem. Comput Oper Res 27:917–934

Amin S (1999) Simulated jumping. Ann Oper Res 84:23–38

Anstreicher K, Brixius N, Gaux JP, Linderoth J (2002) Solving large quadratic assignment problems on computational grids. Math Program 91:563–588

Anstreicher KM, Brixius NW (2001) A new bound for the quadratic assignment problem based on convex quadratic programming. Math Program 89:341–357

Armour GC, Buffa ES (1963) A heuristic algorithm and simulation approach to relative location of facilities. Manag Sci 9:294–309

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

Bos J (1993) Zoning in forest management: a quadratic assignment problem solved by simulated annealing. J Environ Manag 37:127–145
Buffa ES, Armour GC, Vollmann TE (1962) Allocating facilities with CRAFT. Harv Bus Rev 42:136–158
Burer S, Vandenbussche D (2006) Solving lift-and-project relaxations of binary integer programs. SIAM J Optimiz 16:726–750
Burkard R, Rendl F (1984) A thermodynamically motivated simulation procedure for combinatorial optimization problems. Eur J Oper Res 17:169–174
Burkard RE (1990) Locations with spatial interactions: the quadratic assignment problem. In: Mirchandani PB, Francis RL (eds) Discrete location theory. Wiley, New York, pp 387–437
Burkard RE, Cela E (1999) Linear assignment problems and extensions. In: Pardalos P, Du D-Z (eds) Handbook of combinatorial optimization. Springer, Dordrecht, pp 75–149
Burkard RE (2013) Quadratic assignment problems. In: Pardalos P, Du D-Z (eds) Handbook of combinatorial optimization, 2nd edn. Springer, New York, pp 2741–2814
Burkard RE, Offermann J (1977) Entwurf von schreibmaschinentastaturen mittels quadratischer zuordnungsprobleme. Math Method Oper Res 21:121–132
Cantú-Paz E (2001) Migration policies, selection pressure, and parallel evolutionary algorithms. J Heuristics 7:311–334
de Carvalho Jr SA, Rahmann S (2006) Microarray layout as a quadratic assignment problem. In: Huson D, Kohlbacher O, Lupas A, Nieselt K, Zell A (eds) Proceedings of the German conference on bioinformatics, vol 83. Gesellschaft für Informatik, Bonn, pp 11–20
Cela E (1998) The quadratic assignment problem: theory and algorithms. Kluwer Academic Publishers, Dordrecht
Connoly D (1990) An improved annealing scheme for the QAP. Eur J Oper Res 46:93–100
Cordeau JF, Gaudioso M, Laporte G, Moccia L (2006) A memetic heuristic for the generalized quadratic assignment problem. INFORMS J Comput 18:433–443
Coxeter HSM (1973) Regular polytopes. Dover Publications, New York
Cung VD, Mautor T, Michelon P, Tavares AI (1997) A scatter search based approach for the quadratic assignment problem. In: Proceedings of the IEEE international conference on evolutionary computation and evolutionary programming (ICEC'97), Indianapolis, pp 165–170
Dickey JW, Hopkins JW (1972) Campus building arrangement using topaz. Transp Res 6:59–68
Drezner T, Drezner Z (2005) Genetic algorithms: mimicking evolution and natural selection in optimization models. In: Bar-Cohen Y (ed) Biomimetics—biologically inspired technologies. CRC Press, Boca Raton, pp 157–175
Drezner T, Drezner Z (2006) Gender specific genetic algorithms. INFOR Inform Syst Oper Res 44:117–127
Drezner Z (1975) Problems in non-linear programming (the allocation problem). Ph.D. thesis, The Technion, Haifa
Drezner Z (1980) DISCON—a new method for the layout problem. Oper Res 28:1375–1384
Drezner Z (1987) A heuristic procedure for the layout of a large number of facilities. Manag Sci 33:907–915
Drezner Z (2002) A new heuristic for the quadratic assignment problem. J Appl Math Decis Sci 6:163–173
Drezner Z (2003) A new genetic algorithm for the quadratic assignment problem. INFORMS J Comput 15:320–330
Drezner Z (2005a) A distance based rule for removing population members in genetic algorithms. 4OR-Q J Oper Res 3:109–116
Drezner Z (2005b) The extended concentric tabu for the quadratic assignment problem. Eur J Oper Res 160:416–422
Drezner Z (2006) Finding a cluster of points and the grey pattern quadratic assignment problem. OR Spectr 28:417–436
Drezner Z (2008a) Extensive experiments with hybrid genetic algorithms for the solution of the quadratic assignment problem. Comput Oper Res 35:717–736

Drezner Z (2008b) Tabu search and hybrid genetic algorithms for quadratic assignment problems. In: Jaziri W (ed) Tabu search, in-tech, pp 89–108. Available free on: http://books.i-techonline.com

Drezner Z (2010) On the unboundedness of facility layout problems. Math Method Oper Res 72:205–216

Drezner Z, Marcoulides GA (2003) A distance-based selection of parents in genetic algorithms. In: Resende MGC, de Sousa JP (eds) Metaheuristics: computer decision-making. Kluwer Academic Publishers, Boston, pp 257–278

Drezner Z, Misevičius A (2013) Enhancing the performance of hybrid genetic algorithms by differential improvement. Comput Oper Res 40:1038–1046

Drezner Z, Suzuki A (2010) Covering continuous demand in the plane. J Oper Res Soc 61:878–881

Drezner Z, Zemel E (1992) Competitive location in the plane. Ann Oper Res 40:173–193

Drezner Z, Hahn PM, Taillard ÉD (2005) Recent advances for the quadratic assignment problem with special emphasis on instances that are difficult for meta-heuristic methods. Ann Oper Res 139:65–94

Drezner Z, Misevičius A, Palubeckis G (2014) Exact algorithms for the solution of the grey pattern quadratic assignment problem. In review

Duman E, Uysal M, Alkaya AF (2012) Migrating birds optimization: a new metaheuristic approach and its performance on quadratic assignment problem. Inform Sci 217:65–77

Elshafei AN (1977) Hospital layout as a quadratic assignment problem. Oper Res Q 28:167–179

Fischetti M, Monaci M, Salvagnin D (2012) Three ideas for the quadratic assignment problem. Oper Res 60:954–964

Fleurent C, Ferland J (1994) Genetic hybrids for the quadratic assignment problem. In: Pardalos P, Wolkowicz H (eds) Quadratic assignment and related problems, DIMACS series in discrete mathematics and theoretical computer science, vol 16. American Methematical Society, Providence, pp 173–187

Fox BR, McMahon MB (1991) Genetic operators for sequencing problems. In: Rawlins G (ed) Foundations of genetic algorithms. Morgan-Kaufmann, San Mateo, pp 284–300

Francis RL, McGinnis LF Jr, White JA (1992) Facility layout and location: an analytical approach, 2nd edn. Prentice Hall, Englewood Cliffs

Gambardella L, Taillard ÉD, Dorigo M (1999) Ant colonies for the quadratic assignment problem. J Oper Res Soc 50:167–176

Geoffrion AM, Graves GW (1976) Scheduling parallel production lines with changeover costs: practical application of a quadratic assignment/lp approa ch. Oper Res 24:595–610

Gilmore P (1962) Optimal and suboptimal algorithms for the quadratic assignment problem. J SIAM 10:305–313

Glover F (1977) Heuristics for integer programming using surrogate constraints. Decis Sci 8:156–166

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

Glover F, Laguna M (1997) Tabu search. Kluwer Academic Publishers, Boston

Hahn P, Grant T (1998) Lower bounds for the quadratic assignment problem based upon a dual formulation. Oper Res 46:912–922

Hahn P, Krarup J (2001) A hospital facility problem finally solved. J Intell Manuf 12:487–496

Hahn PM, Kim BJ, Guignard M, Smith JM, Zhu YR (2008) An algorithm for the generalized quadratic assignment problem. Comput Optim Appl 40:351–372

Hahn PM, Zhu YR, Guignard M, Hightower WL, Saltzman MJ (2012) A level-3 reformulation-linearization technique-based bound for the quadratic assignment problem. INFORMS J Comput 24:202–209

Heffley DR (1977) Assigning runners to a relay team. In: Ladany SP, Machol RE (eds) Optimal strategies in sports. North-Holland, Amsterdam, pp 169–171

Hilbert D, Cohn-Vossen S (1956) Geometry and the imagination (english translation of Anschauliche Geometrie, 1932). Chelsea Publishing Company, New York

Hillier FS, Connors MM (1966) Quadratic assignment problem algorithms and the location of indivisible facilities. Manag Sci 13:42–57

Holland JH (1975) Adaptation in natural and artificial systems. University of Michigan Press, Ann Arbor

Koopmans TC, Beckmann MJ (1957) Assignment problems and the location of economic activities. Econometrica 25:53–76

Laporte G, Mercure H (1988) Balancing hydraulic turbine runners: a quadratic assignment problem. Eur J Oper Res 35:378–381

Lawler EL (1973) Optimal sequencing of a single machine subject to precedence constraints. Manag Sci 19:544–546

Lee CG, Ma Z (2005) The generalized quadratic assignment problem. Department of Mechanical and Industrial Engineering, University of Toronto, MIE OR Technical Reports (TR2005-01)

Li Y, Pardalos PM (1992) Generating quadratic assignment test problems with known optimal permutations. Comput Optim Appl 1:163–184

Li Y, Pardalos PM, Resende MGC (1994) A greedy randomized adaptive search procedure for the quadratic assignment problem. In: Pardalos PM, Wolkowicz H (eds) Quadratic assignment and related problems, DIMACS series in discrete mathematics and theoretical computer science, vol 16, American Mathematical Society, Providence, pp 237–261

Loiola EM, de Abreu NMM, Boaventura-Netto PO, Hahn P, Querido T (2007) A survey for the quadratic assignment problem. Eur J Oper Res 176:657–690

Love RF, Wong JY (1976) Solving quadratic assignment problems with rectangular distances and integer programming. Nav Res Logist Q 23:623–627

Misevičius A (2003) A modified simulated annealing algorithm for the quadratic assignment problem. Informatica 14:497–514

Misevičius A (2011) Generation of grey patterns using an improved genetic-evolutionary algorithm: some new results. Inf Technol Control 40:330–343

Misevičius A (2012) An implementation of the iterated tabu search algorithm for the quadratic assignment problem. OR Spectr 34:665–690

Misevičius A, Blonskis J (2005) Experiments with tabu search for random quadratic assignment problems. Inf Technol Control 34:237–244

Misevičius A, Guogis E (2012) Computational study of four genetic algorithm variants for solving the quadratic assignment problem. In: Skersys T, Butkienè R, Butleris R (eds) Communications in computer and information science (CCIS). Proceedings of 18th international conference on information and software technologies ICIST 2012, vol 319. Springer, Berlin, pp 24–37

Misevičius A, Tomkevičius A, Karbauskas J (2006) Stagnation-protected tabu search variants for unstructured quadratic assignment problems. Inf Technol Control 35:363–370

Misevičius A, Guogis E, Stanevičienè E (2013) Computational algorithmic generation of high-quality colour patterns. In: Skersys T, Butkienè R, Butleris R (eds) Communications in computer and information science (CCIS). Proceedings of 19th international conference in information and software technologies ICIST 2013. Springer, Berlin, pp 285–296

Misevičius A (2004) An improved hybrid genetic algorithm: new results for the quadratic assignment problem. Knowl-Based Syst 17:65–73

Misevičius A (2005) A tabu search algorithm for the quadratic assignment problem. Comput Optim Appl 30:95–111

Misevičius A (2008) Restart-based genetic algorithm for the quadratic assignment problem. In: Bramer M, Coenen F, Petridis M (eds) Research and development in intelligent systems. Proceedings of AI-2008, the 28th SGAI international conference on innovative techniques and applications of artificial intelligence. Springer, London, pp 91–104

Misevičius A, Rubliauskas D (2009) Testing of hybrid genetic algorithms for structured quadratic assignment problems. Informatica 20:255–272

Misevičius A, Rubliauskas D, Barkauskas V (2009) Some further experiments with the genetic algorithm for the quadratic assignment problem. Inf Technol Control 38:325–332

Nugent C, Vollman T, Ruml T (1968) An experimental comparison of techniques for the assignment of facilities to locations. Oper Res 16:150–173

Nyberg A, Westerlund T (2012) A new exact discrete linear reformulation of the quadratic assignment problem. Eur J Oper Res 220:314–319
Nyström M (1999) Solving certain large instances of the quadratic assignment problem: Steinberg's examples. Technical report, California Institute of Technology. http://resolver.caltech.edu/CaltechCSTR:2001.010
Okabe A, Suzuki A (1987) Stability of spatial competition for a large number of firms on a bounded two-dimensional space. Environ Plann A 16:107–114
Oliveira CAS, Pardalos PM, Resende MGC (2004) GRASP with path-relinking for the quadratic assignment problem. In: Ribeiro CC, Martins SL (eds) Efficient and experimental algorithms. Springer, Berlin/Heidelberg, pp 237–261
Pierce JF, Crowston WB (1971) Tree-search algorithms for quadratic assignment problems. Nav Res Logist Q 18:1–36
Pierskalla WP (1967) The tri-substitution method for the three-dimensional assignment problem. CORS J 5:71–81
Pierskalla WP (1968) The multidimensional assignment problem. Oper Res 16:422–431
Ramakrishnan KG, Resende M, Ramachandran B, Pekny J (2002) Tight QAP bounds via linear programming. In: Pardalos PM, Migdalas A, Burkard R (eds) Combinatorial and global optimization. World Scientific Publishing, Singapore, pp 297–303
Rendl F (2002) The quadratic assignment problem. In: Drezner Z, Hamacher H (eds) Facility location: applications and theory. Springer, Berlin
Rendl F, Sotirov R (2007) Bounds for the quadratic assignment problem using the bundle method. Math Program 109:505–524
Resende M, Ramakrishnan K, Drezner Z (1995) Computational experiments with the lower bound for the quadratic assignment problem based on linear programming. Oper Res 43:781–791
Roupin F (2004) From linear to semidefinite programming: an algorithm to obtain semidefinite relaxations for bivalent quadratic problems. J Comb Optim 8:469–493
Sahni S, Gonzalez T (1976) P-complete approximation problems. J ACM 23:555–565
Schaffer JD, Caruana RA, Eshelman LJ (1989) A study of control parameters affecting online performance of genetic algorithms. In: Schaffer JD (ed) Proceedings of the 3rd international conference on genetic algorithms. Morgan Kaufmann, San Mateo, pp 51–60
Sherali HD, Adams WP (1990) A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. SIAM J Discret Math 3:411–430
Sherali HD, Adams WP (1998) A reformulation-linearization technique for solving discrete and continuous nonconvex problems. Springer, Berlin
Skorin-Kapov J (1990) Tabu search applied to the quadratic assignment problem. ORSA J Comput 2:33–45
Steinberg L (1961) The backboard wiring problem: a placement algorithm. SIAM Rev 3:37–50
Suzuki A, Drezner Z (1996) The p-center location problem in an area. Locat Sci 4:69–82
Suzuki A, Okabe A (1995) Using Voronoi diagrams. In: Drezner Z (ed) Facility location: a survey of applications and methods. Springer, New York, pp 103–118
Szabo PG, Markot M, Csendes T, Specht E (2007) New approaches to circle packing in a square: with program codes. Springer, New York
Taillard ÉD (1991) Robust tabu search for the quadratic assignment problem. Parallel Comput 17:443–455
Taillard ÉD (1995) Comparison of iterative searches for the quadratic assignment problem. Locat Sci 3:87–105
Taillard ÉD (1998) Fant: fast ant system. Technical report, Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale, iDSIA Technical Report IDSIA-46-98
Taillard ÉD (2000) An introduction to ant systems. In: Laguna M, González-Velarde J (eds) Computing tools for modeling, optimization and simulation. Wiley, New York, pp 131–144
Talbi EG, Roux O, Fonlupt C, Robillard D (2001) Parallel ant colonies for the quadratic assignment problem. Futur Gener Comput Syst 17:441–449
Tate D, Smith A (1995) A genetic approach to the quadratic assignment problem. Comput Oper Res 22:73–83

Thonemann U, Bölte A (1994) An improved simulated annealing algorithm for the quadratic assignment problem. Technical report, working paper, School of Business, Department of Production and Operations Research, University of Paderborn, Germany

Wilhelm M, Ward T (1987) Solving quadratic assignment problems by simulated annealing. IIE Trans 19:107–119

Wu Y, Ji P (2008) Solving the quadratic assignment problems by a genetic algorithm with a new replacement strategy. Int J Comput Intell 4:225–229

Chapter 14
Competitive Location Models

H.A. Eiselt, Vladimir Marianov, and Tammy Drezner

Abstract This chapter first provides a review of the foundations of competitive location models. It then traces subsequent developments through the decades under special consideration of customer behavior. After developing a general framework for customers' decision making, the main results are put into this framework. The conclusion outlines a number of areas, in which existing models can be refined and made more realistic.

Keywords Hotelling models • Nash equilibria • von Stackelberg solutions

14.1 The Basic Model: The First 50 Years

Competitive location models were first discussed by Hotelling (1929) in his seminal paper. It has spawned hundreds of contributions (for a summary until the early 1990s, see Eiselt et al. 1993) that investigate many different aspects of the basic model. A recent summary of Hotelling-style models was provided by Eiselt (2011), for details we refer to that work. This chapter will first introduce the basic model, followed by an outline of some of the main components of competitive location models. We then discuss the main aspects and types of consumer behavior, and then review the work on competitive location models under special consideration of customer behavior.

The basic model is easy to describe: consider a line segment, a so-called "linear market," which Hotelling referred to as "main street," along which customers are

H.A. Eiselt (✉)
Faculty of Business Administration, University of New Brunswick, Fredericton, NB, Canada E3B 5A3
e-mail: haeiselt@unb.ca

V. Marianov
Department of Electrical Engineering, Pontificia Universidad Católica de Chile, Santiago, Chile
e-mail: marianov@ing.puc.cl

T. Drezner
College of Business and Economics, California State University-Fullerton, Fullerton, CA 92834, USA
e-mail: tdrezner@fullerton.edu

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_14

uniformly distributed. (The often-mentioned "ice cream vendors on the beach" were actually introduced by Lösch 1954.) Each customer has a fixed and inelastic demand for a given homogeneous good. Duopolists are now attempting to independently enter the market, offering identical products. The competitors are profit maximizers, and they attempt to achieve their objective by determining their respective locations and prices; first both competitors choose their respective locations, followed by the simultaneous choice of prices. It is assumed that both competitors employ mill (or f.o.b.) pricing (a pricing policy in which customers pay a price set by the facility and take care of the transportation themselves) and that transportation costs between customers and facilities are linear. Customers will patronize the facility that offers the good for the lowest full price, i.e., the smallest sum of mill price and transportation costs. For simplicity, it is commonly assumed that the costs of the firms have been normalized to zero.

Already in his original paper, Hotelling did not restrict himself to the aforementioned "main street" with customers in search for inexpensive physical goods from brick-and-mortar retailers. One of the nonphysical applications he mentioned was what we today refer to as brand positioning, viz., the location of a brand in some feature space. More specifically, Hotelling used the example of ciders offered by two firms, whose single distinguishing characteristic is their respective sweetness. Given that a brand is sweeter (more sour) if it is located more to the right (left) side of the market segment, the two firms will determine optimal locations and prices so as to maximize their respective profits.

Similar, albeit with a marked difference, is the political positioning model that was also mentioned in Hotelling's original paper. The idea was very simply for each of two political parties to each locate their own candidate, so as to maximize the number of votes (i.e., the number of customers, or the market share) that the candidate would obtain. The line segment was used to mimic the traditional left–right scale in politics, voters (i.e., their "ideal points," which symbolize their most favored position on the line) were again assumed to be uniformly distributed on the line segment, and the candidates would not have any inherent stand on the issues, they would simply position themselves at a point, where it would win them the largest number of votes. However, in contrast to all other previously mentioned applications, there are no prices in this model.

The main focus of Hotelling's original paper is the existence (or the lack) of a stable solution, i.e., an equilibrium. Hotelling asserts that an equilibrium would exist with both firms locating next to each other at the center of the market. This result is often dubbed the "principle of minimum differentiation," in reference to products or political candidates being very similar to each other. Even though in a footnote, Hotelling cautions that his result would not hold in highly competitive situation (which is precisely what occurs when the two firms locate very close to each other), he presented his agglomeration result as his major finding. Other authors, such as Lerner and Singer (1937) and Eaton and Lipsey (1975) obtained different results, but their contributions were based on Hotelling-style models albeit with fixed and equal prices. Hotelling's original result was not disputed until d'Aspremont et al. (1979) demonstrated 50 years later that no equilibrium exists in Hotelling's model. In

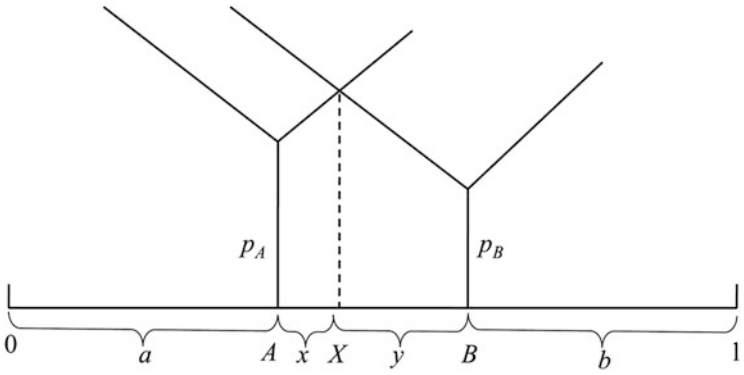


Fig. 14.1 Hotelling's duopoly on a linear market

order to follow the argument, first consider a graphical representation of Hotelling's scenario as shown in Fig. 14.1. Here, the linear market extends from 0 to 1, and the locations of the two competitors are shown as A and B, respectively. They charge mill prices p_A and p_B, respectively, and transportation costs are linear, resulting in full prices to the customers shown in the two "V" shaped functions. The two functions intersect at some point X, which is usually referred to as the *marginal customer*, i.e., the customer who pays the same *full price* (i.e., the mill price plus transportation costs) purchasing from firm A as he does purchasing from firm B. As a matter of fact, the function that describes the full price for all customers on the line segment is the lower envelope of the two "V"-shaped functions. Furthermore, the market can now be subdivided into the following parts: The first piece of length a is firm A's *hinterland*, which A captures in its entirety. Similarly, the stretch b on the right is firm B's hinterland, which is captured by B. The remaining area is the *competitive region* between firms A and B. (The terms "hinterland" and "competitive region" appear to have been introduced by Smithies 1941.) This is subdivided into parts x and y, such that x is the part in which customers can purchase more cheaply from firm A, while in y, customers can purchase the good more cheaply from firm B.

This allows us to determine the market shares of the two firms simply as $M(A) = a + x$ for firm A and $M(B) = b + y$ for firm B. This depiction of the scenario also permits us to examine the two forces that govern the process. The *market share force* pushes the two facilities towards each other. The reason is that—given that his opponent does not react, at least temporarily—a facility can move towards its competitor and, in doing so, not lose market in its own hinterland, while gaining in the competitive region. This force applies, as long as customers do not have finite (and reasonably low) *reservation prices*, i.e., an upper bound on the full price they are able or willing to pay for the good. On the other hand, there is the competitive *pricing force* that pushes the two facilities apart. The reason is that if the two firms locate very close to each other, whatever price one of them sets, his competitor can

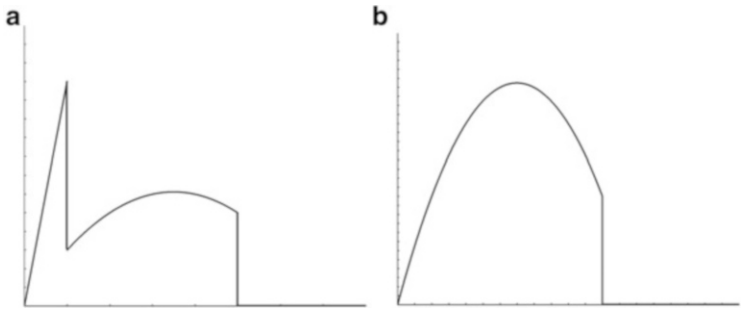


Fig. 14.2 Profit functions in Hotelling's model with linear transportation costs

undercut him slightly and thus capture the entire market. This results in facilities moving apart so as to position themselves in a region with less competitive pressure.

The obvious question is whether or not there is a locational arrangement and a price structure, which represents a stable solution, i.e., an equilibrium. Temporarily holding the location of both and the price of one of the competitors, say, B, constant, Fig. 14.2a, b shows competitor A's profit function π in the case of firms A and B locating close to each other (Fig. 14.2a) or a significant distance apart (Fig. 14.2b).

First consider Fig. 14.2a. From left to right, A's profit function is linearly increasing for low prices p_A (as firm B is cut out and A's profit increases proportional to the price); then, as p_A increases, at some point, B is no longer cut out, there is a marginal customer in the competitive region, and A's profit function is an inverted ellipse. As p_A increases further, there is a point, at which it is sufficiently high so that firm B cuts out firm A, and thus A's profit drops to zero. Notice that there are two local maxima, one at the first breakpoint from the left, and the second in the domain of the quadratic piece of the function. In Fig. 14.2b, the linearly increasing part is valid only for negative prices, which are nonsensical in this application. Other than that, the function is similar to that in Fig. 14.2a, but with a single maximum.

d'Aspremont et al. (1979) first demonstrated that Hotelling's model does not possess an equilibrium in pure strategies, i.e., as long as each player chooses exactly one strategy, rather than randomize. They then demonstrated that an equilibrium was restored in the model if we were to use a quadratic, rather than a linear, transportation cost function. Later, Gabszewicz et al. (1986) pointed out that the lack of the existence of equilibria in Hotelling's model is due to the lack of quasiconcavity of the profit functions of the duopolists (see again Fig. 14.2a). Figure 14.3a, b shows again competitor A's profit π, given a quadratic, rather than linear transportation cost function: Fig. 14.3a for competitors' locations that are close to each other, and Fig. 14.3b for locations far apart. Note that the functions are both quasiconcave.

In general, many competitive location models have shown major signs of instability: Hotelling's original model with variable prices and linear cost functions has no equilibrium, the same model with quadratic transportation costs has one—with firms located at opposite ends of the market. Hotelling's model with a

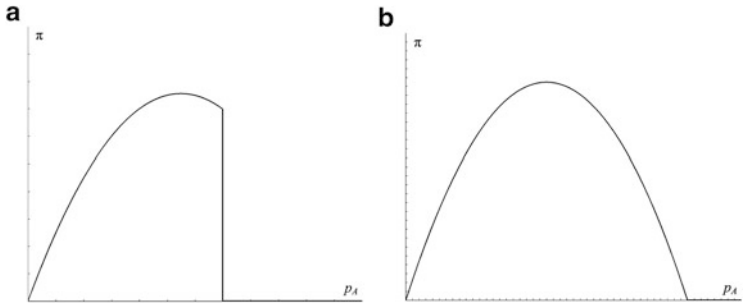


Fig. 14.3 Profit functions in Hotelling's model with quadratic transportation costs

linear-quadratic cost function (see, e.g., Gabszewicz and Thisse 1986, or Anderson 1988) does not have equilibria, as long as the linear part, no matter how small, exists. Hotelling's model with fixed and equal prices (see, e.g., Lerner and Singer 1937 or Eaton and Lipsey 1975) has an equilibrium with minimal differentiation, while the same model with three firms has no equilibrium; the duopoly with fixed and unequal prices, regardless how small the difference between the prices, has no equilibrium.

Consider now the locational arrangement that minimizes the total transportation costs to the customers. Using the notational convention in Fig. 14.1 and unit transportation costs t, the total transportation costs to all customers can be written as

$$\begin{aligned} TC &= t\left[\int_{\Phi=0}^{A}(A-\Phi)\,d\Phi+\int_{\Phi=A}^{X}(\Phi-A)\,d\Phi+\int_{\Phi=X}^{B}(B-\Phi)\,d\Phi+\int_{\Phi=B}^{1}(\Phi-B)\,d\Phi\right] \\ &= t\left[3A^2/4+3B^2/4-AB/2-B+1/2\right] \end{aligned}$$

Partial differentiation $\frac{\partial TC}{\partial A}=0$ and $\frac{\partial TC}{\partial B}=0$ results in the optimal points $A=1/4$ and $B=3/4$, a configuration at which the total transportation costs are $t/8$. In contrast, central agglomeration results in transportation costs of $t/4$, i.e., costs that are twice as high. As the point $(A, B)=(1/4, 3/4)$ minimizes the total transportation costs (which are, given mill pricing, borne by the customers), this point is often referred to as *social optimum*.

Before investigating the key elements of competitive location models, we would like to draw attention to some surveys of the subject. Brown (1989) provides a critique of Hotelling's work and points out various directions, which would make the original model more realistic. Eiselt et al. (1993) provide a taxonomy and a short evaluation of the literature up to that point. Plastria (2001) looks at the optimization aspect of the subject, while Drezner and Eiselt (2002) focus on customer behavior and its consequences on the solution. More recently, Kress and Pesch (2012) surveyed the subject, but concentrate on problems on networks, while Drezner (2014) surveys problems in the plane.

14.2 Elements of Competitive Location Models

The subject of competitive location models, as pioneered by Hotelling, has become a rich research area. Since research has moved into many different directions, it is useful to classify models, e.g., by using the taxonomy proposed by Eiselt et al. (1993). Rather than describe it in detail, we will outline its major components here.

One aspect of all location models, competitive or not, is the choice of *space*. In contrast to regular, noncompetitive, location models, many authors have used much simplified spaces in their models: starting with Hotelling's original linear market, they have also investigated circular markets, which may appear rather contrived at first glance, but are designed to avoid the "end-of-line effects" of bounded linear markets.

Measures of distances are no issue when devising models in a single dimension, but they are, as soon as models in two or more dimensions are investigated. While some authors favor gauges in noncompetitive location models (see, e.g., Durier and Michelot 1985, or Plastria 1992) most contributions in the literature that look at continuous location models in the plane have used Minkowski distances, most prominently Manhattan, Euclidean, and Chebyshev distances.

A similar situation prevails in networks. Measures of distances in trees are not an issue, as, by definition, there is only one path between each pair of points. However, in general networks one could, at least theoretically, use any distance that best models reality. Assuming not only rational, but also cost-minimizing behavior, virtually all authors in the field have chosen shortest path distances. Assuming complete information, one could choose traffic choice models and assume that customers take not the shortest route with respect to distances, but the shortest route with respect to time; or that not all customers use the same route selection strategy all the time. This would suggest itself particularly in highly congested (urban) areas. One concept that is used extensively by authors who deal with network models is known as *node property* or *Hakimi property*. It is based on Hakimi's work (1964) on network location properties, in which he proved that in some models, at least one optimal solution locates all facilities at the nodes of a network.

The second component concerns the *number of players* and facilities that are to be located. Traditionally, papers included duopolists who locate a single facility each, so that the terms "firm" and "facility" (the entity to be located) were synonymous. This is, of course, no longer the case once we include multiple firms or multiple facilities to be located by each of the planners. Here, we will use the game-theoretic term players for the (independently operating) firms, and "facilities" for what they are locating. The number of facilities that one or more of the players wish to locate may be preselected or unspecified. In the latter case, the cost or profit function of a player includes fixed costs for opening a facility at a site.

The third component of competitive location models concerns the pricing policy. One important feature of Hotelling's original model was that he investigated competition in location *and* prices. A more general model would let players also choose their pricing policy. In particular, we typically distinguish between a variety

of different pricing policies. Among the most prominent such policies is *mill pricing*, where players set prices at the source, which are not necessarily the same at all of their facilities. Customers will then purchase the product at the facility they have chosen to patronize and pay for the transport costs themselves and separate from the payment to the firm for the goods. Almost all retail facilities use this principle. A special case of mill pricing is *uniform pricing*, a policy, in which the facility planner sets the same price at all of his facilities. This policy was used by the "Motel 6" chain in the 1980s, until they chose to charge different prices at different locales to better reflect their own cost structure.

Another principle is uniform delivered pricing. In this pricing policy, facility planners will deliver the goods to their customers for a fixed "full price" regardless of customers' locations. Domestic mail is a typical example of this type of pricing policy. Clearly, in such a policy, customers that are located close to the facility from which they receive the goods, will subsidize those who are located farther away. A special case of this policy is "zone pricing," a policy, in which the firm has subdivided their market area into zones, such that a uniform delivered price is charged in each zone. Typical examples are the outdoor store L.L. Bean that sells canoes for one delivered price east of the Mississippi, and another price west of the river, or postal services that typically charge one rate for domestic mail and (at least) one for international mail. Spatial price discrimination is a policy that charges customers a full price according to the customer's location. Its applications have been severely limited by the Robinson-Patman Act of 1936, even though it does provide some benefits to the customers; see, e.g., Anderson et al. (1992). Note that uniform delivered prices and spatial price discrimination are boundary cases of zone pricing; the former in case there is only one zone, and the latter in case each point in space represents its own zone. Many contributions, especially those from the operations research community, assume that prices are universal and fixed, which is the case in legislated pricing or producer-administered mandatory prices.

The fourth component concerns the *rules of the game* the players adhere to. In essence, this feature describes how individual players (re-) act. Consider the simple case of pure location competition. In the latter case, players could simultaneously choose their strategies, i.e., decide on the locations of their facilities. If at this point, none of the players has an incentive to unilaterally change his position, we say that a Nash (or Cournot–Nash) equilibrium has been obtained. Such a situation indicates some stability. Note that all players have, at least potentially, the same information available to them, even though perceptions may differ, indicating some symmetry among players.

Things are getting somewhat more involved, if players have not only locations, but also prices as variables. In such a case, we can employ a refinement of Nash equilibria, viz., Selten's (1975) *subgame perfection*. Loosely speaking, a subgame perfect equilibrium exists, if every subgame of a given game is a Nash equilibrium. Applied to our type of problem, players may choose a "first location, then price" strategy (see, e.g., Anderson and de Palma 1992), i.e., all payers simultaneously choose their locations, and in a second phase, they simultaneously choose their prices. Many authors have chosen this route. At this point, we need to define the

concepts of *pure and mixed strategies*. A pure strategy prescribes a certain course of action (i.e., a decision) for a decision maker, while a mixed strategy will provide a schedule of decision, associated with probabilities that indicate with what likelihood a decision maker should use this strategy. The work by Caplin and Nalebuff (1991) outlines conditions under which a pure-strategy price equilibrium exists in a locational game, while Dasgupta and Maskin (1986), who deal with discontinuous payoff functions, describe conditions for the existence of mixed strategies.

A full sequential strategy has one player, the so-called *leader*, locate first, followed by all other players, the *followers*, which locate later. This asymmetric situation has originally been described by the economist von Stackelberg (1943). The leader, when choosing his locations, will have to guard against the followers. If all players have the same objective and the same perception of the demand structure, this means that the leader will use a strategy to maximize the minimal market share or profit he will obtain. On the other hand, the followers will have a chance to observe the action of the leader and then react accordingly, meaning that they solve a conditional optimization problem, in which they maximize their own market share or profit, given that the leader has already located. Notice that the problem of the follower is much easier to solve mathematically, as it is a simple optimization problem. The problem of the leader, however, is a bilevel optimization problem, as it requires the solution of the follower's problem as an input parameter.

The last major descriptor of competitive location models concerns *customer behavior*. As a matter of fact, this aspect is the main leitmotif of this paper. The first major distinction between different classes of models is between demand allocation models and customer choice models. As the name suggests, in allocation models the firm decides which facility is allocated to a customer. A typical example would be the delivery of furniture to customers, who will receive the goods from whatever warehouse the firm decides to deliver from. (Note that, strictly speaking, the purchase of, say, a sofa, typically involves a mix of allocation and choice models: when customers drive to a store to purchase the sofa is a choice model, while the actual delivery of the sofa is an allocation model.) In scenarios of customer choice, on the other hand, customers choose which facility or firm they want to deal with. Often, the two models are referred to shipping and shopping. This paper deals exclusively with customer choice models.

The manner in which customers choose which facility they patronize, is the main subject of this contribution. The next section will provide a framework for this decision. At this point, suffice it to say that while many, or even most, papers use the "patronize the closest facility" (or cheapest, in case prices are different and mill pricing is assumed), other models have been suggested. For instance, some models include a (single-dimensional) parameter that measures the attractiveness of a facility in contrast to other, competing facilities. Furthermore, an important and fairly recent strand of research uses probabilistic choice rules, according to which customers at the same location do not all behave in the same way. Similarly, it is able to capture the fact that a customer, even if he and all of the competing facilities remain in the same positions, will not always patronize the same facility.

14.3 Consumer Behavior in Competitive Location Models

Consumer behavior is one of the most important aspects in any user-focused models, yet it is crucial to many such models. Some references are Raiport and Sviokla (1994), who identified content, context, and infrastructure as major determinants of customer behavior, Song et al. (2001) and Giudici and Passerone (2002), who use data mining in their analyses of identifying changes in consumer behavior, and Liou (2009), who presents decision rules that foster customer retention in the airline industry.

The three-stage process below presents a decision-making framework that customers use when making their choices. We will discuss the individual stages and demonstrate how they encompass the rules and assumptions made in the literature.

Stage 1 is the *evaluation stage*. In it, customers determine utilities to each of the stores. For the purpose of this paper, we assume that customers actually have complete and correct information, an assumption that may be justified by Internet searches or similar fact-finding processes, together with past experience with the facilities. The utilities created in this stage will be based on all components that typical customers deem important. In the retail context, this may include, but not be restricted to, the price charged at the facility, the distance to the facility, the parking at the facility, the friendliness of the staff, and others. Formally, we can define u_{ij} as the utility a customer at site i has (for simplicity, we will refer to "customer i") for goods or purchased at a facility at site j (called "facility j" for short). Furthermore, we define d_{ij} as the distance between customer i and facility j, while t denotes the unit transportation cost, i.e., the conversion from distance to money. We also need to define p_j as the price charged by facility j, and the basic attractiveness A_j of facility j. The basic attractiveness is a composite parameter that includes different measures, such as floor space of a retail establishment (as a proxy expression for variety), the quality of service, and other features. It is not important to find an exact aggregate measure, it is only important to find an expression that captures the differences between facilities. For simplicity, we will restrict ourselves to a single homogeneous product, such as a brand that can easily be compared between facilities. As an aside, some firms make such comparisons difficult by assigning different model numbers to the same product, one for department stores, and a different number to the product, when it is sold through specialty retail outlets.

The simplest (deterministic) utility function is

- UD1a: $u_{ij} = -td_{ij}$,

i.e., the utility of customer i regarding facility j equals the negative distance between them. Hence, maximizing the utility, such a customer will patronize the facility closest to him. Such a utility function has been used by early contributors, such as Lerner and Singer (1937), Eaton and Lipsey (1975), and later by operations researchers such as Hakimi (1983), ReVelle (1986), Serra et al. (1999).

An extension is the utility function

- UD1b: $u_{ij} = -p_j - td_{ij}$

Maximizing such a utility is equivalent to minimizing the full price of the good, i.e., the mill price plus the transportation costs. Hotelling's own contribution falls into this category, and so do the papers by Serra and ReVelle (1999) and Pelegrín et al. (2006). Note that the utility UD1a is a special case of the utility UD1b with zero prices (or prices that are equal at all existing facilities).

Consider now the utility function

- UD1: $u_{ij} = R_i - p_j - td_{ij}$,

where R_i denotes the *reservation price* customer i assigns to one unit of the good in question, an upper bound customers are prepared to pay for one unit of the good. Given that, the utility is an expression of the amount of money that the customer "saved," i.e., the amount that he was prepared to, but did not have to, spend on a unit of the product. Some authors refer to R_i as the valuation of the product, other refer to it as income, while still others think of it as the budget. In all cases, $R_i - p_j - td_{ij}$ is an expression of the money that was available for the purpose, but did not have to be paid for the product. It is apparent that the utility functions UD1a and UD1b are special cases of the function UD1: Given equal reservation prices $R_i = R_k$, $i \neq k$, maximizing the utility UD1 reduces to UD1b, which, in turn, reduces to UD1a for fixed and equal prices p_j. One important feature of the utility function UD1 is that, in case the utility u_{ij} is nonpositive, it allows customer i to refrain from making any purchases.

Finally, there exists a variety of other deterministic utility functions used by some authors. Among them is Lane (1980), who uses a Cobb–Douglas-style function that expresses the utility as the product of three components: a measure of a characteristic raised to a power, another measure of the facility raised to some power, and the available income of the individual. Neven (1987) frames his discussion in the context of brand positioning, and his utility function is the difference between a (very high) reservation price, and the price plus the squared of the customer-facility distance (which, in this context, is actually the difference between the customer's ideal point and the actual feature of the product). Finally, Kohlberg (1983) uses a utility function that includes the sum of travel time and waiting time, a utility that is important in the context of facilities that feature congestion, such as health-care facilities.

- UD1c: $u_{ij} = R_i - p_j - td_{ij} - W_i$,

where W_i denotes the waiting time. One pertinent example in the context of health services is found in Marianov et al. (2008).

Another utility function incorporates not only distances, which are present in all spatial models—after all, they are what makes a model "spatial"—but also the "attractiveness" of the facilities. As already briefly alluded to above, this one-dimensional measure attempts to capture differences between facilities the way they are perceived by customers: floor space as a proxy for selection (even though the

models under consideration just deal with a single homogeneous good), friendliness of staff, parking, lighting, temperature, cleanliness of the facility, and many others. A simple utility function that incorporates the basic attractiveness of facility j as the parameter A_j is

- UD2a: $u_{ij} = \frac{A_j}{d_{ij}^{\lambda}}$

with some decay parameter λ. For $\lambda = 2$, the relation reverts to the well-known gravity model, first proposed by Reilly (1931) for the determination of trading areas. This function has been used by authors, such as Aboolian et al. (2007), Drezner and Drezner (1997), Eiselt and Laporte (1991), and Suárez-Vega et al. (2014), the last using the slightly more general function "basic attractiveness divided by some increasing continuous function of distance." Clearly, given the absence of prices, these models assume that prices are fixed and equal among facilities.

An alternative treatment that involves an attractiveness parameter is

- UD2b: $u_{ij} = A_j e^{-\beta d_{ij}}$

with some parameter $\beta > 0$ that indicates the customers' sensitivity to differences in distances. Aboolian et al. (2008) use a function of this type, but go one step beyond: their base attraction A_j is a negative exponential function of the price charged at the facility.

Consider now utility functions that include probabilistic components. There are considerably fewer probabilistic location models than there are deterministic models. The probabilistic counterpart of the above deterministic function UD1 is

- UP1: $u_{ij} = R_i - p_j - td_{ij} + \varepsilon_i \mu$,

where ε_i is, usually, a Weibull-distributed random variable, while μ is typically interpreted as a coefficient of heterogeneity of customer tastes.

On the other hand, a probabilistic version of the utility function UD2a is

- UP2: u_{ijk},

defined as the utility a customer at site i has for feature k of facility j. This multidimensional version of the attraction function leads to the probabilistic allocation rule AP1.

Stage 2 in the decision-making process involves the *allocation* of a customer's demand. The most natural thing to use would be the deterministic allocation rule

- AD1: winner-take-all,

which allocates all of customer's demand to the facility he is most attracted to. Most of the contributions in the literature follow this rule. Actually, if the utility function is assumed to include all of a customer's wishes, this rule would be the only logical choice. However, even when considering a single customer, he may opt logically for a facility that is second-best or has an even lower ranking based on its utility. The reason could be that the customer, having patronized on facility, wants some variety, even though it is probably not as good. Alternatively, if a customer point represents actually a group of customers (meaning that customer i is actually

an aggregate, typically of a census tract or some other group of customers), some members among the group may have different rankings and prefer what, on average, is a higher-ranking facility.

This heterogeneity of customer tastes can be dealt with in different ways. One such possibility is to use a

- AD2: proportional allocation.

This allocation rule will allocate a customer's demand according to the relative utility a customer has for a facility. For instance, the proportion of customer i's demand to facility j according to Hakimi's (1990) "proportional" rule equals $u_{ij} \Big/ \sum_k u_{ik}$. As an example, if a customer faces a duopoly, for whose facilities he has computed utilities of 3 and 7, respectively, he will satisfy 30 and 70 % of his total demand at the two respective facilities. Hakimi (1990) also designed a hybrid rule based on AD1 and AD2. He refers to it as a "partially binary" allocation. According to this rule, customers consider only the closest facility or branch of each of the competing firms, and they then distribute their demand proportionally among those branches. Suárez-Vega et al. (2004) investigated AD1, AD2, and the aforementioned hybrid in detail.

Consider now probabilistic allocation functions. A natural extension of Reilly's (1931) argument of attraction functions was Huff's (1964) allocation function, which allocates a proportion of a customer's demand to a firm based on the firm's attractiveness and its distance to the customer,

- AP1a: $p_{ij} = \dfrac{A_j / d_{ij}^{\lambda}}{\sum_k A_k / d_{ik}^{\lambda}}$.

Huff suggested the selection of a location from a pre-specified set of locations, whereas Drezner (1994a, 1995) proposed a model for finding the best location anywhere in the plane. A multidimensional generalization of this idea was proposed by Nakashani and Cooper (1974), the so-called multiplicative competitive interaction model, or *MCI* for short. Assuming that u_{ijk} denotes the utility customer i has for feature k of store j, let p_{ij} denote the probability that a customer at site i makes a purchase at store j. The parameter α reflects how sensitive is p_{ij} to feature k. The *MCI* model then asserts that

- AP1: $p_{ij} = \dfrac{\prod_k u_{ijk}^{\alpha_k}}{\sum_j \prod_\ell u_{ij\ell}^{\alpha_\ell}}$.

Following the arguments of McFadden (1974), the use of the probabilistic utility function UP1 leads to the demand allocation rule

- AP2: $p_{ij} = \dfrac{e^{(R_i - p_j - td_{ij})/\mu}}{\sum_k e^{(R_i - p_k - td_{ik})/\mu}}$,

Note that whereas any of the deterministic utility function could be followed by any of the allocation functions, the allocation function AP2 is a direct consequence of the utility function UP1.

Finally, in the third stage in the decision-making process, customers determine the quantity that they are going to purchase from the chosen facility/facilities. Most authors opt for the quantity choice rule

- Q1: fixed,

in which the quantity customers purchase is fixed. This is typically justified by asserting that the good in question is essential. While such an assumption is convenient, there are, actually, relatively few essential goods in real life: butter can be replaced by margarine, private transportation can—at least within reason—be replaced by public transportation; potatoes could be replaced by pasta, and so forth. Yet, true essential goods exist, such as electric power (which cannot be replaced in the short run), or medical care. Typical examples for the use of this rule include almost all contributions in the literature, starting with Hotelling (1929), Eaton and Lipsey (1975), and d'Aspremont et al. (1979) to Drezner and Drezner (1997), Fernández et al. (2007), Braid (2013), and others.

A very general alternative rule is

- Q2: $q_{ij} = f(p_j + td_{ij}, u_{ij})$,

where q_{ij} denotes the quantity customer i purchases at facility j. This rule states that the quantity that customer i purchases from facility j is a function of the full price to be paid for purchases at that facility and of the utility customer i achieves from purchases at facility j. While a customer's utility is likely to include the full price as one of its components, the quantity purchased by a customer is often assumed to depend on the (full) price of the product, rather than on a customer's utility. The early contribution by Rothschild (1979) uses a negative exponential distribution to relate a customer's demand and the customer-facility distance, while Aboolian et al.'s (2008) work includes not only distance, but also price, in their negative exponential relation. The contributions by Penn and Kariv (1989) and Matsumura and Shimizu (2006) assume that the demand at a point is the difference between a constant and the travel distance, and the difference between a constant and the price paid for the product. Both cases are designed so as to express the amount of money a customer has left over after this purchase.

Once customers have gone through the three stages of their decision-making process, they have decided how much to purchase and whom to purchase it from. This can then be used as input by the competing planners of the facilities. Drezner et al. (1996) analyzed an anomaly in the decision making process that occurs if customers reevaluate their purchasing decision along the way to the chosen facility. The authors also delineated areas in which this phenomenon occurs.

14.4 Results for Different Behavioral Assumptions

This section is organized along the lines of customer behavior. Each part will examine one customer choice rule, followed by results in the literature regarding Nash equilibria, followed by von Stackelberg solutions. The review in this section will be organized along the lines of the customer choice rules outlined in the previous section.

14.4.1 UD1a, Linear Market, Nash Equilibria

Stevens (1961) appears to have been the first to use game theory to reestablish Hotelling result of minimal differentiation for fixed and equal prices. Recognizing the complexity of the problem described in Hotelling's (1929) paper, some contributors decided to simplify matters. Eaton and Lipsey (1975) used fixed and equal prices. While this assumption appears somewhat contrived, it is usually justified by legislated pricing for essential goods. With this assumption, customer choice rule UD1a (the "closest" rule) is applied. Given this assumption, Hotelling's result of minimal differentiation is reestablished, as by moving towards its opponent, a firm gains customers in the competitive region and does not lose customers in its hinterland. The authors also extend the analysis to more than two firms. In particular, they determine that for more than five firms, multiple equilibria exist, and the only case without equilibria is the instance with three facilities. In particular, the two outside facilities will push inwards so as to gain additional market shares, thus squeezing the market of the inside firm to zero. This firm will counteract by "leapfrogging" to the outside, become an outside facility itself, and start moving inwards. Teitz (1968) referred to this behavior as "dancing equilibria." Shaked (1975) investigates the usual Hotelling model with fixed and equal prices, but three facilities that employ mixed strategies. It turns out that an equilibrium exists, in which all facilities randomize their strategies in the central half of the market.

In a follow-up paper, Shaked (1982) investigates the Hotelling model with three firms locating one facility each, with fixed and equal prices, allowing mixed strategies. It turns out that all firms will chose locations in the central half of the market with equal probability. Cancian et al. (1995) consider a Hotelling model with directional constraints, i.e., customers can only walk in one direction towards the firm they want to patronize. The authors determine that with random arrival times of the customers and two or more facilities, no equilibrium exists.

14.4.2 UD1a, Linear Market, von Stackelberg Solution

The first author to introduce sequential (and final) location decisions into the discussion appears to have been Hay (1976). However, it was the contribution of

Prescott and Visscher (1977) that popularized the methodology and the results. In one of their examples, the authors look at a duopoly on a linear market—the simplest possible case—and determine that the leader will locate at the center of the market, while the follower will locate next to the leader, thus resulting in central agglomeration. The authors then extend their analysis to the case of three firms. After considering many cases and subcases (see, e.g., Younies and Eiselt 2011), it is determined that one of the outcomes (arguable the most likely one) is that the three facilities locate at 1/4, 3/4, and 1/2 of the market, capturing 3/8, 3/8 and 1/4 of the market, respectively. The fact that the first two facilities to locate earn 50 % more than the last entrant into the market is, however, troublesome: having established that it takes capability and incentive to be a leader (see, e.g., Younies and Eiselt 2011), we can consider the second and third firms to enter the market as followers. However, why would any follower accept being the third rather than the second entrant, if the latter course of action is much more profitable? A similar result had already been obtained by Teitz (1968), who considered duopolists, so that the location leader would locate two facilities, while the location follower would locate a single facility. He suggested "conservative optimization," i.e., a minimax strategy. While the leader locates his two facilities at 1/4 and 3/4 of the market, the follower will locate his single facility anywhere between the leader's facilities.

An interesting extension is provided by Thisse and Wildasin (1995), who locate private facilities alongside a centrally located public facility. Households have incomes, which they spend on trips to the facilities and paying land rent. In the first stage of the game, all firms locate first, followed by stage two, in which customers locate. The result is that high travel costs yield maximal differentiation, while low travel costs result in minimal differentiation. Bhadury (1996) considers a Hotelling model on the line with fixed and equal mill prices, in which the leader does not have perfect information regarding the follower's variable costs. For a general demand distribution, the author shows that market failure is possible (i.e., the leader may not wish to locate any facilities) and that a greedy strategy is not bad (optimal for an atomistic leader, i.e., one who wishes to locate only a small number of facilities). Osborne and Pitchik (1986) allow the demand distribution to be not necessarily uniform. Allowing mixed strategies, the result for a three-firm problem has all three firms randomize over the central half of the market. Dasci and Laporte (2005) allow facilities to have different cost functions. The paper is novel in that it does not deal with exact facility locations, but with the density of retails branches that are located.

14.4.3 UD1a, Plane, Nash Equilibrium

In two-dimensional space, Okabe and Aoyagi (1991) attempt to prove a conjecture by Eaton and Lipsey (1975) in the two-dimensional plane. With fixed demand and equal mill prices, customers patronize the closest facility. In the infinite two-dimensional plane with Euclidean distances and an infinite number of independent

firms, the market area of each of the firms is a cell in a Voronoi diagram. Each firm attempts to maximize the area of its Voronoi cell. Voronoi cells are in global optimum with the hexagonal pattern. It is noted that results in one-and two-dimensional spaces are markedly different: the pairing in one dimension does not carry over to the two-dimensional plane. Another attempt in the two dimensional plane was reported by Okabe and Suzuki (1987). The authors use the same concept as in the previous paper, but locate finite numbers of facilities (32–256) in a bounded market the shape of a square. Global optimization techniques are sequentially and repeatedly applied. The result is a honeycomb-type pattern that, however, self-destructs again and rebuilds. The instability is likely to be the result of "boundary effects" that distort the results.

Aoyagi and Okabe (1993) consider a Hotelling model in the plane with totally inelastic demand, identical facilities, and customers who purchase the good from the closest facility. Customers are assumed to be located in a compact and convex subset Z of the two-dimensional Euclidean plane. The authors demonstrate that for $n = 2$, an equilibrium exists if and only if the market is point-wise symmetric with respect to some point in Z. The firms will then locate at that point. For three facilities, no global equilibrium exists except maybe in the case of a equilateral triangle.

14.4.4 UD1a, Plane, von Stackelberg Solution

The first author to discuss competitive location problems in the plane given location leaders and followers appears to have been Drezner (1981, 1982). His contribution first considers the simple case, in which each firm locates a single facility in the presence of n demand points. The follower's best location is arbitrarily close to that of the leader. Sorting of angles from the leader's point to the demand points yields an $O(n \log n)$ algorithm for the follower's problem. The leader's problem (given he locates one facility and expects the follower to do the same) is shown to be solvable in $O(n^4 \log n)$ time. In case a minimum separation of some prespecified distance R is required between leader and follower, the complexity of the two problems is still $O(n \log n)$ and $O(n^5 \log n)$, respectively. Other cases include the problem in which the leader locates one, the follower $r > 1$ facilities. This problem is easy: the leader is wedged in and his optimal strategy is to locate right on the point with the largest demand, as that is all he will get. If the leader locates $p > 1$ facilities and the follower locates one facility, then the follower's problem can be solved in $O(n^2 \log n)$ time.

Shigehiro et al. (1995) consider a duopoly with firms A and B in a bounded subset of the two-dimensional plane. Given fixed and equal prices, both firms are market share maximizers. Given demand at grid points and the one of A's two facilities being already located, firm B locates a single facility, followed by firm A locating its second facility. It turns out that firm A will locate its second facility next to it competitor's facility, thus re-establishing the pairing of facilities known from one-dimensional markets. An algorithm for the centroid problem is also described.

Infante-Macias and Muñoz-Perez (1995) discuss medianoid locations in the plane with customer demand occurring at discrete points, and Manhattan distances are used. A given parameter specifies how much closer a new facility must be to a customer to be considered comparable, i.e., equally desirable. For the location of a single new facility, the paper describes an $O(n^3)$ algorithm, for a given number p of new facilities, an $O(n^5)$ algorithm is suggested.

14.4.5 UD1a, Networks, Nash Equilibria

Bhadury and Eiselt (1995) investigate duopoly models with fixed and equal prices on tree networks. They describe locational Nash equilibria in case co-location (i.e., the location of both facilities at the same node) is permitted or not, and they describe a measure of stability of the equilibrium, rather than applying the usual equilibrium-no equilibrium dichotomy. In another paper, the same authors (Eiselt and Bhadury 1998) discuss the reachability of Nash equilibria (assuming that at least one such equilibrium exists) on trees. Starting with arbitrary locations of the duopolists, they apply sequential and repeated short-term optimization to investigate whether or not an equilibrium will be reached. The answer is it will, provided an appropriate tie-breaking rule is employed. Eiselt and Laporte (1993) describe conditions, under which a three-facility problem on a tree has agglomerated, dispersed, and no equilibria.

14.4.6 UD1a, Networks, von Stackelberg Solution

Among the early contributions, Slater's (1975) work stands out. In it, the author introduces leader and follower, respectively, does, however, not make the connection to von Stackelberg's work. The paper proves that on a tree network, the location leader will locate at the median. In his contribution, Hakimi (1983) first introduces von Stackelberg games by referring to the locations of the leader(s) of the sequential game as *centroids* (based on their maximin objective), while the locations of the follower(s) are termed *medianoids* (as their objective is of the "minisum" type). In particular, if the leader has already located p facilities in a pattern denoted by X_p, and if the follower is poised to locate r facilities, the follower's problems is an $(r|X_p)$ medianoid. On the other hand, if a leader wants to locate p facilities, knowing/assuming that the follower will locate r facilities, we talk about an $(r|p)$ centroid. Hakimi discusses a number of results of special cases regarding the node property, i.e., the question whether or not at least one optimal location pattern naturally has locations at the nodes of the given network. In addition, he proves the *NP*-hardness of $(r|X_1)$ medianoid of general networks as well as the *NP*-hardness of the $(1|p)$ centroid. In the same year, Megiddo et al. (1983) show a polynomial $O(n^2r)$ algorithm for the $(r|X_p)$ medianoid problem on trees. Penn and Kariv (1989)

require facilities to be located at the nodes of the tree, but allow a customer's demand to be linearly decreasing in the distance to the closest facility. Both firms are assumed to locate a single facility. Characterizations of the solutions, especially with respect to the median(s) of the tree are described. Hansen and Labbé (1988) present a polynomial algorithm for the (1|1) centroid problem on tree networks. García et al. (2003) follow the analysis of Eiselt (1992, Annals) and determine all von Stackelberg solutions on a tree with parametric, but possibly different, prices. They also discuss the "first entry paradox" (see Ghosh and Buchanan 1988), according to which the leader in a von Stackelberg game would typically have the advantage.

ReVelle (1986) was the first to formulate the highly influential MAXCAP problem on networks, i.e., the problem, in which the follower locates facilities. By modifying the objective, he reduced the formulation to a p-median problem. In follow-up papers, Serra and ReVelle (1994, 1995) present the PRECAP problem that solves the leader's $(r|p)$ centroid problems. The authors design heuristic algorithms for the (bilevel) problem of the leader, and report computational experience. The main contribution in the Hakimi (1990) book chapter is the introduction of three allocation rules: binary (i.e., winner-take-all), partially binary (a customer distributes his demand proportional to the inverse distances to the closest facilities of the two firms), and the (fully) proportional rules, in which customers allocate their demand inversely proportional to the distances to the facilities. The authors also presents results with these allocation rules with respect to the node property. Suárez-Vega et al. (2004) expand on Hakimi's discussion of the three allocation rules for essential and unessential demand at the nodes of the network. The authors also derive finite dominating sets, including those for concave capture functions. Serra et al. (1999) discuss the usual MAXCAP problem, but with an additional constraint that ensures that each facility has at least a market share of a certain size. This is done so as to guarantee the viability of the firm. Some computational testing with two rules is provided; on rule, which checks viability first, then locates and reallocated demand, and the second rules that does not do the checking. It appears that Rule 2 has some advantages.

Spoerhase and Wirth (2008) tackle the notoriously difficult problem of $(r|p)$ centroids. In order to obtain any results (as Beckmann 1972 stated: "As everyone knows, in location theory one is forced to work with simple assumptions in order to get any results at all"), they restrict themselves to paths and trees. Along similar lines, Eiselt (1998) investigates a von Stackelberg problem on a tree, given that the perceptions of leader and follower regarding the demands at the nodes are different. Solutions to the bimatrix game (in which each player has full knowledge about the perception of his opponent) and the hypergame (in which neither competitor knows about the perception of his competitor) are characterized. In general, if a firm can assume that its competitor has researched the demand diligently, it can gain little by finding out about the exact perception of its competitor. Marianov et al. (1999) extend the MAXCAP to the location of hubs by a follower firm, assuming that passengers choose the airline which offers the shortest route (distance) between their origin and destination. Marianov and Taborga (2001) address the problem of locating public health centers competing with private ones for affluent customers,

assuming that the closest center captures the demand. Later, Marianov et al. (2004) extend these results to facilities with waiting lines.

14.4.7 UD1b, Linear Market, Nash Equilibria

Consider now models that employ the customer choice rule UD1b, i.e., models in which customers patronize the least expensive facility. Hotelling's original model belongs into that group, which, with its linear transportation costs, does not exhibit an equilibrium. This was pointed out by d'Aspremont et al. (1979) who also demonstrated that as soon as quadratic transportation costs are used, an equilibrium does exist with maximum differentiation, i.e., the two facilities locate at opposite ends of the market. Anderson (1988) provided further insight into the case: he demonstrated that in case of linear-quadratic transportation cost functions, i.e., cost functions that have a quadratic and a linear component, equilibria only exist, if there is no linear component and the cost function is purely quadratic. Hamoudi and Moral (2005) extend the analysis and investigate linear-quadratic transportation cost functions with different parameters, which result in convex and concave transportation cost functions, respectively. The authors then define profit functions for the two cases. Because a price equilibrium does not exist for all pairs of locations, the authors delineate pairs of locations, for which such an equilibrium does exist. It turns out that the region, in which price equilibria exist in the concave case is complete enclosed in the region, in which equilibria exist in the convex case.

Tabuchi and Thisse (1995) analyze Hotelling's model with a quadratic transport cost function and triangular customer density. Again, a subgame-perfect equilibrium is sought. It turns out that no symmetric location equilibrium exists. Instead, asymmetric equilibria exist at $\left(0, \frac{\sqrt{33}-3}{2\sqrt{2\sqrt{33}+2}}\right)$ and $\left(1-\frac{\sqrt{33}-3}{\sqrt{2\sqrt{33}+2}}, 1\right)$, i.e., (0, 0.3736) and (0.2527, 1), given that we restrict facility locations to the inside of the market. Cremer et al. (1991) locate n facilities on a linear market. Given quadratic transportation costs and the usual Hotelling assumptions (including the "first simultaneous choice of location, then simultaneous choice of mill prices"), the model includes m public and $n-m$ private firms. While private firms maximize their individual profits, public firms maximize the social surplus, which, with the assumption of inelastic demand, reduces to the minimization of transportation costs. For $n=2$, one public and one private firm perform best. The two facilities will locate at the social optimum of $1/4$ and $3/4$, respectively. For $n=3$ and one public facility, profits of the private firms are higher and general welfare is lower than in the all-private case. With two public facilities, the social optimum is reached. Some additional combinations of public and private facilities are also investigated.

An important strand of research considers the original Hotelling model, but allows mixed strategies on prices and pure strategies for the location subgame. Among the earlier attempts is the contribution by Osborne and Pitchik (1987), who determine that facilities will locate at about 0.27 away from the ends of the market

of unit length. Matsumura and Matsushima (2009) use heterogeneity in the form of different production costs, and if those result in pure strategy equilibria not to exist, then mixed strategy equilibria are used. Location equilibria with minimal and maximal differentiation appear each with probability of $1/2$.

Anderson (1987) showed that in the "first location, then price" two-stage game if facility A were to lead in the first-stage location game, then it would be best for its opponent B to be a leader in the second-stage pricing game. As a result, firm A would locate at the center at the market, while firm B will locate at 0.131 (or, symmetrically, at 0.869). Anderson and Neven (1989) use the usual Hotelling assumptions, including duopolists on a linear market, mill pricing and "first location, then price" competition, but allow customers to purchase goods from both firms according to some loss function and the use of a quadratic transportation cost function. The result is maximal differentiation with the duopolists locating at the two ends of the market. In another contribution, the same authors (Anderson and Neven 1991) employ spatial price discrimination in a two stage "first location, than quantity" procedure. The result is an equilibrium with minimum differentiation. The authors also demonstrate that for more than two firms, given linear transportation costs and a regularity condition, all firms will locate at the center of the market. Such agglomeration is often observed in practice, see, e.g., Marianov and Eiselt (2014). Hamilton et al. (1989) describe a Hotelling model with spatial price discrimination and a linear price–quantity relation. The authors compare the results of Cournot (i.e., quantity) and Bertrand (i.e., price) competition. Throughout, Cournot prices are higher than those in Betrand competition, and aggregate welfare (i.e., total surplus–total transport costs) is higher under Bertrand than under Cournot.

Anderson et al. (1997) drop the assumption of uniform demand and consider logconcave demand functions, coupled with quadratic transportation costs. It turns out that if customers are more spread out, prices are higher, and that symmetric demand densities lead to symmetric locations of firms. Bester et al. (1996) reexamine d'Aspremont et al.'s (1979) Hotelling game without coordination (firm A is assumed to locate to the left of firm B) and allow mixed strategies. An infinite number of mixed-strategy Nash equilibria exists, and without coordination, the result of maximum differentiation is invalidated. Eaton (1972) follows Smithies (1941) by considering a model, which includes a linearly sloping price-demand function. The author also uses a modified zero conjectural variation assumption, according to which a firm will react unless undercut. In case of a short market, the result will be agglomeration of the firms, as the length of the market grows, duopoly locations approach the social optimum. Behavior in case of a triopoly is similar: as the length of the market grows, agglomeration forces get weaker. The paper by Kohlberg and Novshek (1982) examines a similar model.

There are a few contributions that examine spaces similar to a line: Eaton's (1976) model allows free entry on a circle, Kats's (1995) model locates duopolists on a circular market, whereas Tsai and Lai (2005) investigate the case of a market, in which customers are distributed along the sides of a triangle, and Braid (1989, 2013) looks at the case of intersecting roadways, i.e., intersecting lines.

14.4.8 UD1b, Plane, Nash Equilibria

Hurter and Lederer (1985) appear to have been among the few investigators to look at the subgame-perfect Nash equilibrium on the plane. Their contribution includes different cost functions for the firms and transportation costs that are proportional to Euclidean distances. Firm are supposed to locate in a given convex set. The authors show that there are no peripheral equilibrium locations. They also demonstrate that the locations that minimize the social costs for serving the entire market are a proper subset of equilibrium locations. Similarly, Tabuchi (1994) locates two firms in the two-dimensional space and uses quadratic transportation costs. The paper determines that for any convex set, there are no interior locational Nash equilibria. The author then determines that in a rectangle, Nash equilibrium has the facilities locate on opposite sides of the rectangle at their respective midpoints. If the rectangle is very long, the Nash equilibrium is unique.

This is not the same as d'Aspremont's et al. (1979) result, as while this result shows maximum differentiation in one direction, it has minimum differentiation in the other. Lederer and Hurter (1986) consider customers located in a subset of the two-dimensional plane with some typically nonuniform demand distribution and firms facing different production and transportation costs. Firms use spatial price discrimination and customer purchase goods from the cheapest source (a number of tie-breaking rules is specified). The resulting "location, then price" game has an equilibrium, and it is shown that identical firms (i.e., those with different production and transportation costs) do not co-locate. The analysis is then extended to nonidentical forms that locate on a disk, and again, there is no co-location.

14.4.9 UD1b, Networks, Nash Equilibria

Lederer and Thisse (1990) examine a competitive network location model, in which firms determine their respective locations and chosen technologies in stage 1, and the prices in stage 2. The authors use spatial price discrimination. In the usual backward recursion, the paper proves that for all first stage location and technology choices, the second stage pricing game has an equilibrium. The socially optimal location and technology choices of the first stage are also a Nash equilibrium. However, locational Nash equilibria may exist that are not socially optimal. An important feature is that if the transport cost function is concave, then the equilibrium locations will satisfy the node property. Labbé and Hakimi (1991) also use delivered pricing and, in addition, a linear price–quantity relation. The two-stage game locates facilities in stage 1, and determined quantities in stage 2. It turns out that for any fixed pair of locations, the quantity game has an equilibrium. If it is required that it is always profitable to supply any market of the graph with a positive quantity of goods, then a location equilibrium exists at the nodes of the graph. If this condition is not satisfied, then either a locational Nash equilibrium does not exist, or it exists on the edges of the graph.

14.4.10 UD1, Linear Market, Nash Equilibria

Among the earliest papers to follow Hotelling's lead is the work by Lerner and Singer (1937). The authors keep Hotelling's linear market and the assumption on linear transportation costs, but introduced a finite reservation price, and assert that each firm assumes that its competitor's location and price is fixed, and a firm only reacts if undercut. In such a case, equilibria do exist. The authors also extend their analysis to spatial price discrimination, which results in social optima. The contribution by Economides (1986) is most interesting, as it includes Hotelling's (1929) and d'Aspremont et al.'s (1979) results as special cases. The utility function includes a budget and the utility inherent in the product. The transportation costs are the facility—customer distance raised to some power α. The main result is that for α less than about 1.26 (which includes Hotelling's original case with $\alpha = 1$), no subgame-perfect Nash equilibrium exists, whereas for α greater than about 1.26, it does exist (which includes d'Aspremont et al.'s case of $\alpha = 2$). More specifically, for $\alpha \in [1.26, 1.6667]$, the equilibrium locations are strictly interior, while for $\alpha \geq 1.6667$, they are at the endpoints of the market.

Zhang (1995) discusses the case of a duopoly with quadratic transportation costs and reservation prices, in which decision makers make their decisions in three phases: locate first, then decide whether or not to adopt a price-matching policy, and then determine the price. The paper shows that if both players use price matching, high reservation prices lead to a unique Nash equilibrium "with tacit collusion on prices." Equilibrium locations for high reservation prices lie at the center of the market (minimum differentiation). Not surprisingly, they find that price matching reduces price competition. The paper of Smithies (1941), which has spawned many followers, discusses a Hotelling model with elastic demand and reservation prices. The author appears to have been the first to use "push" and "pull" forces (see also Eiselt and Laporte 1995). He also found that higher transportation costs lead to less competition, and as unit transportation costs increase, firm will move farther apart. Finally, the interesting contribution by Guo and Lai (2014) adds an online dealer to the brick-and-mortar duopolists. While customers purchasing from the latter, face the usual transportation costs, consumers who deal with the online firm have waiting/inconvenience cost. The authors demonstrate that an equilibrium does indeed exist given a relation between the unit transportation costs and the unit inconvenience cost.

14.4.11 UD1, Linear Market, von Stackelberg Solution

Bonanno's (1987) model examines location, which an incumbent can use to deter future entry of competitors. His model uses quadratic transportation costs, fixed setup costs for new stores and finite reservation prices. The proposed three-stage procedure has the incumbent decide how many stores to open, followed by the

potential entrant who must decide whether or not to enter and, if yes, where to locate his store (the choices of the follower are limited to zero or one store as to ensure tractability), followed by price competition. Given high setup costs, the leader is a monopolist and further entry is blocked. For moderate setup costs, the incumbent locates two stores at the social optimum, and entry is deterred. For even lower setup costs, entry can no longer be deterred by the incumbent.

Meza and Tombak's (2009) model uses uniform distribution, "sufficiently high" reservation prices, quadratic transportation costs, and potentially different production costs. The paper suggests a three-stage model, in which timing (of entry), location, and price are determined. The low-cost firm is the leader. It is possible for a higher-priced firm that is driven from the market, to re-enter at a later stage. With a small difference in costs, firms enter the market immediately with maximal differentiation. For a somewhat larger cost difference, the low-cost leader enters immediately, soon followed by the higher-cost firm, still maintaining maximal differentiation. For an even larger cost difference, the low-cost leader locates at an interior point, followed by its competitor that locates as far away as possible from the leader. With a very high cost difference, the low-cost leader locates at the center of the market and effective blocks all further entry.

14.4.12 UD1, Plane, Nash Equilibria

The paper by Irmen and Thisse (1998) considers a duopoly in d-dimensional real space with weighted squared Euclidean distances. Customers have a utility function that includes a reservation price, the product's price, and the sum of weighted distances between customer and the firm (the customer's ideal point and the product features, as this model is discussed in feature space). The key result is that if there is a main characteristic of the product, then there is a unique equilibrium in the location game, in which the two products exhibit maximum differentiation in that feature, while otherwise being identical. The authors cite an interesting application of their result in the news magazines *Time* and *Newsweek*, whose main difference is in the cover story. The similarity of this result and that by Tabuchi (1994) should also be noted.

14.4.13 UD2a, Linear Market, Nash Equilibria

The contribution by Eiselt (1991) appears to have been the first to use attraction function of the type "facility attractiveness divided by an increasing function of distance" for the purpose of locating competitive facilities. It is shown that as long as the weights are unequal, no equilibrium exists. The author then allows repeated sequential relocation. It turn out that facilities shuttle but converge towards fixed points whose location depends exclusively on the weights: if weights are similar,

the fixed points are close to center, otherwise they are close to the boundaries of the market. The paper then introduces fixed and variable relocation costs, which are subsequently used to force an equilibrium.

14.4.14 UD2a, Plane, von Stackelberg Solution

Drezner (1994b) locates a single new facility in the Euclidean plane with a winner-take-all allocation rule. For each customer, the paper determines a circle around the customer location, so that any facility located inside that circle will capture the customer. Such circles are then constructed for all customer points. This is then used to optimally locate a new facility with given attraction.

14.4.15 UD2a, Network, Nash Equilibria

Eiselt and Laporte (1991) investigate the existence of locational Nash equilibria on a tree, given an attraction function of the type facility attraction divided by distance to some power greater or equal to one. When the base attractions of the facilities are equal, equilibria always exist with either both facilities at the median of the tree (in case co-location is permitted) or with one facility at the median and the other adjacent to it in the largest subtree spanned by the median. For unequal base attractions, if co-location is permitted and the winner-take-all allocation rule applies, then an equilibrium never exists; otherwise (i.e., with co-location permitted and an allocation proportional to the attractions and in case location at the same vertex is prohibited), equilibria may or may not exist.

14.4.16 UD2A, Network, von Stackelberg Solution

von Stackelberg problems in networks enjoy quite some popularity among operations researchers. The main reasons are their relative tractability (the problems can, at least in their basic form, be formulated as integer linear programming problems). This is very much in contrast to the leader's problem, which is a bilevel integer programming problem. Suárez-Vega et al. (2007) employ an attraction function, defined as facility weight divided by an increasing concave function of the distance. Customers purchase proportionally from the facilities they are most attracted to, *provided* they are attracted to them by a measure that exceeds a minimally acceptable threshold. The authors describe a finite dominating set. They deal with the case of a single new facility, but the results generalize to multiple facilities (even though the computations will be more complex). Benati (2003) does not fix the number of facilities the follower is going to locate. Customer

behavior is modeled by a function that relates a customer's attraction to a facility to the sum of this customer's attractions to all facilities. This leads to a concave fractional problem, which is solved by a branch-and-bound method and heuristic concentration techniques.

14.4.17 UD2b, Network, von Stackelberg Solution

Aboolian et al. (2008) investigate a follower problem on a network with an exponential attraction function. In order to capture a customer's demand, the follower must be more attractive than the incumbent by a positive constant. The variable production costs are the same everywhere, and the fixed location costs are location-dependent. Co-location is not permitted. The model is loosely based on work by Serra and ReVelle (1999). The node property does not hold. The authors conjecture that there is a finite dominating set, but are unable to determine it in this nonlinear integer program. Marianov et al. (2008) replace the distance by travel time, and add waiting time as a competitive factor.

Consider now results relating to the probabilistic choice rules introduced in the previous section. Most papers are written by economists, who are mainly interested in the existence of Nash equilibria on a linear market.

14.4.18 UP1, Linear Market, Nash Equilibria

In all of these contributions, the parameter μ can be interpreted as the heterogeneity of the customer tastes with respect to the product under consideration. de Palma et al. (1987a) use fixed and equal prices and unit transportation costs t (in a linear cost function) in their triopoly model. Their main result is that for small values of μ/t, there are no symmetric equilibria. As the value of μ/t increases, there are symmetric dispersed equilibria, a further increase results in dispersed and agglomerated equilibria, while for large values of μ/t, only agglomerated equilibria exist. de Palma et al. (1985) consider the usual "first location, then price" game with a linear transport cost function, and n facilities located on a linear market of length L. The key result is that for large values of μ/tL, there is clustering of the facilities at equilibrium, while small values of μ/tL lead to dispersion. Braid (1988) locates n firms on a line segment, on which the demand occurs at five even spaced the facilities. de Palma et al. (1987b) discuss a duopoly under delivered pricing in their model with linear transportation costs with parameter t. Under sufficient heterogeneity (i.e., $\mu > t/8$), a centrally agglomerated location-price equilibrium exists. The result generalizes to n firms.

Finally in this category, we find the contribution by Anderson et al. (1992), which compares the three main pricing strategies in a duopoly setting. Transportation costs are assumed to be linear, and social surplus is defined as the sum of customer

surplus and the profits of both firms. Starting with small values of the heterogeneity factor μ, there is no equilibrium for mill pricing, and as μ increases, there are first symmetric dispersed equilibria, and finally, for large values of μ, there is a unique centrally agglomerated equilibrium. The case of uniform delivered demand just has no equilibrium for small μ, and centrally agglomerated equilibria for larger values of μ, and spatial discriminatory pricing has equilibria everywhere: outside the quartiles for very small values of μ that move towards a central agglomeration for sufficiently large values of μ.

14.4.19 UP1, Plane, Nash Equilibria and von Stackelberg Solutions

Choi et al. (1990) frame their discussion in the context of product positioning. Customers have a stochastic utility function that results in a logit model, and firms maximize their profit. It is known that as long as the profit functions are pseudo-concave, the game has a Nash equilibrium. The paper uses variational inequalities to analyze computational aspects. The key contribution is a von Stackelberg game with one leader and multiple followers. The solution of a von Stackelberg game in continuous space cannot be a Nash equilibrium, as is often the case in discrete spaces.

14.4.20 UP1, Network, Nash Equilibria

de Palma et al. (1989) investigate a very general model, in which n firms compete with each other, and each locates n_i facilities. Customers first choose a firm they want to patronize, and then they patronize the closest facility of that firm. (Note the similarity of this rule and Hakimi's "partially binary" choice rule.) The main result is that if consumer tastes are "sufficiently heterogeneous," then firm i will locate its n_i facilities at the n_i-median. If a stronger condition on taste heterogeneity is satisfied, then the resulting pattern—all firms locate their facilities at the n_i-medians—is the unique noncooperative Nash equilibrium. A special case is when all firms have the same number of facilities to locate, in which case all firms will locate their facilities at the same nodes, a case of minimum differentiation.

14.4.21 UP1, Network, von Stackelberg Solution

Benati (1999) discusses a maximum capture problem in the presence of heterogeneous customers. Given fixed demand, fixed and equal prices, as well as p leaders on

the market whose locations are known, The paper demonstrates that the follower's objective function is submodular, and that, given appropriate redefining of the problem's parameters, the problem can be formulated as an r-median model.

14.4.22 UP2, Plane, von Stackelberg Solution

Drezner et al. (2002) discuss a medianoid problem in the plane, in which customers' choices are modeled probabilistic and are based on attraction functions. The follower's objective is to minimize the probability that the new facility's revenue falls short of a given threshold. The optimal locations tend to markedly differ from those that are the result of the maximization of the expected market share, especially in those cases, in which the probability of failure is relatively small.

14.4.23 UP2, Network, von Stackelberg Solution

The main contribution of the work by Serra and Colomé (2001) is the comparison of various customer choice models. The basic setting includes fixed demand at the nodes of a network, one homogeneous good, and two profit-maximizing firms with identical cost structures. There are presently q facilities on the market. One new firm enters the market and attempts to locate p new facilities. Customer behavior is modeled as follows. Model 1 is the usual all-or-nothing assumption based on the closest facility, while Model 2 is a multiplicative competitive interaction Model (Nakashani and Cooper 1974), which assumes that the proportion of demand of customer i captured by facility j equals 1/customer-facility distance raised to the power of a parameter that indicates a customer's sensitivity with respect to distance, divided by the sum of such expressions, taken over all facilities. Model 3 is the standard proportional model, and Model 4 assumes partially binary preferences. It turns out that the simple Model 1 appears to be most robust, meaning that it has never more than an 8 % deviation from the solution that is based on the correct customer behavior.

14.4.24 Summary, Extensions, and Outlook

This chapter has described the basic Hotelling model, outlined its major components, described a three-stage procedure that models customer behavior, and has surveyed the literature regarding results of different models. While many different features have been included, most models, while they have some explanatory power, lack many facets of customer decision-making.

The most prominent difference between actual and assumed customer behavior involves the customers' trips to the chosen facility. In particular, all competitive models assume that customers make their individual purchases on a *special-single-purpose trip*, while this type of trip appears fairly rare in practice (with the exception of those trips related to work or emergency). However, a significant proportion of trips are multistop or multipurpose, since for some types of products consumers perform comparison shopping, visiting more than one facility selling the same item; or use the same trip to purchase more than one type or good. This is particularly true in a situation with high costs of fuel and long commuting distances.

One alternative is a *planned multipurpose trip with full information*. In such a case, a customer has set out with a plan, full knowledge about what to purchase at the individual stores (based, e.g., on advertisements or on-line information) and the distances between home base and individual stores (based on past experience). Typically, such a trip resembles a traveling salesman tour; for a good recent reference, see, e.g., Applegate et al. (2007). Planning multi-purpose shopping trips has been shown to foster the agglomeration of facilities; see, e.g., Marianov and Eiselt (2014).

A much more difficult extension concerns *trips without full information*. The main aspect of this single- or multi-purpose trip involves *feature search*. On such a trip, a customer will first patronize a store, obtain information about the features of the desired product (often, but not exclusively, its price), and will then decide, whether to purchase the product, or continue to some other store in order to potentially obtain a better deal. Such a search will incur certain costs (in terms of transportation costs and time), while expecting potential advantages in terms of better features, such as a lower price, better quality, or additional features. How long such searches will be will certainly depend, at least in part, on the amount of money involved and on the utility of a continued search, as compared to that of an immediate purchase on the basis of the information gathered up to this point. Houses, vehicles, furniture and similar high-priced items are typically purchased in this manner. Narula et al. (1983) present a model that includes price search, while Braid's (1996) noncompetitive location model that locates a main facility that has the desired product, and branch facilities, which have the product with a given probability. Customers can obtain information by means of phone search and visit search, respectively.

An interesting strand of research involves *flow capturing*, or *flow interception models* has been developed by Hodgson (1990), Berman et al. (1995), and Berman and Krass (1998). These models replace the assumption of customers making single trips to the chosen facility by assuming that they make purchases on their way to work. Considering work as one part of shopping, this model is a multipurpose shopping model with one fixed stop (work). Competing facilities will attempt to maximize their capture of the flow of customers to work. One of the main issues in these models involves the avoidance of double counting, i.e., customers who have made a purchase at one facility, have their demand drop to zero and they will not make another purchase on their trip. Typical applications for this type of behavior include child care facilities and gas stations.

Additional behavioral patterns involve window shopping and showrooming (the practice of getting advice and information about a product at local stores and the subsequent purchase at a presumably cheaper no-frills internet dealer). The latter behavior has already caused some problems among local stores, even though the aforementioned detrimental effects may be, at least partially, offset by the fact that customers typically obtain detailed technical information online, alleviating the local store from having (expensive) specialized sales staff.

Acknowledgments This paper was in part supported by a grant from the Natural Sciences and Engineering Research Council of Canada. This support is gratefully acknowledged. The authors would also like to thank an anonymous referee for his detailed comments that helped improve the exposition.

References

Aboolian R, Berman O, Krass D (2007) Competitive facility location model with concave demand. Eur J Oper Res 181(2):598–619

Aboolian R, Berman O, Krass D (2008) Optimizing pricing and location decisions for competitive service facilities charging uniform price. J Oper Res Soc 59(11):1506–1519

Anderson SP (1987) Spatial competition and price leadership. Int J Ind Organ 5:369–398

Anderson SP (1988) Equilibrium existence in the linear model of spatial competition. Economica 55(220):479–491

Anderson SP, de Palma A (1992) Spatial equilibrium with footloose firms. J Reg Sci 32(3):309–320

Anderson SP, Neven DJ (1989) Market efficiency with combinable products. Eur Econ Rev 33:707–719

Anderson SP, Neven DJ (1991) Cournot competition yields spatial agglomeration. International Economic Review 32:793–808

Anderson SP, de Palma A, Thisse J-F (1992) Social surplus and profitability under different spatial pricing policies. South Econ J 58:934–949

Anderson SP, Goeree JK, Ramer R (1997) Location, location, location. J Econ Theory 77:102–127

Aoyagi M, Okabe A (1993) Spatial competition of firms in a two-dimensional bounded market. Reg Sci Urban Econ 23:259–289

Applegate DL, Bixby RE, Chvátal V, Cook WJ (2007) The traveling salesman problem: a computational study. Princeton series in applied mathematics, Princeton University Press, Princeton

Beckmann MJ (1972) Spatial Cournot oligopoly. Papers Reg Sci Assoc 28:37–47

Benati S (1999) The maximum capture problem with heterogeneous customers. Comput Oper Res 26:1351–1367

Benati S (2003) An improved branch & bound method for the uncapacitated competitive location problem. Ann Oper Res 122:42–58

Berman O, Krass D (1998) Flow intercepting spatial interaction model: a new approach to optimal location of competitive facilities. Locat Sci 6:41–65

Berman O, Hodgson MJ, Krass D (1995) Flow-interception problems. In: Drezner Z (ed) Facility location: a survey of applications and methods. Springer, NewYork, pp 389–426

Bester H, de Palma A, Leininger W, Thomas J, von Thadden E-L (1996) A noncooperative analysis of Hotelling's location game. Games Econ Behav 12:165–186

Bhadury J (1996) Competitive location under uncertainty of costs. J Reg Sci 36(4):527–554

Bhadury J, Eiselt HA (1995) Stability of Nash equilibria in locational games. Recherche opérationnelle/Oper Res 29(1):19–33

Bonanno G (1987) Location choice, product proliferation and entry deterrence. Rev Econ Stud 54:37–45
Braid RM (1988) Heterogeneous preferences and non-central agglomeration of firms. Reg Sci Urban Econ 18:57–68
Braid RM (1989) Retail competition along intersecting roadways. Reg Sci Urban Econ 19:107–112
Braid RM (1996) The optimal locations of branch facilities and main facilities with consumer search. J Reg Sci 36(2):217–234
Braid RM (2013) The location of firms on intersecting roadways. Ann Reg Sci 50:791–808
Brown S (1989) Retail location theory: the legacy of Harold Hotelling. J Retail 65(4):450–470
Cancian M, Bills A, Bergstrom T (1995) Hotelling location problems with directional constraints: an application to television news scheduling. J Ind Econ 43:121–124
Caplin A, Nalebuff B (1991) Aggregation and imperfect competition: on the existence of equilibrium. Econometrica 59:25–60
Choi CS, DeSarbo WS, Harker PT (1990) Product positioning under price competition. Manage Sci 36(2):175–199
Cremer H, Marchand M, Thisse J-F (1991) Mixed oligopoly with differentiated products. Int J Ind Organ 9:43–53
D'Aspremont C, Gabszewicz JJ, Thisse J-F (1979) On Hotelling's 'Stability in competition.'. Econometrica 47:1145–1150
Dasci A, Laporte G (2005) A continuous model for multistore competitive location. Oper Res 53(2):263–280
Dasgupta P, Maskin E (1986) The existence of equilibrium in discontinuous economic games, I: theory. Rev Econ Stud 53:324–354
De Palma A, Ginsburgh V, Labbé M, Thisse J-F (1985) The principle of minimum differentiation holds under sufficient heterogeneity. Econometrica 53(4):767–781
De Palma A, Ginsburgh V, Thisse J-F (1987a) On existence of locational equilibria in the 3-firm Hotelling problem. J Ind Econ 36:245–252
De Palma A, Pontes JP, Thisse J-F (1987b) Spatial competition under uniform delivered pricing. Reg Sci Urban Econ 17:441–449
De Palma A, Ginsburgh V, Labbé M, Thisse J-F (1989) Competitive location with random utilities. Transp Sci 23:244–252
Drezner Z (1981) On a modified one-center model. Manage Sci 27(7):848–851
Drezner Z (1982) Competitive location strategies for two facilities. Reg Sci Urban Econ 12:485–493
Drezner T (1994a) Optimal continuous location of a retail facility, facility attractiveness, and market share, an interactive model. J Retail 70:49–64
Drezner T (1994b) Locating a single new facility among existing, unequally attractive facilities. J Reg Sci 34(2):237–252
Drezner T (1995) Competitive facility location in the plane. In: Drezner Z (ed) A chapter in facility location: a survey of applications and methods. Springer, New York, pp 285–300
Drezner T (2014) A review of competitive facility location in the plane. Logist Res 7(1):1–12
Drezner T, Drezner Z (1997) Replacing continuous demand with discrete demand in a competitive location model. Naval Res Logist 44:81–95
Drezner T, Eiselt HA (2002) Consumers in competitive location models. In: Drezner Z, Hamacher H (eds) Facility location: applications and theory. Springer, Berlin, pp 151–176
Drezner T, Drezner Z, Eiselt HA (1996) Consistent and inconsistent rules in competitive facility choice. J Oper Res Soc 47:1494–1503
Drezner T, Drezner Z, Salhi S (2002) Solving the multiple competitive facilities location problem. Eur J Oper Res 142:138–151
Durier R, Michelot C (1985) Geometrical properties of the Fermat–Weber problem. Eur J Oper Res 20:332–343
Eaton BC (1972) Spatial competition revisited. Can J Econ 5(2):268–278
Eaton BC (1976) Free entry in one-dimensional models: pure profits and multiple equilibria. J Reg Sci 16(1):21–33

Eaton BC, Lipsey RG (1975) The principle of minimum differentiation reconsidered: some new developments in the theory of spatial competition. Rev Econ Stud 42(1):27–49
Economides NS (1986) Minimal and maximal product differentiation in Hotelling's duopoly. Econ Lett 21:67–71
Eiselt HA (1991) Different pricing policies in Hotelling's duopoly model. Cahiers du CERO 33:195–205
Eiselt HA (1992) Hotelling's duopoly on a tree. Ann Oper Res 40:195–207
Eiselt HA (1998) Perception and information in a competitive location model. Eur J Oper Res 108:94–105
Eiselt HA (2011) Equilibria in competitive location models. In: Eiselt HA, Marianov V (eds) Chapter 7 in foundations of location analysis. Springer Science + Business Media, New York, pp 139–162
Eiselt HA, Bhadury J (1998) Reachability of locational Nash equilibria. OR Spektr 20:101–107
Eiselt HA, Laporte G (1991) Locational equilibrium of two facilities on a tree. Oper Res 25(1):5–18
Eiselt HA, Laporte G (1993) The existence of equilibria in the 3-facility Hotelling model on a tree. Transp Sci 27(1):39–43
Eiselt HA, Laporte G (1995) Objectives in location problems. In: Drezner Z (ed) Facility location: a survey of applications and methods, Springer, New York, pp 151–180
Eiselt HA, Laporte G, Thisse J-F (1993) Competitive location models: a framework and bibliography. Transp Sci 27(1):44–54
Fernández P, Pelegrín B, Dolores M, Pérez G, Peeters PH (2007) A discrete long-term location-price problem under the assumption of discriminatory pricing: formulations and parametric analysis. Eur J Oper Res 179:1050–1062
Gabszewicz JJ, Thisse J-F (1986) Spatial competition and the location of firms. In: Gabszewicz JJ, Thisse J-F, Fujita M, Schweizer U (eds) Location theory. Harwood Academic Publishers, Chur, pp 1–71
Gabszewicz JJ, Thisse J-F, Fujita M, Schweizer U (1986) Location theory. Harwood Academic Publishers, Chur
García Pérez MD, Pelegrín PB (2003) All Stackelberg location equilibria in the Hotelling's duopoly model on a tree with parametric prices. Ann Oper Res 122:177–192
Ghosh A, Buchanan B (1988) Multiple outlets in a duopoly: a first entry paradox. Geograph Anal 20:111–121
Giudici P, Passerone G (2002) Data mining of association structures to model consumer behavior. Comput Stat Data Anal 38:533–541
Guo W-C, Lai F-C (2014) Spatial competition with quadratic transport costs and one online firm. Ann Reg Sci 52(1):309–324
Hakimi SL (1964) Optimum locations of switching centres and the absolute centres and medians of a graph. Oper Res 12:450–459
Hakimi SL (1983) On locating new facilities in a competitive environment. Eur J Oper Res 12:29–35
Hakimi LS (1990) Locations with spatial interactions: competitive locations and games. In: Francis RL, Mirchandani PB (eds) Discrete location theory. Wiley, New York, pp 439–478
Hamilton JH, Thisse J-F, Weskamp A (1989) Spatial discrimination: Bertrand vs Cournot in a model of location choice. Reg Sci Urban Econ 19(1):87–102
Hamoudi H, Moral MJ (2005) Equilibrium existence in the linear model: concave versus convex transportation costs. Papers Reg Sci 84(2):201–219
Hansen P, Labbé M (1988) Algorithms for voting and competitive location on a network. Transp Sci 22(4):278–288
Hay DA (1976) Sequential entry and entry-deterring strategies in spatial competition. Oxf Econ Pap 28(2):240–257
Hodgson MJ (1990) A flow capturing location-allocation model. Geogr Anal 22:270–279
Hotelling H (1929) Stability in competition. Econ J 39:41–57
Huff DL (1964) Defining and estimating a trading area. J Mark 28:34–38

Hurter AP Jr, Lederer PJ (1985) Spatial duopoly with discriminatory pricing. Reg Sci Urban Econ 15:541–553
Infante-Macias R, Muñoz-Perez J (1995) Competitive location with rectilinear distances. Eur J Oper Res 80:77–85
Irmen A, Thisse J-F (1998) Competition in multi-characteristic spaces: Hotelling was almost right. J Econ Theory 78:76–102
Kats A (1995) More on Hotelling's 'Stability in competition'. Int J Ind Organ 13:89–93
Kohlberg E (1983) Equilibrium store locations when consumers minimize travel time plus waiting time. Econ Lett 11:211–216
Kohlberg E, Novshek W (1982) Equilibrium in a simple price-location model. Econ Lett 9:7–15
Kress D, Pesch E (2012) Sequential competitive location on networks. Eur J Oper Res 217:483–499
Labbé M, Hakimi SL (1991) Market and locational equilibrium for two competitors. Oper Res 39(5):749–756
Lane WJ (1980) Product differentiation in a market with endogenous sequential entry. Bell J Econ 11(1):237–260
Lederer PJ, Hurter AP Jr (1986) Competition of firms: discriminatory pricing and location. Econometrica 54(3):623–640
Lederer PJ, Thisse J-F (1990) Competitive location on networks under delivered pricing. Oper Res Lett 9:147–153
Lerner AP, Singer HW (1937) Some notes on duopoly and spatial competition. J Polit Econ 45(2):145–186
Liou JJH (2009) A novel decision rules approach for customer relationship management of the airline market. Exp Syst Appl 36:4374–4381
Lösch A (1954) The economics of location, 2nd rev. edn. Yale University Press, New Haven
Marianov V, Eiselt HA (2014) Agglomeration in competitive location. Ann Oper Res (forthcoming), doi:10.1007/s10479-014-1704-5
Marianov V, Taborga P (2001) Optimal location of public health centres which provide free and paid services. J Oper Res Soc 52:391–400
Marianov V, Serra D, ReVelle C (1999) Location of hubs in a competitive environment. Eur J Oper Res 114:363–371
Marianov V, Ríos M, Taborga P (2004) Finding locations for public service centers that compete with private centers: effects of congestion. Pap Reg Sci 83(4):631–648
Marianov V, Ríos M, Icaza MJ (2008) Facility location for market capture when users rank facilities by travel and waiting times. Eur J Oper Res 191(1):32–44
Matsumura T, Matsushima N (2009) Cost differentials and mixed strategy equilibria in a Hotelling model. Ann Reg Sci 43:215–234
Matsumura T, Shimizu D (2006) Cournot and Bertrand in shipping models with circular markets. Pap Reg Sci 85(4):585–598
McFadden D (1974) Conditional logit analysis of qualitative choice behavior. In: Zarembka P (ed) Frontiers in econometrics. Academic, New York
Megiddo N, Zemel E, Hakimi SL (1983) The maximum coverage location problem. SIAM J Algebr Discrete Methods 4:253–261
Meza S, Tombak M (2009) Endogenous location leadership. Int J Ind Organ 27:687–707
Nakashani M, Cooper LG (1974) Parameter estimation for a multiplicative competitive interaction mode-least squares approach. J Market Res XI:303–311
Narula SC, Harwitz M, Lentnek B (1983) Where shall we shop today? A theory of multiple-stop, multiple-purpose shopping trips. Pap Reg Sci Assoc 53:159–173
Neven DJ (1987) Endogenous sequential entry in a spatial model. Int J Ind Organ 5:419–434
Okabe A, Aoyagi M (1991) Existence of equilibrium configurations of competitive firms on an infinite two-dimensional space. J Urban Econ 29:349–370
Okabe A, Suzuki A (1987) Stability of spatial competition for a large number of firms on a bounded two-dimensional space. Environ Plann A19:1067–1082

Osborne MJ, Pitchik C (1986) The nature of equilibrium in a location model. International Economic Review 27(1):223–237
Osborne MJ, Pitchik C (1987) Equilibrium in Hotelling's model of spatial competition. Econometrica 55(4):911–922
Pelegrín B, Fernández P, Suárez R, García MD (2006) Single facility location on a network under mill and delivered pricing. IMA J Manage Math 17(4):373–385
Penn M, Kariv O (1989) Competitive location in trees: parts I and II. Working paper
Plastria F (1992) On destination optimality in asymmetric distance Fermat–Weber problems. Ann Oper Res 40:355–369
Plastria F (2001) Static competitive facility location: an overview of optimization approaches. Eur J Oper Res 129:461–470
Prescott EC, Visscher M (1977) Sequential location among firms with foresight. Bell J Econ 8:378–393
Raiport JJ, Sviokla JJ (1994) Managing in the market-space. Harv Bus Rev 72(5):141–150
Reilly WJ (1931) The law of retail gravitation. Knickerbocker Press, New York
ReVelle CS (1986) The maximum capture or "Sphere of influence" location problem: Hotelling revisited on a network. J Reg Sci 26(2):343–358
Rothschild R (1979) The effect of sequential entry on choice of location. Eur Econ Rev 12:227–241
Selten R (1975) Re-examination of the perfectness concept for equilibrium points in extensive games. Int J Game Theory 4:22–55
Serra D, Colomé R (2001) Consumer choice and optimal locations models: formulations and heuristics. Pap Reg Sci 80:439–464
Serra D, ReVelle C (1994) Market capture by two competitors: the preemptive location problem. J Reg Sci 34(4):549–561
Serra D, ReVelle C (1995) Competitive location in discrete space. In: Drezner Z (ed) Facility location: a survey of applications and methods. Springer, Berlin
Serra D, ReVelle C (1999) Competitive location and pricing on networks. Geogr Anal 31(1):109–129
Serra D, ReVelle C, Rosing K (1999) Surviving in a competitive spatial market: the threshold capture model. J Reg Sci 39(4):637–652
Shaked A (1975) Non-existence of equilibrium for the two-dimensional three-firms location problem. Rev Econ Stud 42(1):51–56
Shaked A (1982) Existence and computation of mixed strategy Nash equilibrium for 3-firms location problem. J Ind Econ 31(1–2):93–96
Shigehiro S, Shiode S, Hiroaki I, Teraoka Y (1995) A competitive facility location problem. In: Fushimi M, Tone K (eds) Proceedings of APORS '94, Singapore, pp 251–257
Slater PJ (1975) Maximin facility location. J Res Natl Bur Stand 79B:107–115
Smithies A (1941) Optimum location in spatial competition. J Polit Econ 49:423–439
Song HS, Kim JK, Kim SH (2001) Mining the change of customer behavior in an internet shopping mall. Exp Syst Appl 21(3):157–168
Spoerhase J, Wirth H-C (2008) $(r|p)$ centroids problems on paths and trees. Theor Comput Sci 410(47–49):5128–5137
Stevens B (1961) An application of game theory to a problem in location strategy. Pap Reg Sci Assoc 7:143–157
Suárez-Vega R, Santos-Peñate DR, Dorta-González P (2004) Competitive multi-facility location on networks: the $(r|X_p)$-medianoid problem. J Reg Sci 44(3):569–588
Suárez-Vega R, Santos-Peñate DR, Dorta-González D (2007) The follower location problem with attraction thresholds. Pap Reg Sci 86(1):123–137
Suárez-Vega R, Santos-Peñate DR, Dorta-González D (2014) Location and quality selection for new facilities on a network market. Ann Reg Sci 52(2):537–560
Tabuchi T (1994) Two-stage, two-dimensional spatial competition between two firms. Reg Sci Urban Econ 24:207–227
Tabuchi T, Thisse J-F (1995) Asymmetric equilibria in spatial competition. Int J Ind Organ 13:213–227

Teitz MB (1968) Locational strategies for competitive systems. J Reg Sci 8(2):135–138
Thisse J-F, Wildasin DE (1995) Optimal transportation policy with strategic locational choice. Reg Sci Urban Econ 25:395–410
Tsai J-F, Lai F-C (2005) Spatial duopoly with triangular markets. Pap Reg Sci 84(1):47–59
von Stackelberg H (1943) Grundlagen der theoretischen Volkswirtschaftslehre (translated as: The theory of the market economy). W. Hodge & Co. Ltd., London
Younies H, Eiselt HA (2011) Sequential location models. In: Eiselt HA, Marianov V (eds) Chapter 8 in foundations of location analysis. Springer, Berlin
Zhang ZJ (1995) Price-matching policy and the principle of minimum differentiation. J Ind Econ 43:287–299

Chapter 15
Location-Routing and Location-Arc Routing

Maria Albareda-Sambola

Abstract This chapter overviews the most relevant contributions on location-routing problems. Although there exist many different models where location and routing decisions must be made in an integrated way, the chapter focuses on the so-called classical location-routing problems without entering into the details of other related problems that might be included in the location-routing area from a more general point of view. Reflecting the imbalance in the existing literature and available approaches, the case of problems with node routing is treated in detail throughout the chapter, while results concerning arc routing problems are concentrated in a single section.

Keywords Discrete location-routing • Heuristics • Mathematical programming

15.1 Introduction

Combined location-routing problems (LRPs) are location problems where the service to customers is provided by a fleet of vehicles in less-than-truckload routes. That is, more than one customer can be served in one vehicle route from a facility. Therefore, the cost of servicing a customer in a solution of a location-routing problem does not only depend on the facility it is assigned to, but also on the route followed by the vehicle that services it. As happens with pure vehicle routing problems, a basic distinction needs to be made when referring to LRPs, depending on whether the customers are associated with nodes or links of the underlying network. In the first case, in order to provide service to a customer, a vehicle has to visit the corresponding node, whereas in the second case, the vehicle has to traverse the corresponding link. Most of the literature on LRPs is in fact devoted to node routing LRPs and only a few references are concerned with solving some variant with arc routing. For this reason, the name *location-routing problem* is commonly used to refer to problems where customers are located at the nodes, whereas the term *location-arc routing problem* (LARP) is used when customers are located on

M. Albareda-Sambola (✉)
Universitat Politècnica de Catalunya, BarcelonaTech, Barcelona, Spain
e-mail: maria.albareda@upc.edu

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_15

the links of the network. In both cases, the need to design vehicle routes to evaluate the cost of a set of facilities adds an extra level of difficulty to these problems which are, in general, $\mathcal{NP}$-hard.

The first works addressing LRPs date back to the 1960s (e.g. Von Boventer 1961 and Maranzana 1964). However, it was not until the end of the 1980s, when a solid knowledge on both pure location and routing problems was achieved, that location-routing became a really active field of research. The most common approach in the first references addressing this type of problems was to make locational and routing decisions in two separate steps, although it is well-known that this is most likely to yield suboptimal solutions, as shown in Salhi and Rand (1989). For this reason, more recent references address both decisions simultaneously.

LRPs arise as a natural extension of both, location and vehicle routing problems. Moreover, there are several settings where LRPs appear naturally. For example, Schittekat and Sörensen (2009) study the optimization problem arising in some automotive companies that use third-party logistics partners for the distribution of spare parts and model it as a large scale LRP. Other examples of real applications where extensions of the LRP need to be solved are given in Ahn et al. (2012), where the authors present a LRP with profits faced by NASA while planning planetary surface exploration, or in Samanlioglu (2013) where hazardous waste management of a Turkish region is dealt with by solving a multiobjective LRP.

Although there exist papers dealing with planar LRPs (see, for instance, Manzour-al-Ajdad et al. 2012 or Salhi and Nagy 2009), most of the studies concerning LRPs deal with discrete location problems. As a consequence, this chapter will only consider this type of LRPs. Moreover, it does not pretend to be a complete survey of all works in the literature addressing discrete LRPs, and only presents the state of the art methods and the tools that have proven to be the most suitable ones to tackle LRPs. For a complete recent survey on works concerned with LRPs the reader is referred to Prodhon and Prins (2014). The reader can also find a taxonomy of location-routing models and the related literature in Borges Lopes et al. (2013). Earlier works are surveyed in Nagy and Salhi (2007). Given the little attention that LARPs have received, this chapter is also mostly concentrated on LRPs with node routing, and the most relevant issues concerning LARPs are gathered in a single section.

The remainder of this chapter is organized as follows. Section 15.2 provides a formal definition of the considered problems, together with the notation that will be used throughout the chapter. The next two sections describe the main scientific contributions on LRPs; Sect. 15.3 explores the different types of LRP formulations, together with the most relevant valid inequalities used in exact methods, whereas Sect. 15.4 is concerned with heuristic algorithms. The main findings regarding LARPs are outlined in Sect. 15.5, and 15.6 concludes the chapter.

15.2 Problem Definition and Notation

Let J be a set of customers and I a set of locations where facilities can be placed. For each candidate location $i \in I$, let f_i be the cost of setting up a facility at i, and for each arc (i, j) with $i, j \in I \cup J$, let ℓ_{ij} be its length or cost. The basic variant of the LRP consists of choosing a set of locations from I and defining closed routes starting and ending at one of these facilities such that each customer is visited by exactly one of the routes, subject to side constraints. The goal is to minimize the total cost, which typically includes the sum of facility set-up costs plus a traveling cost. We also denote by G the underlying graph of an LRP instance formed by the set of vertices $V = I \cup J$ and the set of links $E = E_{IJ} \cup E_I$, where E_{IJ} contains all links connecting one facility with one customer, and E_J contains all links connecting two different customers. In what follows, both, directed and undirected formulations will be presented. For ease of notation, E will be used indistinctly to denote the set of (directed) arcs (i, j) or the set of (undirected) edges $\{i, j\}$. For any set of nodes $S \subseteq V$, E_S will denote the set of links with both endpoints in S.

If a weight w_j is associated with each customer $j \in J$, capacity constraints can be considered by imposing a maximum weight Q delivered by a vehicle or a maximum weight q_i delivered from each facility $i \in I$. From now on, Q will be referred to as the vehicle capacity, and q_j as the facility capacity and, for each set of customers $S \subseteq J$, $w(S)$ will denote the total weight of customers in S: $w(S) = \sum_{j \in S} w_j$. LRPs considering either type of constraint, or both of them, are referred to as Capacitated LRPs (CLRPs). Additionally, many papers consider fixed vehicle utilization costs, g, and a limited size fleet indexed in set K. Figure 15.1 depicts an LRP solution.

Further considerations and characteristics of the main elements of the problem (number of facilities to locate, types of customers, size and characteristics of the

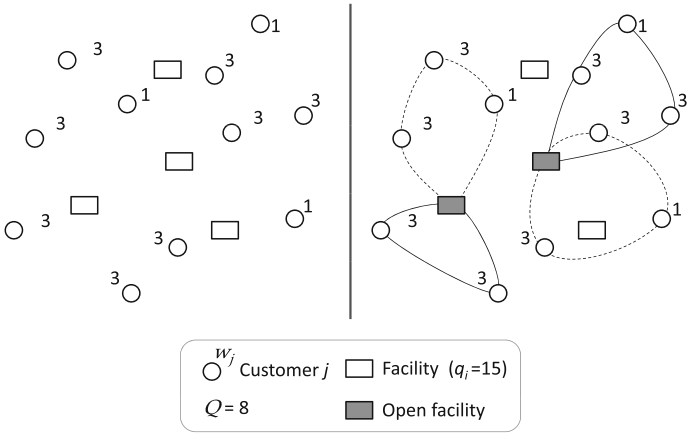


Fig. 15.1 Example of an LRP solution

vehicle fleet, time horizon, etc.) give rise to a large variety of LRPs. A comprehensive recent classification, following the ideas already presented in Laporte (1988) can be found in Borges Lopes et al. (2013).

The main difficulty when modeling LRPs through mathematical programming formulations is to ensure that each vehicle tour is connected to exactly one facility; that is, there are no closed tours visiting only customers, and there are no paths connecting two different facilities. Therefore, incorporating the design of vehicle routes within facility location problems entails a relevant additional level of difficulty. Furthermore, as some authors argue, facility location is most often a strategic decision, while vehicle routing is operational. These facts have discouraged many researchers from considering combined LRPs. However, although routing decisions can be readjusted relatively often once the facilities are established, the possible configurations of the routes are strongly conditioned by these locations. Therefore, if locations are chosen without taking into account the routing component of the final system, initial savings in the facilities set up costs may not compensate for large losses in distribution in the long run. Consider, for instance, the extreme situation depicted in Fig. 15.2. In this example, assume that the capacity of any of the two candidate facilities (black squares) is sufficient to serve all customers (white circles), and there is only one vehicle available at each location, also with a large enough capacity. If one single location is to be chosen and routing costs are ignored (i.e. if an uncapacitated facility location problem is considered in this setting) obviously, the facility will be located at 2. However, if a tour needs to be defined to serve all the customers once this facility is set, its cost will be $2M + (11\pi M)/6 \simeq 7.76M$. On the other hand, if the facility is set at node 1, a better route, with cost $2\pi M \simeq 6.28M$ can be defined. Since distribution is most often a repetitive activity, this extra routing cost for having chosen facility location 2 will be incurred regularly and, after some time, these accumulated extra costs can be larger than the initial possible savings in set up costs.

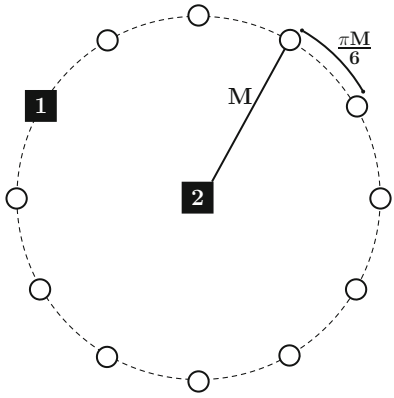


Fig. 15.2 Influence of facility location on the routing costs

15.3 Formulations and Exact Algorithms

The available exact algorithms for solving LRPs rely on mathematical programming formulations of the problem. Most of these formulations have been developed around the existing formulations for discrete facility location problems and multi-depot vehicle routing problems. Since the early formulations of Golden et al. (1977) and of Perl and Daskin (1985), several LRP formulations have been studied. CLRPs have received particular attention, since they are amongst the most basic LRPs. This section will concentrate on these problems.

As mentioned above, the main difficulty when developing a formulation for an LRP model is to guarantee that each route will start and end at one facility and neither closed loops visiting only customers, nor paths connecting two different facilities will be formed. For this reason, to a large extent, the developments concerning formulations for LRP models are strongly related with the literature on capacitated vehicle routing problems, especially, on multi-depots problems. As happens in these problems, one can assume, without loss of generality, that an optimal solution exists in which no edge of E_I is used more than twice and the only edges used twice, if any, belong to E_{IJ}. This is actually the case of problem instances in which the edge lengths satisfy the triangle inequality. Any instance can in fact be easily transformed into an equivalent one satisfying this property, by replacing the actual length of each edge with the length of a shortest path connecting its endpoints.

Broadly speaking, the existing formulations for the LRP can be classified in either of two families. On the one hand, one can find the so-called flow formulations, where different sets of variables are used to determine the set of located facilities and to describe the vehicle routes. On the other hand, one can find set covering formulations, where one single variable is defined associated with each feasible vehicle route. To a large extent, the appropriate solution method depends on the formulation employed; while branch-and-cut approaches are the most suitable for flow formulations, set covering formulations are in general better suited for algorithms based on column generation. The most recently presented algorithms combine column generation and cut generation methods.

15.3.1 Flow Formulations

Within the flow formulations, different models can be distinguished according to two criteria: the number of indices of the variables used to define the vehicle routes (including or not a third index to identify which vehicle uses a given link), and the nature of these variables, known as commodity flow variables when they consider the quantity of goods traveling on every link and as vehicle flow variables when they only indicate whether it is used or not.

An early example of a three-index vehicle flow formulation is that of Perl and Daskin (1985). In fact, this reference defines a three-layer problem with suppliers, distribution centers and customers where, in addition to the characteristics of the basic LRP, the authors consider variable costs associated with the throughput at each distribution center, and extra constraints limiting the length of the routes. The proposed formulation, simplified by excluding these extra considerations, is described next. To this end, the following binary variables will be used:

- For each $i \in I$, y_i indicates whether a facility is established at i.
- For each $i \in I, j \in J$, x_{ij} indicates whether customer j is served from facility i
- For each $(i, j) \in E$ and $k \in K$, z_{ijk} indicates whether vehicle k uses arc (i, j).

Using the above variables, a three index vehicle flow formulation for the LRP is detailed next:

$$\text{(LRP1) minimize} \quad \sum_{i \in I} f_i y_i + \sum_{k \in K} \sum_{(i,j) \in E} \ell_{ij} z_{ijk} \tag{15.1}$$

$$\text{subject to} \quad \sum_{k \in K} \sum_{i \in V} z_{ijk} = 1 \qquad j \in J \tag{15.2}$$

$$\sum_{j \in J} w_j \sum_{i \in V} z_{ijk} \leq Q \qquad k \in K \tag{15.3}$$

$$\sum_{j \in J} w_j x_{ij} - q_i y_i \leq 0 \qquad i \in I \tag{15.4}$$

$$\sum_{k \in K} \sum_{i \in S} \sum_{j \in V \setminus S} z_{ijk} \geq 1 \qquad I \subseteq S \subset V \tag{15.5}$$

$$\sum_{j \in V} z_{ijk} - \sum_{j \in V} z_{jik} = 0 \qquad k \in K, i \in V \tag{15.6}$$

$$\sum_{i \in I} \sum_{j \in J} x_{ijk} \leq 1 \qquad k \in K \tag{15.7}$$

$$\sum_{t \in J} z_{itk} + \sum_{t \in V} z_{jtk} - x_{ij} \leq 1 \qquad i \in I, j \in J, k \in K \tag{15.8}$$

$$y_i \in \{0, 1\} \qquad i \in I \tag{15.9}$$

$$x_{ij} \in \{0, 1\} \qquad i \in I, j \in J \tag{15.10}$$

$$z_{ijk} \in \{0, 1\} \qquad (i, j) \in E, k \in K. \tag{15.11}$$

Constraints (15.2) mean that each customer is reached by one vehicle route, while constraints (15.3) and (15.4) are vehicle and plant capacity constraints, respectively. Additionally, constraints (15.4) guarantee that customers will be served from opened

facilities. Connectivity constraints (15.5) ensure that each vehicle route includes a facility, while flow conservation constraints (15.6) ensure that z variables do indeed define routes, and constraints (15.7) mean that these routes visit one single facility. Finally, constraints (15.8) force the x and z variables to take consistent values.

Formulations of this type tend to be rather large, on the one hand, because they have an exponential number of connectivity constraints and, on the other hand, because they have $O(|V|^3)$ variables. Connectivity constraints, as well as additional valid inequalities, have traditionally been dealt with by using cutting plane procedures, such as branch-and-cut. However, even after relaxing connectivity constraints, the size of the formulations remains too large for solving realistic size instances.

As an alternative, several authors have worked on formulations where vehicle flow variables z do not include the third index to identify which vehicle uses each arc. In fact, early works addressing the particular cases of the LRP with one single depot or one single route per depot, such as Laporte and Nobert (1981) or Laporte et al. (1983) already used this type of approach.

A very successful example of this type of formulations is presented in Belenguer et al. (2011). In this case, the authors propose an undirected formulation that uses the following variables:

- For each $i \in I$, y_i indicates whether a facility is established at i.
- For each edge $\{i, j\} \in E$, z^1_{ij} indicates whether edge $\{i, j\}$ is used exactly once in the solution.
- For each edge $\{i, j\} \in E_{IJ}$, z^2_{ij} indicates whether edge $\{i, j\}$ is used twice in the solution.

Note that, as mentioned above, it can be assumed that the only edges that can be traversed twice in an optimal solution belong to E_{IJ} and, therefore, variables z^2 are only defined for those edges.

Additionally to the above variables, the following notation is used. For each set of customers $S \subseteq J$, $\kappa(S)$ is a lower bound on the minimum number of vehicles needed to serve the aggregate demand of all customers in set S. The most commonly used bound in this type of formulations is

$$\kappa_1(S) = \left\lceil \frac{1}{Q} \sum_{j \in S} w_j \right\rceil .$$

However, instead of $\kappa_1(S)$ some authors have used the optimal value of the bin packing problem defined by the weights of the customers in S, and bin size equal to the vehicle capacity, Q. In what follows, this second bound will be referred to as $\kappa_2(S)$.

The formulation proposed in Belenguer et al. (2011) is

$$\text{(LRP2) minimize} \sum_{i\in I} f_i y_i + \sum_{\{i,j\}\in E} \ell_{ij} z^1_{ij} + \sum_{\{i,j\}\in E_{IJ}} 2\ell_{ij} z^2_{ij} \tag{15.12}$$

$$\text{subject to} \sum_{i\in I} 2z^2_{ij} + \sum_{i\in V\setminus\{j\}} z^1_{ij} = 2 \qquad j \in J \tag{15.13}$$

$$z^1_{ij} + z^2_{ij} \le y_i \qquad i \in I, j \in J \tag{15.14}$$

$$\sum_{i,j\in S} z^1_{ij} \le |S| - \kappa(S) \qquad S \subseteq J \tag{15.15}$$

$$\sum_{s\in S}\sum_{j\in J\setminus S} z^1_{sj} + \sum_{t\in I\setminus\{i\}}\sum_{s\in S}(z^1_{ts} + 2z^2_{ts}) \ge 2 \qquad i \in I, S \subset J; w(S) > q_i \tag{15.16}$$

$$z^1_{jt} + \sum_{s\in S}(z^1_{sj} + z^1_{st}) + \sum_{s,u\in S} z^1_{su} + \sum_{i\in I'} z^1_{ij} + \sum_{i\in I\setminus I'} z^1_{it} \le |S| + 2 \qquad S \subset J, I' \subset I; j,t \in J \setminus S \tag{15.17}$$

$$\sum_{i\in I}(z^1_{ij} + z^2_{ij}) \le 1 \qquad j \in J \tag{15.18}$$

$$y_i \in \{0,1\} \qquad i \in I \tag{15.19}$$

$$z^1_{ij} \in \{0,1\} \qquad \{i,j\} \in E \tag{15.20}$$

$$z^2_{ij} \in \{0,1\} \qquad \{i,j\} \in E_{IJ}. \tag{15.21}$$

The original formulation includes an extra term in the objective function to account for fixed costs for the use of vehicles. Although this term has not been included here, these costs can be easily included in the above formulation by suitably modifying the lengths ℓ_{ij} for each $\{i,j\} \in E_{IJ}$.

In this formulation, constraints (15.13) are the degree constraints, which force each customer to be visited by some route. Constraints (15.14) are imposed in order to ensure that no route is rooted at a closed facility. Constraints (15.15) play two major roles. On the one hand, they forbid solutions with subtours which are not linked to any facility. On the other hand, they ensure that the vehicle capacities are not exceeded. Note that only z^1 variables are involved in these constraints since each z^2 variable is associated with one complete facility-customer-facility tour, which will not violate the vehicle capacity constraints in any feasible LRP instance. Facility capacities are imposed through constraints (15.16): if a set of customers S cannot be fully served from a given facility i because of its capacity, then at least one customer in S must be visited by a vehicle route rooted at a different facility and, therefore, at least two edges must be used that link set S with customers

outside it, or to some facility different from i. Additionally, since individual routes are not identified using 2-index variables, it is necessary to explicitly forbid tours connecting two different facilities. This is done by means of the so-called path elimination constraints (15.17). Additional constraints (15.18) are needed to forbid paths connecting two facilities through one single customer. The path elimination constraints are similar to the chain-barring constraints introduced by Laporte et al. (1988).

Using this formulation enriched with some families of valid inequalities, Belenguer et al. (2011) were able to solve within less than 2 h instances of up to 50 customers and five potential facilities.

15.3.2 Set-Partitioning Formulations

Set partitioning formulations for the LRP were introduced much later than flow formulations. Indeed, papers addressing this type of formulations have appeared relatively recently, in parallel with similar formulations for vehicle routing problems. The first such formulation was presented in Berger et al. (2007); the slightly different formulation presented in Akca et al. (2009) was later used in Baldacci et al. (2011) and further strengthened by Contardo et al. (2014a).

In order to present this type of formulations, some extra notation is required. Variables now correspond to the possible vehicle routes that are feasible with respect to the vehicle capacity and serve more than one customer. These routes will be indexed in $\Gamma = \cup_{i \in I} \Gamma_i$, where Γ_i gathers the routes starting from facility i. The return trips from a facility to a single customer will be dealt with separately. For each route $r \in \Gamma$, we will denote by ℓ_r the total length of the route, by w_r its total demand and, for each edge $\{i, j\} \in E$, the coefficient a_{ijr} will denote the number of times edge $\{i, j\}$ is used in route r. Note that coefficients a_{ijr} are binary if route r is elementary, but can take larger values if non-elementary routes are allowed.

The formulation exploited by Contardo et al. (2014a) uses the following binary variables:

- For each $i \in I$, y_i indicates whether a facility is established at i.
- For each $i \in I$ and $j \in J$, z^2_{ij} indicates whether a return trip from facility i to customer j is part of the solution.
- For each route $r \in \Gamma$, λ_r indicates whether route r is used.

$$\text{(LRP3) minimize} \sum_{i \in I} f_i y_i + \sum_{r \in \Gamma} \ell_r \lambda_r + \sum_{\{i,j\} \in E_{IJ}} 2\ell_{ij} z^2_{ij} \tag{15.22}$$

$$\text{subject to} \sum_{r \in \Gamma} \sum_{i \in V} a_{ijr} \lambda_r + \sum_{i \in I} 2z^2_{ij} = 2 \qquad j \in J \tag{15.23}$$

$$\sum_{r\in\Gamma_i}\sum_{\{j,s\}\in E}(w_j+w_s)a_{jsr}\lambda_r+\sum_{j\in J}2w_j z_{ij}^2\le 2q_i y_i \quad i\in I \tag{15.24}$$

$$y_i\in\{0,1\} \quad i\in I \tag{15.25}$$

$$z_{ij}^2\in\{0,1\} \quad \{i,j\}\in E \tag{15.26}$$

$$\lambda_r\in\{0,1\} \quad r\in\Gamma. \tag{15.27}$$

Here, constraints (15.23) ensure that each customer is either visited once by one of the selected routes, or in a round trip from a facility. Facility capacities are stated by constraints (15.24). For ease of notation, in these constraints, an artificial demand $w_i = 0$ is defined for each facility i.

Of course, in order to take advantage of this formulation it is essential to use a method based on column generation since the number of λ variables is exponential. Therefore, a crucial issue when developing exact solution methods based upon this formulation is the pricing problem. Here, the pricing problem consists of finding negative cost vehicle routes in Γ. It belongs to the family of resource constrained shortest path problems, which have been the focus of an abundant literature, mostly because they appear as pricing problems in many column generation algorithms where vehicle routes are involved (see, for instance, Desrochers et al. 1992; Feillet et al. 2007; Righini and Salani 2008).

In Contardo et al. (2014a), which has been the most successful work so far, the authors allow for solutions that contain cycles, as long as they contain at least three nodes. For this case, to guarantee that even if Γ contains non-elementary routes, these routes will not be part of a solution of LRP3, the authors replace the degree constraints (15.23) by their following stronger variant, the strengthened degree constraints:

$$\sum_{r\in\Gamma}\sum_{k:\{j,k\}\in E}a_{jkr}\lambda_r+\sum_{i\in I}z_{ij}^2\ge 1 \quad j\in J. \tag{15.28}$$

On top of the efficiency of the algorithm used in the pricing problem, most set partitioning based exact algorithms for the LRP also rely on the addition of valid inequalities to tighten the bounds obtained during the branching process. In particular, Baldacci et al. (2011) proved that all valid inequalities developed for flow formulations can be transformed into valid inequalities for the set partitioning formulation presented above, since, thanks to the distinction between routes visiting one or more customers made in the variables definition, the following equalities hold:

$$z_{ij}^1=\sum_{r\in\Gamma}a_{ijr}\lambda_r \qquad \forall\{i,j\}\in E. \tag{15.29}$$

Additionally to this equivalence, when adapting valid inequalities originally stated for flow formulations to set-partitioning formulations, some authors have used the following result, first established in Laporte et al. (1985) in the context of vehicle routing problems. Many of the valid inequalities derived for two-index formulations for vehicle routing problems are concerned with a combination of connectivity and capacity issues. In these cases, arguments of the type "at least κ vehicles are needed to satisfy the demand of all customers in $S \subset J$" result in constraints of the form "the border of S is crossed, at least, 2κ times", that is, the sum of flows on edges with a single endpoint in S must be at least 2κ. In these constraints, the number of routes visiting S is overestimated using the flow in the cut-set of S, since there is no way to compute the exact number of routes that visit S using the flow variables. When equivalence (15.29) is used to derive valid inequalities for LRP3 from these valid inequalities, the coefficient of each λ_r variable for a given set S is the number of times route r traverses the border of S. Bearing in mind the rationale behind the constraints, one can see that, actually, these coefficients can be changed to take value 2 if route r visits at least one customer in S, and 0 otherwise. In general, this results in stronger valid inequalities.

15.3.3 Valid Inequalities

It is impractical to list all the valid inequalities that have been more or less successfully used for LRPs. Actually, most of the valid inequalities that have been developed for vehicle routing problems have been adapted later for the case of LRPs and in many cases, families of inequalities have been gradually strengthened or extended. In what follows, we present a selection of the most recent families. For more detailed information on these cuts and their evolution, the reader is referred to Belenguer et al. (2011) and Contardo et al. (2013) for flow formulations, and to Baldacci et al. (2011) and Contardo et al. (2014a) for set partitioning formulations.

15.3.3.1 y-Strengthened Capacity Cuts (y-SCC)

For $S \subset J$, and a route $r \in \Gamma$, let the binary parameter ξ_{rS} take value 1 if route r visits at least one customer in S, and 0 otherwise. Given $S' \subset S$ such that $\kappa_1(S') = \kappa_1(S)$, the following inequalities are valid:

$$\sum_{r \in \Gamma} \xi_{rS} \lambda_r + \sum_{i \in I} \sum_{j \in S \setminus S'} z_{ij}^2 \geq \kappa_1(S).$$

This family of constraints is a strengthening proposed in Contardo et al. (2014a) of the previous y-Capacity Cuts derived in Belenguer et al. (2011).

15.3.3.2 Set Partitioning Effective Strengthened Facility Capacity Inequalities (SP-ESFCI)

As mentioned above, the main difficulty when modeling vehicle routes is to ensure the connectivity of the solutions, especially in capacitated problems. When locational decisions must also be made, ensuring connectivity and capacity satisfaction entails an extra degree of complexity. Most of the known valid inequalities focus on vehicle capacities and rarely take facility capacities into account. SP-ESFCI aim at putting facility capacity constraints in relation with the locational variables.

To this end, we need to extend the definition of κ_1 to take into account a set of facilities. Given a set of customers $S \subset J$ and a set of facilities $H \subset I$, we define $\kappa_1(S,H) = \max\left\{0, \left\lceil \frac{w(S)-\sum_{i\in H} q_i}{Q} \right\rceil\right\}$ as a lower bound on the number of vehicle routes rooted at facilities outside H, needed to serve all customers in S, even if all facilities in H provided their service to customers in S. Then, for $S' \subset S \subset J$, and $i \in H \subset I$ with $\kappa_1(S \setminus S', H) = \kappa_1(S,H)$, the following inequality is valid:

$$\sum_{i\in I\setminus H}\sum_{r\in\Gamma_i}\xi_{rS}\lambda_r + \sum_{i\in I\setminus H}\sum_{j\in S\setminus S'} z^2_{ij} \geq \kappa_1(S, H\setminus\{i\}) + y_i\Big(\kappa_1(S,H) - \kappa_1(S,H\setminus\{i\})\Big). \tag{15.30}$$

The main idea behind these constraints is similar to that of the y-SCC inequalities, but now, the constraint takes two different shapes depending on whether facility i is opened or not.

15.3.3.3 Strengthened Framed Capacity Inequalities (SFrCI)

Moving back to vehicle capacities, we find the following valid inequalities, which have been successively improved since some early papers on vehicle routing.

Given a subset of customers $S \subset J$, partitioned into disjoint subsets $\mathscr{S} = \{S_1, \ldots, S_t\}$ ($S = \cup_{s=1}^t S_s$), we denote by $\kappa_3(S|\mathscr{S})$ the optimal value of the bin packing problem defined as follows. For each set S_s in $\mathscr{S}$, we define $\kappa_1(S_s)$ items of size Q, except for the last one, which will have a size equal to $w(S) - (\kappa_1(S)-1)Q$, and we define bin capacities equal to Q. Then, the SFrCI corresponding to frame $(S, \mathscr{S})$ is:

$$\sum_{r\in\Gamma}\xi_{rS}\lambda_r + \sum_{s=1}^{t}\sum_{r\in\Gamma}\xi_{rS_s}\lambda_r \geq \kappa_3(S|\mathscr{S}) + \sum_{s=1}^{t}\kappa_1(S_s). \tag{15.31}$$

These inequalities generalize and reinforce the capacity inequalities, which force that the number of routes that visit a given set of customers S is at least $\kappa_1(S)$. Note that when no location decisions have to be made, in the presence of degree constraints, capacity constraints are equivalent to subtour elimination constraints (15.15). Indeed, when for a given set $S \subset J$, $\mathscr{S}$ only contains one set,

the corresponding SFrCI constraint is a capacity constraint (in this case, $\kappa_3(S|\mathscr{S}) = \kappa_1(S)$). So, the two terms in the left-hand side of (15.31) are identical, the two terms in the right-hand side are also equal, and the inequality becomes:

$$\sum_{r\in\Gamma} \xi_{rS}\lambda_r \geq \kappa_1(S),$$

which is the basic expression of the capacity constraint.

As is the case for other sets of inequalities, the framed capacity inequalities (FrCI) where originally developed for two-index flow formulations and later adapted to the set-partitioning formulation by using Eq. (15.29), and reinforced by modifying the coefficients of the λ_r variables as explained in the last section. The FrCI for formulation LRP2 corresponding to $(S, \mathscr{S})$ is

$$\sum_{j\in S}\sum_{k\in V\setminus S} z^1_{jk} + 2\sum_{i\in I}\sum_{j\in S} z^2_{ij} + \sum_{s=1}^{t}\sum_{j\in S_s}\left(\sum_{k\in V\setminus S_s} z^1_{jk} + 2\sum_{i\in I} z^2_{ij}\right)$$
$$\geq 2\left(\kappa_3(S|\mathscr{S}) + \sum_{s=1}^{t}\kappa_1(S_s)\right). \tag{15.32}$$

To illustrate that FrCI (and, therefore, SFrCI) is a broader set of inequalities that can be stronger than the combination of capacity constraints for the individual sets S_s, Fig. 15.3 gives an example of a fractional solution with $\mathscr{S} = \{S_1, \cdots, S_4\}$, where the capacity constraints for each of the S_s sets are satisfied, but the overall FrCI constraint is violated. In this figure, customers are numbered from 1 to 7 and w_i is given inside each customer. Note that, in this example, we have $S = \cup_{s=1}^{4} S_s$, $w(S) = 20$ and $Q = 7$, so that $\kappa_1(S) = 3$. Thus, the capacity constraint for set S is satisfied, since the total flow in edges with one endpoint in S equals its lower bound, $2 \cdot 3 = 6$. Also, for each set in the partition, $w(S_s) < Q$, so that $\kappa_1(S_s) = 1$ and the

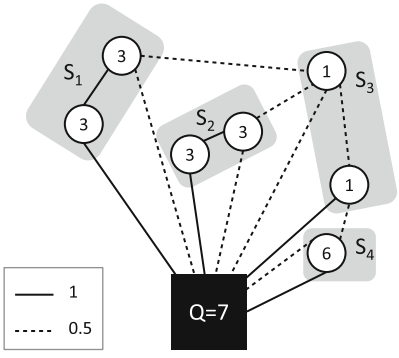


Fig. 15.3 Example of unsatisfied FrCI

z-degree of S_s is 2 or larger in all cases. In contrast, the evaluation of constraint (15.32) gives

$$6 + (2 + 2 + 3 + 2) \geq 2(4 + 1 + 1 + 1 + 1),$$

which is clearly not satisfied. Here, note that in the computation of $\kappa_3(S|\mathscr{S})$, four items were defined, with sizes 6, 6, 2 and 6, respectively, and the bin capacity was set to 7.

The example of Fig. 15.3 also provides some insight in the way how the variable definition in set partitioning formulations such as LRP3 forbids some fractional solutions that are sometimes encountered when using flow formulations. Indeed, the solution of the figure can be obtained in a relaxation of formulation LRP2, but it is impossible to obtain it from formulation LRP3, since it cannot be decomposed as the (fractional) combination of vehicle routes which are feasible with respect to the vehicle capacity constraint.

15.4 Heuristic Algorithms

Many heuristics have been devised for different variants of LRPs. It is not the goal of this chapter to enumerate and explore all these contributions. Instead, we concentrate on the tools that have been most useful in those heuristics.

In the design of heuristics for LRPs it is very difficult to ignore the fact that the problem combines decisions of two completely different natures: the location of the facilities and the design of vehicle routes. Indeed, even solution methods based on the use of neighborhoods tend to distinguish between the neighborhoods that affect the set of facilities (add, drop or swap) and those that are typically used in vehicle routing problems. A clear example of this fact is the variable neighborhood search (VNS) heuristic recently proposed in Jarboui et al. (2013) for an LRP with capacitated facilities and uncapacitated vehicles or the granular tabu search heuristic presented in Willmer Escobar et al. (2013) for an LRP where both vehicles and depots are capacitated. Possible exceptions are some algorithms based on the construction of giant tours that encode both types of decisions, so that tour modifications can alter both, facility locations and vehicle routes. Examples of this type of algorithm are those of Yu et al. (2010) or Contardo et al. (2014b).

A commonly accepted classification for heuristic methods for LRPs, due to Nagy and Salhi (2007), includes four categories, depending on how the interaction between these decisions is taken into account in the design of heuristics.

- *Sequential methods* split the problem into its subproblems. First they solve the location problem, using estimates of the routing costs that only take into account the distances between customers and facilities and, they then solve the routing problems defined at each opened facility with its assigned customers. Although Srivastava and Benton (1990) show that this type of methods, that are typically

quite fast, can produce pretty good solutions for some types of instances, in general, they tend to have a rather poor behavior, and most authors moved fast to other types of heuristics.

- *Clustering-based methods* partition the set of customers into clusters and then they either locate a depot for each cluster and solve a vehicle routing problem afterwards, or solve an auxiliary traveling salesman problem for each cluster before locating the depots. Barreto et al. (2007) present a method of this type and also analyze different clustering criteria in this context. A more recent example of this type of method is the constructive procedure considered in the two-phase method of Willmer Escobar et al. (2013) for the capacitated LRP. With their algorithm, the authors have provided the currently best known solutions for many of the existing benchmark instances (with up to 200 customers and 20 facilities) but at a high computational cost since some instances required more than 10 min of CPU time.

- *Iterative methods* can be seen as an evolution of sequential methods, where several iterations of a sequential method are performed, and the information obtained at each iteration is used to guide the methods used for choosing the locations and designing the vehicle routes built at the next one. The algorithm proposed in Prins et al. (2007) falls in this category. Using their algorithm, the authors could find very good solutions (proven to be optimal in several cases) for instances with up to 200 customers and 20 facilities, and the CPU time exceeded 1 min in only a reduced subset of the considered instances.

- In *hierarchical methods* the problem is considered in a more integrated way, without splitting its components. However, the two decisions are not considered to be equally important; facilities location is regarded as the main problem decision and vehicle routes design as a secondary one. Many contributions fit in this category (Albareda-Sambola et al. 2005; Ting and Chen 2013), especially the most recent ones, since they tend to yield better results. Indeed, the results obtained in Ting and Chen (2013) are comparable with those of Willmer Escobar et al. (2013), although solutions are slightly worse in general terms, this is compensated by somehow smaller CPU times.

Finally, one can also find in the literature one approximation algorithm for the LRP in Harks et al. (2013). The proposed algorithm builds a solution by combining the solutions to two auxiliary problems: and uncapacitated facility location problem, and a minimum spanning tree. For this algorithm, they prove an approximation factor of 4.38.

15.5 Location Arc Routing

LARPs are typically defined on graphs $G = (V, E)$ that can be either directed, undirected or, in the most general case, mixed. In G, a set $I \subset V$ of selected nodes where facilities may be established is given, together with a selected subset of links $R \subseteq E$, known as required arcs or edges, which must be traversed to receive some service. Common applications of LARPs include garbage collection, road maintenance and postal delivery. For details on these applications, the reader is referred to Ghiani and Laporte (2001).

In contrast to the volume of the literature on LRPs with node routing, LARPs have been addressed only in a few references. This is due in part, to the difficulty of these problems, but also to the fact that several strategies have been devised to transform arc routing problems into node routing problems by suitably modifying the underlying graph (see, for instance Pearn et al. 1987; Baldacci and Maniezzo 2006; Longo et al. 2006). However, significant differences exist between the structures of the routes depending on whether service is provided at the nodes or on the links. These differences suggest that, as happens with pure routing problems, specific approaches for either type of problem may yield more efficient algorithms.

The most relevant difference between routes in node and arc routing is that in node routing problems one can assume, without loss of generality, that no node will be visited more than once, and the only links that may be traversed twice are those connecting one facility with one customer, allowing thus for routes visiting one single customer. In contrast, in arc routing problems, even required links may be traversed more than once in optimal solutions. Also, the set of required arcs induces a family of connected components of G which, as happens in pure arc routing problems, play an important role in determining which links are susceptible of being used more than once.

The first paper addressing a LARP is probably that of Levy and Bodin (1989) in which a problem with uncapacitated vehicles arising in the USA postal services was solved. To this end, the authors split the problem into its components and solve them sequentially, following the scheme (1) location of facilities, (2) allocation of required edges to facilities, and (3) route design.

Uncapacitated LARPs were also studied in Ghiani and Laporte (1998). One of the first consequences of having uncapacitated vehicles is that, when the triangle inequality holds, only one route needs to be built for each open facility. Moreover, the authors show that, in this case, optimal solutions exist where all the required edges belonging to the same connected component are served in the same route, which allows to transform this particular LARP into different arc routing problems, depending on whether the number of depots to locate is bounded or not. Applying a branch-and-cut algorithm to these problems, the authors solve uncapacitated LARP instances on graphs with up to 200 nodes. Since then, no exact algorithm for any LARP variant has been proposed, and only heuristic algorithms for different variants can be found in the literature. Actually, two mixed integer programming formulations for capacitated LARPs were proposed by Doulabi and Seifi (2013):

one for the general case, and a second one for the particular case where one single depot has to be located. Another formulation is also presented in Borges Lopes et al. (2014). However, these papers do not explore the possibility of solving these formulations exactly, possibly because they all use flow variables, with up to four indices in some cases, and therefore, they are rather large.

Bearing in mind the evolution of the formulations for the capacitated arc routing problem (CARP), one might expect set partitioning formulations to allow for more efficient solution methods. Indeed, the most successful algorithms for the CARP so far, proposed by Bode and Irnich (2012) and Bartolini et al. (2013), both rely on set partitioning formulations for this problem. In any case, further research is still needed before reaching efficient exact methods for solving general LARPs. Although it is true that research on the CARP has been very fruitful in the past years, the subproblem obtained from a LARP when the set of facilities to open is fixed is a CARP with multiple depots, which has hardly been studied, and for which only heuristic algorithms exist (see, for instance, Amberg et al. 2000).

In the case of heuristic methods, the original approaches relying on the sequential solution of the different subproblems of a LARP have evolved with a recent focus on the use of metaheuristics. Doulabi and Seifi (2013) propose a simulated annealing heuristic which, at each iteration, proceeds following an allocation-routing-location scheme: it first builds a routing solution then tries to improve the depot locations. More recently, Borges Lopes et al. (2014) have proposed and compared several heuristics combining tabu search, variable neighborhood search, and GRASP for which they also test different constructive heuristics. According to their computational experiments, the combination of tabu search and GRASP provides the best results. With this combination, they find optimal or near optimal solutions in less than a minute, for instances with up to 140 nodes and 190 required edges. They also propose a set of benchmark instances for future comparisons.

In contrast to the scarce literature available on the LARP, a relatively large variety of related problems have been studied. This is the case, for instance, of the capacitated arc routing problem with intermediate facilities presented in Ghiani et al. (2001). In this case, no location decisions need to be made, and a single depot is considered, like in the CARP, but several facilities are available in the network where a vehicle can unload the demand collected at the required edges before the loaded demand exceeds the vehicle capacity.

Other examples are the capacitated arc routing problem with refill points or the synchronized arc and node routing problem, presented in Amaya et al. (2007) and Salazar-Aguilar et al. (2013), respectively. In these cases, an additional fleet of vehicles is available to refill the main fleet, and the locations where these vehicles meet each other need to be determined when designing their respective routes. These problems differ in the types of routes performed by the vehicles used to replenish the service vehicles.

A recent paper on the directed profitable location rural postman problem (Arbib et al. 2014) also deserves a mention. This is an uncapacitated LARP where required arcs have associated profits and the decision maker can choose whether or not to serve any of them, taking into account the differences between the profit generated

and the cost of reaching the arcs. Using a branch-and-cut algorithm, the authors can solve to optimality instances involving up to 140 nodes and 190 required arcs.

15.6 Conclusions

This chapter has summarised some the most relevant research contributions on LRPs and LARPs. As it has been shown, the different research directions followed in the study of formulations and exact algorithms for LRPs have finally converged to one single proposal, which has been able to incorporate most of the relevant contributions in the field so far. In the case of heuristic algorithms, the research activity has recently been reactivated, giving raise to several competitive algorithms in the last years. The most successful approaches involve one or several metaheuristics, and the current activity in this area gives the impression that relevant further improvements can be expected in the near future.

In contrast, research on LARPs is still in its early stages. Exact algorithms have only been proposed for very particular cases, and even in the case of heuristics the literature is rather scarce. Keeping in mind the evolution followed by the research on LRPs, especially in what concerns exact algorithms, further research is still required on arc routing problems with multiple depots before it is possible to devise efficient algorithms for solving LARPs.

References

Ahn J, de Weck O, Geng Y, Klabjan D (2012) Column generation based heuristics for a generalized location routing problem with profits arising in space exploration. Eur J Oper Res 223:47–59

Akca Z, Berger RT, Ralphs TK (2009) A branch-and-price algorithm for combined location and routing problems under capacity restrictions. In: Proceedings of the eleventh INFORMS computing society meeting, Charleston, pp 309–330

Albareda-Sambola M, Díaz JA, Fernández E (2005) A compact model and tight bounds for a combined location-routing problem. Comput Oper Res 32:407–428

Amaya A, Langevin A, Trépanier M (2007) The capacitated arc routing problem with refill points. Oper Res Lett 35:45–53

Amberg A, Domschke W, Voß S (2000) Multiple center capacitated arc routing problems: a tabu search algorithm using capacitated trees. Eur J Oper Res 2000:360–376

Arbib C, Servilio M, Archetti C, Speranza MG (2014) The directed profitable location rural postman problem. Eur J Oper Res 236:811–819

Baldacci R, Maniezzo B (2006) Exact methods based on node-routing formulations for undirected arc-routing problems. Networks 47:52–60

Baldacci R, Mingozzi A, Wolfler Calvo R (2011) An exact method for the capacitated location-routing problem. Oper Res 59:1284–1296

Barreto S, Ferreira C, Paixão J, Souza Santos B (2007) Using clustering analysis in a capacitated location-routing problem. Eur J Oper Res 179:968–977

Bartolini E, Cordeau J-F, Laporte G (2013) Improved lower bounds and exact algorithm for the capacitated arc routing problem. Math Program 137:409–452

Belenguer JM, Benavent E, Prins C, Prodhon C, Wolfler Calvo R (2011) A branch-and-cut method for the capacitated location-routing problem. Comput Oper Res 38:931–941
Berger RT, Coullard CR, Daskin MS (2007) Location-routing problems with distance constraints. Transp Sci 41:29–43
Bode C, Irnich S (2012) Cut-first branch-and-price-second for the capacitated arc-routing problem. Oper Res 60:1167–1182
Borges Lopes R, Ferreira C, Sousa Santos B, Barreto S (2013) A taxonomical analysis, current methods, and objectives on location-routing problems. Int T Oper Res 20:795–822
Borges Lopes R, Plastria F, Ferreira C, Sousa Santos B (2014) Location-arc routing problem: heuristic approaches and test instances. Comput Oper Res 43:309–317
Contardo C, Cordeau J-F, Gendron B (2013) A computational comparison of flow formulations for the capacitated location-routing problem. Discret Optim 10:263–296
Contardo C, Cordeau J-F, Gendron B (2014a) An exact algorithm based on cut-and-column generation for the capacitated location-routing problem. INFORMS J Comput 26:88–102
Contardo C, Cordeau J-F, Gendron B (2014b) A GRASP+ILP-based metaheuristic for the capacitated location-routing problem. J Heuristics 20:1–38
Desrochers M, Desrosiers J, Solomon MM (1992) A new optimization algorithm for the vehicle routing problem with time windows. Oper Res 40:342–354
Doulabi SHH, Seifi A (2013) Lower and upper bounds for location-arc routing problems with vehicle capacity constraints. Eur J Oper Res 224:189–208
Feillet D, Gendreau M, Rousseau L-M (2007) New refinements for the solution of vehicle routing problems with branch and price. INFOR 45:239–256
Ghiani G, Laporte G (1998) Eulerian location problems. Networks 34:291–302
Ghiani G, Laporte G (2001) Location-arc routing problems. OPSEARCH 38:151–159
Ghiani G, Improta G, Laporte G (2001) The capacitated arc routing problem with intermediate facilities. Networks 37:134–143
Golden BL, Magnanti TL, Nguyen HQ (1977) Implementing vehicle routing algorithms. Networks 7:113–148
Harks T, König FG, Matschuke J (2013) Approximation algorithms for capacitated location routing. Transp Sci 47:3–22
Jarboui B, Houda D, Hanafi S, Mladenović N (2013) Variable neighborhood search for location routing. Comput Oper Res 40:47–57
Laporte G (1988) Location-routing problems. In: Golden BL, Assad AA (eds) Vehicle routing: methods and studies. North-Holland, Amsterdam, pp 163–197
Laporte G, Nobert Y (1981) An exact algorithm for minimizing routing and operating costs in depot location. Eur J Oper Res 6:224–226
Laporte G, Nobert Y, Pelletier P (1983) Hamiltonian location problems. Eur J Oper Res 12:82–89
Laporte G, Nobert Y, Desrochers M (1985) Optimal routing under capacity and distance restrictions. Oper Res 33:1050–1073
Laporte G, Nobert Y, Arpin D (1988) An exact algorithm for solving a capacitated location-routing problem. Ann Oper Res 6:293–310
Levy L, Bodin LD (1989) The arc oriented location routing problem. INFOR 27:74–94
Longo H, Aragão MP, Uchoa E (2006) Solving capacitated arc routing problems using a transformation to the CVRP. Comput Oper Res 33:1823–1837
Manzour-al-Ajdad SMH, Torabi SA, Salhi S (2012) A hierarchical algorithm for the planar single-facility location routing problem. Comput Oper Res 39:461–470
Maranzana F (1964) On the location of supply points to minimize transport costs. Oper Res Quart 15:261–270
Nagy G, Salhi S (2007) Location-routing: issues, models and methods. Eur J Oper Res 177: 649–672
Pearn WL, Assad AA, Golden BL (1987) Transforming arc routing into node routing problems. Comput Oper Res 14:285–288
Perl J, Daskin MS (1985) A warehouse location-routing problem. Transp Res B-Meth 19:381–396

Prins C, Prodhon C, Ruiz A, Soriano P, Wolfler Calvo R (2007) solving the capacitated location-routing problem by a cooperative lagrangean relaxation-granular tabu ssearch heuristic. Transp Sci 41:470–483

Prodhon C, Prins C (2014) A survey of recent research on location-routing problems. Eur J Oper Res 238:1–17

Righini G, Salani M (2008) New dynamic programming algorithms for the resource constrained elementary shortest path problem. Networks 51:155–170

Salazar-Aguilar MA, Langevin A, Laporte G (2013) The synchronized arc and node routing problem:application to road marking. Comput Oper Res 40:1708–1715

Salhi S, Nagy G (2009) Local improvement in planar facility location using vehicle routing. Ann Oper Res 167:287–296

Salhi S, Rand GK (1989) Effect of ignoring routes when locating depots. Eur J Oper Res 39: 150–156

Samanlioglu F (2013) A multi-objective mathematical model for the industrial hazardous waste location-routing problem. Eur J Oper Res 226:332–340

Schittekat P, Sörensen K (2009) Supporting "3PL" decisions in the automotive industry by generating diverse solutions to a large-scale location-routing problem. Oper Res 57:1058–1067

Srivastava R, Benton WC (1990) The location-routing problem: considerations in physical distribution system design. Comput Oper Res 17:427–435

Ting CJ, Chen CH (2013) A multiple ant colony optimization algorithm for the capacitated location routing problem. Int J Prod Econ 141:34–44

Von Boventer E (1961) The relationship between transportation costs and location rent in transportation problems. J Reg Sci 3:27–40

Willmer EJ, Linfati R, Toth P (2013) A two-phase hybrid heuristic algorithm for the capacitated location-routing problem. Comput Oper Res 40:70–79

Yu VF, Lin SW, Lee W, Ting CJ (2010) A simulated annealing heuristic for the capacitated location routing problem. Comput Ind Eng 58:288–299

Chapter 16
Location and Logistics

Sibel A. Alumur, Bahar Y. Kara, and M. Teresa Melo

Abstract Facility location decisions play a critical role in designing logistics networks. This chapter provides some guidelines on how location decisions and logistics functions can be integrated into a single mathematical model to optimize the configuration of a logistics network. This will be illustrated by two generic models, one supporting the design of a forward logistics network and the other addressing the specific requirements of a reverse logistics network. Several special cases and extensions of the two models are discussed and their relation with the scientific literature is described. In addition, some interesting applications are outlined that demonstrate the interaction of location and logistics decisions. Finally, new research directions and emerging trends in logistics network design are provided.

Keywords Forward logistics network design • Reverse logistics network design • Models • Applications

16.1 Introduction

Logistics network design (LND) and facility location decisions are closely interrelated. The latter are prompted by the need either to build a new logistics network or to re-design a network that is already in place. When a company enters new markets or grows into new product segments, a new logistics network has to be designed. However, "green field" projects are less frequent compared with re-design initiatives. Changing market and business conditions compel a company to modify

S.A. Alumur (✉)
Department of Management Sciences, University of Waterloo, Waterloo, ON, Canada
e-mail: sibel.alumur@uwaterloo.ca

B.Y. Kara
Department of Industrial Engineering, Bilkent University, Ankara, Turkey
e-mail: bkara@bilkent.edu.tr

M.T. Melo
Business School, Saarland University of Applied Sciences, Saarbrücken, Germany
e-mail: teresa.melo@htwsaar.de

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_16

the physical structure of its logistics network from time to time. Major drivers of network re-design projects comprise variations in the demand pattern and its spatial distribution as well as increased cost pressure and service requirements. Moreover, mergers, acquisitions, and strategic alliances also trigger the expansion or reconfiguration of a logistics network in order to exploit the benefits and synergies of integrating the acquired operations. Typically, re-design activities take the form of opening new facilities (e.g., to be closer to new markets) and closing existing facilities (e.g., to consolidate operations). As highlighted by Ballou (2001) and Harrison (2004), well-conceived re-design decisions can result in a 5–15 % reduction of the overall logistics costs, with 10 % being often achieved.

The (re-)design of a logistics network is a complex undertaking. It concerns not only determining the number, size, and capacity of facilities (e.g., plants and warehouses) to be operated but it also involves planning and integrating a manifold of logistics functions that such facilities will perform. These functions range from procurement of raw materials, transformation of these materials into semi-finished and end products, and the delivery of finished products to customers through one or several distribution stages. Depending on the industrial context, strategic decisions may also concern the collection and recovery of product returns.

This chapter provides a holistic approach to strategic network planning by integrating facility location decisions with decisions relevant to the configuration of a logistics network. The integrated view will be illustrated by two general modeling frameworks for designing forward and reverse logistics networks.

The remainder of the chapter is organized as follows. Section 16.2 presents a comprehensive model for logistics networks with forward flows. Due to its generic features, the model applies to a wide range of situations. Its relation with other models proposed in the literature is established and extensions are discussed. Section 16.3 focuses on reverse logistics network design (RLND) and introduces a generic mathematical formulation for the design of a multi-purpose reverse logistics network. Furthermore, some special cases and extensions of the proposed model are presented. Section 16.4 addresses various representative applications of forward and reverse LND problems from different areas. Finally, in Sect. 16.5 future research directions are discussed.

16.2 A General Logistics Network Design Model

We introduce a base model that captures the main features of an LND problem. The starting point is either a potential framework for a new network structure or an existing network whose physical structure is to be re-designed. To this end, a general network typology, as depicted in Fig. 16.1, is considered. Any number of facility layers and any system of transportation channels can be modeled. The network entities are categorized in so-called *selectable* and *non-selectable* facilities. The former group includes a set of facilities already in place, that could be closed, and a set of potential locations for establishing new facilities. In contrast, non-selectable

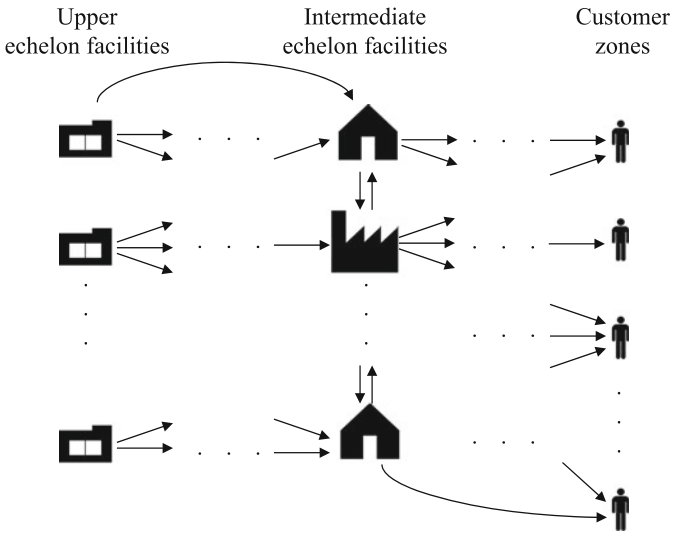


Fig. 16.1 General structure of a logistics network

facilities comprise facilities that are not subject to location decisions. Typically, such facilities include suppliers as well as existing plants and/or warehouses that should be maintained. In addition, customer zones are viewed as special members of this set as they have demand requirements for multiple commodities. As shown in Fig. 16.1, no restrictions are imposed on the availability of transportation channels for the flow of materials through the network. In particular, direct commodity flows from upstream sources to customer zones (or to facilities not immediately below in the hierarchy) are possible as well as flows between facilities in the same echelon. In this rather general network typology, procurement, production, distribution, and customer service decisions are to be made along with facility location and sizing decisions. The mathematical model in Sect. 16.2.2 captures the aforementioned features. The required notation is first introduced in Sect. 16.2.1. Several special cases and extensions are discussed in Sect. 16.2.3.

16.2.1 Notation and Definition of Decision Variables

Table 16.1 introduces the index sets that are used in the base model. In addition to the various types of network entities, also multiple commodities are considered, ranging from raw materials and intermediate products to finished goods. Moreover, different kinds of resources may be available for manufacturing and handling commodities.

Table 16.2 describes input parameters related to logistics operations. Multi-stage production processes can be taken into account through bills-of-materials (BOMs). In this case, the relationships between components and parent items are defined

Table 16.1 Index sets

Symbol	Description	Index symbol
N	Set of potential locations for new facilities	i
E	Set of existing facilities that could be closed	i
I	Set of selectable facilities, $I = N \cup E$	i
J	Set of non-selectable locations (e.g., customer zones)	$j,\ j'$
L	Set of all entities, $L = I \cup J$	$\ell,\ \ell'$
P	Set of products	$p,\ q$
M, H	Set of manufacturing, resp. handling, resources	$m,\ h$

Table 16.2 Logistics parameters

Symbol	Description
$d_{\ell p}$	Demand of location $\ell \in L$ for product $p \in P$ (typically, $d_{ip} = 0$ for $i \in I$)
$\alpha_{\ell qp}$	Number of units of product $q \in P$ required to manufacture one unit of product $p \in P$ ($q \neq p$) at facility $\ell \in L$
$\mu_{\ell mp}$	Number of units of resource $m \in M$ required to manufacture one unit of product $p \in P$ at facility $\ell \in L$
$\lambda_{\ell hp},\ \hat{\lambda}_{\ell hp}$	Number of units of resource $h \in H$ required to handle one unit of product $p \in P$ upon its arrival at, resp. shipment from, facility $\ell \in L$
$KM_m,\ KH_h$	Capacity of manufacturing resource $m \in M$, resp. handling resource $h \in H$
$EM_m,\ EH_h$	Maximum increase in capacity of manufacturing resource $m \in M$, resp. handling resource $h \in H$

by given parameters. Capacities of service facilities are modeled in a general way through manufacturing and handling resources. Three different relation types are considered. In a *many-to-one* relationship, several resources are available at the same facility. Some resources may be product-specific (e.g., a machine dedicated to a given item) while others may be shared by multiple commodities (e.g., production line or order picking system). A *one-to-one* association corresponds to the classical way of modeling capacity in facility location models (e.g., storage space in a warehouse). *One-to-many* relationships can also be modeled, although these are less common. This could be the case, for example, of a team of experts responsible for several production lines in different facilities. Resource availability can be increased at additional expense, e.g., through overtime work or leasing extra storage space. Resource consumption is described by specific parameters. In the case of handling resources, the same type of equipment (e.g., forklift truck) may be required with different intensity to unload incoming goods at a facility and load goods to be shipped from the same facility.

Table 16.3 summarizes all facility and logistics costs. Facility costs are related to establishing new facilities and closing existing facilities, and typically reflect economies of scale. In addition, facility operating costs represent, for example, business overhead costs such as staff and security costs. Logistics costs are incurred for purchasing items from external sources (e.g., procurement of raw materials),

Table 16.3 Cost parameters

Symbol	Description
FC_i	Fixed setup cost of establishing a new facility in location $i \in N$
SC_i	Fixed cost of closing existing facility $i \in E$
OC_ℓ	Fixed cost of operating facility $\ell \in L$
$BC_{\ell p}$	Unit cost of buying product $p \in P$ at facility $\ell \in L$ from an external source
$PC_{\ell p}$	Unit cost of producing product $p \in P$ at facility $\ell \in L$
$TC_{\ell\ell' p}$	Unit cost of transporting product $p \in P$ from facility $\ell \in L$ to facility $\ell' \in L$ $(\ell \neq \ell')$
MC_m, HC_h	Unit cost of expanding manufacturing resource $m \in M$, resp. handling resource $h \in H$
$DC_{\ell p}$	Unit penalty cost for not serving demand of facility $\ell \in L$ for product $p \in P$

Table 16.4 Decision variables

Symbol	Description
y_i	1 if the selectable facility $i \in I$ is operated, 0 otherwise
$s_{\ell p}$	Quantity of product $p \in P$ purchased at facility $\ell \in L$ from an external source
$z_{\ell p}$	Quantity of product $p \in P$ manufactured at facility $\ell \in L$
$x_{\ell\ell' p}$	Quantity of product $p \in P$ shipped from facility $\ell \in L$ to facility $\ell' \in L$ $(\ell \neq \ell')$
w_m, $\overline{w}_h$	Number of extra capacity units of manufacturing resource $m \in M$, resp. handling resource $h \in H$
$u_{\ell p}$	Quantity of unsatisfied demand of location $\ell \in L$ for product $p \in P$

for manufacturing commodities, and for distributing multiple products through the network. The latter costs may also include charges for handling goods at the source facility and at the destination facility (e.g., order picking and warehousing costs). Furthermore, additional costs are considered for resource expansion. Penalty costs are also incurred for failing to meet customer demand. These costs represent the additional expense for outsourcing unfilled demand.

Finally, strategic decisions on facility location and logistics operations are ruled by the variables in Table 16.4.

16.2.2 A Mixed-Integer Linear Programming Model

Under the assumption that all inputs are known non-negative quantities, the logistics network (re-)design problem can be formulated as a mixed-integer linear program (MILP) as follows.

The objective function (16.1) describes the aim of the decision-making process, namely to identify the network configuration with the least total cost. To this end, fixed costs associated with opening, closing, and operating facilities are considered. The latter include a fixed cost term for maintaining facilities that are not subject to

location decisions (i.e., $\sum_{j\in J} OC_j$). Variable costs account for resource expansion and for material procurement, production, and distribution. In addition, penalty costs are incurred to unfilled demand.

$$
\begin{aligned}
(P_1)\quad \text{Minimize}\quad & \sum_{i\in N} FC_i\, y_i \\
& + \sum_{i\in E} SC_i\,(1-y_i) + \sum_{i\in I} OC_i\, y_i + \sum_{j\in J} OC_j + \sum_{m\in M} MC_m\, w_m \\
& + \sum_{h\in H} HC_h\, \overline{w}_h + \sum_{\ell\in L}\sum_{p\in P} BC_{\ell p}\, s_{\ell p} + \sum_{\ell\in L}\sum_{p\in P} PC_{\ell p}\, z_{\ell p} \\
& + \sum_{\ell\in L}\sum_{\ell'\in L\setminus\{\ell\}}\sum_{p\in P} TC_{\ell\ell' p}\, x_{\ell\ell' p} + \sum_{\ell\in L}\sum_{p\in P} DC_{\ell p}\, u_{\ell p} \qquad (16.1)
\end{aligned}
$$

$$
\begin{aligned}
\text{subject to}\quad & s_{\ell p} + \sum_{\ell'\in L\setminus\{\ell\}} x_{\ell'\ell p} + z_{\ell p} = \\
& \sum_{q\in P} \alpha_{\ell pq}\, z_{\ell q} + \sum_{\ell'\in L\setminus\{\ell\}} x_{\ell\ell' p} + d_{\ell p} - u_{\ell p}, \quad \ell\in L,\ p\in P \qquad (16.2)
\end{aligned}
$$

$$
\sum_{\ell\in L}\sum_{p\in P} \mu_{\ell mp}\, z_{\ell p} \le KM_m + w_m, \quad m\in M \qquad (16.3)
$$

$$
\begin{aligned}
& \sum_{\ell\in L}\sum_{p\in P} \lambda_{\ell hp}\, s_{\ell p} + \sum_{\ell\in L}\sum_{\ell'\in L\setminus\{\ell\}}\sum_{p\in P} \left(\hat{\lambda}_{\ell hp} + \lambda_{\ell' hp}\right) x_{\ell\ell' p} \\
& \le KH_h + \overline{w}_h, \quad h\in H \qquad (16.4)
\end{aligned}
$$

$$
0 \le w_m \le EM_m, \quad m\in M \qquad (16.5)
$$

$$
0 \le \overline{w}_h \le EH_h, \quad h\in H \qquad (16.6)
$$

$$
0 \le u_{\ell p} \le d_{\ell p}, \quad \ell\in L,\ p\in P \qquad (16.7)
$$

$$
0 \le s_{ip} \le \mathscr{M}\, y_i, \quad i\in I,\ p\in P \qquad (16.8)
$$

$$
0 \le z_{ip} \le \mathscr{M}\, y_i, \quad i\in I,\ p\in P \qquad (16.9)
$$

$$
0 \le x_{i\ell p} \le \mathscr{M}\, y_i, \quad i\in I,\ \ell\in L\setminus\{i\},\ p\in P \qquad (16.10)
$$

$$
0 \le x_{\ell ip} \le \mathscr{M}\, y_i, \quad \ell\in L\setminus\{i\},\ i\in I,\ p\in P \qquad (16.11)
$$

$$
s_{jp} \ge 0,\ z_{jp} \ge 0,\ x_{jj' p} \ge 0, \quad j, j'\in J\ (j\ne j'),\ p\in P \qquad (16.12)
$$

$$
y_i \in \{0,1\}, \quad i\in I. \qquad (16.13)
$$

Constraints (16.2) are the usual flow balance equations. The inbound flow of an item to a facility consists of procuring or producing the item at the facility or

receiving it from other locations. The outbound flow results from using the product as a raw material to manufacture other commodities, distributing the item to other facilities, or serving demand in case the location is a customer zone. Inequalities (16.3), resp. (16.4), guarantee that the usage of manufacturing, resp. handling, resources does not exceed the available capacity. Constraints (16.5)–(16.6) stipulate that capacity expansions must be within given limits. Constraints (16.7) rule the maximum amount of unsatisfied demand. Inequalities (16.8)–(16.11) ensure that procurement, production, and distribution activities only occur at operating facilities. A sufficiently large constant $\mathscr{M}$ is used in these constraints which can be adjusted depending on each specific situation. Typically, $\mathscr{M}$ is replaced by the maximum quantity that can be processed by a facility with respect to all product types. Finally, constraints (16.12) are non-negativity conditions for the logistics operations in non-selectable locations, while constraints (16.13) are binary requirements for the location variables.

Although the above problem is NP-hard, being a generalization of the simple plant location problem (see Krarup and Pruzan 1983), Melo et al. (2008) could solve medium and large-sized randomly generated instances to optimality with general purpose optimization software within reasonable time. To analyze the quality of the MILP formulation, the linear relaxation bound was also compared with the optimal solution of the tested instances. In general, a relatively small gap could be observed. These findings have important practical implications, since managers often need to base their decisions on the results of several scenarios. Hence, for a company to be able to perform "what-if" analysis and thereby identify good quality (or even optimal) solutions with an acceptable level of computational effort is a major step towards better decision support.

16.2.3 Special Cases and Model Extensions

Historically, researchers have focused relatively early on the design of distribution systems with at most two facility layers (e.g., plants and warehouses). In these simple networks, decisions were mostly confined to facility location and distribution operations. The contribution by Geoffrion and Graves (1974) is such an example. In recent years, the trend has been towards the development of more comprehensive models that integrate location decisions with supplier selection, production planning, technology acquisition, inventory management, transportation mode selection, and vehicle routing, just to mention some important logistics functions considered in this area (see Melo et al. 2009 for a comprehensive review). In many cases, the proposed models combine strategic decisions (e.g., location and capacity choices) with tactical decisions (e.g., inventory and transportation management) or even operational decisions (e.g., vehicle routing). Usually, the interplay of different planning levels can only be captured at the cost of increased model complexity. This will be illustrated in Sect. 16.4 by three applications.

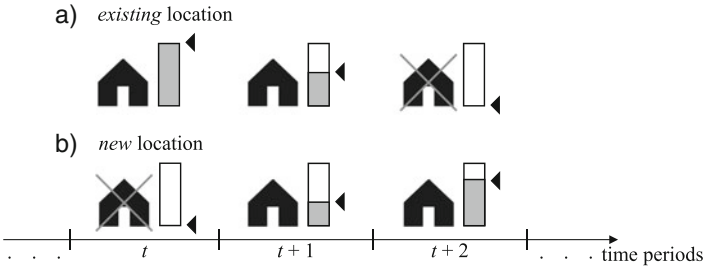


Fig. 16.2 Facility sizing over multiple periods (the *crossed symbol* indicates a closed facility)

The generic formulation (P_1) comprises some of the aforementioned features and it can also be adapted or extended to include further aspects relevant to LND. For example, it is easy to add single-sourcing requirements to (P_1) to ensure that the demand of each customer zone for a particular product is entirely satisfied from a unique facility. A straightforward extension of (P_1) is also to embed the (re-)design of a logistics network in a multi-period planning horizon. Such a setting is meaningful since the establishment of new facilities is typically a long-term project involving time-consuming activities and requiring the commitment of substantial capital resources. In this case, strategic decisions can be constrained by the budget available in each time period. Logistics decisions will be in turn impacted by the location choices. Fleischmann et al. (2006) and more recently Correia et al. (2013) included this feature in their dynamic network design models.

A multi-period setting is also appropriate for planning the re-design of a logistics network that is already in place. In this context, existing facilities may have their capacities expanded, reduced or even moved to new sites over several time periods as illustrated in Fig. 16.2 (the bars in the figure next to the facilities indicate their size). In turn, new facilities can be established through successive sizing. A gradual transfer of production and/or storage capacities from existing locations to new sites ensures a smooth implementation of relocation plans and avoids logistics operations from being disrupted. Melo et al. (2006, 2012, 2014) proposed several models and heuristics for this special form of network re-design.

In the mathematical model (P_1) all inputs (i.e., logistics and cost parameters) are taken as known quantities. As noted by Melo et al. (2009), most of the research dedicated to LND problems focuses on deterministic formulations. This is explained by the complexity posed by many of these problems and the serious computational hurdle that arises when the problem size becomes large. In the last two decades, increasing attention has been given to the development of new models that incorporate the uncertainties inherent to decision-making in LND (see Klibi et al. 2010). This is the case, for example, of the multi-echelon LND problem addressed by Santoso et al. (2005). Uncertainty is captured with respect to supply and demand quantities, resource capacities, and processing as well as transportation costs. Recently, Huang and Goetschalckx (2014) developed a scenario planning approach for a similar problem focusing on solution robustness. The goal is to obtain

a network configuration such that the solution values do not substantially vary over different scenarios. Several authors also included stochastic problem characteristics in a multi-period setting such as Aghezzaf (2005), Pan and Nagi (2010), and Nickel et al. (2012).

A further relevant aspect in strategic network design is the integration of location decisions with inventory management. Demand uncertainty and risk pooling play an important role in this context. Inventory decisions concern working inventories at storage locations (i.e., the amounts of products that have been ordered from suppliers but not yet requested by customers) and safety stocks. The latter are intended as a buffer against stockouts during ordering lead times. Shen (2005), Ozsen et al. (2009), and Shu (2010) study the trade-off between inventory, transportation, and fixed costs to locate warehouses and allocate customers. Combining inventory management and location decisions into a single model often results in mixed-integer non-linear programming formulations that can only be solved for small problem instances. Recently, Tancrez et al. (2012) developed a heuristic procedure that is able to solve large-scale multi-echelon location-inventory problems comprising plants, distribution centers, and customers.

Finally, the growth in globalization has led to the emergence of global supply chains, that is, worldwide networks of suppliers, manufacturers, distribution centers, and retailers. Consequently, the integration of financial considerations with location and logistics decisions has gained increasing importance in network design. Financial factors comprise, among others, taxes, duties, tariffs, exchange rates, and transfer prices. Meixell and Gargeya (2005) discuss various contributions in this area while Wilhelm et al. (2005) propose a comprehensive model for the design of a logistics network under the North American Free Trade Agreement (NAFTA).

16.3 A General Reverse Logistics Network Design Model

Reverse logistics refers to all operations involved in the return of products and materials from a point of use to a point of recovery or proper disposal. The purpose of recovery is to recapture value through options such as reusing, repairing, refurbishing, remanufacturing, and recycling. Reverse logistics includes the management of the return of end-of-use or end-of-life products as well as defective and damaged items, or packaging materials, containers, and pallets.

Major driving forces behind reverse logistics activities include economical factors, legislations, and environmental consciousness. As stated by De Brito and Dekker (2004), companies become active in reverse logistics because they can make a profit and/or because they are forced to focus on such functions, and/or because they feel socially motivated. These factors are usually intertwined. For example, a company can be compelled to reuse a certain percentage of components in order to achieve a recovery target set by the legislation. This will lead to a decrease in the cost of purchasing components and in waste generation. Jayaraman and Luo (2007) suggest that proper management of reverse logistics operations can lead to greater

profitability and customer satisfaction, and at the same time be beneficial to the environment.

Many actors are involved in the design and operation of a reverse logistics network. Even though extended producer responsibilities present in the legislations in various countries give the responsibility of recovering used products to original equipment manufacturers, governments need to establish the necessary infrastructure. Responsibilities can be shared among different parties, such as producers, distributors, third-party logistics providers, or municipalities, in designing and operating the reverse logistics networks.

In a reverse logistics network, end-of-life or end-of-use products can be generated at private households and at commercial, industrial, and institutional sources, which are referred to as generation points. Products are usually collected at special storage facilities called collection or inspection centers. Products are then sent for proper recovery through reusing, repairing, refurbishing, remanufacturing, or recycling. Inspected or recovered products and components can then be sold to suppliers, to (re)manufacturing facilities, or to customers in the secondary market. A generic reverse logistics network is depicted in Fig. 16.3.

Unlike forward logistics networks, where demand occurs at the lower echelon facilities, in reverse networks demand (for recovery) arises at the upper echelon facilities. However, a reverse logistics network is not a mirror image of a forward network. In addition to the typical forward supply chain actors, different actors and facilities are involved in reverse logistics networks, such as disposers, remanufacturers, and the secondary market. Moreover, unlike forward networks, which are mostly driven by economical factors, there are further factors motivating the establishment of reverse logistics networks such as environmental laws.

In Sect. 16.3.2, a generic mathematical formulation for the design of a multi-purpose reverse logistics network is presented. The required notation and the

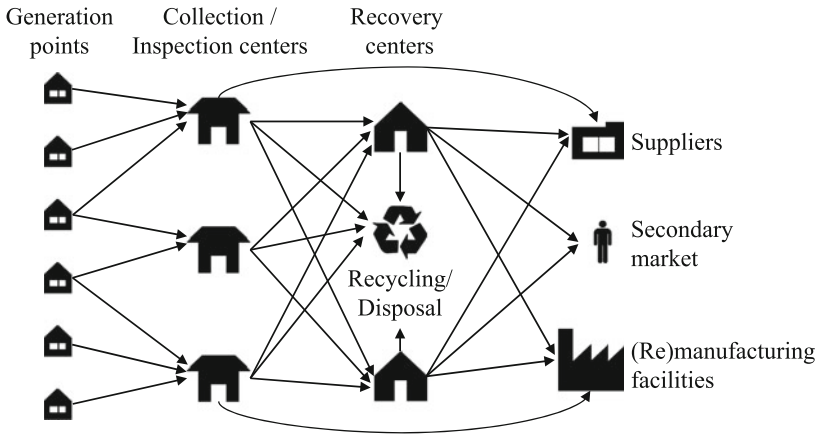


Fig. 16.3 A generic reverse logistics network

decision variables are first defined in Sect. 16.3.1. Some special cases and possible extensions of the proposed model are additionally discussed in Sect. 16.3.3.

16.3.1 Notation and Definition of Decision Variables

The notation used in the generic RLND model is analogous to the notation introduced in Sect. 16.2.1 for the forward LND model. Similar to the forward network design problem, multiple commodities are considered in the configuration of the reverse logistics network. These are represented by the set P, which may include used, inspected, repaired, or refurbished products, components, or raw materials. In order to represent a different state (inspected, repaired, refurbished, etc.) of a certain item, a different product type needs to be defined within the set P. Table 16.5 describes all index sets that are required for modeling the RLND problem.

The set of available recovery options may include conventional options, such as repair, refurbish, and recycle as well as other options such as inspection, disassembly, selling to suppliers, to the secondary market or to external (re)manufacturing facilities, and disposal. Even though the latter options may not be regarded as recovery alternatives, in order to provide a generic model incorporating all the decisions present in real-life reverse logistics networks, they are included in the set R. Observe that some recovery options may be operated by third-party logistics providers. These external facilities belong to the set J_r. Moreover, it is assumed that generation points are also included in this set of non-selectable facilities.

Table 16.6 introduces the required parameters. Transitions between the stages of products and reverse BOMs are taken into account by the parameter β. For example, a damaged product can be converted into a repaired product through the recovery option repair, or a used product can be disassembled into its components at a disassembly facility. Each recovery option has a given capacity which can be expanded at selectable facilities. Revenues may be obtained through some recovery options, e.g., by selling products or components to recycling facilities, to the

Table 16.5 New index sets

Symbol	Description	Index symbol
R	Set of recovery options (e.g., repair, refurbish, recycle)	r
N_r	Set of potential locations for recovery option $r \in R$	i
E_r	Set of existing facilities with recovery option $r \in R$	i
I_r	Set of selectable facilities with recovery option $r \in R$, $I_r = N_r \cup E_r$	i
J_r	Set of non-selectable locations with recovery option $r \in R$ (e.g. secondary market, disposal)	$j,\ j'$
L	Set of all locations, $L = \cup_{r \in R}\ (I_r \cup J_r)$	$\ell,\ \ell'$

Table 16.6 New parameters

Symbol	Description
$g_{\ell p}$	Amount of product $p \in P$ generated at location $\ell \in L$
β_{rqp}	Number of units of product $p \in P$ obtained by processing one unit of product $q \in P$ ($q \neq p$) using recovery option $r \in R$
$KR_{r\ell}$	Capacity of recovery option $r \in R$ at location $\ell \in L$
ER_{ri}	Maximum increase in capacity for recovery option $r \in R$ at location $i \in I_r$
RT_{rp}	Recovery target for product $p \in P$ with recovery option $r \in R$
$RE_{r\ell p}$	Revenue from recovering one unit of product $p \in P$ with recovery option $r \in R$ at location $\ell \in L$ (e.g., revenue from recycling or from the secondary market)
$RC_{r\ell p}$	Cost of recovering one unit of product $p \in P$ with recovery option $r \in R$ at location $\ell \in L$
FC_{ri}	Fixed setup cost of establishing recovery option $r \in R$ at location $i \in N_r$
SC_{ri}	Fixed cost of closing recovery option $r \in R$ at existing facility $i \in E_r$
$OC_{r\ell}$	Fixed cost of operating recovery option $r \in R$ at location $\ell \in L$
EC_{ri}	Unit cost of expanding capacity of recovery option $r \in R$ at location $i \in I_r$

Table 16.7 New decision variables

Symbol	Description
y_{ri}	1 if recovery option $r \in R$ is operated at the selectable facility $i \in I_r$, 0 otherwise
$v_{r\ell p}$	Amount of product $p \in P$ recovered with recovery option $r \in R$ at location $\ell \in L$
w_{ri}	Number of extra capacity units established for recovery option $r \in R$ at location $i \in I_r$

secondary market or to external (re)manufacturing facilities. Some recovery options may also incur costs as in the case of product disposal.

Finally, Table 16.7 describes the decision variables. The RNLD model also uses the flow variables $\mathbf{x}$ introduced in Table 16.4.

16.3.2 A Mixed-Integer Linear Programming Model

With the notation introduced in the previous section, the reverse logistics network (re-)design problem can be formulated as an MILP as follows. The objective function (16.14) maximizes the total profit. It sums the revenues obtained from various recovery options (e.g., by sending products to recycling facilities, by selling products to the secondary market) and subtracts the total cost of establishing and operating the network. The latter comprises the cost of recovering products at facilities, setting up new recovery options at facilities, closing existing recovery options, operating new and existing recovery options at facilities, transporting products, and expanding the capacities of recovery options. Observe that a fixed

cost term is also included in (16.14) to account for the operation of non-selectable facilities.

Equalities (16.15) are the flow balance constraints. For each location and product, the total inflow comprises the amount of product generated at that location, the total amount of product obtained after processing various items, and the total amount of product shipped to this location from other locations. The total inflow is equal to the total outflow which includes the total amount of product recovered at that location and the total amount of product shipped to other locations. Constraints (16.16) ensure that the recovery target for each product category and recovery option is met. Recovery targets are usually stipulated by legislations for different types of recovery options. Inequalities (16.17)–(16.19) are the capacity constraints. Constraints (16.17) guarantee that the total amount of recovered products at the selectable facilities does not exceed the total capacity. Similar conditions are set at non-selectable facilities by inequalities (16.18). Constraints (16.19) restrict the expansion of capacity at selectable facilities to be within given limits. Similar to the forward LND model, constraints (16.20)–(16.21) impose that products can only be shipped from operated facilities. Lastly, conditions (16.22)–(16.24) set the domains of the decision variables.

$$
\begin{aligned}
(P_2)\quad \text{Maximize}\quad & \sum_{r\in R}\sum_{\ell\in L}\sum_{p\in P} RE_{r\ell p}\, v_{r\ell p} \\
& - \sum_{r\in R}\sum_{\ell\in L}\sum_{p\in P} RC_{r\ell p}\, v_{r\ell p} - \sum_{r\in R}\sum_{i\in N_r} FC_{ri}\, y_{ri} \\
& - \sum_{r\in R}\sum_{i\in E_r} SC_{ri}\,(1-y_{ri}) - \sum_{r\in R}\sum_{i\in I_r} OC_{ri}\, y_{ri} - \sum_{r\in R}\sum_{j\in J_r} OC_{rj} \\
& - \sum_{\ell\in L}\sum_{\ell'\in L\setminus\{\ell\}}\sum_{p\in P} TC_{\ell\ell' p}\, x_{\ell\ell' p} - \sum_{r\in R}\sum_{i\in I_r} EC_{ri}\, w_{ri} \qquad (16.14)
\end{aligned}
$$

$$
\begin{aligned}
\text{subject to}\quad & g_{\ell p} + \sum_{r\in R}\sum_{q\in P} \beta_{rqp}\, v_{r\ell q} + \sum_{\ell'\in L\setminus\{\ell\}} x_{\ell'\ell p} = \\
& \qquad \sum_{r\in R} v_{r\ell p} + \sum_{\ell'\in L\setminus\{\ell\}} x_{\ell\ell' p}, \quad \ell\in L,\ p\in P \qquad (16.15)
\end{aligned}
$$

$$
\sum_{\ell\in L} v_{r\ell p} \geq RT_{rp}, \quad r\in R,\ p\in P \qquad (16.16)
$$

$$
\sum_{p\in P} v_{rip} \leq KR_{ri}\, y_{ri} + w_{ri}, \quad r\in R,\ i\in I_r \qquad (16.17)
$$

$$
\sum_{p\in P} v_{rjp} \leq KR_{rj}, \quad r\in R,\ j\in J_r \qquad (16.18)
$$

$$
0 \leq w_{ri} \leq ER_{ri}\, y_{ri}, \quad r\in R,\ i\in I_r \qquad (16.19)
$$

$$0 \leq x_{i\ell p} \leq \mathscr{M} \sum_{r \in R} y_{ri}, \quad i \in \cup_{r \in R} I_r,\ \ell \in L \setminus \{i\},\ p \in P \tag{16.20}$$

$$0 \leq x_{\ell ip} \leq \mathscr{M} \sum_{r \in R} y_{ri}, \quad \ell \in L \setminus \{i\},\ i \in \cup_{r \in R} I_r,\ p \in P \tag{16.21}$$

$$x_{jj'p} \geq 0, \quad j, j' \in \cup_{r \in R} J_r\ (j \neq j'),\ p \in P \tag{16.22}$$

$$v_{r\ell p} \geq 0, \quad r \in R,\ \ell \in L,\ p \in P \tag{16.23}$$

$$y_{ri} \in \{0, 1\}, \quad r \in R,\ i \in I_r. \tag{16.24}$$

The proposed model is generic in the sense that it includes multiple types of products and components at different stages (inspected, repaired, refurbished, etc.). Moreover, it considers reverse BOMs and transitions between the stages of products through various recovery options. The problem is modeled with a profit oriented objective function accounting for the revenues from different recovery options in addition to costs.

In terms of problem complexity, the above RLND model has similar attributes to the forward network design problem (P_1). Moreover, general purpose optimization software (e.g., CPLEX or Gurobi) can be used to solve (P_2). However, for large-sized instances there may be a need for customized algorithms and heuristics.

16.3.3 Special Cases and Model Extensions

The generic model (P_2) can be easily tailored to different applications. A reverse logistics network design application for the collection and recovery of waste electrical and electronic equipment is detailed in Sect. 16.4.4.

The term *closed-loop supply chain* refers to a network comprising both forward and reverse flows. Figure 16.4 depicts the structure of such a network. The cost of processing a return flow in a supply chain designed by considering only forward flows can be much higher than processing a flow in the forward direction. Thus, supply chain networks that include flows in the reverse direction should be designed by integrating forward and reverse logistics activities. The models introduced in Sects. 16.2.2 and 16.3.2 are readily extendible to the design of closed-loop supply chains. The interested reader is referred to Krikke et al. (2003), Easwaran and Üster (2009), and Salema et al. (2010) for exemplary studies determining the locations of facilities within closed-loop supply chain networks.

As emphasized in Sect. 16.2.3, the dynamic nature of the (re-)design problem should not be disregarded. Multi-period models in RLND were proposed, for example, by Lee and Dong (2009), Salema et al. (2010), and Alumur et al. (2012).

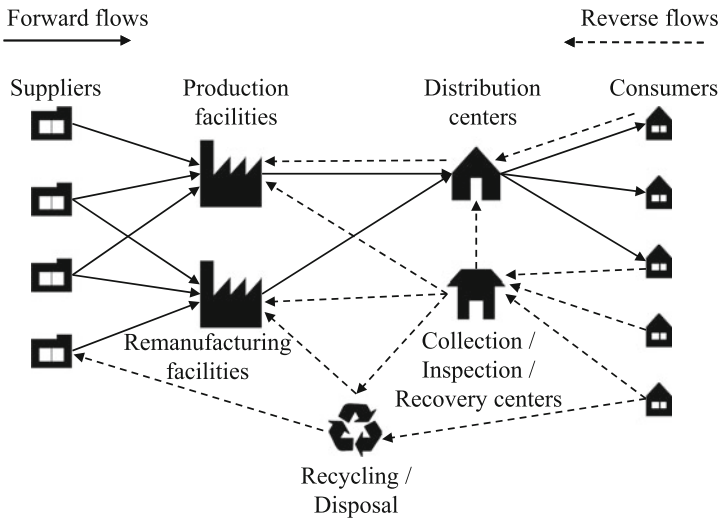


Fig. 16.4 A closed-loop supply chain network

A distinguishing feature of RLND problems is that various sources of uncertainty in supply arise at the upper echelon facilities (e.g., uncertainty in the amount and in the quality of returned products). There are some studies addressing uncertainty issues in the context of RLND such as Realff et al. (2004), Listeş and Dekker (2005), Listeş (2007), Salema et al. (2007), El-Sayed et al. (2010), and Fonseca et al. (2010).

As discussed at the beginning of Sect. 16.3, major driving forces in reverse logistics networks include not only economical factors, but also legislations and environmental consciousness. Thus, in addition to the actors involved in forward logistics networks, actors such as municipalities, foundations, third-party logistics providers, and disposers, are involved in designing and operating reverse logistics networks. Multiple actors lead to decision problems with multiple objectives. Even though there are some studies that consider the multi-objective nature of this design problem (e.g., Pati et al. 2008, Fonseca et al. 2010, Tari and Alumur 2014), this issue requires further attention.

For other extensions and special cases on RLND, the interested reader is referred to the reviews by Fleischmann et al. (2004), Bostel et al. (2005), Akçalı et al. (2009), and Aras et al. (2010).

16.4 Applications

The aim of this section is to demonstrate the richness in LND through presenting applications from various areas including organ transportation in addition to classical areas. The general form of the models described in Sects. 16.2 and 16.3

allows them to be applied to an LND problem of a manufacturer as well as of a logistics service provider under appropriate set, parameter, and variable definitions.

In this section, four applications from different sectors are discussed. Section 16.4.1 presents the network design problem of a global beverage company. Many companies utilize logistics service providers in their distribution networks. In Sect. 16.4.2 an application from this area is provided. Section 16.4.3 is devoted to an atypical application in LND arising in organ transportation. The problem has additional features resulting from the nature of the good being transported. Finally, Sect. 16.4.4 illustrates an application for waste electrical and electronic equipment.

16.4.1 Logistics Network Design of a Beverage Company

Beverage companies usually operate bottling factories in which the required materials are mixed, bottled, and then packaged to be shipped to end users. Global companies usually need to import some of the input materials, like flavors and syrups, to guarantee the same quality worldwide. Moreover, ingredients may also be provided by local suppliers. Thus, inbound logistics involves both international and national shipments to the manufacturing plant. In turn, the outbound flow from the plant comprises bottled and packaged beverages ready to consume. The flow of end products may also be targeted at neighboring countries, thus involving again national and international shipping. The schematic representation of the logistics network, which is a specialized version of Fig. 16.1, is given in Fig. 16.5.

The main decisions in this LND problem include the location of new distribution centers (DCs) and the choice of transportation channels for the inbound and outbound flows of these DCs. As can be seen from Fig. 16.5, the manufacturer may choose to operate additional DCs closer to the customs area to ease the overall customs process. Certain beverages are not produced in every country. Thus, there is a bottled beverage flow from the customs area towards DCs for those products that are not manufactured in a country. Shipments to international customers (via the customs) mainly consist of products that are produced in the local country and they will constitute the in-country product flow in the LND problems of other countries.

Observe here that, in addition to finding the locations of DCs and deciding on the transportation structures to use, the LND problem also includes routing decisions for deliveries to the customers (see the dashed lines in Fig. 16.5). Typically, a global beverage company resorts to logistics service providers to handle the distribution of orders to end users. The service provider operates its own logistics network, which will be detailed in the next subsection. Apart from location and routing decisions, a typical beverage company also questions:

- the level of inventories at the DCs,
- the need for consolidation; some examples include consolidation on the route and consolidation at the facility,

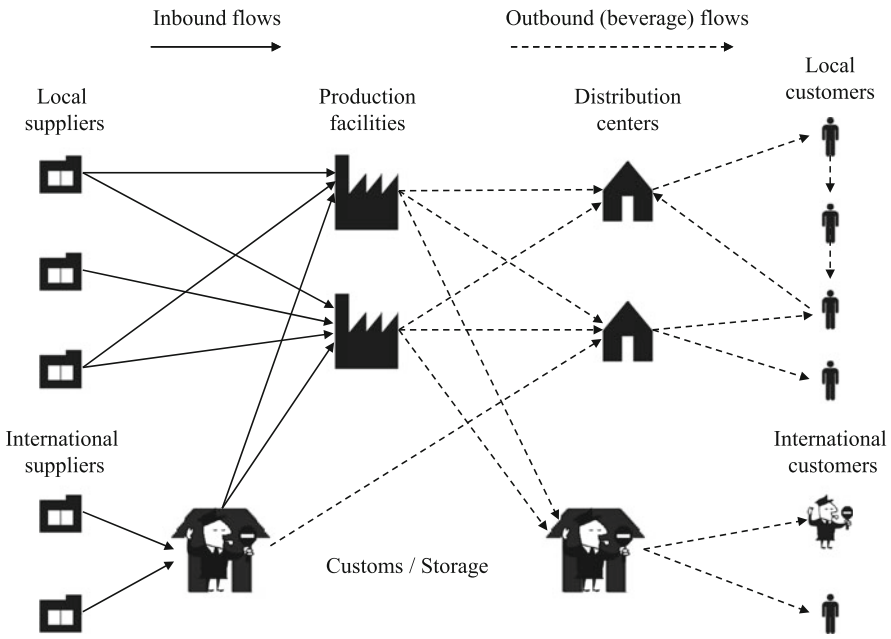


Fig. 16.5 Logistics network of a beverage company

- the transportation mode to be utilized (especially from plants to DCs rail transportation is a valid option).

A beverage company is also engaged in reverse logistics activities through the return of empty flagons to the manufacturing plant. Typically, a logistics service provider combines the delivery of beverages to customers with the collection of empty refillable beverage containers.

16.4.2 Logistics Network Design of a Logistics Service Provider Company

LND is a crucial problem for logistics service provider (LSP) companies since they offer warehousing and transportation services to multiple manufacturers having specific requirements. A typical LSP company generally operates based on yearly contracts, each defining the level of integration to be provided to the customer. This can range from basic services, which mainly handle the transportation aspect of the overall distribution network, to integrated logistics activities, which can even include packaging, labeling, and customs clearance type of services. The design of the network of such a company is, of course, influenced by the level of integration. Nevertheless, a typical LSP usually operates several DCs and the

number of DCs is based on the geographical span and on the promised service levels. Since the logistics network of the service provider company does not include inbound shipments towards production plants, a generic network is composed of production facilities, DCs, and customers (cf. Fig. 16.1). The main decisions to be made in the LND problem include the location of DCs and the choice of appropriate transportation structures.

Consolidation is a crucial aspect in the distribution network of an LSP company. Especially in small geographical regions, say in urban areas, companies try to consolidate customer orders into full truckload shipments. As a result, delivery and/or collection vehicles serve many customers on each route they travel.

Typically, an LSP company operates a few DCs and delivery vehicles travel from/to DCs to service customers. In the upper echelon of the network products flow from factories or central warehouses to DCs. Thus, such a company may consolidate shipments in both stages of the network. Different modes of transportation may be used for bulk transportation from upper echelon facilities.

By nature, LSPs offer services to many companies. Depending on their yearly contracts, the same DC may be used for more than one customer. This type of consolidation brings out the importance of warehouse management activities. Hence, the costs of operating DCs may grow with increasing capacity utilization.

Usually, the type of service offered by such a company is one-way: from the plant or DC towards the customers. This results in empty vehicles returning to the DCs. Providing service to more than one company may actually help in filling vehicles on their return trips. An LSP company usually works with a fleet of vehicles which are not dedicated to any DC or customer zone. Depending on the origin and destination of the demand, vehicles are assigned dynamically.

LSP companies often choose to specialize their services based on the sector of activity of their customers. Some examples include service providers for the automotive industry or the cold chain, parcel delivery companies, etc. The generic distribution network needs to be specialized depending on the application dynamics of the sector where the service provider operates. For example, for cargo delivery companies consolidation (hubbing) is very important in the design of the network (see e.g.,Tan and Kara 2007, Yaman et al. 2007, and Alumur and Kara 2008).

16.4.3 Logistics Network Design for Organ Transportation

In this section, an atypical application of distribution logistics is discussed, namely the design of a network for organ transportation. Due to the nature of the "product" that flows through the network, this problem has specific features. It cannot be simply considered as a cold chain application, mainly because it is not possible to re-freeze and store organs. The organ which is harvested from a donor has to be implanted into the recipient's body within the so-called *ischemia time*, which represents the time that an organ can be safely secured without fresh blood circulation. Thus, in this area, apart from logistics costs, delivering in a timely

manner is more important and so the logistics network is designed mainly based on delivery time requirements.

Since the organ cannot be stored, DCs or warehouses are not considered in the distribution network. Once an organ is donated, a search is conducted for the recipient with the best match and then the organ is transported to the hospital of the recipient. The most important aspect is to find the best match and send the organ in a timely manner so that the donated organ (which is definitely a very scarce resource) is not wasted. Search for potential recipients and organ transportation are under the jurisdiction of regional coordination centers (RCCs) operated by the government. Each RCC is responsible for a region, and any organ donated to an RCC is usually transferred into a recipient's body in the same region.

In this context, the LND problem consists of finding the best locations for RCCs so that the regions covered by them are balanced in terms of their donor-recipient ratio and the transportation of organs in each region is possible within the ischemia time. For this type of networks, donors represent the supply side and the hospitals performing organ transplants (and where the recipients are registered) are the demand points. Examples of this type of centralized organ transportation networks include Bruni et al. (2006), Kong et al. (2010), Beliën et al. (2013), and Çay and Kara (2014). We remark that in this application area the location of an RCC mainly determines a region. Shipment consolidation at an RCC is not allowed since the transportation of an organ from a donor to a recipient is a dedicated trip carried out, for example, by helicopter.

16.4.4 Reverse Logistics Network Design for Waste Electrical and Electronic Equipment

The Waste Electrical and Electronic Equipment (WEEE) Directive of the European Commission (2002/96/EC) sets collection, recovery, and recycling targets for all types of electrical and electronic goods. The achievement of the targets for each product category is calculated according to the total amount of WEEE that goes through specific recovery options. Original equipment manufacturers are held responsible for financing the collection, treatment, recovery, and disposal of their products.

The Directive enforces a separate collection for WEEE. For this purpose, appropriate facilities should be set up for collection. These facilities accumulate the returns, either dropped off by the product holders or picked up by the collectors. After collection, the returns can be sent to recycling and proper disposal, or to inspection and disassembly centers. The inspected products can be disassembled into components in these centers or sold to external facilities. The returns that are deemed non-remanufacturable through inspection are recycled or disposed of. In the event that the original equipment manufacturer decides to establish remanufacturing

facilities, then suitable components can be re-used in such facilities to obtain new products that can be sold to the secondary market.

The RLND problem under the WEEE Directive focuses on determining the locations and capacities of collection and inspection centers, on deciding if it is profitable to establish remanufacturing facilities, on setting the amount of products or components to send to different recovery options, to recycling and disposal, and on fixing the flow of products and components through the facilities in the network (see e.g., Alumur et al. 2012).

16.5 Conclusions

This chapter highlighted the importance of integrating location decisions with other decisions relevant to the design of forward and reverse logistics networks. Although much work has been published addressing LND problems, emphasis has been mostly given to a subset but not all of the features that such comprehensive projects often require. Hence, several research directions still require intensive research. In particular, models addressing the design of multi-commodity, multi-echelon networks through determining the timing of facility locations, expansions, contractions, and relocations over an extended time horizon have received less attention than their static counterpart.

Traditionally, LND has been dominated by economic aspects leading to the network configuration that either minimizes total cost or maximizes total profit. The generic models presented in Sects. 16.2.2 and 16.3.2 illustrate these features. Sustainable LND is an emerging research area that aims at capturing the trade-offs between costs on facility location and logistics functions and their environmental footprint. Due to the growing awareness on environmental issues, companies have recognized the need to create environmentally friendly logistics systems to mitigate the negative environmental impact of their business activities. This calls for the development of models with multiple and conflicting objectives. For example, Chaabane et al. (2012) formulate a bi-objective LND model involving the minimization of network design costs and the minimization of green gas emissions. The latter criterion is part of a longer list of environmental factors that should be considered, according to Chen et al. (2014), together with social and economic factors when deciding on the location of manufacturing facilities.

Humanitarian logistics has also become a new research field involving LND. Döyen et al. (2012) integrate facility location decisions with transportation, inventory management, and shortage policies in a two-echelon model. Uncertainty on the location and intensity of a natural disaster is explicitly incorporated into the model. The integration of different sources of uncertainty (e.g., customer demand, product return in the context of reverse logistics) with network design decisions is also a research direction requiring further attention.

Finally, it goes without saying that LND has given rise and will continue to provide a rich variety of problems. LND presents a challenging area for future research

on the development of mathematical models and optimization methodologies. More and more organizations recognize the importance of an efficient and agile logistics network for responding to changes in the business environment and enabling future growth. Therefore, LND will play an even greater role for companies in all industries striving to deliver outstanding supply chain performance.

References

Aghezzaf E (2005) Capacity planning and warehouse location in supply chains with uncertain demands. J Oper Res S 56:453–462

Akçalı E, Çetinkaya S, Üster H (2009) Network design for reverse and closed-loop supply chains: an annoted bibliography of models and solution approaches. Networks 53:231–248

Alumur S, Kara B (2008) Network hub location problems: the state of the art. Eur J Oper Res 190:1–21

Alumur SA, Nickel S, Saldanha da Gama F, Verter V (2012) Multi-period reverse logistics network design. Eur J Oper Res 220:67–78

Aras N, Boyacı T, Verter V (2010) Designing the reverse logistics network. In: Ferguson M, Souza G (eds) Closed loop supply chains: new developments to improve the sustainability of business practices, chap 5. CRC Press, Boca Raton, pp 67–98

Ballou RH (2001) Unresolved issues in supply chain network design. Inf Syst Front 3:417–426

Beliën J, De Boeck L, Colpaert J, Devesse S, Van den Bossche F (2013) Optimizing the facility location design of organ transplant centers. Decis Support Syst 54:1568–1579

Bostel N, Dejax P, Lu Z (2005) The design, planning, and optimization of reverse logistics networks. In: Langevin A, Riopel D (eds) Logistics systems: design and optimization, chap 6. Springer, New York, pp 171–212

Bruni ME, Conforti D, Sicilia N, Trotta S (2006) A new organ transplantation location-allocation policy: a case of Italy. Health Care Manag Sci 9:125–142

Çay P, Kara BY (2014) Organ transportation logistics: a case for Turkey. Technical Report, Department of Industrial Engineering, Bilkent University, Ankara

Chaabane A, Ramudhin A, Paquet M (2012) Design of sustainable supply chains under the emission trading scheme. Int J Prod Econ 135:37–49

Chen L, Olhager J, Tang O (2014) Manufacturing facility location and sustainability: a literature review and a research agenda. Int J Prod Econ 149:154–163

Correia I, Melo T, Saldanha da Gama F (2013) Comparing classical performance measures for a multi-period, two-echelon supply chain network design problem with sizing decisions. Comput Ind Eng 64:366–380

De Brito MP, Dekker R (2004) A framework for reverse logistics. In: Dekker R, Fleischmann M, Inderfurth K, Van Wassenhove LN (eds) Reverse logistics: quantitative models for closed-loop supply chains, chap 1. Springer, Berlin, pp 3–27

Döyen A, Aras N, Barbarosoğlu G (2012) A two-echelon stochastic facility location model for humanitarian relief logistics. Optim Lett 6:1123–1145

Easwaran G, Üster H (2009) Tabu search and benders decomposition approaches for a capacitated closed-loop supply chain network design problem. Transp Sci 43:301–320

El-Sayed M, Afia N, El-Kharbotly A (2010) A stochastic model for forward-reverse logistics network design under risk. Comput Ind Eng 58:423–431

Fleischmann B, Ferber S, Henrich P (2006) Strategic planning of BMW's global production network. Interfaces 36:194–208

Fleischmann M, Bloemhof-Ruwaard JM, Beullens P, Dekker R (2004) Reverse logistics network design. In: Dekker R, Fleischmann M, Inderfurth K, Van Wassenhove LN (eds) Reverse logistics: quantitative models for closed-loop supply chains, chap 4. Springer, Berlin, pp 65–94

Fonseca MC, García-Sánchez A, Ortega-Mier M, Saldanha da Gama F (2010) A stochastic bi-objective location model for strategic reverse logistics. TOP 18:158–184

Geoffrion AM, Graves GW (1974) Multicommodity distribution system design by Benders decomposition. Manag Sci 20:822–844

Harrison TP (2004) Principles for the strategic design of supply chains. In: Harrison T, Lee H, Neale J (eds) The practice of supply chain management: where theory and application converge, chap 1. Springer, New York, pp 3–12

Huang E, Goetschalckx M (2014) Strategic robust supply chain design based on the Pareto-optimal tradeoff between efficiency and risk. Eur J Oper Res 237:508–518

Jayaraman V, Luo Y (2007) Creating competitive advantages through new value creation: a reverse logistics perspective. Acad Manag Perspect 21:56–73

Klibi W, Martel A, Guitouni A (2010) The design of robust value-creating supply chain networks: a critical review. Eur J Oper Res 203:283–293

Kong N, Schaefer AJ, Hunsaker B, Roberts MS (2010) Maximizing the efficiency of the U.S. liver allocation system through region design. Manag Sci 56:2111–2122

Krarup J, Pruzan PM (1983) The simple plant location problem: survey and synthesis. Eur J Oper Res 12:36–81

Krikke HR, Bloemhof-Ruward JM, Van Wassenhove LN (2003) Concurrent product and closed-loop supply chain design with an application to refrigerators. Int J Prod Res 41:3689–3719

Lee DH, Dong M (2009) Dynamic network design for reverse logistics operations under uncertainty. Transp Res E-Log 45:61–71

Listeş O (2007) A generic stochastic model for supply-and-return network design. Comput Oper Res 34:417–442

Listeş O, Dekker R (2005) A stochastic approach to a case study for product recovery network design. Eur J Oper Res 160:268–287

Meixell MJ, Gargeya VB (2005) Global supply chain design: a literature review and critique. Transp Res E-Log 41:531–550

Melo MT, Nickel S, Saldanha da Gama F (2006) Dynamic multi-commodity capacitated facility location: a mathematical modeling framework for strategic supply chain planning. Comput Oper Res 33:181–208

Melo MT, Nickel S, Saldanha da Gama F (2008) Network design decisions in supply chain planning. In: Buchholz P, Kuhn A (eds) Optimization of logistics systems: methods and experiences, chap 1. Praxiswissen, Dortmund, pp 1–19

Melo MT, Nickel S, Saldanha da Gama F (2009) Facility location and supply chain management: a review. Eur J Oper Res 196:401–412

Melo MT, Nickel S, Saldanha da Gama F (2012) A tabu search heuristic for redesigning a multi-echelon supply chain network over a planning horizon. Int J Prod Econ 136:218–230

Melo MT, Nickel S, Saldanha da Gama F (2014) An efficient heuristic approach for a multi-period logistics network redesign problem. TOP 22:80–108

Nickel S, Saldanha da Gama F, Ziegler HP (2012) A multi-stage stochastic supply network design problem with financial decisions and risk management. Omega 40:540–524

Ozsen L, Daskin M, Coullard C (2009) Facility location modeling and inventory management with multisourcing. Transp Sci 43:455–472

Pan F, Nagi R (2010) Robust supply chain design under uncertain demand in agile manufacturing. Comput Oper Res 37:668–683

Pati RK, Vrat P, Kumar P (2008) A goal programming model for paper recycling system. Omega 36:405–417

Realff MJ, Ammons JC, Newton DJ (2004) Robust reverse production system design for carpet recycling. IIE Trans 36:767–776

Salema MI, Barbosa-Póvoa AP, Novais AQ (2007) An optimization model for the design of a capacitated multi-product reverse logistics network with uncertainty. Eur J Oper Res 179:1063–1077

Salema MI, Barbosa-Póvoa AP, Novais AQ (2010) Simultaneous design and planning of supply chains with reverse flows: a generic modelling framework. Eur J Oper Res 203:336–349

Santoso T, Ahmed S, Goetschalckx M, Shapiro A (2005) A stochastic programming approach for supply chain network design under uncertainty. Eur J Oper Res 167:96–115
Shen ZJ (2005) A multi-commodity supply chain design problem. IIE Trans 37:753–762
Shu J (2010) An efficient greedy heuristic for warehouse-retailer network design optimization. Transp Sci 44:183–192
Tan PZ, Kara BY (2007) A hub covering model for cargo delivery systems. Networks 49:28–39
Tancrez JS, Lange JC, Semal P (2012) A location-inventory model for large three-level supply chains. Transp Res E-Log 48:485–502
Tari I, Alumur SA (2014) Collection center location with equity considerations in reverse logistics networks. Technical Report, Department of Industrial Engineering, TOBB University of Economics and Technology, Ankara
Wilhelm W, Liang D, Rao B, Warrier D, Zhu X, Bulusu S (2005) Design of international assembly systems and their supply chains under NAFTA. Transp Res E-Log 41:467–493
Yaman H, Kara BY, Tansel BC (2007) The latest arrival hub location problem for cargo delivery systems with stopovers. Transp Res B-Met 41:906–919

Chapter 17
Stochastic Location Models with Congestion

Oded Berman and Dmitry Krass

Abstract In this chapter we describe facility location models where consumers generate streams of stochastic demands for service, and service times are stochastic. This combination leads to congestion, where some of the arriving demands cannot be served immediately and must either wait in queue or be lost to the system. These models have applications that range from emergency service systems (fire, ambulance, police) to networks of public and private facilities. One key issue is whether customers travel to facilities to obtain service, or mobile servers travel to customer locations (e.g., in case of police cars). For the most part, we focus on models with static (fixed) servers, as the underlying queueing systems are more tractable and thus a richer set of analytical results is available. After describing the main components of the system (customers, facilities, and the objective function), we focus on the customer-facility interaction, developing a classification of models based on the how customer demand is allocated to facilities and whether the demand is elastic or not. We use our description of system components and customer-response classification to organize the rich variety of models considered in the literature into four thematic groups that share common assumptions and structural properties. For each group we review the solution approaches and outline the main difficulties. We conclude with a review of some important open problems.

Keywords Congestion • Facility location • Mobile and immobile servers • Queuing • Stochastic demand

17.1 Introduction

The class of facility location models that is the main focus of the current chapter make the following key assumptions:

1. Customers generate *stochastic* stream of demands, typically assumed to be a Poisson process, or, more generally a renewal process.

O. Berman (✉) • D. Krass
Rotman School of Management, University of Toronto, Toronto, ON, Canada M5S 2C8
e-mail: berman@rotman.utoronto.ca; krass@rotman.utoronto.ca

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_17

2. Facilities contain resources (often called "servers") that have *limited* capacity and *stochastic service times*.
3. Customer-facility interactions happen as the result of *customers traveling to facilities* to seek service, i.e., our primary focus is on the "fixed" or "immobile" server models (in the "mobile server" case, servers travel to customers to provide service).
4. Due to stochastic arrivals of customer demands at the facilities, stochastic service times, and limited capacities, facilities may experience periods of *congestion* where not all arriving demands can be served immediately. Customers that arrive when the system is busy may either enter a queue or leave without getting service. This behavior will result in either *queues*, or *lost demands*, or both.

Applications of these models range from public service facilities such as hospitals, medical clinics and government offices, to private facilities such as retail stores or repair shops.

We note that assumptions listed above specifically exclude a number of interesting and important classes of related location models (some of these are treated in other chapters in the current volume). First, there are many models that incorporate capacity limitations in a deterministic, rather than stochastic, manner. These include models seeking to ensure that there is sufficient average capacity to provide adequate service, models that try to design a system that should perform well even under stochastic conditions by equalizing loads between facilities, and models that handle possible congestion indirectly by requiring certain reserve capacity at the facilities. All of these can be regarded as deterministic approximations of the underlying stochastic system. While this deterministic approach leads to large technical simplifications and, as a result, much easier computations, the roughness of the approximation is usually impossible to estimate a priori. This may lead to systems with poor levels of customer service (at some of the facilities), and is typically not appropriate in cases where understanding and controlling potential congestion is important.

Second, there are some models where facilities are modeled as reliability, rather than queueing, systems, i.e., a facility may "fail" with certain probability in some periods, at which point it cannot provide service to customers (who are typically assumed to try to seek service from non-failed facilities). These models do incorporate stochastic demands explicitly. Moreover, "failure" periods may be regarded as representing periods of congestion at the facilities when new customer arrivals are blocked. Thus, these models are closer to the systems we study. However, the key difference is that "reliability" models treat the blockage probability as exogenous to the system (a typical assumption is that each facility may fail with certain probability at any time, where such probability is a system parameter), while models where facilities are represented as queues treat the probability of blockage as endogenous, i.e., it is a direct outcome of other decisions such as capacity allocation and customer-facility interactions. Thus, reliability models can only be regarded as approximations for the systems we are interested in.

Third, there is an important class of models where servers are assumed to be "mobile", i.e., servers travel to customers rather than customers traveling to facilities. Examples of the underlying systems include emergency services (fire, ambulance, police) as well as repairmen making house calls. These models are close "cousins" of the fixed-server models as they do include most of the same components: stochastic demand streams, stochastic service times, congestion/queuing behavior. However, these models also include additional significant levels of complexity, such as dynamic dispatching and routing of servers, where servers can be repositioned between facilities, re-routed before completion of the call, etc. The underlying queuing models are analytically intractable, even if the facility locations are assumed fixed, leading to various approximation-based approaches. In contrast, the queuing systems underlying models with fixed servers are often (though not always) analytically tractable, allowing for more (theoretically) precise solutions in many cases. We refer the reader to a survey by Berman and Krass (2002) and to a more recent survey on emergency systems planning by Ignolfsson (2013) for more details on models with mobile servers. We note that the technical distinction between models with fixed and mobile servers does not lie in the server mobility per se, but rather in how the underlying queuing network is modeled (in fact, some of the models described in this chapter have been applied in mobile server contexts). We will provide more precision for this distinction below, once the underlying technical framework is properly introduced.

The field of *Stochastic Location models with Congestion and Immobile Servers* (SLCIS), the main focus of this chapter, has seen a rather explosive growth over a relatively recent time period. As noted in Berman and Krass (2002), by the early 2000s, only a handful of papers on SLCIS could be found. However, by 2006 over 20 contributions were listed in the comprehensive review by Boffey et al. (2006) (we are only counting the papers that meet the assumptions for SLCIS models discussed earlier). In the last eight years, this number has roughly doubled. It is our intent to review the current state of the field, as well as to systematize the many variants of SLCIS models that have been proposed.

We note that much of the recent work has been on models with elastic demand—i.e., where the intensity of customer demands depends on the quality of the service provided by the facilities. In this regard it is important to mention a review by Brandeau et al. (1995) that describes early foundation for much of this work.

As with most other location models, one could focus on cost minimization or on net revenue (profit) maximization. Cost minimization is more appropriate when the revenues are either not well-defined (e.g., in the case of public health facilities), or are assumed to be exogenous to the model (e.g., when customer demand levels and prices are fixed). While most SLCIS models in the literature are formulated with the cost minimization objective, profit optimization is more general and is much more natural when demand is elastic. Therefore, we will assume this objective type in our general formulation in the following section.

The remainder of this chapter is organized as follows. We start by describing the main model components in Sect. 17.2. These components include customers, facilities, and the objective function of the model. A crucial part of any SLCIS

model is the set of assumptions made about how customers and facilities interact, specifically how customer demand is "allocated" to facilities and how much of the potentially available demand is "captured". These issues are explored in detail in Sect. 17.3, where we also introduce a classification of SLCIS models based on the types of customer response. All model components come together in Sect. 17.4 where we formulate a "general" SLCIS model and review the main features that are typically included in various sub-classes. In Sect. 17.5 we provide an overview of SLCIS models discussed in the literature, providing a unifying structure organized around four main "themes". We also discuss the key challenges that arise for different model classes and computational approaches that have been developed. In the last section we discuss conclusions and suggestions for future research.

17.2 Key Model Components

As noted earlier, SLCIS models describe the system consisting of customers, facilities and their interactions. We start by describing each of these components in more detail.

17.2.1 Customers

Customers are assumed to be located in a set J, with customer location $j \in J$ capable of generating a demand stream with maximum intensity of $\lambda_j^{\max}$ per unit time. In the vast majority of models described in the literature, J is assumed to be a discrete set, often conceptualized as the set of nodes of some underlying network $G = (J, A)$, where A is the set of links. Other common alternatives in location (but not in SLCIS) literature include J being a sub-region of the real plane R^2, or consisting of both links and nodes of a network G. The most general SLCIS setting we are aware of is given in Baron et al. (2008), where J is a bounded sub-space of R^N and can contain a mixture of discrete points and continuous regions. To keep the presentation as transparent as possible, we will retain the common assumption that J is discrete and $n = |J|$ is the number of customer demand points, which we will frequently refer to as "nodes".

Let u_j represents the *utility* derived by customers at node $j \in J$ from services offered by the facilities. The demand stream generated by j is assumed to be a Poisson process with rate $\lambda(u_j) \in [0, \lambda_j^{\max}]$. We will postpone the description of utility functions until Sect. 17.3.1, since other system components need to be defined first. However, we can already identify two different classes of SLCIS models: the *elastic demand* models, where $\lambda(u_j)$ is a non-constant function, i.e., $\lambda(u_j) \neq \lambda_j^{\max}$ for some values of u_j, and the *inelastic demand* models where the demand rate is

assumed to be constant and equal to $\lambda_j^{\max}$. As a shorthand, we will use $\lambda_j = \lambda(u_j)$ to represent the demand rate of customer node $j \in J$. The inter-arrival times of the demand processes generated by different customer locations are assumed to be independent.

We should also note that while it is tempting to relax the Poisson assumption for the demand process, this must be done with care as the facilities see aggregate demands from different customer locations, i.e., a superposition of the demand processes. In order to apply standard queueing results to the facilities, the demand process seen by each facility must be a renewal process. While the superposition of Poisson processes is Poisson, which is obviously a renewal process, in general, the superposition of renewal processes is not a renewal process. This quickly leads to a loss of tractability for the models. Thus, except for some trivial extensions, the Poisson assumption for demand streams appears unavoidable (one interesting exception occurs when customer demand space is continuous, rather than finite, in which case facilities see Poisson arrivals under much loser conditions—see Baron et al. (2008) for the development and required assumptions). However, there is no problem, at least from the analytical point of view, in assuming that the demand process at each node $j \in J$ is not time-homogenous, i.e., that the demand rate is a function of time. To simplify the presentation, we will stick with the time-homogenous assumption.

An important implicit assumption in all SLCIS models we are aware of is that all customers generate "identical" demands (in terms of service requirements), i.e., that the streams of demand are indistinguishable once they reach the facility.

17.2.2 Facilities

Customer demands are serviced by the *facilities* that contain *service resources* (or "servers"). All aspects related to the facilities, including their number, locations, and the amount/types of resources allocated to them can, potentially, be treated as decision variables in the model. In describing the system dynamics below we will initially treat the values of these variables as having already been determined, but will relax this assumption when describing model formulations later.

We will assume that facility locations must belong to some set I and that at most $m \geq 0$ facilities can be located; we will use $i \in I$, to represent the location (site) of facility i. By far, the most common assumption in SLCIS literature is that set I is discrete, i.e., that all potential locations for the facilities have already been enumerated. In this case, we can assume without loss of generality that $I \subset J$ (since any point in I not containing customers can be treated as a customer demand point with the maximum demand rate equal to 0). Other options, include $I \subset R^2$, leading to *continuous* SLCIS models (see, for example, Brimberg and Mehrez 1997; Brimberg et al. 1997), or $I \subset J \cup A$ for a network G, leading to *network*

SLCIS models (see, e.g., Berman et al. 2014). Unless stated otherwise, we will generally assume I to be discrete.

To take advantage of the discreteness of I we will follow the typical convention in location modeling and define $y_i \in \{0, 1\}$ to be a binary indicator variable with the value 1 if a facility is open at site $i \in I$, and 0 otherwise. To ensure that the total number of open facilities does not exceed m we require:

$$\sum_{i \in I} y_i \leq m. \tag{17.1}$$

If a facility is opened at $i \in I$ (i.e. $y_i = 1$), it must be allocated some service capacity $\mu_i > 0$, which can be thought of as the average processing rate. We will assume that $\mu_i = 0$ whenever $y_i = 0$, which can be assured by using the constraints

$$\mu_i \leq M y_i, \quad i \in I, \tag{17.2}$$

where M is the maximum possible processing capacity that can be assigned to a facility.

As noted in Baron et al. (2008), there are two standard approaches to represent facility capacity in queuing environment: as a "single-server" facility where the capacity level can take on any value in some interval $\mu_i \in [0, \mu^{\max}]$, where $\mu^{\max}$ is the maximum practical capacity level, or as a "multi-server" facility housing $\kappa_i \geq 0$ parallel servers each with fixed capacity μ^0, where $\kappa_i \in \{0, \ldots, k\}$ is an integer, $\mu_i = \kappa_i \mu^0$ is the processing capacity of facility i, and k is the maximum number of servers that can be stationed at a facility (with $\mu^{max} = k\mu_0$).

While there are some important differences between the single-server and multi-server models (these will be touched on later) our bias is to favor the single-server representation. It is more transparent, typically leads to cleaner analytical results, and seems more practical as well: a typical facility will house a variety of processing resources and discrete "servers" may be hard to identify. For example, a medical clinic will often house doctors, nurses, examination rooms, X-ray machines, etc. While it is sensible for a planner to think of processing capacity of a clinic in terms of patients per hour (and how this processing capacity changes when certain resources are added or removed), it is harder to think of the clinic containing κ distinct servers (are these doctors? nurses? rooms?). Thus, unless stated otherwise, each facility will be assumed to house a single "server" with capacity μ.

The service times at each facility are assumed to be stochastic. More specifically, following Baron et al. (2008), we assume First Come First Serve (FCFS) service discipline and that service requirements (which can be thought of as the amount of work required to process one customer request) are independent and identically distributed random variables with a cumulative distribution function (CDF) $\mathcal{F}_S(w)$, and a well-defined moment generating function (MGF) $G_S(\eta)$. We also assume that the mean service time $E[S] = 1$—this assumption is made with no loss of generality as it simply rescales service times. Note that in this framework, since μ_i represents the service rate of facility i, the mean service time is $1/\mu_i$ and it is not hard to show that the distribution of service times is given by $F_S(\mu_i w)$ with MGF $G_S(\eta/\mu_i)$.

We define x_{ij} to be *demand allocation* decision variables, specifying what portion of demand from customer node $j \in J$ is directed to facility $i \in I$. We will initially assume that demand allocations are binary, with the value of 1 if the demand stream generated by customer node j is directed to facility i, and 0 otherwise. The key underlying assumption is that once the decisions about the number of facilities, their locations y_i and the service capacities μ_i for $i \in I$ are made, the demand allocations x_{ij} can be determined; the exact mechanism for determining demand allocations depends on the underlying assumptions about system dynamics and is described later. Mathematically, we assume that x_{ij} satisfies the following set of constraints

$$\sum_{i \in I} x_{ij} \leq 1, \; j \in J \tag{17.3}$$

$$x_{ij} \leq y_i, \; i \in I, \; j \in J \tag{17.4}$$

$$x_{ij} \in \{0, 1\}, \; i \in I, \; j \in J \tag{17.5}$$

These constraints are quite standard in location models: (17.3) ensures that at most 100 % of customer demand from j is allocated to the facilities, (17.4) prevents allocating a customer to an unopened facility, and (17.5) enforces the binary assumption for the allocations.

The integrality of x_{ij} reflects the "single sourcing" assumption made in most SLCIS models, requiring each customer node to be assigned to at most one facility. An alternative is to allow "multisourcing", in which case x_{ij} is allowed to be continuous, by replacing (17.5) with its linear relaxation. We also note that constraints (17.3)–(17.5) represent "minimal" requirements on x_{ij}; they are often supplemented by other constraints describing the mechanisms by which allocation of customers to facilities is made.

We allow for the possibility that the demand from j is not assigned to any facility, i.e., $\sum_{i \in I} x_{ij} = 0$, which we interpret as the case of "*intentionally*" *lost demand*, i.e. demand that could have been captured but was lost at the system planning stage, usually due to insufficient overall system capacity. We note that even when $x_{ij} = 1$ some demand from i may be lost due to congestion at facility J - this portion can be regarded as "unintentionally" lost demand, since the system did attempt to provide service to customers at i. The amount of lost demand is typically controlled via a penalty cost or constraints—we will return to these when we discuss specific model formulations below. For each facility i we define the set $N_i = \{j \in J | x_{ij} = 1\}$, which represents the *service region* of facility i (clearly $N_i = \emptyset$ when $y_i = 0$).

Observe that once λ_i and x_{ij} are known, the demand rate facing an open facility i is a Poisson process with rate

$$\Lambda_i = \sum_{j \in N_i} \lambda_j = \sum_{j \in J} \lambda_j x_{ij}. \tag{17.6}$$

As mentioned earlier, the Poisson property results from the fact that superposition of Poisson processes is also a Poisson process. Moreover, the demand streams faced

by different facilities are independent of each other. Thus, each facility $i \in I$ acts as a stand-alone queueing system with Poisson arrivals and general service times, i.e., an $M/G/1$ (or $M/G/\kappa_i$) queue with service rate μ_i.

System stability (i.e., ensuring that queue lengths are finite) requires that

$$\Lambda_i \leq \mu_i, i \in I, \tag{17.7}$$

which acts as a constraint on capacity assignment decisions. In addition, the framework defined above allows us to express the key performance characteristics of the facilities, such as the steady-state system waiting time $W_i = W(\Lambda_i, \mu_i)$ (this includes both queueing and service times), and the steady-state number of customers in the system $L_i = L_i(\Lambda_i, \mu_i)$, both of which are random variables whose distributions can, in principle, be obtained. We will come back to these quantities when we discuss system costs and service-level constraints in the next section.

It may also be useful to require that each facility face some minimum demand rate Λ^{min} in order to ensure that it can be operated economically; sometimes these minimum demand rates are imposed by regulators for public service facilities (see, e.g., Zhang et al. 2010). These constraints take the form

$$\Lambda_i \geq \Lambda^{min} y_i, i \in I. \tag{17.8}$$

We note that many models make additional assumptions regarding the operations of facilities. For example, the assumption that the distribution of service times is exponential is quite common (though likely not very realistic in many real-life systems; e.g., see the discussion in Boffey et al. 2006). Some authors (e.g., Boffey et al. 2010) assume limited buffer space at the facilities. We will delay the discussion of these additional aspects until Sect. 17.5. For the moment we regard each facility as an infinite-buffer $M/G/1$ or $M/G/\kappa$ queue.

Remark The fact that each facility (once location, capacity and customer allocation decisions are made) can be viewed as an independent queueing system is the main characteristic distinguishing immobile from mobile server models; in mobile server models the systems operated by different facilities cannot be decoupled. This is because in these models the typical assumption is that server assignments are dynamic, i.e., depend on the state of the system. Thus a server from a given facility may service demands from customers at point j under some conditions, but not under others. This leads to a system which is not, in general, separable, and where servers located at different facilities must be treated as distinguishable. Such queueing networks are analytically intractable even when all location, capacity and allocation decisions are made. Thus, all modeling approaches involve strong approximations and/or descriptive/simulation components (e.g., the Hypercube model proposed by Larson (1974) is frequently used as the modeling foundation).

In contrast, SLCIS models decompose into a set of queues with Poisson arrivals—systems for which strong analytical results (both exact and approximate)

are available. We emphasize that this tractability rests in the static nature of customer-to-facility allocations (the demand allocations are determined once and then remain in force for all states of the system). Thus, SLCIS models where customers decide which facility to visit based on the current state of the system (e.g., based on posted information about current waiting times), or where other dynamic customer allocation mechanisms may be present, are likely to be closer (in terms of tractability and solution approaches) to models with mobile servers. On the other hand, models with mobile servers where static and non-intersecting service regions are assumed for all facilities (effectively assuming away dynamic customer reallocation) are quite similar to SLCIS models; many of the mobile server models reviewed in Berman and Krass (2002) fall into this group. Thus, instead of differentiating stochastic location models with mobile vs. immobile servers, it would be more accurate to differentiate models with dynamic vs. static customer assignments.

17.2.3 Costs, Revenues, and Constraints

To complete the description of the system it remains to specify two components: (1) the mechanisms by which customers are "allocated" to the facilities, expressed by the variables x_{ij} (which would also determine the actual demand rates $\lambda_j, j \in J$), and (2) the overall system costs and constraints assuring acceptable service levels. We will postpone the discussion of (1) until Sect. 17.3, focusing on the costs and constraints in the current section and treating values of the key location, allocation, capacity assignment and demand level decisions $\{y_i, x_{ij}, \mu_i, \lambda_i\}, i \in I, j \in J$ as fixed.

17.2.3.1 Travel Cost and Coverage Constraints

We assume that for each customer $j \in J$ and potential facility location $i \in I$ a distance metric $d(i, j)$ is defined, satisfying the regular properties of distance. The travel cost function $TC(d)$ for $d \geq 0$, representing the cost of traveling distance d is assumed to be non-decreasing and non-negative. This yields the System Travel Cost of

$$STC = \sum_{j \in J} \sum_{i \in I} TC(d(i, j))\lambda_j x_{ij}, \tag{17.9}$$

where we assume that constraint (17.4) ensures that customers are only assigned to open facilities. This expression merely states that the system travel cost is the sum of travel costs of all customers to their assigned facilities. We note that a frequent assumption is that the travel cost is a linear function of distance. More generally, since both J and I are discrete, one could simply redefine the distance measure

to be $d'(i,j) = TC(d(i,j))$ for all $j \in J$, $i \in I$ and use this new measure in place of the original one. Thus, after suitably redefining distances and without loss of generality, we can write

$$STC = \sum_{j \in J} \sum_{i \in I} \beta d(i,j) \lambda_j x_{ij}, \tag{17.10}$$

where $\beta > 0$ is a parameter relating the travel cost to other terms in the objective function (defined below). We will use this linear form in place of (17.9) from this point on.

Of course, a possible concern with the previous expression is that the short travel cost of one customer will be added to the long travel cost of another, resulting in the total quantity that may look reasonable, but will still provide poor service to some customers. To assure that no customer faces an unreasonably long travel distance, one can impose *coverage constraints*:

$$\sum_{i \in I} d(i,j) x_{ij} \leq R \text{ for all } j \in J, \tag{17.11}$$

where $R > 0$ is the "coverage radius", i.e., the maximum allowed travel distance for a customer to be "covered" by a facility (this constraint should be interpreted as referring to the "adjusted " distance measure that incorporates the travel cost, as discussed above). We note that most SLCIS models will include either (17.10) or (17.11); while, in principle, both can be used in the same model, such usage is rare.

17.2.3.2 Congestion Costs and Service Level Constraints

While travel-related costs are present in all classes of location models covered in the current volume, the congestion-related costs and constraints are, of course, a defining feature of the stochastic location models with congestion. As discussed earlier, the two common performance measures in a queueing system operated by each open facility $i \in I$ are the system waiting time W_i (recall that this includes the service time; a closely related measure is W_i^q which only covers the waiting time in queue) and the number of customers in the system L_i, which are random variables with certain steady-state distributions. The most common way to define congestion costs is in terms of expectations of these quantities, $\overline{W}_i$ and $\overline{L}_i$, respectively. Since the two are related by Little's Law, we will focus on the former (which is also more commonly used). For an $M/G/1$ queue, the expression for the mean waiting time in the system $\overline{W}$ can be found in any standard reference on queuing (see, e.g., Gross and Harris 1985, p. 255):

$$\overline{W} = \overline{W}^q + \frac{1}{\mu} = \frac{1+\gamma^2}{2} \frac{\rho}{1-\rho} \frac{1}{\mu} + \frac{1}{\mu} \tag{17.12}$$

where $\overline{W}^q$ is the expected time in queue, $\rho = \lambda/\mu$ is the utilization ratio and γ^2 is the squared coefficient of variation for service times, given by $\gamma^2 = \sigma^2\mu^2$, where σ^2 is the variance of service times. Each term in the expression for $\overline{W}^q$ has an intuitive interpretation. Recall that we are assuming Poisson arrivals, which have coefficient of variation equal to 1, and thus the term $\frac{1+\gamma^2}{2}$ represents the average squared coefficient of variation for arrival and service processes, often called the "variability factor" (for exponential service this term equals to 1). The second term, $\rho/(1-\rho)$ can be interpreted by recalling that ρ is the probability that the server is busy and thus $(1-\rho)$ is the probability that an arriving demand goes straight into service. The ratio can thus be interpreted as the length of the busy period measured in units of the length of the free period. The last term is simply the average service time per customer, sometimes known as the "scale effect" to recognize that as more capacity is assigned to the system, the average service time per customer declines. Thus

$$\overline{W}^q = [\text{Variability Factor}]\left[\frac{\text{Prob system busy}}{\text{Prob system free}}\right][\text{Scale Effect}]. \tag{17.13}$$

The expression for $\overline{W}$ simply adds the expected service time to the above.

Remark As noted earlier, two popular ways to represent the queueing system at a given facility are as either single-server $M/G/1$ queue with capacity μ, where μ is a decision variable, or as a multi-server $M/G/\kappa$ system where each of the κ servers has capacity μ^0 and κ is the decision variable. If we set $\kappa\mu^0 = \mu$, i.e., require both systems to have the same processing capacity, we can ask to what extent are these systems "equivalent"? Can the simpler $M/G/1$ system be used as an approximation of harder-to-analyze $M/G/\kappa$ one?

Equations (17.12) and (17.13) can be used to analyze the relationship between these two systems. First note that the coefficient of utilization ρ is the same when $\mu = \kappa\mu^0$. While no closed-form expression for $\overline{W}$ is known for the multi-server $M/G/\kappa$ case, a popular approximation (see e.g., Hopp and Spearman 2000, p. 273) is:

$$\overline{W} = \overline{W}^q + \frac{1}{\mu^0} \approx \frac{1+\gamma^2}{2}\frac{\rho^{\sqrt{2(\kappa+1)}-1}}{1-\rho}\frac{1}{\kappa\mu^0} + \frac{1}{\mu^0}, \tag{17.14}$$

which is very similar to (17.12): focusing on the expression for $\overline{W}^q$, we see that the only difference is that ρ in the numerator of (17.12) is replaced with $\rho^{\sqrt{2(\kappa+1)}-1}$ in (17.14). In fact, the latter approximates the probability that all servers are busy in the $M/G/\kappa$ system. Thus, each term in the intuitive interpretation (17.13) of $\overline{W}^q$ has the same interpretation for both systems. The only difference in the expected waiting times is that $M/G/1$ system is busy more frequently (since $1 > \rho > \rho^{\sqrt{2(\kappa+1)}-1}$), thus yielding larger values of $\overline{W}^q$. On one hand, the relative difference in $\overline{W}^q$ can be quite large (it approaches 100 % as $\rho \to 0$). On the other hand, this difference should be small when ρ is close to 1 and waiting times in both systems are significant,

while when ρ is small, the waiting times in both systems are quite small and the large relative difference may not be of practical significance. Thus, as a rough approximation, $M/G/1$ system can be used in place of $M/G/\kappa$ when the expected waiting times are of primary interest.

However, when the primary measure of interest is the expected total time in the system $\overline{W}$, one has to be more careful. When the system is highly utilized, i.e., ρ is close to 1, the main determinant of $\overline{W}$ is the waiting time and the previous argument applies. However, when the system utilization is lower, the expected service time will play a large role. Since it is $1/\mu^0$ for $M/G/\kappa$ and $1/\mu = \kappa/\mu^0$ for $M/G/1$, the difference is quite large and approximation is no longer appropriate. Thus, with respect to $\overline{W}$, the approximation can only be justified in the heavy utilization case.

Turning our attention back to the $M/G/1$ system, we would like to rewrite (17.12) in terms of decision variables in our model. This is not difficult to do, and with a little algebraic manipulation we obtain the following expression for the expected waiting time at an open facility $i \in I$:

$$\overline{W}_i = \overline{W}_i^q + \frac{1}{\mu_i} = \frac{(1+\gamma^2)\Lambda_i}{2\mu_i(\mu_i - \Lambda_i)} + \frac{1}{\mu_i} \tag{17.15}$$

with Λ_i given by (17.6). We assume that $\overline{W}_i = 0$ if there is no facility at i.

Several comments are in order. First, we treat γ^2 as an intrinsic model parameter, rather than a decision variable, i.e., we assume that the coefficient of variation of service times is fixed in advance. While this is certainly the case when a specific distribution of service times is assumed (e.g., for $M/M/1$ queues $\gamma^2 = 1$), there is, in principle, no reason why this should not be a decision parameter in the system. For example, if the decision on how much capacity to install in facility i also deals with *what kind* of capacity to install, then the coefficient of variation γ could well be affected, as well as μ_i: service systems with higher level of automation may have lower γ, while more manual processes may have higher γ (of course the resulting values may be different at different facilities, so γ_i notation would have to be used). Another case where γ may be a decision variable is when customers at different nodes have different service time variabilities, in which case the allocation decisions x_{ij} may well influence the total demand Λ_i and the variability of service times γ_i as well as μ_i. Nevertheless, we are not aware of any SLCIS model that treats this parameter as a decision variable; in fact the value of the coefficient of variation is assumed to be identical at all facilities, which is reflected in our usage of γ without a subscript.

Second, observe that $\overline{W}_i$ (and $\overline{W}_i^q$) is decreasing in μ_i, increasing in Λ_i and convex with respect to both μ_i and Λ_i whenever system stability conditions (17.7) hold. These properties are exploited in many SLCIS models that follow.

Let $WC(w)$ represent the "waiting cost", i.e. the cost incurred by customers waiting w units of time (henceforth we assume that waits include service times, i.e. use measure W defined earlier; an equivalent treatment can be developed by focusing on waiting times in queue only, i.e. W^q). As with the travel costs, we

assume that $WC(w)$ is non-negative and non-decreasing, noting that many models make the simplifying assumption that the waiting cost is proportional to w. The total expected waiting cost in the system can now be expressed as

$$SWC = \sum_{j \in J} \sum_{i \in I} WC(\overline{W}_i) x_{ij}. \tag{17.16}$$

In view of non-linear dependence of the expected waiting time $\overline{W}_i$ on the decision variables, SWC is a non-linear function even when the waiting cost is assumed to be linear.

We note that since the waiting cost is only incurred by customers who are assigned to some facility, we should also add a penalty term for customers that are not assigned to any facility (i.e., not served)—otherwise the model may have an incentive to not assign customers even if service capacity is available. The "intentionally lost demand" customers may be represented in the revenue term described later (i.e., they are treated as an opportunity cost of lost revenue). Alternatively they can be represented by a term $p \sum_{j \in J} \left(1 - \sum_{i \in I} x_{ij}\right)$ which may be added to the SWC expression above, where p represents the penalty for choosing to not service a customer.

There are two potential issues with using (17.16) as the *sole* measure of service quality (in terms of waiting times) at the facilities. First, as with the system travel cost, a small value of SWC does not necessarily ensure that all customers are receiving adequate service—a small expected waiting time at one facility may "hide" a large expected waiting time at another. Thus, one may want to add the constraints (these are traditionally stated in terms of waiting time, rather than system time; we follow this tradition):

$$\overline{W}_i^q \leq EW, \quad i \in I, \tag{17.17}$$

where EW represents the acceptable maximum waiting time at any facility.

Second, the *expected* waiting time may not be sufficient to express the desired service quality; we may wish to ensure that most customers experience no waiting at all or that the probability of "long" waits is sufficiently low. For this we need to consider a constraint of the form

$$P(W_i^q > T) \leq \alpha_T, \quad i \in I, \tag{17.18}$$

where $P(\cdot)$ is the steady-state distribution of W_i^q, $T > 0$ is the specified threshold for the waiting times, and $\alpha_T \in (0, 1)$ is the maximum acceptable probability of waits longer than T at any facility. For example, α_0 represents the maximum acceptable proportion of customers that must wait for service at any facility.

Both (17.17) and (17.18) above are examples of *Service level Constraints* (SCs) that are quite common in SLCIS models. Since (17.17) refers to the expected behavior of the system, while (17.18) refers to the probability of occurrence of

certain (undesirable) events, we will refer to the former as the "Mean SC" and the latter as the "Probabilistic SC". While the Mean SC is easily expressed in terms of the decision variables by substituting (17.15) into (17.17), the Probabilistic SC requires an expression for the steady-state distribution of the waiting time, which is not generally available. One option is to make additional assumptions about the distribution of service times (e.g., assuming $M/M/1$ or $M/E_k/1$ queues at the facilities) since steady-state distributions of waiting times have been derived for many common systems. Another option is to use an approximation. The one we follow here is based on Baron et al. (2008). Assume that the service constraints (17.18) are specified and let

$$V(T, \alpha_T) = -\frac{ln(\alpha_T)}{T};$$

observe that since $ln(\alpha_T) < 0$, this is a positive constant that is decreasing in α_T and in T. Then (under certain mild technical assumptions), constraint (17.18) is satisfied whenever

$$G_S(\frac{V(T, \alpha_T)}{\mu_i})(\Lambda_i - 1) \leq V(T, \alpha_T), \tag{17.19}$$

where $G_S(\cdot)$ is the MGF of service times defined earlier. Recall that $G_S(\eta)$ is an increasing function for $\eta > 0$, implying that the left-hand side of (17.19) is decreasing in μ_i. This is quite intuitive: when T or α_T are decreased, the probabilistic SC becomes tighter, requiring more capacity at the facility. In fact, as $V(T, \alpha_T)$ becomes larger, satisfying (17.19) requires more capacity μ_i.

This leads to a general view of service constraints: for any arrival rate Λ_i at facility $i \in I$ one can define a minimum capacity level $\bar{\mu}(\Lambda_i)$ such that SC holds if and only if

$$\mu_i \geq \bar{\mu}(\Lambda_i), \tag{17.20}$$

where $\bar{\mu}(\Lambda_i)$ is computed (perhaps numerically) from (17.17), (17.18), or (17.19). Of course, an equivalent view is to specify a function $\bar{\Lambda}(\mu)$, which is just an inverse of $\bar{\mu}(\Lambda)$, so that SC holds whenever

$$\Lambda_i \leq \bar{\Lambda}(\mu_i), \tag{17.21}$$

i.e., for a given capacity level μ_i there is a maximal arrival rate $\bar{\Lambda}(\mu_i)$ for which an adequate service level can be provided by facility i. This view extends to other definitions of SCs (e.g., instead of using waiting time one could use L or another service level measure)—the only thing that changes is the way functions $\bar{\mu}(\Lambda)$ and $\bar{\Lambda}(\mu)$ are computed.

We note that system stability conditions imply that $\bar{\mu}(\Lambda) > \Lambda$ (equivalently $\bar{\Lambda}(\mu) < \mu$) and the difference $\bar{\mu}(\Lambda) - \Lambda$ may be interpreted as the amount of the "capacity cushion" (capacity in excess of the minimal possible level) needed

to ensure adequate service given the arrival rate Λ. For many systems and many specifications of service level constraints it has been shown that this amount grows proportionately to $\sqrt{\Lambda}$, i.e.

$$\bar{\mu}(\Lambda) \approx \Lambda + Q\sqrt{\Lambda} \tag{17.22}$$

for some constant Q (see, e.g., the discussion in Castillo et al. 2009). The derivations in Whitt (1992) suggest that, under many conditions, a good interpretation for Q is provided by

$$\sqrt{2}Q \approx \sqrt{\gamma^2 + 1}P(W > 0),$$

where γ is the coefficient of variation of arrivals. Thus, $\sqrt{2}Q/\sqrt{\gamma^2 + 1}$ is approximately equal to the probability of waiting, a natural service level measure. To summarize, when the probability of waiting is used as the service-level measure, the constraint

$$P(W_i > 0) \leq \alpha_0, \; i \in I$$

holds if

$$\mu_i \geq \bar{\mu}(\Lambda_i) \approx \Lambda_i + \left[\sqrt{\frac{\gamma^2 + 1}{2}}\alpha_0\right]\sqrt{\Lambda_i}, \; i \in I. \tag{17.23}$$

Similar expressions can be derived with for service level measures where the threshold for waiting time is set above 0.

As noted earlier, incidence of long waits can be controlled through service level constraints and/or explicit waiting cost terms in the objective function. While, in principle, both can be used in the same SLCIS model, it is far more common to use one or the other. In models where only service level constraints are used, these constraints will be tight in an optimal solution (since capacity is costly). If, in addition, the demand is assumed to be inelastic, Λ_i is a linear function of the decision variables x_{ij}. In this case a significant simplification is achieved by using the previous expression: setting the SC as an equality, we can eliminate decision variables μ_i from the model, replacing them with the right-hand side of (17.23).

17.2.3.3 Facility Costs

We assume that the decision to open a facility at $i \in I$ incurs two types of costs: the *fixed cost* FC_i, which depends on the characteristics of the location i, and the *variable cost* $VC(\mu_i)$, which depends on the amount of capacity μ_i allocated to the facility. The function $VC(\mu)$ is assumed to be non-decreasing and non-negative with $VC(0) = 0$; concavity of $VC(\mu)$ is a frequently made assumption, reflecting

economies of scale. With these definitions, the System Facility Cost is defined as follows:

$$SFC = \sum_{i \in I} FC_i y_i + \sum_{i \in I} VC(\mu_i) \tag{17.24}$$

17.2.3.4 Revenues and Overall Objectives

We assume that each customer that is served brings in a revenue r to the system (for public service applications, we can treat r as a "system benefit" parameter). The total expected revenue can be expressed as

$$SR = r \sum_{i \in I} \Lambda_i = r \sum_{j \in J} \lambda_j \sum_{i \in I} x_{ij}. \tag{17.25}$$

In principle, the parameter r can be treated as a decision variable—the price charged by the decision-maker for service. However, in the vast majority of SLCIS literature this term is treated as an exogenous parameter (Tong 2011 and Berman et al. 2014 being the exceptions). Since treating prices as decision variables introduces significant new complications, we will generally treat r as constant in the model.

We also observe that when demand is inelastic (i.e., $\lambda_j = \lambda_j^{\max}$ for all $j \in J$) and when the constraints require that all customers must be served (i.e., $\sum_{i \in I} x_{ij} = 1, j \in J$), it is easy to see that $SR = r \sum_{j \in J} \lambda_j^{\max}$, which is a constant. In this case, the revenue term in the objective can be dropped, leading to a pure cost minimization case. Even in models where some customers may not be served, but the demand is inelastic, it is common to use cost minimization with a penalty term, which can be interpreted as opportunity cost for unserved customers.

To summarize, the overall objective for a general SLCIC model is given by

$$\text{maximize}\,[SR - STC - SWC - SFC]\,,$$

where the respective components are defined by (17.25), (17.10), (17.16), and (17.24). We note that in most specific models described in the literature, only a subset of the terms above is present, the rest being implicitly controlled by constraints (e.g., in the presence of service level constraints, the *SWC* term is often dropped).

Most of the terms above depend on demand allocations x_{ij} and demand rates λ_j, which have not yet been described. This is the subject of the following section.

17.3 Customer Response: Demand Levels and Allocations

In this section we discuss the two remaining key issues in SLCIS models: the mechanism determining the allocation of customer demand to facilities, represented by x_{ij} variables, and the amount of demand λ_j generated by customers at $j \in J$.

In location modeling two approaches for allocating customer demand to facilities are generally considered: *directed choice*, where the same decision-maker determining the number and locations of the facilities also has the power to assign customers to the facilities in a way that will optimize the model objective, and *user choice* where customers self-assign to facilities based on maximization of their own utility functions, which may not be aligned with the overall model objective. For example, a common customer utility function is the travel distance. Thus, in a user choice environment, each customer will select the closest facility, while in the directed choice case a customer may be assigned to a further facility even when a closer one is open (if such assignment reduces the overall facility cost).

The same framework can be applied to the SLCIS models. However it may be more useful to also classify the models in terms of the assumed customer reaction. We differentiate four classes of models:

Type NR: Models with no customer reaction: customers do not control the demand allocations and the demand rates are fixed (directed choice with inelastic demand)

Type AR: Models with allocation-only reaction: customers select utility-maximizing facilities, but the demand rates are fixed (user choice with inelastic demand)

Type DR: Models with demand rate-only reaction: customer do not control the demand allocations but do determine the demand rates (directed choice with elastic demand)

Type FR: Models with full customer reaction: customers control both, the allocation of demand (by selecting the utility-maximizing facilities) and the demand rates (user choice with elastic demand).

This classification is summarized in Table 17.1.

The *NR models* correspond to the standard directed choice assumptions in the literature: the values of the assignment variables x_{ij} are entirely controlled by the decision-maker and must only satisfy the basic constraints (17.3)–(17.5). One may also interpret such models as describing a "social optimum" (also known as "first best solution" in economics)—the customers will accept whatever assignments are needed to optimize the overall system objective, even if that means that some of

Table 17.1 Model classification by customer response

	Demand allocation	
	Decision-maker	Customer
Inelastic demand	NR	AR
Elastic demand	DR	FR

them may have to travel to more distant and more congested facilities than the ones available in their immediate neighborhood. On the other hand, since the objective function combines the costs borne by the decision-maker (facility costs *SFC*) with those borne by the customers (travel cost *STC* and waiting cost *SWC*), the interests of both parties should be "balanced" in the solution. Customer demand is assumed to be inelastic, with $\lambda_j = \lambda_j^{\max}$ for all $j \in J$. Since customer utility has no effect in this model, there is no need to define it. We note that x_{ij} are usually assumed to be binary in NR models (though it is easy to construct examples showing that higher objective values may be possible with fractional assignments). This is due to the concern that enforcing fractional demand allocations is likely impractical in most contexts. Thus, in NR models only the "minimal" constraints (17.3)–(17.5) need to be imposed on demand allocations: the decision-maker is free to choose any allocation that satisfies these constraints.

The other three model types assume some form of customer reaction in the form of utility-maximizing behavior. The description of the utility mechanism is provided next.

17.3.1 Customer Utility Functions

Recall that u_j is the utility derived by customer $j \in J$ from the service provided by the facilities. Note that there are two costs borne by the customer: travel and waiting. Suppose a customer experiences travel distance d (as before we assume that distances have been redefined to represent travel costs) and expected system waiting time w. Let the utility $U(d, w)$ be a non-increasing function of d and w. To relate u_j to $U(d, w)$ we assume that the total utility derived by customer j is only affected by the facilities this customer actually visits, letting

$$u_j = \sum_{i \in I} U(d(i, j), \overline{W}_i) x_{ij}, \tag{17.26}$$

Note that this definition remains valid even when the single-sourcing assumption is relaxed. In this case, x_{ij} represents the proportion of time facility i is used by customer j and u_j can be interpreted as the resulting *expected* utility. Observe also that if a customer does not receive service from any facility, $x_{ij} = 0$ for all $i \in I$ and $u_j = 0$.

Perhaps the most natural specification for the utility function $U(d, w)$ is the linear form

$$U^L(d, w) = -(\tau_d d + \tau_w w), \tag{17.27}$$

where $\tau_d, \tau_w > 0$ are the relative weights on travel distance and waiting time, respectively. When $\tau_w = 1$, the parameter τ_d can be interpreted as the average travel speed, so that $\tau_d d$ is the average travel time, and the right-hand side of (17.27)

represents the negative of the total expected time spent by the customer in the system (until the end of service).

There are two other common specifications of $U(d, w)$. The simpler one is

$$U^D(d, w) = -\tau_d d, \tag{17.28}$$

i.e., customer's utility is simply proportional to the traveling distance (representing the travel cost) and is independent of the waiting time. This is a very popular specification form appearing (often implicitly) in numerous SLCIS models. While the lack of dependence on w may seem counterintuitive, it is usually justified by assuming that customers do not have advance knowledge of waiting times at the facilities and thus must make their decisions based on travel times only. This justification is not entirely convincing sine in a steady-state system some learning about expected waiting times should, presumably, occur. Alternative justification is that the waiting costs are dominated by the travel costs. Perhaps more importantly, as will be seen below, specification (17.28) avoids many technical complications that occur when a more general utility structure is used and can thus be treated as an approximation.

Another natural specification is the log-linear form

$$U^E(d, w) = exp(-\tau_d d - \tau_w w), \tag{17.29}$$

which is quite similar to (17.27) with the advantage of the utility being non-negative, convex and bounded by 1. Note that $U^E(d, w) = 1$ when $d = w = 0$, i.e., when the customer incurs neither travel nor waiting cost, and $U^E(d, w) \to 0$ as $d, w \to \infty$. This makes it convenient to interpret $U^E(d, w)$ as *the proportion of maximum available demand realized from customer j if this customer is faced with travel distance d and expected wait w*. This interpretation will be useful when describing elastic demand models below.

Finally, we note that a utility function can be defined in terms of service measures other than the expected waiting time $\overline{W}$ — one can use the probability of waiting $P(W^q > 0)$, or any other performance measure of the queuing system operated at the facilities. The specifications of the utility can also be generalized to incorporate other decision variables, such as the price charged by the facility operator for service (see Berman et al. 2014 for an example).

17.3.2 SLCIS Models with Customer Reaction

Once a utility function is specified, it should be possible to specify the customer reaction as well. At a first glance, this seems fairly straightforward: a SLCIS model with customer reaction can be viewed as a bi-level game, where the decision-maker first specifies the number, locations and capacities of the facilities (i.e., values of m, y_i and μ_i for $i \in I$) and then each customer selects a utility-maximizing strategy. Unfortunately, as we will see shortly, complications quickly arise. This has to do,

primarily, with the fact that customer utility is a function of the waiting time $\overline{W}_i$, which is not directly controlled by the decision-maker, but rather arises as a result of joint actions of the decision-maker and *all* customers: the former determines facility locations and capacities μ_i, while the latter determine the demand rates Λ_i. This gives rise to traffic equilibrium conditions, where the actions of one customer (adjusting their demand rate λ_j and/or demand allocation x_{ij}) change the waiting times at the facilities and thus affect the utilities of all other customers. Thus, not only is there a bi-level game being played between the decision-maker and the customers, but there is also a simultaneous non-cooperative game being played between the customers themselves. Moreover, the response functions in the latter are rather complicated, which may lead to lack of equilibria (if customers are restricted to simple strategies), or to multiple equilibria, not to mention serious difficulties in computing these equilibria. We discuss these issues briefly below, referring the interested reader to more general references on spatial equilibria like Nagurney (1999).

17.3.2.1 AR: Models with Allocation-Only Reaction

In this type of models, it is assumed that the demand rate of each customer node is fixed a priori, with $\lambda_j = \lambda_j^{\max}$ for all $j \in J$. However, the customers determine their demand allocations, i.e., the values of x_{ij} variables, in a utility-maximizing fashion. For concreteness, we will assume the linear specification of the utility function $U^L(d, w)$ given by (17.27), though much of the discussion extends to alternative specifications as well.

We first consider the original "single-sourcing" assumption. Since the customer will allocate all of their demand to a utility-maximizing facility, $x_{ij} = 1$ implies that

$$U^L\left(d(i,j), \overline{W}_i\right) \geq U^L\left(d(k,j), \overline{W}_k\right) \quad \text{for all } k \in I \text{ with } y_k = 1,$$

which, assuming for simplicity that $\tau_w = \tau_d = 1$ in (17.27), is equivalent to

$$d(i,j) + \overline{W}_i \leq d(k,j) + \overline{W}_k \quad \text{if } y_k = 1, k \in I.$$

Recalling that Λ_i is given by (17.6) and $\overline{W}_i$ by (17.15), this leads to the following equilibrium conditions that must be satisfied by allocations x_{ij}:

$$d(i,j) + \overline{W}_i \leq [d(k,j) + \overline{W}_k] y_k + M(1 - x_{ij}), \quad i,k \in I, j \in J \tag{17.30}$$

$$\overline{W}_i = \frac{(1+\gamma^2)\Lambda_i}{2\mu_i(\mu_i - \Lambda_i)} + \frac{y_i}{\mu_i + M(1-y_i)}, \quad i \in I \tag{17.31}$$

$$\Lambda_i = \sum_{j \in J} \lambda_j^{\max} x_{ij}, \quad j \in J \tag{17.32}$$

$$\sum_{i \in I} x_{ij} \leq 1, \qquad j \in J \tag{17.33}$$

$$x_{ij} \leq y_i, \qquad i \in I, j \in J \tag{17.34}$$

$$x_{ij} \in \{0, 1\}, \qquad i \in I, j \in J \tag{17.35}$$

where M is a suitably large constant. We assume that some finite limit can be imposed on the expected waiting time $\overline{W}_i$ at any facility and that $M \geq d(i, j) + \overline{W}_i$ for all j and i.

Of course a trivial solution to this system is to have $x_{ij} = 0$ for $j \in J, i \in I$ (which also implies $\overline{W}_i = 0$ for all $i \in I$), i.e., to have complete loss of all customer demand. Clearly, we are interested in non-trivial solutions where at least some customers choose to obtain service. On the other hand, the system may not have enough capacity to serve all customers. We therefore make the following definition.

Definition 17.1 A subset of customer nodes $J' \subset J$ is **serviceable** if

$$\sum_{j \in J'} \lambda_j^{\max} \leq \sum_{i \in I} \mu_i .$$

A subset J' is **fully served** if $\sum_{i \in I} x_{ij} = 1$ for all $j \in J'$, i.e. if (17.33) holds as equality for all $j \in J'$.

This definition simply assures that there is sufficient capacity to serve any serviceable subset. We are interested solutions where at least some serviceable subsets of J are fully served. Unfortunately, the system (17.30)–(17.35) may have no such solutions.

Example 17.1 Consider a network with one customer node j and two facility nodes 0, 1 both of which contain facilities, i.e., $y_0 = y_1 = 1$. Assume further that $\mu_0 = \mu_1 > \lambda_j^{\max}$, and thus $J = \{j\}$ is serviceable. Assume $d(j, 0) = d(j, 1)$. Then, since $W_i = 0$ if $x_{ij} = 0$ and $W_i > 0$ when $x_{ij} = 1$ for $i = 0, 1$, there is no feasible solution to the system (17.30)–(17.35). Indeed, if customers at j select facility i, it creates non-zero waiting time at that facility, making the other facility a utility-maximizing choice. Other similar examples of non-existence of equilibria with binary allocation vectors are easy to construct.

The underlying reason for the phenomena illustrated above is that single-sourcing strategies create discontinuities (a facility receives either all of customer's demand, or none of it), while the existence of equilibria typically requires continuity of the underlying functions. Indeed, intuitively it is clear that in the previous example equilibrium allocations are achieved if the customers at j visit each facility with equal frequency. This, of course, requires the relaxation of the single-sourcing assumption, allowing x_{ij} to take on fractional values, which are interpreted as visit frequencies. In addition to replacing (17.35) with its linear relaxation, the equilibrium-defining inequality (17.30) has to be adjusted as follows.

Recall the definition of u_j given by (17.26), which is now interpreted as the expected utility for customers at $j \in J$ given a fractional allocations vector $x_{ij}, j \in J, i \in I$ (we emphasize that the waiting times are affected by the allocations of all customers, not just the ones at j). We seek allocations under which no customer can improve their utility by making unilateral changes. It follows that the equilibrium utilities $u_j^*, j \in J$ must satisfy

$$d(i,j) + \overline{W}_i \begin{cases} = -u_j^* \text{ if } x_{ij} > 0; \\ \geq -u_j^* \text{ if } x_{ij} = 0 \end{cases}$$

(recall that we are assuming linear utilities which are equal to the negative of total travel and waiting times). These conditions can be represented by replacing (17.30) with the following non-linear complementarity conditions:

$$d(i,j) + \overline{W}_i \geq v_j, \ j \in J, i \in I \tag{17.36}$$

$$(d(i,j) + \overline{W}_i - v_j)x_{ij} = 0, \ j \in J, i \in I \tag{17.37}$$

$$v_j \geq 0, \ j \in J \tag{17.38}$$

where $v_j = -u_j^*$, representing the equilibrium "disutility" for customers at $j \in J$, is included in the model as a new decision variable. We will refer to a solution of the system (17.31)–17.38) as *Customer Flow Equilibrium*.

The following result follows directly from Theorem 5.4 of Ashtiani and Magnanti (1981) by continuity of $U\left(d(i,j), \overline{W}_i(\boldsymbol{x})\right)$ for all $j \in J, i \in I$, where $\boldsymbol{x}$ is a fractional allocation vector with components x_{ij}.

Theorem 17.1 *For any values of $y_i \in \{0,1\}$ and $\mu_i \geq 0$ such that $\mu_i \leq My_i$, if a subset $J' \subset J$ is serviceable, then there exists at least one customer flow equilibrium $x_{ij}, j \in J, i \in I$ under which J' is fully served.*

In particular, if the system has the capacity to service all of customer demand, i.e., J is serviceable, at least one customer flow equilibrium must exist under which all customers are served.

The discussion and the result above is quite general: in particular, it extends to models with elastic demand (i.e., models of type FR discussed below). Additionally, in place of the expected waiting time for an $M/G/1$ queue, a general measure of "congestion" can be used with the only requirements that it is strictly increasing, twice differentiable, non-negative and convex (recall that all capacity decisions are considered to be fixed in this section). These requirements are clearly satisfied by most performance measures for queueing systems, including multi-server and limited-buffer queues. We refer the reader to Brandeau et al. (1995) for a discussion of these more general settings.

It is important to realize that the customer flow equilibrium may not be unique. In fact, there may be multiple allocation vectors satisfying the equilibrium conditions for a particular fully served subset of customer nodes. For an example, consider

adding a second identical customer node j' to the system in Example 1. Now, if customers at both nodes are assigned to different facilities: $x_{ij} = 1, x_{(1-i)j} = 0$, $x_{ij'} = 0, x_{(1-i)j'} = 1$ for $j = 0, 1$, we have two different equilibria. In fact, there may be infinitely many equilibria: any assignment satisfying

$$x_{ij} = \alpha, x_{(1-i)j} = 1 - \alpha, x_{ij'} = 1 - \alpha, x_{ij'} = \alpha, \quad \alpha \in [0, 1]$$

is also an equilibrium. In principle, different equilibrium allocation vectors may lead to different values of the objective function in the underlying SLCIS model, creating uncertainty as to which solution will actually arise. However, all equilibria are "similar" in certain key aspects, as shown in the following theorem based on the result provided in Brandeau and Chiu (1994):

Theorem 17.2 *For any two customer flow equilibria under which a subset $J' \subset J$ is fully served, the values of $\Lambda_i \quad i \in I$ (total demand seen at each facility) and $v_j, \ j \in J$ (equilibrium disutility of each customer node) are the same.*

This theorem implies that, under a sensible specification of the objective function, where the total travel and waiting cost for each customer node is a function of v_j, all equilibria will give rise to the same values of the objective.

While the previous results show that AR models with multi-sourcing demand allocations are well-posed, there is an important issue concerning computational tractability of system (17.31)–(17.38). Even for fixed facility locations and capacities, solving the customer flow equilibrium conditions is far from easy. While certain numerical approaches (described in Nagurney 1999) do exist, they are computationally heavy even for moderate-size problems (see Tong 2011). Often, to get reasonable algorithmic efficiency one has to make simplifying assumptions about the system, e.g., assuming $M/M/1$ queues simplifies (17.31), making the system much more solvable—see Zhang et al. (2010) who were able to compute equilibria for a system with $|J| \approx 500$ and $|I| \approx 40$ (note that their model also had elastic demands, which likely increased computational complexity). Keeping in mind that computing customer flow equilibrium is only a subproblem of an SLCIS model, embedding this computation in an overall exact optimization procedure is nearly impossible. Hence both of the papers cited above resort to search heuristics for the upper level (location and capacity allocation decisions).

In view of the difficulties involved in using the customer flow equilibrium approach above, it is natural to think of model simplifications. We mention three such approaches. One is to keep the single-sourcing assumption in spite of the possible non-existence of equilibria (see Zhang et al. 2009). The reason this may be reasonable is that, as mentioned earlier, nonexistence is a result of discontinuity—when re-assignment of a single customer alters the waiting times at the facility for the remaining customers. It is reasonable to assume that for realistic problem instances, this should not be an issue: as the number of customers and customer nodes grows, no single assignment should exert a significant impact on waiting times at the facilities. Thus, asymptotically, single-sourcing equilibria should emerge. Indeed, Zhang et al. (2009) did not report issues with nonexistence of

equilibria when solving realistic-size problem instances for mammography clinics in Montreal, Canada. The obvious advantage of the single-sourcing approach is that the system (17.30)–(17.35) is much easier to solve and can be embedded as part of constraints in a larger SLCIS model.

The second approach is to use distance-only utilities $U^D(d)$ given by (17.28). Since these are independent of waiting times, the existence of customer flow equilibria is no longer an issue; utility-maximizing behavior by customers merely implies that once facility locations are specified, each customer travels to the closest facility, replacing (17.30) with

$$d(i,j) \le d(k,j)y_k + M(1-x_{ij}), \quad i,k \in I, j \in J, \tag{17.39}$$

which leads to significant simplifications (obviously, single-sourcing assumption can be retained here as well).

Yet another alternative to customer flow equilibrium is to use market share allocation approach, as discussed in Sect. 17.3.2.4 below.

17.3.2.2 DR: Models with Demand-Only Reaction

In this model class, the decision-maker has the control of the demand allocation vector $\boldsymbol{x}$, however, the demand $\lambda_j = \lambda(u_j)$ for customer node $j \in J$ is assumed to be a function of the utility u_j realized by customers at j. Following Brandeau et al. (1995) we assume that

$$\lambda_j = \lambda_j^{\max} h(u_j),$$

where, as defined earlier, $\lambda_j^{\max}$ is the maximum possible demand rate at node j and $h(u) \in [0,1]$ is a strictly decreasing, twice differentiable function with $h(0) = 1$ and $h(u) \to 0$ as $u \to u_j^{min}$, where u_j^{min} is the lower bound on the utility for customers at j (e.g., if utilities are scaled to be non-negative, then we can set $u_j^{min} = 0$). Thus, $h(u_j)$ can be interpreted as the percentage of the maximum available demand at j that is "captured" by the system; it is often called the "participation rate".

Recall that by (17.26), the utility u_j is a function of the waiting time and travel distance faced by customers at j. As in the case of *NR* models, we will assume that x_{ij} is binary, motivated by the same considerations as before: when customer demand allocations are dictated by the decision-maker, rather than by an equilibrium condition of the previous section, enforcing fractional assignments is typically unrealistic. Thus, assuming all customers at j will be served (as will be shown below, this assumption holds automatically in DR models), $x_{ij} = 1$ for exactly one $i = i(j) \in I$. Then, we have

$$\lambda_j(d(i(j),j), \overline{W}_{i(j)}) = \lambda_j^{\max} h(U(d(i(j),j), \overline{W}_{i(j)})), \quad j \in J. \tag{17.40}$$

One example of a functional form of h that satisfies the required assumptions is the exponential utility U^E given by (17.29), leading to the popular "exponential decay" demand specification:

$$\lambda_j(d(i(j), j), \overline{W}_{i(j)}) = \lambda_j^{\max} \exp(-\tau_d d(i(j), j) - \tau_w \overline{W}_{i(j)}), \quad j \in J. \quad (17.41)$$

While this expression is assumed in several published DR models, most of the results below apply to more general functional forms as well. Observe that (17.40) implicitly defines an equilibrium condition: the left-hand side depends on the waiting time $\overline{W}_{i(j)}$ at facility $i(j)$, which is a function of demand $\Lambda_{i(j)} = \sum_{j \in J} \lambda_j x_{i(j),j}$ seen by this facility. Thus, (17.40) should be seen as a system of $|J|$ equations that must be solved to yield the actual demand rates; this system decouples into subsystems consisting of all customers $j \in J$ with $i(j) = i$ for each facility i with $y_i = 1$. Thus, even though the allocation variables x_{ij} are fixed (or, rather, set by the decision-maker) for DR models, the issues related to existence and uniqueness of equilibria must be dealt with. The following result is based on Berman et al. (2014), where it is established for the case where price r is also a decision variable.

Theorem 17.3 *For any given facility location, capacity, and demand allocations y_i, μ_i, x_{ij} for $i \in I, j \in J$, there exist unique equilibrium arrival rates $\lambda_j(d(i(j), j), \overline{W}_{i(j)})$ and waiting times $\overline{W}_i$.*

Note that, unlike the case for AR models, this result holds with binary demand allocations x_{ij} (it obviously extends to the fractional allocations as well). As illustrated in Aboolian et al. (2012), as well as in Berman and Kaplan (1987), computation of the equilibrium demand is relatively simple in this case, based on the fixed-point iteration approach.

An interesting feature of the DR model is that it is self-regulating: as waiting times become longer at the facilities, customer demand is automatically reduced. Thus, the system stability is assured by (17.40) without the need for explicit constraints (17.7). Moreover, even though customer assignments are "dictated" by the decision-maker through the specification of x_{ij}, assigning customer j to a more distant or more congested facility leads to lower demand λ_j, with the resulting loss of revenue. Thus, the model assures that customer assignments must take customer utilities into consideration, while avoiding the complexities of full traffic equilibrium treatment. In fact, Aboolian et al. (2012) report (based on computational experiments) that optimal solutions where some customers are *not* assigned to their utility-maximizing facility are quite rare, though they do occur.

The behavior of DR model involves an interesting feedback loop: as the service offered by the facilities is improved (by locating the facilities closer to customer nodes, or allocating more capacities to the facilities), the customers respond by generating more demand (positive feedback), which leads to increased congestion at the facilities, leading to reduced demand (negative feedback). Thus one could legitimately ask whether models with elastic demand may lead to counter-intuitive

results where service improvements result in a net loss of demand. Fortunately, this is not the case as shown in the following result from Berman et al. (2014):

Theorem 17.4 *For $j \in J$, let $\lambda_j(d_j, w_j)$ be the equilibrium demand rate when the travel time is d_j and the expected waiting time is w_j. Then λ_j is non-increasing in d_j and w_j (strictly decreasing when the utility function is strictly decreasing in the corresponding parameter).*

Thus, with a reasonably behaved utility function, when the service offered to customers at $j \in J$ is improved in terms of either travel distance or waiting time, or both, the demand rate increases, leading to higher revenue for the decision-maker (for this customer node). Since nodes that are currently not served (i.e., with $\sum_i x_{ij} = 0$) can be treated as having the travel distance that is so high that the demand rate is negligibly close to 0, the decision to serve these nodes by assigning them to *any* open facility can be treated as reducing the travel distance. This leads to the following result:

Corollary 17.1 *In the elastic demand case, there exists an optimal solution to SLCIS where every demand node is served.*

17.3.2.3 FR: Full Response Models

In this model class, the customer response to facility location and capacity allocation decisions includes both the level and the allocation of demand. Thus, the equilibrium values of x_{ij} and λ_j are described by a system that includes flow equilibrium conditions (17.36)–(17.38), as well as the elastic demand equilibrium (17.40). The existence and uniqueness of equilibria are assured by the following corollary:

Corollary 17.2 *The equilibrium existence and uniqueness results of Theorems 17.1 and 17.2 extend to the FR model class.*

The reader can refer to Brandeau et al. (1995) for further details; note that the uniqueness result has the same limitations as for the AR models (i.e., uniqueness can only be guaranteed with respect to the values of the objective, provided the objective function is suitably defined). Also, just as in AR models, this corollary requires fractional allocation vectors x_{ij}.

The computation of equilibrium solutions presents even more challenges than for AR models. This has lead to an alternative specification of demand allocation vectors described in the following section.

17.3.2.4 FR and AR Models with Proportional Allocations: Market Share Models

Our development of AR and FR models was based on the assumption that customers allocate their demand in a utility-maximizing fashion. As we have seen, this

assumption leads to flow equilibrium-type conditions with the ensuing structural and computational difficulties. An alternative approach is based on the assumption that customers allocate their demand among many (possibly all) facilities in proportion to the utility derived from these facilities. Essentially, each customer node $j \in J$ is viewed as a "market" with facilities competing for the shares of this market. These models, that are axiomatically rooted in the stochastic utility theory, have generated a large body of literature, particularly in economics and marketing; in the latter they are accepted as the dominant model for customer choice in the presence of many substitutable alternatives (e.g., predicting market share of a particular brand when many other brands are available).

In the competitive location literature these models have appeared under many names, including "competitive interaction models", "Huff-type models", "gravity models", "multinomial logit models", "market-share models". While there are minor specification differences between these, the basic structure remains the same; we refer the reader to the recent review by Berman et al. (2009a).

Since SLCIS models of AR and FR type can be regarded as bi-level games played between the decision-maker and the customers, proportional allocation mechanism can be applied to the SLCIS context as well (in effect, it specifies the solution to the non-cooperative game played between customers once the decision-maker's strategy is specified). The specification is quite simple: for customers at $j \in J$ and (open) facility at $i \in I$, the demand allocation is given by

$$x_{ij} = \frac{U(d(i,j), \overline{W}_i) y_i}{\sum_{k \in I} U(d(k,j), \overline{W}_k) y_k}, \tag{17.42}$$

where the numerator represents the utility derived from facility i and the denominator is the total utility derived by customers at j from all open facilities. Note that if there are any pre-existing competitive facilities that may attract customer demand, they should be included as an extra sum $\sum_{k \in C} U(d(k,j), \overline{W}_k)$ in the denominator, where C is the set of competitive facilities. To simplify the exposition, we will assume no competitive facilities in the remainder of the current section.

This specification implies that the demand allocations are fractional, and the demand rate from j attracted by facility i is (as before) $\lambda_j x_{ij}$, where λ_j is elastic for FR models and inelastic in AR case.

Note that from Eq. (17.42) it follows that market shares add up to 1, i.e., all available demand from j is served. This may be unrealistic if none of the available facilities provide good service to j. The easy modification is to introduce a "dummy" facility 0, representing "no service", and letting $U(d(0,j), \overline{W}_0) = u_{j0}$— a constant representing the utility value of not getting served (e.g., the customer may choose to consume a different product). The popular Multinomial Logit (MNL) specification (McFadden 1974) employs exponential utilities, leading to

$$x_{ij} = \frac{\exp(-\tau_d d(i,j) - \tau_w \overline{W}_i) y_i}{\sum_{k \in I} \exp(-\tau_d d(k,j) - \tau_w \overline{W}_k) y_k}, \tag{17.43}$$

where weights τ_d, τ_w can be estimated from the available consumer demand allocation data using the MNL methodology.

The advantage of the proportional allocation approach is that the values of x_{ij} are directly computable from (17.42) or (17.43) without having to solve the cumbersome flow equilibrium equations. Nevertheless, it is important to recognize that an equilibrium condition is implicit in the definition above, even in case of models with inelastic demand: the expressions for x_{ij} above are functions of waiting times $\overline{W}_i$, which, in turn, are functions of x_{ij}. Thus, (17.42) together with waiting time specification (17.15) and facility-level demand specification (17.6) form a system of non-linear equations. A solution to this system represents an equilibrium demand allocations and waiting times. In case of FR models, one also has to add the elastic demand specification (17.40) and the equilibrium solution includes the demand rates at each customer node. Thus, the issues of existence and uniqueness of the equilibrium must be addressed. These were examined in some detail by Lee and Cohen (1985). The existence follows directly from standard fixed-point results and the continuity of x_{ij} in (17.42) and is based on Theorem 1 in Lee and Cohen (1985):

Theorem 17.5 *There exists an equilibrium solution* $(x_{ij}, \overline{W}_i, \lambda_j), i \in I, j \in J$ *to the proportional allocation model.*

Lee and Cohen (1985) also examine uniqueness and stability of equilibria, where stability refers to whether a system where customers start with some arbitrary demand allocations, evaluate their utilities and then re-allocate according to (17.42) will naturally reach an equilibrium. They derive sufficient conditions for both uniqueness and stability. In the context of our AR and FR models, their results imply the following:

Theorem 17.6

1. *For AR models with proportional allocation the equilibrium is unique and stable*
2. *For FR models with proportional allocation the equilibrium is unique and stable if*

$$1 \geq \frac{u_j}{\lambda_j} \frac{\partial \lambda_j}{\partial u_j}, \text{ for all } j \in J$$

where $u_j = \sum_{i \in I} U(d(i,j), \overline{W}_i) y_i$ *is the utility derived by customers at* j *from all open facilities.*

The condition in part (2) above states that the elasticity of demand from node j with respect to the utility provided by all facilities must not exceed 1. As shown in Lee and Cohen (1985) this holds automatically when the demands are given by (17.41), as well as by many other common specifications of demand (we note that weaker, but harder to verify, sufficient conditions are also provided in Lee and Cohen 1985).

We close this section by noting that the analysis in Lee and Cohen (1985) assumes that all location and capacity allocation decisions have already been made.

To the best of our knowledge, no papers on SLCIS models of FR class with proportional demand allocation are available, though there are several publications on AR models (i.e., where demand is inelastic) with proportional allocation. These will be further discussed in Sect. 17.5 below.

17.4 General SLCIS Model Specification

In this section we summarize the discussion in the preceding sections. Putting all the modeling components together allows us to provide the following formulation for the General SLCIS with M/G/1 queues at facilities:

$$\text{maximize } Z =$$

$$r \sum_{j \in J} \lambda_j \sum_{i \in I} x_{ij} \tag{17.44}$$

$$- \sum_{j \in J} \sum_{i \in I} \beta d(i,j) \lambda_j x_{ij} \tag{17.45}$$

$$- \sum_{j \in J} \sum_{i \in I} WC(\overline{W}_i) x_{ij} \tag{17.46}$$

$$- \sum_{i \in I} FC_i y_i - \sum_{i \in I} VC(\mu_i) \tag{17.47}$$

$$\text{subject to } \overline{W}_i = \frac{(1+\gamma^2)\Lambda_i}{2\mu_i(\mu_i - \Lambda_i)} + \frac{y_i}{\mu_i + M(1-y_i)}, \qquad i \in I \tag{17.48}$$

$$[\ \lambda_j \text{ specification for DR and FR models }] \tag{17.49}$$

$$[\ x_{ij} \text{ specification for AR and FR models }] \tag{17.50}$$

$$[\ \text{Coverage Constraints }] \tag{17.51}$$

$$[\ \text{SC Constraints }] \tag{17.52}$$

$$\sum_{i \in I} y_i \leq m \tag{17.53}$$

$$\Lambda_i = \sum_{j \in J} \lambda_j x_{ij}, \qquad i \in I \tag{17.54}$$

$$\sum_{i \in I} x_{ij} \leq 1, \qquad j \in J \tag{17.55}$$

$$x_{ij} \leq y_i, \qquad i \in I, j \in J \tag{17.56}$$

$$\mu_i \geq \Lambda_i, \qquad i \in I, j \in J \quad (17.57)$$

$$x_{ij} \geq 0;\ \mu_i \geq 0;\ y_i \in \{0,1\}; \textit{integer}, \qquad i \in I, j \in J. \quad (17.58)$$

The objective function (17.44)–(17.47) represents the total profit which includes the revenue, travel, congestion, and facility fixed and capacity costs, respectively. Constraints (17.48) define the expected waiting time for M/G/1 queues. These can be substituted with constraints defining other relevant congestion measures, different queueing mechanisms or both. Specifications (17.49) are only relevant for elastic demand models of type DR and FR type; when the demand rate is assumed to be inelastic, one should omit these and set $\lambda_j = \lambda_j^{\max}$. Similarly, specifications (17.50) are only relevant for user-choice models of AR and FR type. Constraints (17.53)–(17.57) enforce the basic interconnections between the decisions variables and are typically present in some form in all models.

To the best of our knowledge, no published work contains all components listed in the general formulation above. The specific SLCIS models considered in the literature typically include only some of the terms in the objective function, differ in terms of the queueing assumptions and performance measures, as well as in which (if any) of the specifications (17.49)–(17.52) to include. The models also differ in terms of the decision variables. While variables y_i and x_{ij} are present in all models we are familiar with (though x_{ij} may be restricted to binary values only), most models will assume that the number of facilities is m and not a decision variable. Many models also assume that all facilities have identical capacity μ, thus dropping the decision variables μ_i as well.

It is clear that the variety of SLCIS models one can define by mixing and matching different parts of the general formulation above is almost unlimited. In the next section we try to bring some structure to the models considered in the literature by grouping them around some common themes and describing the key challenges and solution techniques that have been developed for them.

17.5 SLCIS Models in the Literature: Overview and Classification

Our primary focus (with a few exceptions) is on relatively recent SLCIS models that have appeared since the survey of Boffey et al. (2006).

As noted earlier, the published SLCIS models constitute a rather bewildering pattern of different assumptions, constraints and response mechanisms. However, several common themes do emerge, allowing us to identify four common types of models: Coverage-Oriented, Service-Objective, Balanced-objective, and Explicit Customer Response. These are described in more details in the following sections. The relevant references are summarized in Tables 17.2, 17.3, and 17.4. These tables have the following format: the first column identifies the reference by the list of authors/year of publication; the next two columns identify the Model Class

Table 17.2 Coverage-type and service-objective models

Authors	Cust. Resp./ Model Class	Utility Function	Queueing Model	Flexible # Facilities?	Flexible processing rate μ?	Coverage Constraint / Lost demand	SC	Revenue (Captured Demand)	Travel Time	Congest. Cost	Facility fixed cost	Server Variable Cost	Solution Approach	Model Type/ Comments
O. Berman, D. Krass and J. Wang, 2006	AR	distance	M/M/1/c	Yes	No	Yes	Yes: % blocked calls				Min Total # of facilities		Variety of heuristics including tabu search and random adaptive search	*Type C.* Demand is lost due to coverage and cogestion constraints
H.T. Kakhki and F.M. Moghadas, 2010	NR	N/A	M/G/1	No	No	Yes	Yes	Yes: max covered demand					Exact : Obtain semi-definite relaxation that will provide an UB	*Type C.* No testing or comp. results.
V. Marianov and D. Serra, 1998	NR	N/A	M/M/1, M/M/K	No	No	Yes	Yes	Yes					EXACT Linearized the SLC, leading to a linear MIP	*Type C.*
O. Baron, O. Berman and D. Krass, 2008	AR	distance	GI/G/N, GI/G/1	Yes	Yes	Yes	Yes				Yes	General concave	Decompose the problem into several simpler sub-problems. Developed heuristic based on the equitable facility configurations	*Type C.* Both single and multiple server models considered
O. Berman and Z. Drezner, 2007	AR	distance	M/M/k	No	Yes. Total # of servers bounded	No	No		Yes	Yes			Descent, simulated annealing, Tabu Search and genetic heuristics	*Type S.*
B. Boffey R.D. Galvao and V. Marianov, 2010	NR	N/A	M/Er/1/c	No	No	No	Yes: # blocked		Yes				Turns into Capacitated p-median in M/M/1/N case, solved as MIP (this is for r=1). For general Er, do a greedy-type heuristic	*Type S.*
R. Aboolian, O. Berman and Z Drezner, 2009	AR	distance	M/M/k	No	Yes. Total # of servers bounded	No	No		Yes, minmax	Yes, minmax			Meta-heuristics	*Type S.*
T. Drezner and Z. Drezner, 2011	AR / Prop. Alloc.	distance, exp	M/M/k	No	No	No	No		Yes	Yes			Heuristic (descent, tabu search, simulated annealing, genetic)	*Type S.*
T. Hamaguchi and Nakade, 2010	AR	distance	M/G/1	No	No	No	No		Ignored	Max prob $W < \tau$			Heurristic (greedy + tabu), service times computed exactly via Laplace transform	*Type S.* Maximize probability that waiting time is below τ
V. Marianov, T.B. Boffey and R.D Galvao, 2009	NR	N/A	M/Er/K/c	No	No	No	Yes: # blocked		Yes				Similar to Boffey, Galvao and Marianov, 2010 with SLC estimated via Erlang queues	*Type S.*
V. Marianov and D. Serra, 2008	NR	N/A	M/M/c/K	No	No	No	Demand loss due to blockages		Yes	Yes			Ant colony heuristic	*Type S.* Bi-Obsjective; (1) Travel cost, (2) "Congestion cost" (with a coefficient for the number of customers in the system)
Q. Wang, R. Batta, and C.M. Rump, 2002	AR	distance	M/M/1	No	No (fixed mu)	No	Yes: max utilization rate bounded		Yes	Yes			Heuristics	*Type S.*

Table 17.3 Balanced-objective models

Authors	Cust. Resp./ Model Class	Utility Function	Queueing Model	Flexible # Facilities?	Flexible processing rate μ?	Coverage Constraint / Lost demand	SC	Revenue (Captured Demand)	Travel Time	Congest. Cost	Facility fixed cost	Server Variable Cost	Solution Approach	Comments
R. Aboolian, O. Berman and Z. Drezner, 2008	AR	distance	M/M/k	Yes	Yes	No	No		Yes	Yes	Yes	Yes	Exact algorithm and heuristics	
Castillo,I., A. Ignolfsson and T. Sim, 2009	NR	N/A	MM1, MMk	Yes	Yes	No	No		Yes	Yes	Yes	Yes	EXACT: Eliminated capacity variables, obtaining concave objective, then used Lagrangian Relaxation	
S. Elhedhli, 2006	NR	N/A	M/M/1	Yes	Yes	No	No		Yes	Yes	Yes		EXACT: Linearization approach which eliminates capacity variables and repalces non-linear term in the objective with a family of linear constraints; used column generation	
S. Kim, 2013	NR	N/A	G/G/1	No	No	No	Yes: Total Wait		Yes	Yes	Yes		EXACT Uses clearing function f(mu,W), i.e. throughput at a facility with wait given by W; this allows for linearization of constraint, but f(,) is non-linear; used column	
V. Marianov and M. Rios, 2000	NR	N/A	M/M/1	Yes	No	No	Yes: prob queue below a threshold		Link constructi on cost		Yes		EXACT: linearized the SLC, then solved MIP	Application to the location of ATM switches
S.H.R Pasandideh and A. Chambaria, 2010	NR	N/A	M/M/1/c	Yes (total location cost is bounded)	μ is fixed but buffer size is a decision variable	No	No		Obj 1	Obj 1		Obj 2: Min Ave % idle time per facility	Genetic heuristic	Bi-objective: (1) total waiting time, (2) total % idle at the facilities
N. Vidyarthi and S. Jayaswal, 2013	NR	N/A	M/G/1	Yes	Yes	No	No		Yes	Yes	Yes		EXACT Linearized "nasty" term in obj funciton, leading to a convex problem with exp numberof costraints. Similar to Elhedli, 2006	
Q. Wang, R. Batta, and C.M. Rump, 2004	AR	distance	M/M/1	Yes	Yes	No	Yes: max utilization rate bounded		Yes	Yes	Yes	Yes	Heuristics	Present several different models, but most general one is of "social optimum" type
H. Abouee-Mehrizi, S. Babri, O. Berman and H. Shavand, 2011	AR / Prop. Alloc.	exp	M/M/1/bal king	No	Yes	No	No	Demand loss due to balking, no SLC			Yes	Yes	Tabu search procedure to determine the location of the facilities, exact algorithm to obtain the optimal service rate at each facility, and a heuristic algorithm to obtain the price	Max total profit with linited room capacity for waiting

Table 17.4 Explicit customer response models

Authors	Cust. Resp./ Model Class	Utility Function	Queueing Model	Flexible # Facilities?	Flexible processing rate μ?	Coverage Constraint / Lost demand	SC	Revenue (Captured Demand)	Travel Time	Congest. Cost	Facility fixed cost	Server Variable Cost	Solution Approach	Comments
O. Berman and Z. Drezner, 2006	DR	distance	M/M/1	No	No	No	Yes: Exp. Wait Times	Yes					Single facility: exact O(n^3); Multi-facility: NLIP, heuristic algorithms (ascent algorithm, tabu, sim. annealing)	
O. Berman, R. Huang, S Kim and M.B.C. Menezes, 2007	AR / search	distance	M/M/1/c	No	No	No	Demand loss to blockage, search costs	Demand lost to blockages, search					Heuristics combined with an iterative calibration scheme to estimate the expected demand rate faced by the facilities	Comes closer to mobile server models due to dynamic search behavior by customers
O. Berman and E. Kaplan, 1987	DR	Linear	M/M/1	No: m=1	No	No	No	Yes					Eaxct algorithm (1-facility)	Single facility setting
O. Berman, D. Krass, and D. Tong, 2014	DR	Multipli-cative	G/G/1 and M/M/1	No: m=1 or fixed	Yes	Yes	No	Yes			Yes	Yes	Exact algorithms (1-facility)/ heuristic for m-facility	Demand elastic in travel, wait, and price
V. Marianov, M. Rios and J. Icaza, 2008	AR/ prop. Alloc	exp	M/M/K/c	No	No	No	No	Demand lost due to blockage					Heuristics	Max captured demand (at own facilities)
R. Aboolian, O. Berman and D. Krass, 2012	DR	Exp	M/M/1, M/M/k	Yes	Yes	No	Yes: Wait	Yes				Yes	Exact algorithm and heuristics	Max the profit including a feedback loop between customer demand and congestion
R. Rabieyan and M. Seifbarghy, 2010	AR / prop. Alloc	distance	M/M/1/balking	No	No	Constraint on idle rate at facilities	No	Demand loss due to balking					Three meta-heuristics	Max total benefit that incorporates travel distance and accounts for balking
D. Tong, 2011	FR	Multipli-cative	G/G/1	Yes	Yes	No	No	Yes			Yes	Yes	EXACT for 1-facility, exact and heuristic for m-facility case	Considers several models, including FR models with traffic equilibrium conditions
Y. Zhang, O. Berman, P. Marcotte and V. Verter ,	FR	Linear	M/M/k	Yes: min workload per facility	Yes	No	No	Yes					By-Level optimization model; lower level solved by variational inequalities and upper level heuristically	Max total participation (captured demand)
Y. Zhang, O. Berman, V. Verter, 2009	FR	Linear	M/M/1	Yes: min workload per facility	No	No	No	Yes					Location allocation heuristic(that includes an equilibrium facility–client allocation sets)	Max total participation (captured demand)
Y. Zhang, O. Berman, and V. Verter, 2012	Model 1: FR; Model 2: AR/ Prop. Alloc	distance	M/M/k	Yes: min workload per facility	Yes	No	Yes: Wait. Lost demand in model 2	Lost demand in Model 2					Probabilistic search algorithm and a genetic algorithm.	model 1: MNL (gravity) allocation (based on shortest time), model 2: shortest time allocation

by customer response type, as well as by the utility function used, if applicable. The following three columns indicate the main underlying system assumptions: the nature of the queuing system, and whether the number of facilities and the number of servers are flexible or not. The next two columns identify the presence of coverage and service level constrains. The following five columns indicate the presence of the specific terms in the objective function. The last two columns briefly describe the solution approach and any additional comments.

17.5.1 Coverage-Type Models

Coverage-type models aim to design the system that provides *adequate* service to customers, where adequacy is usually defined through travel distance and congestion delays, which are controlled through coverage and service level constraints, respectively. The defining feature of this model class is the presence of coverage constraints (17.51). The coverage-type models are denoted by "C" in the "Model Type" column of Table 17.2; they include Baron et al. (2008), Berman et al. (2006), Kakhki and Moghadas (2010), Marianov and Serra (1998). These models were among the very first SLCIS models to be considered, dating back to Marianov and Serra (1998), and stem directly from similar models for systems with mobile servers (see Berman and Krass 2002 for an extensive discussion).

Coverage-type models usually assume that it may not be possible to provide adequate service to all customers and thus demand losses may occur. The objective is typically to maximize the "captured" demand, i.e., the total demand of customers who get adequate service. The travel and congestion costs are not included in the objective as these are controlled through the corresponding constraints. Earlier models were of type NR (directed choice); later models tended to be of type AR, but customer allocations were assumed to be only a function of travel distance, i.e., the underlying utility is given by (17.28), avoiding all complications related to equilibrium behaviors. It is interesting to note that even though demand is assumed to be inelastic, the assumption of demand losses can be viewed as (a rather crude) form of demand elasticity—corresponding to the implicit utility function which has a stepwise function form, with customers using service provided by the facilities if coverage and service level constraints are met, and not using it otherwise.

The typical formulation maximizes the objective consisting of (17.44) with $r = 1$ (i.e., the captured demand), subject to constraints (17.51)–(17.56). For models of type AR, one also adds constraints specifying the allocations. These enforce each customer to travel to the closest available facility. These constraints can be specified in various forms; see Berman et al. (2006) for a discussion.

It can be seen that this leads to a formulation which is a linear mixed-integer program (MIP), except for the service level constraints. However, as discussed in Sect. 17.2.3.2, under some conditions, the latter can be linearized. Recall that a general service level constraint can be recast as either (17.20), requiring adequate service capacity at each facility, or (17.21), placing an upper limit on the allowed

arrival rate at each facility. When the capacities μ_i are decision variables, these reformulations remain non-linear. However, if one makes a simplifying assumption that all facilities have identical service rate μ (for multi-server facilities, this implies assuming identical number of servers at all facilities), non-linearities disappear. This is a common assumption in coverage-oriented (and other SLCIS) models: Berman et al. (2006), Kakhki and Moghadas (2010), Marianov and Serra (1998) assume identical and pre-specified service rates at the facilities. Under this assumption, (17.21) takes the form

$$\Lambda_i \leq \bar{\Lambda},$$

where the right-hand side is a constant which depends on the desired service level and is computable in advance. This shows the equivalence of a cover-type SLCIS model with fixed service rates to the capacitated location problems. Such connection is discussed at length in Boffey et al. (2006).

The resulting linear MIP may, in principle, be solved exactly using off-the-shelf software, such as CPLEX. However, as pointed out in Berman et al. (2006), the formulation resulting from the addition of linearized service level constraints and the "closest assignment" constraints tends to be large and not very tight, causing computational difficulties for even moderately-sized instances. This has led Berman et al. (2006) and other authors to develop heuristic approaches.

Finally, we note an important result from Baron et al. (2008), who studied a very general version of the coverage-type SLCIS model, where both the number and the capacities of facilities are decision variables and the facility-related costs are quite general (in their version, all customer demand must be served and the objective is to minimize fixed and variable location costs). They show that, under quite general conditions, the optimal facility configuration is one that ensures that each facility sees (approximately) the same demand, i.e., ideally, $\Lambda_i = \Lambda_k$ should hold for all open facilities $i, k \in I$ (identical demand may not be possible to achieve when customer demand originates from discrete nodes and single-sourcing assumption is made). Once the facility locations are determined, the optimal capacities μ_i can be determined through a separate optimization model.

This result provides an important insight for coverage-type models: when the goal is to ensure "satisfactory" service experience, the optimal design should equalize loads at the facilities. This leads to an "Equitable Location Problem"—a deterministic problem where one seeks to locate a set of facilities so that the attracted demand is distributed as evenly as possible. Such problem was addressed in Baron et al. (2007), Berman et al. (2009b), and Suzuki and Drezner (2009).

17.5.2 Service-Objective Models

Service-objective models seek to design a system that optimizes "customer service" using limited resources. These models are denoted by "S" in the "Model Type"

in Table 17.2, and include Aboolian et al. (2009), Berman and Drezner (2007), Boffey et al. (2010), Drezner and Drezner (2011), Hamaguchi and Nakade (2010), Marianov et al. (2009), Marianov and Serra (2011), and Wang et al. (2002).

Here "limited resources" means that the number of facilities to be located and the total available service capacity are specified through constraints, rather than through the objective function term (17.47). "Customer service" is typically defined as the combination of travel and congestion costs; thus the objective function typically includes terms (17.45) and (17.46). Since the congestion cost term (17.46) only measures the aggregate congestion, some authors (see Boffey et al. 2010; Marianov et al. 2009; Marianov and Serra 2011 and Wang et al. 2002) impose service level constraints to ensure that congestion is controlled at each facility. Service-objective models assume inelastic demand, so the revenue term is missing in the objective as all available customer demand is assumed to be "covered" (even though some models do allow for demand losses due to congestion, these losses are controlled through service level constraints). Thus, all customers must be assigned to facilities and thus constraint (17.55) is specified as equality.

The models of this class are either of NR type (directed assignment, no customer response) or AR type with distance-based utility function (customers travel to the closest open facility). An interesting exception is the use of AR model with proportional allocation and exponential utility (17.29) by Drezner and Drezner (2011) (though they do not comment on the existence and uniqueness of the equilibrium solution, it is in fact assured by the results cited earlier).

While the constraint set for service-objective models is quite similar to that of coverage-oriented models (in fact, it is somewhat simpler since the coverage constraints and, in some cases, service level constraints are missing), inclusion of the congestion term in the objective leads to a non-linear model for which finding exact solutions is problematic. This difficulty is further compounded when the queues at the facilities are of multi-server type and/or have non-Markovian service times: in these cases exact closed-form expressions for the congestion-related performance measures are either not available, or are quite complex, requiring a separate procedure to evaluate the congestion levels for each set of values of the facility location and customer allocation decision variables. For this reason, the proposed solution methods are all heuristic-based, typically employing meta-heuristic approaches such as tabu search, simulated annealing, and genetic algorithms.

Service-objective models become significantly more complicated when capacities of facilities are allowed to be flexible (i.e., when μ_i are not assumed to be identical at all facilities). Most of the published models assume identical capacities, with Aboolian et al. (2009) and Berman and Drezner (2007) being notable exceptions.

17.5.3 Balanced-Objective Models

Balanced-objective models seek to design a system that "balances" the costs incurred by the two main "players" in the system: customers, who bear the travel and congestion costs, and the decision-maker who bears the facility-related costs. Balanced-objective models are listed in Table 17.3 and include the following references: Aboolian et al. (2008), Abouee-Mehrizi et al. (2011), Castillo et al. (2009), Elhedhli (2006), Kim (2013), Marianov and Rios (2000), Pasandideh and Chambaria (2010), Rabieyan and Seifbarghy (2010), Vidyarthi and Jayaswal (2013), and Wang et al. (2004).

One may view balanced-objective models as seeking to achieve some kind of "social optimum"; the objective functions in these models are similar to social welfare functions in economics. Since the objective incorporates customer concerns, the models are typically of NR type: customers accept the directed assignments to optimize "social welfare", even if this leads to assignments that are suboptimal from individual customers' point of view (two references that incorporate customer response are Aboolian et al. 2008 and Abouee-Mehrizi et al. 2011). The demand is assumed to be inelastic. The coverage and service level constraints are typically absent, as service adequacy is addressed by the objective. The objective function typically includes the "customer-borne" cost terms (17.45)–(17.46) representing travel and congestion costs, as well as the "operator-borne" facility costs (17.47). Since most models do not assume any demand losses, the revenue term (17.44) is not included; the exception being Abouee-Mehrizi et al. (2011), who model revenue losses due to balking and thus optimize the net profit. Other distinguishing features of most models of this class are simple constraint sets and the inclusion of flexible capacity at the facilities as the decision variables. The main solution difficulty stems from the non-linearities inherent in the congestion term (third term of the objective function). There are several approaches for either making these terms less complex or linearizing them, leading to interesting exact algorithms. We describe two such approaches below.

The first is based on Castillo et al. (2009). They assume an $M/M/1$ queuing system at the facilities and use the average number of customers in the system $L_i(\Lambda_i, \mu_i)$ as the performance measure at facility i. For $M/M/1$ queue, this can be written as

$$L_i(\Lambda_i, \mu_i) = \frac{\Lambda_i}{\mu_i - \Lambda_i}. \tag{17.59}$$

All costs are assumed to be linear and uniform (i.e., identical for all facilities), leading to the following objective function:

$$\text{minimize } Z = \beta \sum_{j \in J} \sum_{i \in I} d(i,j)\lambda_j x_{ij} + WC \sum_{i \in I} L_i(\Lambda_i, \mu_i) + FC \sum_{i \in I} y_i + VC \sum_{i \in I} \mu_i, \tag{17.60}$$

where WC, FC, VC are the waiting cost, fixed cost and variable cost parameters respectively. This function is minimized subject to constraints (17.53), (17.55) specified as equality, as well as (17.54), (17.56) and (17.57).

Observe that for any specified values of x_{ij} and y_i, the optimal capacity μ_i^* can be determined separately for each facility. Indeed, it is not difficult to show that

$$\mu_i^* = \Lambda_i + \sqrt{\frac{WC}{VC}\Lambda_i}.$$

Observe the similarity of this expression to (17.22) discussed earlier. It also has the same interpretation: the optimal capacity at facility i consists of the minimal level Λ_i, necessary to ensure system stability, and "capacity cushion" which grows with the square root of Λ_i and whose size depends on the ratio of waiting and capacity costs. Substituting the last expression into (17.60) and performing some algebraic manipulations allows us to re-state the objective function as

$$\text{mininize } Z = \beta \sum_{j\in J}\sum_{i\in I} d(i,j)\lambda_j x_{ij} + 2\sqrt{WC \cdot VC}\sum_{j\in J}\sum_{i\in I}\sqrt{\sum_{j\in J}\lambda_j x_{ij}} + FC\sum_{i\in I} y_i,$$

subject to constraints (17.53), (17.56), and (17.55) specified as equality; the variables Λ_i and μ_i are no longer needed.

This is a MIP with a single concave (more specifically, square root) term in the objective. Several methods are available to obtain exact solutions for models of this type, which also arise in location-inventory models, competitive location models and other contexts. One approach, based on Lagrangian Relaxation, is described in Shen (2005); a variant of this is used in Castillo et al. (2009). Another approach, based on piecewise linear approximation of the concave term, is presented in Aboolian et al. (2007).

It should be noted that in view of the discussion preceding (17.22), a similar "trick" for replacing the congestion cost term with a concave form should work for more general queueing systems as well, at least as an approximation.

The second approach for obtaining exact solutions to balanced-type SLCIS is based on Elhedhli (2006). Once again we start with the model whose objective function is given by (17.60) and assume an $M/M/1$ queue at each facility. In addition, it is assumed that processing capacity of a facility must be equal to one of $H+1$ discrete values, i.e., that $\mu_i \in \{0, \mu^1, \mu^2, \ldots, \mu^H\}$ for all $i \in I$, where $\mu^1 < \mu^2 < \ldots < \mu^H$.

Treating the expected queue length L_i as a decision variable, we rewrite (17.59) as

$$\Lambda_i = \frac{L_i}{1+L_i}\sum_{h=1}^{H}\mu^h z_{ih},\ i \in I, \tag{17.61}$$

where z_{ih} is a binary decision variable taking the value of 1 if $\mu_i = \mu^h$ and 0 otherwise, with the constraints $\sum_{h=1}^{H} z_{ih} \leq 1, i \in I$ added to the model. Now consider the function $f(L) = \frac{L}{1+L}$. It is concave, and can thus be represented as the minimum of tangent lines, yielding a linear form. This can be used to represent the expression (17.61) as an infinite set of linear constraints (note that the objective is already linear, in terms of the new variable L_i). The resulting MIP can be solved through a column generation approach. The reader should refer to Elhedhli (2006) for details.

In summary, the simpler structure of balanced-objective models allows for effective exact approaches to be developed. Another interesting observation is that the "location-allocation" and "capacity determination" sub-problems often separate. As noted earlier, these models, being of type NR, may assign individual customers to rather distant facilities. However, since the travel cost is in the objective function, these "undesirable" assignments can be controlled by increasing the corresponding cost coefficients. The computational results in Castillo et al. (2009) suggest that when travel costs are "reasonably" high, the overwhelming majority of customers (over 99 % in the instances solved) are assigned to the closest open facility in the optimal solution.

17.5.4 Explicit Customer Response Models

The final class we consider consists of SLCIS models where "explicit" customer response mechanism is specified, i.e., they are of types AR, DR, or FR. These models are listed in Table 17.4. The demand in these models is generally elastic, though in a few cases elasticity is specified implicitly through demand losses due to blockages. The objective always includes the revenue term (17.44), and may also include the facility cost terms (17.47), unless the number of facilities and servers is given.

While this class of models has received much recent attention, the earliest publications date back to the very beginning of the SLCIS modeling: see Berman and Kaplan (1987). Some of the seminal early work is described in Brandeau et al. (1995).

Many of the technical issues related to this class of models have been covered in Sect. 17.3.2. The problem of determining the optimal location for a single facility (Berman and Drezner 2006; Berman and Kaplan 1987; Tong 2011; Berman et al. 2014) can be solved exactly. However, the treatment of the multi-facility case is generally quite difficult since, as noted earlier, in addition to the non-linear objective function the underlying models include the feedback loop between the customer demand and congestion and/or the equilibrium conditions for facility-client allocations, or both. Thus, heuristic approaches are almost always employed for multi-facility models. These heuristics are usually two-level: at the lower level they incorporate subroutines for computing the equilibrium solutions (using

non-linear optimization techniques) for a given location set. At the upper level they try improvement strategies to the determine a good set of open facilities, often using meta-heuristics. As in the case of balanced-objective models, the determination of the optimal capacity at a facility can often be done through a separate exact optimization procedure, for a given location and customer-allocation scheme.

We illustrate the foregoing discussion with the approach loosely based on Aboolian et al. (2012), who proposed one of the few exact approaches available for Explicit Customer Response models (in fact, the approach outlined below is an improvement on the original methodology). The model is of DR type, i.e., customers accept directed assignments to facilities, responding by reducing their demand when travel and congestion costs increase. Both $M/M/K$ and $M/M/1$ queueing systems can be considered; we will focus on the latter for simplicity. The primary queuing performance measure is the expected waiting time $\overline{W}_i$ at each facility i. While a general concave utility function may be used, we employ the exponential utility (17.29) for transparency, with the elastic demand given by (17.41). The fixed and variable costs are assumed to be uniform, i.e., identical for all locations.

We start by observing that if customers at node $j \in J$ are assigned to facility i, the maximum demand is given by

$$\lambda_{ij}^{\max} = \lambda_j^{\max} \exp(-\tau_d d(i,j)),$$

quantities that can be pre-computed. The resulting model can be formulated as follows:

$$\text{maximize } Z = r \sum_{i \in I} \Lambda_i - FC \sum_{i \in I} y_i - VC \sum_{i \in I} \mu_i \tag{17.62}$$

$$\text{subject to } \quad \overline{W}_i = \frac{y_i}{\mu_i - \Lambda_i} \qquad i \in I \tag{17.63}$$

$$\Lambda_i = \sum_{j \in J} \lambda_{ij}^{\max} \exp(-\tau_w \overline{W}_i) x_{ij} \qquad i \in I \tag{17.64}$$

$$(17.55), (17.56)$$

This reflects the typical structure of DR models: explicit specification of the waiting time and demand, in addition to regular constraints for location models. Note that system stability constraints (17.57) are omitted, as the demand automatically adjusts to the offered capacities.

The next observation is that once customer allocation variables x_{ij} are specified, both the optimal capacities at the facilities and the actual realized customer demands are easy to determine. In fact, the latter only depend on x_{ij} through the total *maximal* demand allocated to each facility:

$$\Lambda_i^{\max} = \sum_{j \in J} \lambda_{ij}^{\max} x_{ij}. \tag{17.65}$$

For each facility i we now solve the following univariate "capacity optimization" model:

$$\begin{aligned} \text{maximize} \quad & r\Lambda_i - VC\mu_i \\ \text{subject to} \quad & \Lambda_i = \Lambda_i^{\max} \exp(-\tau_w \frac{\Lambda_i}{\mu_i - \Lambda_i}) \\ & \mu_i \geq 0. \end{aligned}$$

Aboolian et al. (2012) show that the solution to this model is unique and can be found through a simple univariate search. Note that the solution yields both, the optimal capacity μ_i and the corresponding demand level Λ_i. It is convenient to represent these quantities as functions of the allocated maximum demand: $\mu(\Lambda_i^{\max})$, $\Lambda(\Lambda_i^{\max})$. Substituting these quantities into the original model (17.62)–(17.64), (17.55), (17.56) we obtain

$$\begin{aligned} \text{maximize} \quad & Z = r \sum_{i \in I} \Lambda(\Lambda_i^{\max}) - FC \sum_{i \in I} y_i - VC \sum_{i \in I} \mu(\Lambda_i^{\max}) \\ \text{subject to} \quad & (17.55), (17.56), (17.65), \end{aligned}$$

where the only non-linearities occur in the objective function. By solving the capacity optimization model repeatedly over a range of possible values of $\Lambda_i^{\max}$, we can construct a piecewise linear approximation of the functions $\Lambda(\Lambda_i^{\max})$ and $\mu(\Lambda_i^{\max})$ to any desired level of tolerance. Using these approximations in the model above yields a linear MIP which can be solved using standard off-the-shelf software.

As noted earlier, the separation of capacity optimization and customer allocation problems is a common feature of Explicit Customer-Response models and has been used by a number of authors. However, an important driver of the exact approach outlined above is that the model in Aboolian et al. (2012) is of DR type, i.e., directed assignment and single-sourcing are both assumed. The analysis presented in Aboolian et al. (2012) suggests that neither of these assumptions is very restrictive (echoing the results in Castillo et al. 2009 discussed earlier). It was observed that in the vast majority of instances solved, customers were, in fact, assigned to facilities that minimize their sum of waiting and travel times, i.e., the facilities they would have selected under an FR model. Also, by splitting the original customer nodes into k copies each containing $1/k$ of the original demand, and allowing each of these new nodes to be assigned to a different facility, the impact of the single-sourcing assumption was examined. Again, it turned out that for the instances solved, the violation of this assumption was rare (all copies of the original node were assigned to the same facility in the vast majority of the cases) and when split assignments occurred, they did not have a large impact on the objective function. Intuitively, both effects can be explained by the fact that in DR models the incentives of customers and the decision-maker, while not identical, are well-aligned: by forcing customers to use a less convenient facility, the realized demand (and the revenue) are reduced.

Thus, when designing the system, a design that maximizes customer utilities is often optimal, even though such maximization is not explicitly enforced in the model.

17.6 Conclusions

In this chapter we have focused on a rather specialized sub-field of stochastic location models: problems with congestion and static customer assignments. However, as discussed above, this is a very active and growing field of research. We believe that the key drivers of this growth are that, on the one hand, SLCIS models do capture very important trade-offs and stochastic effects that must be taken into account when designing many real-life systems. On the other hand, these models retain enough structure to enable exact algorithmic approaches and managerial insights that may not be available when more complex models (e.g., models with mobile servers or dynamic customer assignments) are considered.

The variety of SLCIS models considered in the literature is quite bewildering. We have tried to systematize the models along two dimensions: by customer response and demand elasticity (leading to our NR/AR/DR/FR classification), and by the key structural elements of the models, as described in Sect. 17.5. We believe that this classification should be useful to future researchers in this field, both with respect to the importance of clearly spelling out the assumptions for customer behavior and key model objectives, and with regards to realizing what key difficulties may arise for a given model type.

Many open questions remain, as should be clear from the preceding sections. The assumptions made with respect to queueing behavior in many models are quite restrictive and could likely be generalized using the approximation approaches described in Sect. 17.2.3.2. The assumptions underlying NR models or AR models with distance-only utility are questionable and could lead to under-performance of the resulting system (especially with respect to the realized demand). The reliance of many authors on heuristic approaches without the ability to benchmark the resulting solutions versus the optimal ones is not comforting given the strategic nature of decisions underlying SLCIS models. In short, many ways to improve on the existing models remain to be explored. We hope that some of these improvements will be investigated in the next generation of SLCIS models.

Finally we would like to mention that many of the issues that have been explored in the SLCIS context (customer response, elastics demand) are still waiting to be addressed in the models with mobile servers/dynamic customer assignments. As noted earlier, these models involve a different level of complexity, with the underlying queueing systems being much less tractable. Nevertheless, the assumptions regarding customer behavior and response are very important and deserve further study.

References

Aboolian R, Berman O, Krass D (2007) Competitive facility location model with concave demand. Eur J Oper Res 181:598–619

Aboolian R, Berman O, Drezner Z (2008) Location and allocation of service units on a congested network. IIE Trans 40:422–433

Aboolian R, Berman O, Drezner Z (2009) The multiple server center location problem. Ann Oper Res 167:337–352

Aboolian R, Berman O, Krass D (2012) Profit maximizing distributed service system design with congestion and elastic demand. Transp Sci 46:247–261

Abouee-Mehrizi H, Babri S, Berman O, Shavand H (2011) Optimizing capacity, pricing and location decisions on a congested network with balking. Math Method Oper Res 74:233–255

Ashtiani H, Magnanti T (1981) Equilibria on a congested transportation network. SIAM J Algebra Discr 2:213–226

Baron O, Berman O, Krass D, Wang Q (2007) The equitable location problem on the plane. Eur J Oper Res 183:578–590

Baron O, Berman O, Krass D (2008) Facility location with stochastic demand and constraints on waiting time. M&SOM-Manuf Serv Oper 10:484–505

Berman O, Drezner Z (2006) Location of congested capacitated facilities with distance-sensitive demand. IIE Trans 38:213–221

Berman O, Drezner Z (2007) The multiple server location problem. J Oper Res Soc 58:91–99

Berman O, Kaplan E (1987) Facility location and capacity planning with delay-dependent demand. Int J Prod Res 25:1773–1780

Berman O, Krass D (2002) Facility location problems with stochastic demands and congestion. In: Drezner Z, Hamacher H (eds) Facility location: application and theory. Springer, Berlin, Heidelberg, New York, pp 329–371

Berman O, Krass D, Wang J (2006) Locating service facilities to reduce lost demand. IIE Trans 38:933–94

Berman O, Drezner T, Drezner Z, Krass D (2009a) Modeling competitive facility location problems: New approaches and results. In: Oskoorouchi M (ed) Tutorials in operations research, INFORMS, pp 156–181

Berman O, Drezner Z, Tamir A, Wesolowsky G (2009b) Optimal location with equitable loads. Ann Oper Res 167:308–326

Berman O, Krass D, Tong D (2014) Pricing, location and capacity planning on a network under congestion. Working Paper, University of Toronto

Boffey B, Galvão R, Espejo L (2006) A review of congestion models in the location of facilities with immobile servers. Eur J Oper Res 178:643–662

Boffey B, Galvão R, Marianov V (2010) Location of single-server immobile facilities subject to a loss constraint. J Oper Res Soc 61:987–999

Brandeau M, Chiu S (1994) Facility location in a user-optimizing environment with market externalities: Analysis of customer equilibria and optimal public facility locations. Locat Sci 2:129–147

Brandeau M, Chiu S, Kumar S, Grossman T (1995) Location with market extermalities. In: Drezner Z (ed) Facility location, Springer, NewYork, pp 121–150

Brimberg J, Mehrez A (1997) A note on the allocation of queueing facilities in a continuous space using a minisum criterion. J Oper Res Soc 48:195–201

Brimberg J, Mehrez A, Wesolowsky G (1997) Allocation of queuing facilities using a minimax criterion. Locat Sci 5:89–101

Castillo I, Ignolfsson A, Sim T (2009) Social optimal location of facilities with fixed servers, stochastic demand and congestion. Prod Oper Manag 18:721–736

Drezner T, Drezner Z (2011) The gravity multiple server location problem. Comput Oper Res 38:694–701

Elhedhli S (2006) Service system design with immobile servers, stochastic demand, and congestion. M&SOM-Manuf Serv Oper 8:92–97
Gross D, Harris C (1985) Fundamentals of queueing theory, 2nd edn. Wiley, New York
Hamaguchi T, Nakade K (2010) Optimal location of facilities on a network in which each facility is operating as an M/G/1 queue. J Serv Sci Manag 3:287–297
Hopp WJ, Spearman M (2000) Factory physics, 2nd edn. McGraw Hill, New York
Ignolfsson A (2013) EMS planning and management. In: Zaric G (ed) Operations research and health care policy. Springer Science+Business Media, New York, pp 105–128
Kakhki H, Moghadas F (2010) A semidefinite relaxation for the queueing covering location problem with an M/G/1 system. Proceedings of the European workshop on mixed integer nonlinear programming, pp 231–236
Kim S (2013) A column generation heuristic for congested facility location problem with clearing functions. J Oper Res Soc 64:1780–1789
Larson R (1974) A hypercube queuing model for facility location and redistricting in urban emergency services. Comput Oper Res 1:67–95
Lee H, Cohen M (1985) Equilibrium analysis of disaggregate facility choice systems subject to congestion-elastic demand. Oper Res 33:293–311
Marianov V, Rios (2000) A probabilistic quality of service constraint for a location model of switches in ATM communications networks. Ann Oper Res 96:237–243
Marianov V, Serra D (1998) Probabilistic maximal covering location-allocation for congested system. J Reg Sci 38:401–424
Marianov V, Serra D (2011) Location of multiple-server common service centers or facilities, for minimizing general congestion and travel cost functions. Int Reg Sci Rev 34:323–338
Marianov V, Boffey T, Galvão R (2009) Optimal location of multi-server congestible facilities operating as M/Er/m/N queues. J Oper Res Soc 60:674–684
McFadden D (1974) Conditional logit analysis of quantitative choice behavior. In: Zarembka A (ed) Frontiers in econometrics. Academic Press, New York
Nagurney A (1999) Network economics: A variational inequality approach. Kulwer Academic, Boston
Pasandideh S, Chambaria A (2010) A new model for location-allocation problem within queuing framework. J Ind Eng 6:53–61
Rabieyan R, Seifbarghy M (2010) Maximal benefit location problem for a congested system. J Ind Eng 5:73–83
Shen ZJ (2005) Multi-commodity supply chain design problem. IIE Trans 37:753–762
Suzuki A, Drezner Z (2009) The minimum equitable radius location problem with continuous demand. Eur J Oper Res 195:17–30
Tong D (2011) Optimal pricing and capacity planning in operations managemnet. Ph.D. Thesis, University of Toronto, Toronto
Vidyarthi N, Jayaswal S (2013) Efficient solution of a class of location-allocation problems with stochastic demand and congestion. Working paper
Wang Q, Batta R, Rump C (2002) Algorithms for a facility location problem with stochastic customer demand and immobile servers. Ann Oper Res 111:17–34
Wang Q, Batta R, Rump C (2004) Facility location models for immobile servers with stochastic demand. Nav Res Logist 51:138–152
Whitt W (1992) Understanding the efficiency of multi-server service systems. Manag Sci 38: 708–723
Zhang Y, Berman O, Verter V (2009) Incorporating congestion in preventive healthcare facility network design. Eur J Oper Res 198:922–935
Zhang Y, Berman O, Marcotte P, Verter V (2010) A bilevel model for preventive healthcare facility network design with congestion. IIE Trans 42:865–880

Chapter 18
Demand Point Aggregation for Some Basic Location Models

Richard L. Francis and Timothy J. Lowe

Abstract Location problems occurring in urban or regional settings may involve many tens of thousands of "demand points," usually individual residences. In modeling such problems it is common to aggregate demand points to obtain tractable models. We discuss aggregation approaches to a large class of location models, consider various aggregation error measures, and identify some effective measures. In particular, we focus on an upper bounding methodology for the error associated with aggregation. The chapter includes an example application.

Keywords Aggregation • Demand points • Location

18.1 Introduction

Many location problems involve locating services (called *servers*) with respect to customers of some sort (called *demand points*, and abbreviated as DPs). Usually there is travel between servers and DPs, so that travel distances, or (more generally) travel costs, are of interest. Location models represent these travel costs, and solutions to the models can provide locations of the servers of (nearly) minimal cost. For books on location models and modeling, see Daskin (2013), Drezner (1995), Drezner and Hamacher (2002), Francis et al. (1992), Handler and Mirchandani (1979), Love et al. (1988), Mirchandani and Francis (1990), and Nickel and Puerto (2005).

A common difficulty with modeling location problems that occur in urban or regional areas is that the number of DPs may be quite large, since each private residence might be a DP. In this case it may be impossible, and also unnecessary, to include every DP in the corresponding model. Further, the models may be NP-hard to optimize (Kariv and Hakimi 1979). For problems as diverse as those including

R.L. Francis (✉)
Industrial and Systems Engineering, University of Florida, Gainesville, FL, USA
e-mail: francis@ise.ufl.edu

T.J. Lowe
Management Science Department, Tippie College of Business, University of Iowa, Iowa City, IA, USA
e-mail: timothy-lowe@uiowa.edu

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_18

the location of branch banks (Chelst et al. 1988), tax offices (Domich et al. 1991), network traffic flow (Sheffi 1985), and vehicle exhaust emission inspection stations (Francis and Lowe 1992) a popular aggregation approach is used: to suppose every DP in each postal code area or zone of the larger urban area is at the centroid of the postal code area or zone, and to compute distances accordingly. The result is a smaller model to deal with, but one with an intrinsic error. If the modeler wishes to aggregate to have a small number of *aggregate demand points* (abbreviated as ADPs), and also desires a small error, then aggregation becomes a nontrivial matter.

It is tempting to ask the following question: How many ADPs are enough? There are no general answers to this question. This is because there are important tradeoffs in doing aggregation. Aggregation often decreases: (1) data collection cost, (2) modeling cost, (3) computing cost, (4) confidentiality concerns and (5) data statistical uncertainty. The first four items seem self-explanatory; item (5) occurs because aggregation leads to pooled data, which provides larger samples and thus smaller sample standard deviations. The price paid for aggregation is increased model error: instead of working with the actual location model we work with some approximate location model. How to trade off the benefits and costs of aggregation is still an open question. The question is open in part because there is no general agreement on how to measure the aggregation error, and also because there is no accepted way to attach a cost to model aggregation error. To the best of our knowledge, professional judgment is generally used to do the tradeoffs. Francis et al. (2009) provide a survey of various demand point aggregation error measures, and an extensive literature discussion. In fact, much of the early material in this chapter, and Table 18.4, is from that paper.

One can categorize location models as *strategic*, *tactical*, or *operational* in scope. As pointed out by Bender et al. (2001), planar distances are often used for strategic-level location models, and network distances for tactical-level location models. Such models are often converted to equivalent mixed integer programming (MIP) models for solution purposes, using some finite dominating set principle to reduce the set of possible locations of interest to a finite set (Hooker et al. 1991). Thus results to follow for these planar and network models also apply to their MIP representations, including those for the p-median, p-center, and covering location models. These models are the subject matter of Chaps. 2, 4, and 5 respectively. Operational-level location models are not too common (mobile servers are one example), but for such models no aggregation may be best. Note that the scope of the location model may well indicate the degree of aggregation; a more detailed scope requires a more detailed aggregation.

18.2 Terminology and Examples

We suppose that servers and DPs are all either points in the plane, or on some travel network. In either case, there is some well-defined set of server points and DPs, say Ω, and a distance $d(x, y)$ between any two points x, y in Ω. If Ω is a travel

network (assumed undirected) then $d(x, y)$ is usually the length of a shortest path between x and y. For planar problems when $\Omega = R^2$, with $x = (\chi_1, \chi_2)$, $y = (\psi_1, \psi_2)$, $d(x, y)$ is often the $\ell\boldsymbol{p}$-distance: $||x - y||_{\boldsymbol{p}} = [|\chi_1 - \psi_1|^{\boldsymbol{p}} + |\chi_2 - \psi_2|^{\boldsymbol{p}}]^{1/\boldsymbol{p}}$, with $\boldsymbol{p} \geq 1$. Taking $\boldsymbol{p} = 1$ or 2 gives the well-known rectilinear or Euclidean distance, respectively. The limiting case of the $\ell\boldsymbol{p}$-distance as $\boldsymbol{p}$ goes to infinity, denoted by $||x - y||_\infty$, is given by $||x - y||_\infty = \max\{|\chi_1 - \psi_1|, |\chi_2 - \psi_2|\}$, and is called the Tchebyshev distance. The Tchebyshev distance in R^2 is sometimes analytically convenient because it is known (Francis et al. 1992) to be equivalent to the planar rectilinear distance under a 45-degree rotation of the axes. We define the diameter of Ω by $diam\,(\Omega) = sup\,\{d\,(x, y) : x, y \in \Omega\}$, with the understanding that possibly $diam(\Omega) = +\infty$. More generally, Ω can be a metric space (Goldberg 1976) with metric d, but no loss of insight occurs by considering the network and planar cases for Ω.

Suppose we have n DPs, $v_j \in \Omega$, $j = 1, \ldots, n$. Denote the list (or vector) of DPs by $V = (v_1, \ldots, v_n)$. When we aggregate, we replace each DP v_j by some ADP v_j' in Ω, obtaining an ADP list $V' = (v_1', \ldots, v_n')$. While the DPs are usually distinct, the ADPs are not, since otherwise there is no computational advantage to the aggregation. When we wish to model q distinct ADPs, we let Γ denote the set of q distinct ADPs, say $\Gamma = \{\gamma_1, \ldots, \gamma_q\}$. We use the former (latter) ADP notation when the correspondence between DPs and ADPs is (is not) of interest. Usually we have $q \ll n$.

For any positive integer p, let $S = \{s_k, \ldots, s_p\}$ denote any p-server, the set of locations of the p servers, $S \subset \Omega$. (This symbol p is a different symbol from the one defining the $\ell\boldsymbol{p}$-distance.) Denote the location model with the given original DPs by $f(S{:}V)$, and the one with the aggregate DPs by $f(S{:}V')$. The notation $f(S{:}V)$ and $f(S{:}V')$ captures a key idea that *an aggregation is a replacement of V by V′, with the entries of V′ not all distinct*.

For the large class of location models with similar or indistinguishable servers, with only the closest one to each DP assumed to serve the DP, for any p-server $S \subset \Omega$ and DP $v \in \Omega$ we denote by $D(S,v) \equiv \min\{d(s_k, v): k = 1, \ldots, p\}$ the distance between v and a closest element in S. We then define the closest-distance vectors $D\,(S, V) \equiv \left(D\left(S, v_j\right) : j = 1, \ldots, n\right)$, $D\,(S, V') \equiv \left(D\left(S, v_j'\right) : j = 1, \ldots, n\right) \in R_+^n$. Suppose g is some "costing" function with domain R_+^n attaching a cost to $D(S,V)$ and $D(S,V')$. This gives original and approximating location models $f(S{:}V) \equiv g(D(S,V))$ and $f(S{:}V') \equiv g(D(S,V'))$, respectively. Important and perhaps best-known instances of g are the p-median and p-center costing functions, $g(U) = w_1u_1 + \cdots + w_nu_n$, and $g(U) = \max\{w_1u_1, \ldots, w_nu_n\}$ respectively; the w_j are positive constants, often called "weights", and may be proportional to the number of trips between servers and DPs. Thus $f(S{:}V)$ is either the p-median model, $w_1\,D(S,v_1) + \ldots + w_n\,D(S,v_n)$, or the p-center model, $\max\{w_1\,D(S,v_1), \ldots, w_n\,D(S,v_n)\}$. These models originate from Hakimi (1965) (each is called unweighted if all $w_j = 1{:}\, j = 1, \ldots, n$). They are perhaps the two best-known models in location theory. The covering model, a model related to the center model, will be described later in this chapter. Subsequently,

we refer to the p-center, p-median, and covering location model as PCM, PMM, and CLM respectively. These models are NP-hard to minimize (Kariv and Hakimi 1979; Megiddo and Supowit 1984).

Consider several aggregation examples which serve to illustrate our notation and basic aggregation ideas. Let $J = \{1, \ldots, n\}$ denote the set of all DP indices. We suppose, for these examples, that the DPs will be aggregated into two postal code area centroids. Let J_i denote the subset of indices of the DPs in postal area $i = 1, 2$. Let γ_i denote the centroid of postal area $i = 1, 2$. Clearly, the J_i form a partition of J. To aggregate the DPs in the postal code areas into the centroids we replace each v_j with $j \in J_i$, by γ_i for i = 1, 2. Thus $v'_j = \gamma_i$ for $j \in J_i$ and i = 1, 2. Hence V' is now the n-vector of ADPs, and $\Gamma = \{\gamma_1, \gamma_2\}$ is the ADP set.

Example aggregation 1, PMM: $f(S{:}V) = \sum \{w_j D(S, v_j) : j \in J\}$. Let $\omega_1 = \sum \{w_j : j \in J_1\}$, $\omega_2 = \sum \{w_j : j \in J_2\}$. We then have $f(S{:}V') = \sum \left\{w_j D\left(S, v'_j\right) : j \in J\right\} = \sum \{w_j D(S, \gamma_1) : j \in J_1\} + \sum \{w_j D(S, \gamma_2) : j \in J_2\}$ $= \omega_1 D(S, \gamma_1) + \omega_2 D(S, \gamma_2)$. This example illustrates how aggregation error can occur. If only p-servers are of interest (with $p \geq 2$), then taking S to be $\{\gamma_1, \gamma_2\}$ minimizes $f(S{:}V')$ with minimal value of 0, giving a useless underestimation of $\min\{f(S{:}V){:}S\}$.

If there is only one server, $S = \{s\}$, and the $\ell\boldsymbol{p}$-distance is used, then it is known that this 1-median under-approximation is valid for all s. Letting $\omega = \sum\{w_j : j \in J\}$, and $\gamma = \sum\{(w_j/\omega)\, v_j : j \in J\}$ be the centroid of the DPs, so that $f(s{:}V') = \omega ||s - \gamma||_{\boldsymbol{p}}$, it is known (Francis and White 1974) that $f(s{:}V) \geq f(s{:}V')$ for all s. This is an important reason why underestimation can occur for PMM aggregation when few centroid ADPs are used. It is also known that for $\ell\mathbf{p}$ distances (Plastria 2001) the difference $f(s{:}V) - f(s{:}V')$ goes to zero as s gets farther from γ along an infinite ray with one end point at γ. There are good theoretical reasons due to self-canceling error (Plastria 2000, 2001); (Francis et al. 2003) for using centroids as ADPs for the PMM, but none that we know of for the PCM and CLM. Indeed, better choices than centroids are available for the latter two models.

Example aggregation 2, PCM: $f(S{:}V) = \max\{w_j D(S, v_j) : j \in J\}$. Let $w_1^+ = \max\{w_j : j \in J_1\}, w_2^+ = \max\{w_j : j \in J_2\}$. We then have $f(S : V') = \max\left\{w_j D\left(S, v'_j\right) : j \in J\right\} = \max\left\{\max\left\{w_j D\left(S, v'_j\right) : j \in J_1\right\}, \max\left\{w_j D\left(S, v'_j\right) : j \in J_2\right\}\right\} = \max\left\{\max\left\{w_j D\left(S, \gamma_1\right) : j \in J_1\right\}, \max\left\{w_j D\left(S, \gamma_2\right) : j \in J_2\right\}\right\} = \max\left\{w_1^+ D\left(S, \gamma_1\right), w_2^+ D\left(S, \gamma_2\right)\right\}$. Again, if only p-servers ($p \geq 2$) are of interest, then taking S to be $\{\gamma_1, \gamma_2\}$ minimizes $f(S{:}V')$ with minimal value of 0, giving an underestimate of $f(S{:}V)$.

Example aggregation 3, CLM: minimize $|S|$ subject to $D(S, v_j) \leq r_j$, $j \in J$, $S \subset \Omega$, where r_j is a "covering radius" associated with v_j. All but two covering constraints for the aggregated model are redundant. Define $\rho_1 = \min\{r_j : j \in J_1\}$, $\rho_2 = \min\{r_j : j \in J_2\}$. Thus the aggregated model has constraints $D(S, \gamma_1) \leq \rho_1$, $D(S, \gamma_2) \leq \rho_2$, $S \subset \Omega$. This means it takes at most two servers to solve the aggregated model. CLMs and PCMs are known to be closely related (Kolen and

Tamir 1990). We shall see that aggregation results developed for one model often also apply to the other.

These examples of models illustrate two equivalent approaches for representing n DPs with an aggregation of q ADPs. Either we have a partition of the DP index set J into q sets $J_1, \ldots, J_q$ with one ADP per set, or for each v_j there is a replacing ADP v_j', with each v_j' in the set Γ of q distinct ADPs. In either case, three aggregation decisions (Francis et al. 1999) must be made: (1) the number of ADPs, (2) the location of ADPs, (3) the replacement rule: for each v_j, what is v_j'? The (reasonable) replacement rule often used is to replace each DP by a closest ADP. Further, for the aggregation to be computationally useful we require the number of ADPs, q, to be less (usually much less) than the number of DPs, n; also it is reasonable to have $p \ll q$. The authors note that versions of these three aggregation decisions occur in location modeling. Hence results in location theory help in doing DP aggregation, so DP aggregation is a sort of "second-order" location problem to solve prior to solving the original or "first-order" problem.

These three examples may suggest that as more ADPs are used the aggregation error decreases—ideally, if we could use $q = n$ ADPs, we don't have an aggregation error at all. In fact there are classes of location models where the law of diminishing returns applies: aggregation error decreases at a decreasing rate as q increases (Francis et al. 2004a). Thus a very small value of q may cause a very high aggregation error, while a large value of q might give little less error than an appreciably smaller value of q.

18.3 Case Study

This section is based on the work by Dekle et al. (2005), where supplemental information may be found. We refer to the authors of this study as the "team".

FEMA is an acronym for Federal Emergency Management Agency, a national U.S. agency that deals with disasters such as fires, floods, hurricanes, tornadoes, and terrorist attacks. This work stems from a FEMA request to all counties in Florida to identify possible locations for disaster recovery centers (DRCs). FEMA describes a DRC as "a facility established in or nearby the community affected by the disaster, where people can meet face-to-face with representatives from Federal, State, local and volunteer agencies to obtain assistance." For the county this study deals with, Alachua County, FEMA required the identification of at least three DRCs, which could be called upon at very short notice for use in a local disaster. Alachua County had a population of about 219,000 at the time of the study. The east–west and north–south dimensions of the county are about 32 and 30 miles (51.5, 48.3 km) respectively; the land area is about 874 square miles (2,266 km^2).

FEMA provided seven DRC requirements/evaluation criteria. The County accepted all of these requirements, but added four more, including that the proposed DRC locations should be buildings allowing reasonable travel distances to them by potential users. This criterion was the most challenging to satisfy, and led to the

principal objective of the study. The team spent a substantial effort discussing with their Alachua County sponsor possible principal objectives for the study; eventually they agreed upon the following idealized one: minimize the total number of DRCs needed, subject to each county resident being within a specified distance r (called the "radius") of a closest DRC. Thus if $B(s,r)$ denotes the set of all points in the plane whose distance from a given point s is at most r, a requirement meaning that each county resident location must be in at least one $B(s,r)$ for some DRC location s; that is, each resident point in the county must be "covered" by at least one $B(s,r)$ for some DRC location s. Hence the location requirement specifies a "covering" problem (see Chap. 5). It was the belief of the team (eventually confirmed) that if they could solve this idealized problem meeting the location requirement, then they could find nearby locations that would meet all the other requirements.

A natural and important question was how to measure distances between points. Ideally, shortest path distances on the existing road network would have been used, but these were unavailable due to the very limited study budget. Since the county had a largely rectilinear/right-angle road network, the team, with the agreement of its sponsor, settled on the use of rectilinear distances: for any planar points $s=(s_1,s_2)$, $t=(t_1,t_2)$, $d(s,t)=|s_1-t_1|+|s_2-t_2|$ defines the rectilinear distance between s and t.

We refer to resident locations as "demand points", abbreviated as DPs. For any real aggregation location problem, obtaining and dealing with DP data will very probably be a major part of the problem-solving effort. Interaction with the county property appraiser's office elicited the information that principal DP data sources could be obtained from GIS data available in a library, and from the county property appraiser's office. The county DP data was arranged by "parcels" of land. There were about 6,600 parcels, and for each parcel the following information was known: x and y coordinates of the parcel center, the total heated square footage of the parcel buildings, and whether parcel buildings were residential or commercial. The parcel locations were used as residential location/DPs, and as possible DRC sites. As many as 3,900 of the parcels seemed possibly usable for DRC sites, as they had public or commercial buildings whose total usable footage exceeded 2,000 ft^2. Figure 18.1 shows a plot of all the DPs, as well as the aggregated DPs (yet to be discussed).

Covering models are discussed in Chap. 5; they provide a way to compute, for a specified covering radius r, a minimal set of locations, say $S=\{s_1,\ldots,s_k\}$, so that each DP is contained in at least one $B(s_i,r)$. To formulate the covering problem using all the available data as an integer program model would give a constraint matrix with about 6,600 rows and 3,900 columns. The size of this model was beyond the resources available to the team to deal with. The covering algorithm readily available to the team was one in Excel, which could deal with at most 200 variables/columns. The need to somehow aggregate the DP data and the potential site data thus became quite evident.

In a later section we discuss a useful error bound for covering location problems, $\max\left\{d\left(v_j,v_j'\right):j=1,\ldots,n\right\}$, where v_j is the location of DP j, and v_j' is the ADP (aggregate DP) that replaces v_j; the v_j are distinct while the v_j' are not. Choosing the

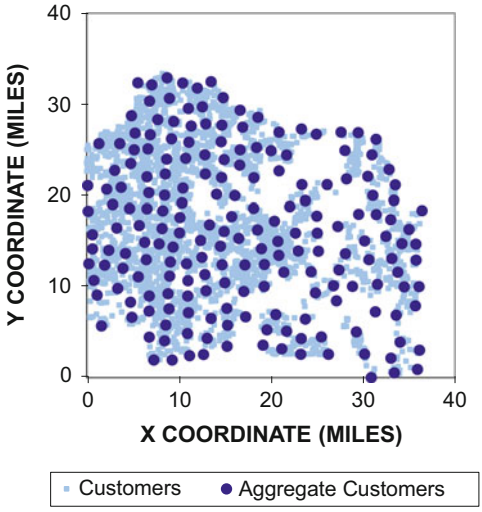


Fig. 18.1 Plot of demand points and aggregate demand points

v_j' to keep this error bound small keeps the covering error small. Note, if there are n distinct v_j', that $\max\left\{d\left(v_j, v_j'\right) : j = 1, \ldots, n\right\}$ may be viewed as the objective function of an n-center problem with DPs v_j and facility locations defined by the v_j'. This observation indicated that it would be reasonable to modify some p-center algorithm to locate the ADPs. As discussed in Dekle et al. (2005), the team used a variation of a Dyer and Frieze (1985) pick-the-farthest (PTF) algorithm to pick the ADPs. Possible center locations were also similarly aggregated. Figure 18.1 illustrates that the algorithm chooses well-dispersed locations. A number of runs of the PTF algorithm were made, and finally solutions were chosen that reduced the number of DPs from 6,600 to 198 and the number of potential DRC sites from 3,900 to 162.

The teams' version of the Dyer–Frieze algorithm works as follows. First, an arbitrary DP is chosen as an ADP. Next, a DP whose closest-distance to an ADP is farthest is then chosen to be an ADP. Continuing, at each iteration a DP is chosen as an ADP whose closest-distance to the collection of ADPs is farthest. This process continues until the closest-distance of every remaining DP to the collection of ADPs is no more than a "control parameter" b. This parameter may be adjusted to provide a computationally manageable number of ADPs. Dyer and Frieze give a low-order implementation of this approach.

Subsequently, the covering location model is solved using the ADPs as DPs; the model formulation guarantees that each ADP will be within the radius r of at least one center. However, original DPs not chosen as ADPs may possibly not be within such a radius r; supposing that v is any such unchosen DP, the algorithm guarantees that some ADP, say v', satisfies $d(v',v) \le b$. Thus for any center s that covers v', $d(s,v) \le d(s,v') + d(v',v) \le r + b$. Hence if b can be kept small then the uncovered DPs will be nearly covered, as was true in this application (see Table 18.1).

Table 18.1 (a) shows how some DRC performance measures changed with various r values for the idealized stage 1; (b) shows similar results for the actual stage 2 results

	(a) Idealized			(b) Actual		
Travel limit r (miles)	10	15	20	10	15	20
Maximum travel distance (miles)	10.9	15.1	20.3	14.0	25.8	26.94
Average travel distance (miles)	4.9	9.1	7.6	4.86	6.76	7.36
% Parcels covered	99.78	99.96	99.92	97.7	89.8	97.4
Average covering violation (miles)	0.184	0.84	0.184	1.05	2.80	2.55

Note that $\max\{d(v_i,v_i'):$ all unchosen DPs $v_i\} \leq b$ when the algorithm terminates, so keeping b small guarantees a small aggregation error. Aggregation error is discussed in the next section.

Once the DPs were aggregated, and the potential DRC sites were also similarly aggregated, the covering location problem could be solved. We call the covering location problem the idealized problem, while we call the one that considers all eleven criteria the actual problem. The team solved the idealized problem first, and then sought good solutions to the actual problem that were "close" to those of the idealized problem. This approach greatly simplified the problem and worked acceptably.

Because of initial uncertainty about an appropriate value of r, the greatest distance any resident should need to travel to a closest DRC, it was decided to treat r as a parameter of the study, try various r values, and then evaluate the resultant solutions. The team eventually chose r values of 10, 15 and 20 miles (16.1, 24.2 and 32.3 km respectively). By solving the idealized covering model with these three r values, solutions were found requiring 8, 4 and 3 DRC's respectively; see Fig. 18.2 for the case of 3 DRCs; note Fig. 18.2 illustrates three $B(s,20)$ regions. The team then proceeded to solve the actual problem by finding potential DRC locations near the idealized solutions which would meet the other evaluation requirements. To aid in this effort, they and the sponsor developed a score card, much like a grade card, on which they could score each potential location considered; most of the buildings considered were schools, churches, recreation centers, or government buildings. Table 18.1a illustrates some DRC performance measures for the solutions to the idealized problem. Discrepancies between Table 18.1a performance measures and the three different radius measures are due to aggregation effects, and can be seen to be quite small. Table 18.1b shows performance measures for the actual problem. There are some bigger discrepancies than in Table 18.1a, but these locations scored well on all the other criteria. Also it was recognized that the proper choice of a radius value r was somewhat subjective.

A number of modeling insights were gained in the course of this study, including the following. (1) Sponsors may not have a principal objective. (2) The choice of a model may be somewhat subjective. (3) Getting and working with all the data can be most of the work in an aggregation/location study. (4) Data aggregation can be essential and helpful. (5) The covering location model solutions were easy to

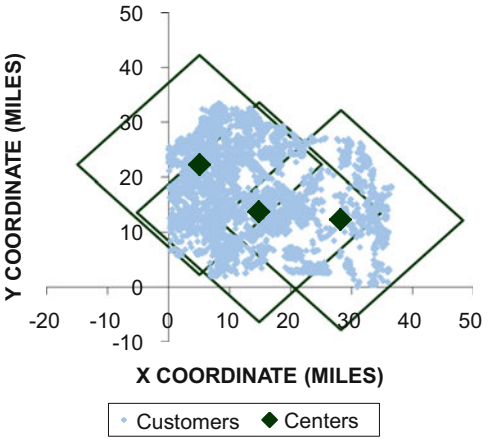


Fig. 18.2 The set of points within 20 miles of three disaster recovery centers (DRCs)

explain to the sponsor, in part due to the figures. (6) The well-dispersed locations of the covering model also had political and geographic redundancy advantages.

The three-location solution to the actual location model for $r = 20$, which covered 97.4 % of the parcels, was accepted by the sponsor. The following is a quote from a letter the sponsor provided to the team.

> The Florida Division of Emergency Management has requested that all county emergency management offices provide at least three sites preidentified as potential DRCs. With completion of this project, Alachua County is now able to comply with this request ...
>
> Overall, this was an outstanding project which has provided the Office of Emergency Management with tangible results. When DRCs must be opened in the future, it will be based upon careful research and problem solving rather than guesses on which locations would be best.

In closing, we remark that this approach easily generalizes to covering problems using network distances, given adequate network data. The approach worked well, and controls the covering error. We recommend its use for aggregating covering location problems, as well as unweighted *p*-center problems.

18.4 Aggregation Error Measures

While there can be other types of error in location models, the one we focus on is *demand point aggregation error*, which result from replacing DPs by ADPs. Thus, instead of actual distances we obtain approximating ones. The use of these approximating ADPs creates error. It is thus important for the location modeler who does the aggregation to be aware of the aggregation error being created. The modeler who does DP aggregation intentionally introduces error into the model. The use of ADPs is the *cause* of the aggregation error, but there are error *effects*—including inaccurate values of the objective function and of server locations, due

to using the approximating distances. It is important to consider both cause and effects in order to get the whole picture. There are a number of ways to measure error effects; further, the magnitude of aggregation effects can depend on the model structure—for the same aggregation, some models can have more error than others. What is clear, in any case, is that the way to minimize DP aggregation error is to not aggregate DPs—certainly this is what we recommend when it is feasible. *The ideal way to aggregate DP data is not to aggregate it*.

If DP data must be aggregated, then we need to consider aggregation error measures. We list and summarize ten such measures in Table 18.2. All these error measures have an ideal value of zero. One simple way to measure aggregation error is to consider *ADP–DP distances*. If these distance values are all zero then ADPs and DPs are identical, so there is no error. Later we establish a relationship between ADP and DP distances and other error measures, including the *distance difference error*. For the PMM, this distance difference error leads to an error we call *DP error*. Like the difference error, the DP error can be negative or positive. Still considering the PMM, note that the *total DP error* $e(S)$ in Table 18.2 satisfies $e(S) = f(S:V') - f(S:V)$, the difference between the aggregated PMM and the original model. Even though no DP error is zero, the total DP error can be zero or nearly zero, since negative errors can cancel out positive errors—this is the concept

Table 18.2 Various demand point aggregation error measures for a location model $f(S{:}V)$. Ideal error measures have value zero for all j and all S

No.	Error name	Error definition
1	ADP–DP distances	$d\left(v'_j, v_j\right), \quad j \in J$
2	Distance difference error	$D\left(S, v'_j\right) - D\left(S, v_j\right), \quad j \in J$, all S
3	DP error, PMM	$e_j(S) = w_j\left[D\left(S, v'_j\right) - D\left(S, v_j\right)\right], j \in J$, all S
4	Total DP error, PMM	$e(S) = \sum\left\{e_j(S) : j \in J\right\}$, all S
5	ABC error for PMM: $J_1, \ldots, J_q$ is a partition of $J = \{1, \ldots, n\}$;$\omega_i \equiv \sum\{w_j : j \in J_i\}$ for $i = 1, \ldots, q$	$abc_i(S) = \omega_i D(S, \gamma_i) - \sum\left\{w_j D\left(S, v_j\right) : j \in J_i\right\}$, all S
6	Absolute error, any location model	$ae(S) = \lvert f(S:V') - f(S:V)\rvert$, all S
7	Relative error, for all S with $f(S{:}V) > 0$	$rel(S) = ae(S)/f(S:V)$, all S
8	Maximum absolute error	$mae(f', f) = \max\{ae(S) : S, S \subset \Omega, \lvert S\rvert = p\}$
9	Error bound eb Ratio error bounds (when $f(S{:}V)$, $f(S{:}V') > 0$)	A number eb with $ae(S) \le eb$ for all S, $\lvert f(S:V')/f(S:V) - 1\rvert \le eb/f(S:V)$, $\lvert f(S:V)/f(S:V') - 1\rvert \le eb/f(S:V')$ for all S
10	Location error	a measure, $diff(S', S^*)$, of the "difference" between p-servers S' and S^*

of *self-canceling error*. Unfortunately self-canceling error only applies to models with an additive cost structure.

Next, consider *ABC errors* for the PMM, due to Hillsman and Rhoda (1978), pioneering aggregation researchers. Note that ABC errors are sums of the DP errors which are organized by the ADPs. Suppose we represent an aggregation by a partition of $J = \{1, \ldots, n\}$, say $J_1, \ldots, J_q$, such that for $i = 1, \ldots, q$, every DP v_j with $j \in J_i$ is aggregated into the ADP γ_i; that is, $v'_j = \gamma_i$ for $j \in J_i$. Thus $\sum\{w_j D(S,v_j): j \in J_i\}$ is replaced in the aggregate model by $\sum\{w_j D(S, \gamma_i): j \in J_i\} = \omega_i D(S,\gamma_i)$, where $\omega_i \equiv \sum\{w_j: j \in J_i\}$. In the parlance of Hillsman and Rhoda, the ABC error illustrates their *Source A* error, which they define actually as $\omega_i D(S,\gamma_i)$. Using $\omega_i D(S,\gamma_i)$ instead of $\sum\{w_j D(S,v_j) : j \in J_i\}$ is a source of error. The special case of Source A error when $\gamma_i \in S$, so that $\omega_i D(S,\gamma_i) = 0$, is their *Source B* error. If $\omega_i D(S,\gamma_i) = 0$, then it is useless as an estimate of $\sum\{w_j D(S,v_j): j \in J_i\}$. The *Source C* error is a related sort of allocation error involving closest-distance definitions. Suppose $s_k \in S$ is closest to γ_i; we might then assume that every $v_j \in J_i$ will be closest to s_k. However, in reality, some $v_j \in J_i$ may be closer to another element of S than s_k. In effect, we would allocate some DPs to a wrong server location that is not closest to them. Note $abc_i(S) = \sum\{e_j(S): j \in J_i\}$ for all i, so total ABC error is $e(S) = f(S:V') - f(S:V)$. ABC error can be negative or positive, again resulting in possible self-cancellation effects. Hillsman and Rhoda recognize and discuss both total error and error self-cancellation.

Now consider any location model $f(S:V)$ with p-server S and its approximation $f(S:V')$. A difficulty with error measures that can be negative or positive is that a smaller error (e.g., −3,000) can be worse than a bigger error (e.g., +3). We can avoid this difficulty by using the (nonnegative) *absolute error*, $ae(S) \equiv |e(S)| = |f(S:V') - f(S:V)|$ defined for all S. This measure is familiar from the calculus for measuring how well one function approximates another. Related to $ae(S)$ is the idea of an *error bound*: a number eb for which $ae(S) \leq eb$ for all S. An equivalent way to define an error bound, using f' and f to denote the functions $f(S:V')$ and $f(S:V)$ respectively, is based on the *maximum absolute error*, $mae(f', f)$, a number which may very well be quite difficult to compute. Any error bound is then an upper bound on the maximum absolute error. Good error bounds may be much easier to compute than the maximum absolute error. A *relative error* can be defined when $f(S:V)$ is always positive: $rel(S) \equiv ae(S)/f(S:V)$, perhaps converted to percent. Depending on the model structure, $ae(S)$ may be large but $rel(S)$ may still be small due to the magnitude of $f(S:V)$. Relative error is not affected by the measurement scale chosen, whereas the preceding error measures are.

Assuming $f(S:V) > 0$ and $f(S:V') > 0$ for all $S \subset \Omega$, the relative error idea gives other equivalent ways of expressing the error bound, for all $S \subset \Omega$:

$$\left|\frac{f(S:V')}{f(S:V)} - 1\right| \leq \frac{eb}{f(S:V)} \iff \left|\frac{f(S:V)}{f(S:V')} - 1\right| \leq \frac{eb}{f(S:V')}.$$

If the model $f(S:V)$ is on a national scale, but aggregation is done on a city/town scale (e.g., $eb = 10$ miles, $f(S:V) = 500$ miles), we could have relatively small ratios

eb/f$(S{:}V)$ and *eb/f*$(S{:}V')$, in which case the model ratios would be nearly one and we would have a good aggregation. By contrast, if the model is on a city/town scale and the aggregation is also on a city/town scale, we might have a poor aggregation. *We need the aggregation scale to be substantially smaller than the model scale in order to have a good aggregation*. This is one reason that aggregation may be of more interest for problems of city/town/regional scope than those of national or international scope.

There is another way to view the use of an aggregation error bound. The error bound allows us to draw conclusions about a family of original models, instead of just one. If the actual location model is $F(S{:}V)$ instead of $f(S{:}V)$, but the error bound applies to both, that is $|f\ (S:V') - F\ (S:V)| \le eb$ and $|f\ (S:V') - f\ (S:V)| \le eb$ for all S, then whatever conclusions we draw about the function f using the error bound inequality also apply to the function F. While we lose accuracy when we aggregate, we gain the ability to draw approximate conclusions about a family of original functions. As a general example of the function F, suppose instead of the DP set $\{v_j{:}\ j \in J\}$ we have a different DP set, say $\{b_j{:}\ j \in J\}$, defining F, while all other model data is the same as for $f(S{:}V)$. If each DP b_j is aggregated into v_j', then each of the functions F and f will be aggregated into the same approximating model, denoted by f'. Further, if also $d\left(v_j, v_j'\right) = d\left(b_j, v_j'\right)$ for $j \in J$, then the methods we present later would provide both F and f', and f and f', with the same error bound. The data for F and f differ, but are sufficiently similar that the aggregation does not detect the differences.

Denote (globally) minimizing solutions to any original and approximating location models $f(S{:}V)$ and $f(S{:}V')$ by S^* and S' respectively. While we usually cannot expect to find S^* if we must aggregate DPs, we can still obtain some information about S^* if we know an error bound *eb* and S'. Geoffrion (1977) proves that if $|f\ (S':V') - f\ (S^*:V)| \le eb$, then $|f\ (S':V) - f\ (S^*:V)| \le 2eb$. Supposing $f(S'{:}V) > 0$, we thus have $\left|1 - f\ (S^*:V)\ /f\ (S':V)\right| \le 2eb/f\ (S':V)$. Hence, if $2eb$ is small relative to $f(S'{:}V)$, we may reasonably accept S' as a good substitute for S^*. We assume henceforth that we can compute S' but not S^*. Note that if we wish to use S' to approximate S^*, then it makes no sense to allow $p \ge q$, for then we can place a new facility at every ADP and may achieve a minimal approximating function value of $f(S'{:}V') = 0$. Certainly it is desirable to have $p \ll q$.

Various authors, cited in Francis et al. (2009), have proposed different types of optimality errors which we list in Table 18.3. The first error can be computed, and indicates how well the approximating function estimates the original function at S'. For large models, the second two errors cannot be computed without knowing S^*. They can be computed for smaller models where S^* can be found without the need to aggregate, or for larger models if one *assumes* the algorithm used to solve the original problem provides S^*. Unless one can be certain that S^* is known, or that some properties of S^* are known, the latter two measures do not seem useful.

Although it is reasonable to use some measure of the difference between $f(S{:}V')$ and $f(S{:}V)$ to represent aggregation error, doing so results in what may well be called the *paradox of aggregation* (Francis and Lowe 1992). Often our principal reason to

Table 18.3 Various types of optimality errors for any location model $f(S{:}V)$

No.	Error name	Error definition
1	Total error at S'	$e(S') = f(S'{:}V') - f(S'{:}V)$
2	Opportunity cost error	$f(S^*{:}V) - f(S'{:}V')$
3	Optimality error	$f(S^*{:}V) - f(S'{:}V)$

Ideal error measures are zero

aggregate is because we cannot afford, computationally, to make many function evaluations of $f(S{:}V)$. We want to aggregate to make the error small; however, algorithms to do this typically require numerous function evaluations of $f(S{:}V)$ and thus cannot be used for this purpose. Usually it is practical, however, to compute error measures for at least a few S, and we certainly recommend doing so whenever possible. For example, given we know V and V', we can use a sampling approach to compute a random sample of size K of p-servers, say $S_1, \ldots, S_K$, compute $f(S_k{:}V')$ and $f(S_k{:}V)$ for each sample element S_k, and then compute a sample error estimate of any error measure of interest.

Location error (Casillas 1987; Daskin et al. 1989) involves some comparison of the p-server locations S^* and S'. There are several difficulties with using this concept. First, if we really knew S^* we would not need to do the aggregation. Second, when $|S^*| \geq 2$, there appears to be no accepted way to define the difference between S^* and S'. Third (assuming we do know S^*), the function $f(S{:}V)$, particularly if it is the PMM function, may well be relatively flat in the neighborhood of S^*, as pointed out by Erkut and Bozkaya (1999). This means we could have some S' with $f(S'{:}V)$ only a little larger than $f(S^*{:}V)$, but S' is "far" from S^*. Fourth, S' and S^* may not be unique global minima. Why are comparisons made between S' and S^*? We speculate they are made in part due to unstated subjective evaluation criteria, or known but unstated supplementary evaluation criteria. As another possible example of the use of location error, we might solve the approximating model with three different levels of aggregation (numbers of ADPs), obtaining three corresponding optimal p-servers say S', S'' and S'''. In this case, differences between successive pairs of these p-servers might be of interest; we might want to know how stable the optimal server locations are as we change the level of aggregation (Murray and Gottsegen 1997) .

Subjective or unstated aggregation error criteria may well be important, but are not well-defined. Thus two analysts can study the same DP aggregation and not agree on whether it is good or not. Further, if a subjective evaluation derives from some visual representation of DPs and ADPs, such an analysis may single out some relatively simple visual error feature that is inappropriate for the actual model structure. For example, a visual analysis could not evaluate the (computationally intense) absolute error for the PMM. Some generally accepted way to measure location error is desirable.

How should we measure the location error $diff(S,Y)$, the "difference" between any two p-servers S and Y? The answer is not simple, because the numbering of the elements of S and of Y is arbitrary, and we must find a way to match up corresponding elements. Further, S and Y are not vectors,

but sets. We propose the use of a method discussed by Francis and Lowe (1992). For motivation, consider the case where for each element s_k of S there is only one "nearby" element of Y, say $y_{k^\wedge}$. In this case we might use either $\max\{d(s_k, y_{k^\wedge}) : k = 1, \ldots, p\}$ or $\sum\{d(s_k, y_{k^\wedge}) : k = 1, \ldots, p\}$ as *diff*(S,Y). More generally, define the p x p matrix $C = (c_{ij})$ with $c_{ij} = d(s_i, y_j)$. Define an assignment (permutation matrix) to be any 0/1 $p \times p$ matrix $Z = (z_{ij})$ having a single nonzero entry of one in each row, and a single nonzero entry of one in each column, and let P denote the set of all such $p!$ assignments (permutation matrices). Define the objective function value $v(Z)$ for every assignment $Z \in P$ by $v(Z) \equiv \max\{c_{ij}\ z_{ij}: Z \in P\}$, so that $v(Z)$ is the largest entry in C for which the corresponding entry in Z is one. Define $\Delta(S,Y) = \min\{v(Z): Z \in P\}$, so that $\Delta(S,Y)$ is the minimal objective function value of the min-max assignment problem with cost matrix C. We propose using $\Delta(S,Y)$ for *diff*(S,Y). There are several good reasons for using $\Delta(S,Y)$. One reason is that it has all the properties of a distance (see Goldberg 1976): **symmetry**: $\Delta(S,Y) = \Delta(Y,S)$; **nonnegativity**: $\Delta(S,Y) \geq 0$ and $\Delta(S,Y) = 0 \iff S = Y$; **triangle inequality**: $\Delta(S,Y) \leq \Delta(S,Z) + \Delta(Z,Y)$ for any p-servers S, Y and Z. Another reason, further explored in Francis et al. (2009), is that it is related to the absolute error. (We could also use the optimal value of the conventional min-sum assignment model for *diff*(S,Y). This optimal value also has all the properties of a distance, but we know of no useful relationship between it and absolute error.) We call the distance Δ the *min-max distance*. Note, for any two p-servers S, $Y \subset \Omega$, $\Delta(S,Y) \leq diam(\Omega)$. Further, when $p = 1$ the min-max distance is just the usual distance, $d(x_1, y_1)$.

Both min–max and min–sum assignment models are well-studied and are efficiently solvable in low polynomial order for any set of real coefficients (Ahuja et al. 1993) . In the assignment models we study, the coefficients typically correspond to distances between points in some geometric spaces, e.g., planar Euclidean or rectilinear cases. For these geometric models significantly more efficient algorithms have become available (Agarwal et al. 1999; Agarwal and Varadarajan 1999; Efrat et al. 2001; Varadarajan 1998).

There are a number of relationships between the error measures of Table 18.2. These relationships, some of which may not be obvious, are a subject of discussion in Francis et al. (2009), where there are also numerical examples of many of the error measures. It also seems worth pointing out that error measures 2 through 7 of Table 18.2 are local error measures, since they depend on S. By contrast, measures 1, 8 and 9 may be considered global error measures.

There is no general agreement on which aggregation error measure is best. Until the research community agrees on one or more error measures, progress in comparing various aggregation approaches, and in building a cumulative body of knowledge, will necessarily be limited. The lack of agreement on error measures also limits progress in trading off aggregation advantages and disadvantages. Further, because comparisons of various aggregation algorithm results should all be based on the same error measures, there is currently little point in developing a data base of DPs that can be used by the profession to test their aggregation methods. We personally recommend the uses of relative error based on absolute error

and/or error bounds, together with ADP-DP distances. The bound in the inequality $\left|1-f\left(S*:V\right)/f\left(S':V\right)\right| \leq 2eb/f\left(S':V\right)$ seems particularly promising.

An alternative to using some low computational order approach to aggregate the original demand point set, and then solving the resulting aggregated location model to optimality, is to use some low computational order metaheuristic approach (Pardalos and Resende 2002; Reeves 1993; Resende and de Sousa 2004) to approximately minimize the original, unaggregated location model. The first approach gives bounds on optimality to the original model. The second approach introduces an additional source of error, since a heuristic is used, but may possibly result in a better solution. Given the current state of the art, which approach is best is not known. Indeed, "best" may not even be well-defined, since there is no generally accepted measure of aggregation error.

18.5 Error Bounds

We have argued that an upper bound on the absolute error is among the best representations and measures of the error associated with an aggregation. We have used the symbol *eb* to represent this upper bound so that with $f(S,V)$ a general location model, $|f\left(S:V'\right)-f\left(S:V\right)| \leq eb$.

Consider now obtaining error bounds for the PMM and PCM, say $\mathrm{eb}_{\mathrm{pmm}}$ and $\mathrm{eb}_{\mathrm{pcm}}$, with these two models defined in Examples 1 and 2 respectively. Both error bounds are direct consequences of the triangle inequality for shortest distances, which holds for all $j \in J$ and all $S \subset \Omega$:

$$\begin{aligned} -d\left(v_j', v_j\right) \leq D\left(S, v_j'\right) - D\left(S, v_j\right) \leq d\left(v_j', v_j\right) \\ \Longleftrightarrow \left|D\left(S, v_j'\right) - D\left(S, v_j\right)\right| \leq d\left(v_j', v_j\right). \end{aligned} \tag{18.1}$$

The *p*-median and the *p*-center models have the following error bounds respectively:

$$eb_{pmm} = \sum \left\{w_j d\left(v_j', v_j\right) : j \in J\right\},\ eb_{pcm} = \max\left\{w_j d\left(v_j', v_j\right) : j \in J\right\}.$$

The error bounds themselves can be viewed as location models; if v_j' is the closest ADP to v_j (which is reasonable), then we have

$$eb_{pmm} = \sum \left\{w_j D\left(\Gamma, v_j\right) : j \in J\right\},\ eb_{pcm} = \max\left\{w_j D\left(\Gamma, v_j\right) : j \in J\right\}.$$

Since it is of interest to have small error bounds when doing aggregation, we can view each of the latter two error bound expressions as a location model, and use

heuristic location minimization algorithms to compute Γ. Thus doing aggregation may be viewed as solving a location problem.

We remark for PMM, if S is restricted to being in a finite set of possible sites, and there are fixed site costs but the sites are not aggregated, then the site fixed costs can be added to the objective function without affecting the error bound.

Francis et al. (2009) give an extensive discussion of the use of the above error bounds for aggregation. The conditions for the PMM error bound to be tight are much stronger than for the PCM error bound to be tight, and this is reflected by better computational experience for the PCM than the PMM. However, computational experience does show that the PMM error bound is well correlated with sample absolute error measures, and that it makes sense to locate ADPs so as to keep the PMM error bound small.

Another location problem of interest is the covering location model, defined by Example 3. Since $D(S,v_j) \le r_j$ is equivalent to $D(S,v_j)/r_j \le 1$, from (18.1) we obtain

$$\left| D\left(S, v_j'\right)/r_j - D\left(S, v_j\right)/r_j \right| \le d\left(v_j', v_j\right)/r_j, \text{ for all } j \in J \text{ and all } S \subset \Omega. \tag{18.2}$$

Thus we obtain n error bounds, one for each original constraint. Clearly, it makes sense to aggregate so as to keep these error bounds small.

Let us now build on (18.2), the basic error bound idea for constraints. Generally, we have location constraints of the form $f_j(S) \le r_j,\ j \in J,\ S \subset \Omega$. Suppose each function $f_j(S)$ is replaced by some approximating function, say $f_j'(S)$, resulting in some constraints that are not distinct for the aggregated model of $f_j'(S) \le r_j,\ j \in J,\ S \subset \Omega$. If we now define functions $f(S)$ and $f'(S)$ by $f(S) \equiv \max\{(1/r_j) f_j(S): j \in J\}$, $f'(S) \equiv \max\left\{\left(1/r_j\right) f_j'(S) : j \in J\right\}$, then the constraints for the two models are equivalent to $f(S) \le 1$ and $f'(S) \le 1$ respectively. Hence we can view $f'(S)$ as an aggregated version of the function $f(S)$, and apply whatever function error measures are of interest. It is known (Francis et al. 2004a, b, c) for example, that if $f_j'(S)$ and $f_j(S)$ have error bound $b_j \left(= d\left(v_j', v_j\right)/r_j \text{ for the CLM}\right)$ for $j \in J$, then $f(S)$ and $f'(S)$ have the (unitless) error bound $eb = \max\{b_j: j \in J\}$. For the CLM, the resulting error bound is identical in form to that for the PCM; hence aggregation methods providing small PCM error bounds also can provide small CLM error bounds, and vice-versa.

When $f(S)$ and $f'(S)$ are any original and aggregated functions with some error bound eb, it follows directly that $f'(S) \le 1 - eb \Rightarrow f(S) \le 1$; $f(S) \le 1 \Rightarrow f'(S) \le 1 + eb$. Thus the constraint $f'(S) \le 1 - eb$ gives a restriction of the original constraint, while $f'(S) \le 1 + eb$ gives a relaxation. Each can be easier to deal with than the original constraint and may be used to compute lower and upper bounds on the optimal objective function value of the original model. Supposing $eb \ll 1$ (which is clearly desirable), feasibility conclusions about one model thus allow us to draw feasibility or "near-feasibility" conclusions about the other model.

Table 18.4 Relaxation and restriction of both the original and aggregated covering location models assuming all $\delta_j < r_j$

Constructing aggregated CLM		
1	Definitions	$\gamma_1, \ldots, \gamma_q$: *the q distinct ADPs* $\delta_j \equiv d(v'_j, v_j),\ j \in J;\ \delta_j < r_j,\ j \in J$ $\beta_i \equiv \min\{r_j - \delta_j : v'_i = v_j\}, i = 1, \ldots, q$ $\rho_i \equiv \min\{r_j + \delta_j : v'_i = v_j\},\ i = 1, \ldots, q$
2	Original covering constraints	$D(S, v_j) \le r_j,\ \ j \in J$, all S
3	Aggregate constraints	$D(S, v'_j) \le r_j,\ \ j \in J$, all S
4	Restrictions of both original and aggregate constraints	$D(S, v'_j) \le r_j - \delta_j,\ j \in J,\ \ \text{all } S \iff$ $D(S, \gamma_i) \le \beta_i,\ i = 1, \ldots, q,\ \ \text{all } S$
5	Relaxations of both original and aggregate constraints	$D(S, v'_j) \le r_j + \delta_j,\ j \in J,\ \ \text{all } S \iff$ $D(S, \gamma_i) \le \rho_i,\ i = 1, \ldots, q,\ \ \text{all } S$

Following Francis et al. (2004b), Table 18.4 illustrates the use of error bounds as discussed to obtain a relaxation and restriction of the aggregated CLM as well as a relaxation and restriction of the original model.

Francis et al. (2004b) used the approach of Table 18.4. They solved to optimality a CLM with almost 70,000 original CLM constraints by solving several aggregated CLMs each with less than 1,000 covering constraints. Their computational experience was usually that the minimal objective function value of the original model was underestimated when solving the approximating model without enough ADPs, which is consistent with the discussion in Sect. 18.2. The case study of Sect. 18.3 uses some of these aggregation ideas.

The error bound $\max\left\{w_j d\left(v'_j, v_j\right) : j \in J\right\}$ for the PCM and CLM for some choice of the w_j including $w_j = 1/r_j$ is quite robust. It applies to an obnoxious facility location model (Francis et al. 2000); Erkut and Neuman 1989) and, when doubled, to a p-center hub location model (Gavriliouk 2003; Ernst et al. 2002a, b).

18.6 Conclusions

For location problems with hundreds of thousands of demand points, aggregation is often essential. This chapter has dealt with the topic of demand point aggregation for location models. We have pointed out that demand point aggregation causes error, and presented some possible ways of measuring this error. Our focus has been on the concept of an error bound, an upper bound on the maximum absolute error due to aggregation. Error bounds are given for three key location models: the p-median model (PMM), the p-center model (PCM) and covering location model (CLM). We have shown that minimizing the error bounds for (PMM) or (PCM) results in a location problem. This is a concept that we have called "the paradox

of aggregation". We have also presented an application of the covering location model to a real public sector location problem in the state of Florida, and have demonstrated error bound analysis for this problem.

Difficulties in computing actual errors lead to the concept of an error bound, and this error bound can be used as a surrogate for the maximum absolute error. In fact, error bounds can be computed for many other location models since many of these models share properties with (PMM), (PCM), or (CLM). In addition, error bound analysis can be extended to more general costing functions g if $f(S) = g(D(S,V))$ and the costing function g is **s**ub**a**dditive and **n**on**d**ecreasing (SAND) (see Francis et al. 2000, 2009).

Based on our work on demand point aggregation for location modeling, we offer the following observations:

1. the work of Hillsman and Rhoda is widely recognized and influential; in particular, self-canceling error is a helpful concept for models with additive structure;
2. there is little average-case analysis of aggregation error;
3. much more research on aggregation for the median problem has been done than for center, covering and other models;
4. progress is definitely being made in understanding aggregation error;
5. aggregation error bounds can be useful, particularly for center and covering models;
6. aggregation error measures used vary greatly, and there is no agreement on how to measure error; hence it is pointless to ask which aggregation algorithm is best, since "best" is not defined.

References

Agarwal PK, Varadarajan KR (1999) Approximation algorithms for bipartite and nonbipartite matchings in the plane. In: 10th ACM-SIAM symposium on discrete algorithms (SODA), pp 805–814

Agarwal PK, Efrat A, Sharir M (1999) Vertical decomposition of shallow levels in 3-dimensional arrangements and its applications. SIAM J Comput 29:912–953

Ahuja RK, Magnanti TL, Orlin JB (1993) Network flows: theory, algorithms, and applications, exercise 12.23. Prentice-Hall, Englewood Cliffs, p 505

Bender T, Hennes H, Kalcsics J, Melo T, Nickel S (2001) Location software and interface with GIS and supply chain management. In: Drezner Z, Hamacher H (eds) Facility location: applications and theory. Springer, Berlin

Casillas PA (1987) Data aggregation and the p-median problem in continuous space. In: Ghosh A, Rushton G (eds) Spatial analysis and location-allocation models. Van Nostrand Reinhold Publishers, New York, pp 227–244

Chelst KR, Schultz JP, Sanghvi N (1988) Issues and decision aids for designing branch networks. J Retail Bank 10:5–17

Daskin MS (2013) Network and discrete location: models, algorithms, and applications, 2nd edn. Wiley, Hoboken

Daskin MS, Haghani AE, Khanal M, Malandraki C (1989) Aggregation effects in maximum covering models. Ann Oper Res 18:115–139
Dekle J, Lavieri M, Martin E, Emir-Farinas H, Francis RL (2005) A Florida county locates disaster recovery centers. Interfaces 35:133–139
Domich PD, Hoffman KL, Jackson RHF, McClain MA (1991) Locating tax facilities: a graphics-based microcomputer optimization model. Manag Sci 37:960–979
Drezner Z (ed) (1995) Facility location: a survey of applications and methods. Springer, Berlin
Drezner Z, Hamacher HW (eds) (2002) Facility location: theory and algorithms. Springer, Berlin
Dyer M, Frieze A (1985) A simple heuristic for the p-center problem. Oper Res Lett 3:285–288
Efrat A, Itai A, Katz MJ (2001) Geometry helps in bottleneck matching and related problems. Algorithmica 31:1–28
Erkut E, Bozkaya B (1999) Analysis of aggregation errors for the p-median problem. Comput Oper Res 26:1075–1096
Erkut E, Neuman S (1989) Analytical models for locating undesirable facilities. Eur J Oper Res 40:275–291
Ernst A, Hamacher HW, Jiang HW, Krishnamorthy M, Woeginger G (2002a) Uncapacitated single and multiple allocation p-hub center problems. Report CSIRO, Melbourne
Ernst A, Hamacher HW, Jiang HW, Krishnamorthy M, Woeginger G (2002b) Heuristic algorithms for the uncapacitated hub center single allocation problem. Report CSIRO, Melbourne
Francis RL, Lowe TJ (1992) On worst-case aggregation analysis for network location problems. Ann Oper Res 40:229–246
Francis RL, White JA (1974) Facility layout and location: an analytical approach, problem 7.25. Prentice-Hall, Englewood Cliffs, p 324
Francis RL, McGinnis LF, White JA (1992) Facility layout and location: an analytical approach, 2nd edn. Prentice-Hall, Englewood Cliffs
Francis RL, Lowe TJ, Rushton G, Rayco MB (1999) A synthesis of aggregation methods for multi-facility location problems: strategies for containing error. Geogr Anal 31:67–87
Francis RL, Lowe TJ, Tamir A (2000) On aggregation error bounds for a class of location models. Oper Res 48:294–307
Francis RL, Lowe TJ, Rayco MB, Tamir A (2003) Exploiting self-canceling demand point aggregation error for some planar rectilinear median problems. Nav Res Logist 50:614–637
Francis RL, Lowe TJ, Tamir A (2004a) Demand point aggregation analysis for a class of constrained location models: a penalty function approach. IIE Trans 36:601–609
Francis RL, Lowe TJ, Tamir A, Emir-Farinas H (2004b) Aggregation decomposition and aggregation guidelines for a class of minimax and covering location models. Geogr Anal 36:332–349
Francis RL, Lowe TJ, Tamir A, Emir-Farinas H (2004c) A framework for demand point and solution space aggregation analysis for location models. Eur J Oper Res 159:574–585
Francis RL, Rayco MB, Lowe TJ, Tamir A (2009) Aggregation error for location models: survey and analysis. Ann Oper Res 167:171–208
Gavriliouk EO (2003) Aggregation in hub location models. M.Sc. thesis, Department of Mathematics, Clemson University, Clemson
Geoffrion A (1977) Objective function approximations in mathematical programming. Math Prog 13:23–37
Goldberg R (1976) Methods of real analysis, 2nd edn. Wiley, New York
Hakimi SL (1965) Optimum location of switching centers and the absolute centers and medians of a graph. Oper Res 12:450–459
Handler GY, Mirchandani PB (1979) Location on networks: theory and algorithms. The MIT Press, Cambridge
Hillsman EL, Rhoda R (1978) Errors in measuring distances from populations to service centers. Ann Reg Sci 12:74–88
Hooker JN, Garfinkel RS, Chen CK (1991) Finite dominating sets for network location problems. Oper Res 39:100–118
Kariv O, Hakimi SL (1979) An algorithmic approach to network location problems: part 1, the p-centers; part 2, the p-medians. SIAM J Appl Math 37:513–560

Kolen A, Tamir A (1990) Covering problems. In: Mirchandani PB, Francis RL (eds) Discrete location theory. Wiley-Interscience, New York, pp 263–304

Love R, Morris J, Wesolowsky G (1988) Facility location: models and methods. North-Holland Publishers, Amsterdam

Megiddo N, Supowit KJ (1984) On the complexity of some common geometric location problems. SIAM J Comput 13:182–196

Mirchandani PB, Francis RL (eds) (1990) Discrete location theory. Wiley-Interscience, New York

Murray AT, Gottsegen JM (1997) The influence of data aggregation on the stability of p-median location model solutions. Geogr Anal 29:200–213

Nickel S, Puerto J (2005) Location theory: a unified approach. Springer, Berlin

Pardalos PM, Resende M (eds) (2002) Handbook of applied optimization. Oxford University Press, Oxford

Plastria F (2000) New error bounds in continuous minisum location for aggregation at the gravity centre. Stud Locat Anal 14:101–119

Plastria F (2001) On the choice of aggregation points for continuous p-median problems: a case for the gravity center. TOP 9:217–242

Reeves C (1993) Modern heuristic techniques for combinatorial problems. Blackwell Scientific Press, Oxford

Resende MGC, de Sousa JP (eds) (2004) Metaheuristics: computer decision-making. Kluwer Academic Press, Boston

Sheffi Y (1985) Urban transportation networks: equilibrium analysis with mathematical programming models. Prentice-Hall, Englewood Cliffs, pp 14–16

Varadarajan KR (1998) A divide and conquer algorithm for min-cost perfect matching in the plane. In: Proceedings 38th annual IEEE symposium on foundations of computer sciences, pp 320−331

Part III
Applications

Chapter 19
Location and GIS

Giuseppe Bruno and Ioannis Giannikos

Abstract The essence of facility location problems is to determine the position of a set of facilities in a given location space in order to provide some service to a set of actors which are supposed to patronize some of these facilities. This implies that the availability of geographically referenced information represents the fundamental prerequisite to model and solve such problems. Considering that Geographic Information Systems (GIS) offer enormous possibilities for integrating, storing, editing, analyzing, sharing and displaying spatial as well as non-spatial information, it is evident that GIS can play a crucial role for supporting decision making in the field of location science. We aim at illustrating and discussing the various linkages and application opportunities between location science and GIS and highlight the ways these two disciplines have influenced each other. Finally, we wish to indicate possibilities for further connections that may materialize in the future.

Keywords GIS • Location analysis • Optimization

19.1 Introduction

Although the study of location science formally began in the early twentieth century with the formulation of the classical Weber problem, its origins can be traced as far back as the seventeenth century when Pierre de Fermat formulated the problem of finding the geometric median of three points. Since the 1960s location science has evolved into a truly multidisciplinary area since it utilizes elements from mathematics, engineering, geography and economics, among other disciplines. The developments that have shaped modern location science may be regarded in two dimensions: (a) the formulation of new models, describing more realistic aspects

G. Bruno (✉)
University of Naples Federico II, Naples, Italy
e-mail: giuseppe.bruno@unina.it

I. Giannikos
University of Patras, Patras, Greece
e-mail: i.giannikos@upatras.gr

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_19

of problems involving the location of objects in space and (b) the proposal of new algorithms that have enabled researchers and practitioners alike to efficiently tackle larger and more complicated problems. A pivotal role in this process of evolution is certainly attributed to the rapid advances in computer technology. These advances facilitated the development of modern Geographic Information Systems (GIS) which have now become an invaluable decision support tool in many planning problems where geographically referenced information must be taken into account. In fact, the emergence of GIS results from the insight that, in order to fully comprehend certain phenomena, it is necessary to associate them with the locations where they occurred. Hence, the need to store, handle and analyze spatial data has been prominent in many disciplines. One of the first documented applications of spatial analysis is the study by Picquet in 1832 in which he represented the 48 districts of the city of Paris by halftone color gradient according to the percentage of deaths by cholera per 1,000 inhabitants. In 1854 John Snow depicted a cholera outbreak in London using points to represent the locations of some individual cases. His study of the spatial distribution of cholera helped to identify the source of the disease, a contaminated water pump, whose handle he disconnected, thus terminating the outbreak. The term "GIS" was initially used by Roger Tomlinson and his colleagues who developed a digital natural resources inventory system for Canada in the 1960s. The system provided capabilities for measurement, digitizing, scanning and overlay, thus enabling the spatial analysis of stored data.

Given that placing objects in some sort of space is the core of location science, several possibilities arose for interaction between location science and GIS. Initially, GIS were seen as an efficient means of handling data and visualizing results of location science problems, resulting in numerous applications where GIS and location models were linked in a loosely coupled way. However, over the last decade, GIS have evolved into highly sophisticated systems offering enormous capabilities for data storage and manipulation. Consequently, a much broader range of possibilities emerged for linking location science models with GIS in ways that fully exploit the analytical capabilities of modern GIS. Our aim in this chapter is to discuss the various linkages between location science and GIS and to highlight the ways these two disciplines have influenced each other. In addition, we wish to indicate possibilities for further connections that may materialize in the future.

This is not the first attempt to analyze the connections between location science and GIS. Church (1999, 2002) and Murray (2010) give notable reviews of the linkages between the two disciplines. Since the two fields and especially GIS continue to develop rapidly, it is important to evaluate how these linkages have evolved over time in comparison to the earlier reviews.

In the discussion of GIS we have chosen not to focus on any particular GIS software package. The reasons for this are twofold. Firstly, our objective is to present the theoretical principles and the functionality of GIS as well as the connections with location science, rather than specific GIS techniques. Secondly, we expect GIS software technology to develop at such a rate that references to particular packages may soon become obsolete.

The rest of the chapter is organized as follows. In the following section we present the principles of GIS and give an overview of their basic functions. We then discuss the main elements of location science and the various types of models arising in the literature. In the next section we analyze some of the possible connections between location science and GIS and discuss how the interaction between the two disciplines has developed over the years. We then present some applications exploiting this interaction and finish with some conclusions and further research suggestions.

19.2 Principles of GIS

Different definitions have been proposed for describing GIS by single authors or scientific and institutional organizations (Chrisman 1999). In broad terms GIS are information systems that integrate, store, edit, analyze, share and display geographic information as well as non-spatial information for supporting decision making. In the practical use of the term, it came to indicate a technology as well as a tool or a way of data acquisition, management, manipulation, analysis and display.

Data lies at the core of any GIS tool. Obtaining accurate, up to date and reliable data is often more difficult or more costly than acquiring a GIS tool itself. Typically, GIS store information as a collection of thematic layers that are linked together by geography. In practical terms, GIS combine *spatial data*, namely data that is in some way referenced to locations on the earth and *attribute data* that can be generally defined as additional information about each of the spatial features. Attribute data is typically represented in tabular format. For instance, in a GIS implementation of a facility location problem, spatial data may refer to the coordinates of the customers and the candidate locations for the facilities and attribute data may refer to the demand of each customer or the fixed cost of each candidate location. Other types of data such as image or multimedia are also becoming relevant in GIS following the rapid advances in technology. Documentation of GIS datasets is known as *metadata*. Metadata contains such information as the coordinate system, when the data was created, last updated, etc.

Spatial data is represented using a vector or raster/image format. The vector data model implies the use of discrete line segments (vectors) and points to represent geographic features. It can represent points, lines and areas. Each point or vertex consists of an X coordinate and a Y coordinate. An area is represented as a sequence of vectors where each vector starts where the previous one ends and where the last vector ends where the beginning vector of the sequence starts, thus enclosing the area in question. The raster data model divides the study area into a regular grid of cells with each cell containing a single value reflecting the dominant property or attribute within the cell. Since most data is captured in a vector format, e.g., by digitizing, data must be converted to the raster structure. This is called *vector–raster conversion*. Most GIS software allows the user to define the raster cell size for vector–raster conversion. It is imperative that the original data scale, e.g., accuracy,

be known prior to conversion. This should determine the cell size of the output raster map during conversion. If the cell size is large, then data may be unnecessarily generalized. On the other hand, if the cell size is too small, an excessive number of cells may be created resulting in a huge amount of data and slower processing times.

Each of these two spatial data models is characterized by certain advantages and disadvantages. In the vector model, data can be represented at its original resolution without generalization. Moreover, since most geographic data is in vector form, no data conversion is required. On the contrary, the location of each vertex needs to be stored explicitly. In addition, continuous data such as sea-depth or elevation cannot be represented easily in vector form. As far as the raster model is concerned, the location of each cell in the raster is implied by its position in the grid which implies that no geographic coordinates need to be stored, other than one reference point, e.g., the top left corner of the grid. On the other hand, it is not easy to represent in a raster model linear features or network structures. For more details on the advantages and disadvantages of these models, see Church and Murray (2009).

Image data may also be used to store remotely sensed imagery, such as satellite scenes or aerial photos. Image data is typically stored in a variety of formats (e.g., .TIFF, .PNG, .JIF, etc.). Most GIS software packages allow the input and display of such formats typically, through conversion into a raster format (and perhaps vector) to be used analytically with the GIS.

Finally, attribute data is typically represented through relational database models where data is organized in tables containing rows and columns. Each row corresponds to a record and each column stores the values of a specific attribute. Most GIS packages offer an internal relational data model as well as support for external relational databases thus enabling the use of large existing datasets.

Most of the early GIS implementations gave greater emphasis on spatial data and tended to ignore the time dimension in data representation. However, the existence of a huge volume of spatial-temporal data and the ever advancing technology have necessitated the extension of traditional models to cater for the temporal dimension as well. The inclusion of time often results in complex, large, and highly varied datasets. At the moment there does not seem to be a standard database model or analytical approach to handle these complex datasets. As reported by de Smith et al. (2013), specialized techniques have been developed for specific cases. Typical examples include the approach employed to capture land-use change (see IDRISI's Land Change Modeler package (IDRISI (2013))), the modeling of coastline advance and retreat (see Ahmad 2011) and the extension of spatial scan statistical procedures to spatio-temporal point data for crime analysis (see Cheng and Adepeju 2013).

19.2.1 GIS Functionality

Since the mid 1990s a large variety of GIS tools has been developed that have been employed for academic as well as commercial purposes. Some of them are generic GIS packages that may be used in different applications whereas others were

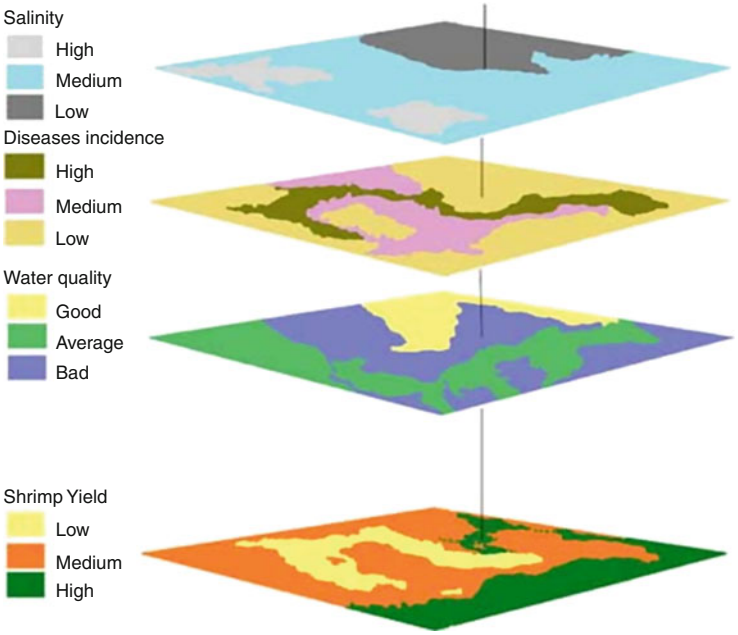


Fig. 19.1 Example of different layers (FAO 2003)

originally developed for a specific purpose (e.g., the processing of data from satellite and aerial surveys) and subsequently evolved into more sophisticated systems. Regardless of their origin or their purpose, all GIS perform a set of basic functions including the *management, transformation, analysis and visual presentation* of spatially referenced information.

The *management* of spatial and attribute data refers to the need to input, store and handle large amounts of data that may come from different sources (e.g., public records, company files, etc.) or that may be available in different formats. As noted by Murray (2010), system management is often related to representation issues and whether a raster and/or vector view of space is adopted. The *transformation* of information reflects the need for georeferencing i.e. for linking each data item to its location in a common coordinate system. This allows different data sets to be linked together based on the fact that they refer to the same location. Each data set constitutes a different layer of information. New layers may then be produced by aggregating, converting or overlaying existing layers with each other, as shown in Fig. 19.1 (FAO 2003). The figure depicts data related to a disease affecting shrimps in the sea and shows how GIS may be used to analyze whether in a certain area there is a relation between the occurrence of this disease in shrimp farming and parameters such as water quality in the ponds or major canals, salinity level, feeding levels, etc.

The *analysis* function enables the application of query, proximity, centrality and other functions to one or more information layers. Finally, the *visual presentation* of data and results has been a core component of all GIS packages and offers tools for the production of digital map, figures and graphic displays. For more details on the structure and the functionality of GIS packages see Murray (2010) and de Smith et al. (2013).

19.2.2 GIS Software

The basic functions described above can be performed by a large variety of GIS packages that has become available to both academic and commercial users over the years. The list is long and rapidly changing. Many of these packages are free while others are available for a small fee to all or selected groups of users. Special reference must be made to the development of open source GIS, which has become a long tradition in the history of GIS, with the appearance of the first package in 1978. In open source applications users may freely access and modify the source code, thus providing the package with an ever increasing range of capabilities. Such projects typically involve a large number of volunteer programmers. Finally, there exist numerous GIS commercial products that are licensed at varying per user prices, from a few hundred to over a thousand US dollars per user.

Access to spatial data as well as advanced mapping and spatial analysis over the Internet is becoming more common. As a result, a wide range of web-based or web-deployed tools has been developed, enabling the representation and analysis of datasets, without the need for local GIS software installation. Following the advances in cloud computing, GIS Cloud has been suggested as an approach to upgrade the conventional GIS applications in order to provide a broad spectrum of services to the users across the globe (Bhat et al. 2011). In fact several leading GIS vendors have already developed GIS Cloud solutions in order to provide on-demand services to their clients.

Detailed lists and reviews of GIS products can be found in Wikipedia and in specialist magazines and websites such as Geoplace (www.geoplace.com) and Geocommunity (www.geocomm.com).

According to de Smith et al. (2013), a frequent criticism of GIS software is that it is over-complicated, resource-hungry and requires specialist expertise to understand and use. Indeed, in many applications, only a handful of the capabilities provided by modern GIS is exploited. As a result, many users prefer to utilize specialized tools for their required analytical work and draw on the strengths of GIS in data management and mapping to provide input/output and visualization functionality. Example approaches include: (i) using high-level programming facilities within a GIS (e.g., macros, scripts, VBA, Python)—in fact, many libraries and add-ins have been developed in this way; (ii) using wide-ranging programmable spatial analysis software libraries and toolsets that incorporate GIS file reading, writing and display, such as R-Spatial; (iii) using general purpose data processing toolsets such

as MATLab, Excel, Python's Matplotlib, Numeric Python (Numpy); or (iv) directly through mainstream programming languages (e.g., Java, C++). The advantage of these approaches is control and transparency, the main disadvantage is that the development of such applications calls for a significant level of expertise and requires ongoing maintenance.

The complexity of GIS implementations and the huge variety of applications imply that it is not easy to develop benchmarks for testing the quality, speed and accuracy of GIS products. As a result, it is up to the user to carefully assess their particular current and future needs and to consider the features of each package (cost, maintainability, transparency, flexibility, etc.) before they adopt a specific product.

Since their appearance in the late 1960s, GIS have evolved tremendously both in terms of the related technology and with respect to the underlying methodology. Their ever increasing use has raised several research questions concerning the development of theories, techniques, data and technology for interpreting the relationships and patterns involving spatial data. In fact, this realization resulted in the introduction of the term "Geographic Information Science" (GIScience) to signify that the systematic study of these issues constitutes a science in its own right (Goodchild 2010). The need to address these issues systematically inspired the establishment of the US University Consortium for Geographic Information Science (www.ucgis.org) which involves more than 60 institutions and defines GIScience as "the development and use of theories, methods, technology, and data for understanding geographic processes, relationships and patterns". Hence, GIS are not merely a tool for decision support but a rapidly changing domain which poses significant challenges for academics and practitioners alike.

19.3 Generalities on Facility Location Problems

In general, the essence of Facility Location Problems (FLPs) is to determine the position of a set of facilities in a given location space in order to provide some service to a set of actors which are supposed to patronize some of the available facilities. These actors correspond to the demand (actual or potential) that must be satisfied. This definition implies the following fundamental ingredients of a FLP (see also Eiselt and Laporte 1995; ReVelle and Eiselt 2005).

Location Space It represents the space where demand points are present and facilities are to be located. It can be a physical space (e.g., a region or a city) or not (e.g., a market or any multi-dimensional space defined by a set of variables). Typically, the dimension of the space is assumed to be sufficiently large to consider facilities dimensionless in such a way that they can be represented as points.

The location space can be considered continuous, discrete or it may be represented by a network. In a continuous space facilities are allowed to be located at any point except within potential "forbidden zones". Continuous space models are

sometimes referred to as *site-generation models* since the generation of appropriate sites is left to the model in hand. On the other hand, in a discrete space facilities may only be located at some predefined points. For this reason discrete space models can also be referred to as *site-selection models* since the choice is limited within a set of known candidates. Using network based models the choice may be restricted to nodes or to any point of the network (node and/or arc). When a simultaneous choice of nodes and arcs is required, the problem is usually referred to as a *network design problem*. An example of this class is the so-called *corridor location problems* where routes of arcs connecting two points have to be located. The characteristics of the location space and the specific application generally drive the adoption of a metric that is used to measure distances between elements of the space (facilities and/or demand points).

Facilities The term *facility* is used to denote an object to be located in order to optimize the interaction with other pre-existing objects. Classical examples of facilities are industrial or commercial structures (e.g., retail outlets, plants, warehouses, bank branches), public services sites (e.g., schools, hospitals, fire stations, waste disposal sites), transportation and logistics infrastructures (e.g., terminals, cross-dockings, metro stations, parking lots). Facilities are usually characterized by attributes such as the number and the type of services they provide, their capacity, their attractiveness, the costs associated with their establishment and operation. Depending on the "intensity" of these attributes, facilities may produce certain "effects" on a set of actors. If these effects are judged as positive, then facilities are defined as "desirable". For instance, this is the case of schools, public service sites or metro stations where users generally wish to be as close to them as possible. Otherwise they are considered "undesirable" as in the case of nuclear or chemical plants, waste disposal sites or incinerators, airport or military installations and so on. There also exist situations where facilities are partly desirable, partly undesirable (e.g., commercial stores) as they produce some positive effects (i.e. accessibility to services) as well as some negative ones (i.e. traffic congestion) on the surrounding area.

A fundamental characteristic of a FLP is the number of new facilities to be located. The simplest case is the *single-facility problem* when the position of only one facility has to be determined, while the more general one is the *multi-facility problem* in which the aim is to simultaneously locate more than one facility. The number of facilities can be either pre-specified or a decision variable of the problem. In the latter case, there may be restrictions on the minimum or the maximum number of facilities to be located. The decision problem can also consider the possibility to shut down existing facilities or to reposition some of the existing ones.

Demand It represents the actors involved in the FLP. Depending on the kind of service provided, they can be defined as customers, users, residents, population centers and so on. Demand can be represented in continuous or in discrete fashion. In the first case the demand area may be partitioned into sub-areas such that within each sub-area it may be assumed that the demand is uniformly distributed.

Otherwise demand may be assumed to be concentrated on discrete points. In any case, it is always possible to transform continuous into discrete demand and vice versa through appropriate procedures. However, during these operations particular attention should be paid to approximations and errors introduced in the model (Current and Schilling 1990; Francis et al. 2002).

When facilities provide different types of services, demand should concern several kinds of services and the corresponding FLPs are referred to as *multi-commodity* problems. Depending on each particular application, demand can be deterministic or stochastic. In both these cases, it can be estimated either by combining current data and/or attributes or by using appropriate forecasting tools.

Interactions Between Elements of a Problem In a FLP mainly two kinds of interactions have to be taken into account: customer–facility interactions and facility–facility interactions. In some applications customer–facility interactions concern how customers patronize their own facilities or how they are "allocated" to facilities. In some cases customers are free to decide on the basis of a utility function which, in general, combines attributes of facilities and distances between customers and facilities while, in other cases, customers are obliged to patronize certain facilities according to given rules. Facility–customer interactions may also concern the determination of the intensity of the effects produced by facilities to the customers. This is typical, for instance, in problems where risks and/or damage generated by obnoxious activities have to be evaluated on the population living in the area around the facility position.

Facility–facility interactions take into account how facilities interact with each other to capture the available demand. In some cases there is competition in order to capture as much of the demand as possible (i.e. commercial stores of different companies). This aspect is also known as cannibalization effect. On the other hand, in some applications facilities are located in such a way that they cooperate in order to assure a certain level of accessibility to the users (i.e. bank offices, public service sites, franchising stores).

Objective Function(s) Location decisions can be made according to different criteria or objective functions whose choice mainly depends on the nature of facilities (desirable or undesirable). In the case of desirable facilities, efficiency is the most commonly used criterion. Efficiency is typically associated to costs, and distance is the most common proxy for costs. For this reason, objective functions are in most cases expressed as functions of distances between customers and facilities, possibly weighted by the demand associated with each customer.

Denoting with p the number of facilities to be located, problems differ according to whether p is pre-defined or a decision variable. In the first case, the minimization of the sum of the weighted distances between demand points and facilities to be located (minisum objective) is the typical objective of the well known class of p-median problems (Cooper 1963; Hakimi 1964; ReVelle and Swain 1970). When p is a decision variable, the objective to be adopted is usually the minimization

of the sum of the fixed setup costs and the variable costs to serve customers from the facilities. This problem is known as *uncapacitated* or *simple facility location problem* (Erlenkotter 1978). However, if efficiency is mainly viewed from the customers' point of view, an alternative measure to be minimized can be represented by the maximum distance between customers and their patronized facilities. In practice this so called minmax objective, typical of the class of *center* problems (continuous or discrete), is focused on customers in the worst condition (Hakimi 1964; Minieka 1970; Goldman 1971; Elzinga and Hearn 1972; Drezner and Wesolosky 1980).

Another classical concept used to measure efficiency is related to the ability of facilities to "cover" demand. More precisely a facility is said to cover a demand point if their mutual distance does not exceed a given "coverage radius" which can be evaluated depending on the specific application. In this context when the number of facilities is specified a priori, the objective consists in positioning them in such a way that they are able to cover as much demand as possible (Maximal Coverage Location Problem) (Church and ReVelle 1974). When the number of facilities represents a decision variable, the problem is to determine the minimum number of facilities whose location ensures the coverage of the overall demand (Set Covering Location Problem) (Hakimi 1965; Toregas et al. 1971).

In the case of undesirable facilities, customers wish that facilities be located as far away from them as possible and objectives may be defined accordingly. More specifically, instead of minisum and minmax objectives used for desirable facility problems, maxsum and maxmin objectives are usually employed to formulate undesirable facilities location problems (Church and Garfinkel 1978; Dasarathy and White 1980; Drezner and Wesolosky 1980). However as the adoption in the model of such objectives (maxsum, maxmin) can lead to very poor solutions from the efficiency point of view, constraints regarding minimum levels of efficiency should also be included.

Another class of interesting problems is based on the so called equality measures. Either in the case of desirable or undesirable facilities, the decision maker may be interested in finding solutions that assure a certain "fairness" in the access to facilities. In order to describe this objective, various expressions have been proposed, based on the minimization of measures related to the distribution of distances between customers and facilities. Examples of such measures include the variance, the mean absolute deviation or the Gini coefficient. For more details, see Marsh and Schilling (1994) and Eiselt and Laporte (1995).

However, it should be underlined that locational decision problems in practice can involve multiple, conflicting and incommensurate evaluation criteria and, in this sense, they are multiobjective in nature. Hence, in order to tackle FLPs formulated using multiple conflicting objectives, appropriate multiobjective techniques are needed, some of which have been reviewed by Current et al. (1990) and Farahani et al. (2010).

Depending on the combinations of the elements characterizing FLPs, a wide range of mathematical models can be defined. Due to this variety, different classification schemes have been proposed in the literature such as the ones suggested

by Francis et al. (1983), Brandeau and Chiu (1989), Eiselt and Laporte (1995), Hamacher and Nickel (1998), ReVelle and Eiselt (2005) and ReVelle et al. (2008).

19.4 Linkages Between Location Science and GIS

Location science problems have been studied in various forms for hundreds of years. On the other hand, GIS was not developed to solve location science problems as such. Their primary purpose was to collect, store, manage, manipulate, display and analyze spatial data. In fact, for a long period of time, the two fields seemed to develop almost independently. However, as practical problems became more sophisticated, it emerged that GIS offered excellent possibilities to handle the spatial data needed to solve these problems. As a result, GIS were initially viewed as a tool to provide data to location science models and to visually present their results. Indeed, there are numerous applications where GIS and location science models were combined in a *loosely coupled* way in the sense that spatial as well as attribute data were extracted from the GIS to be used by an already defined location science model. The model was then solved by commercial software or some special purpose procedure and the results were imported back into the GIS for visual presentation. In this setting, data requirements were determined primarily by the location science model and the main task was to consider the data structures utilized by the GIS and the location science component and to develop a procedure for exchanging data and results files between them. Examples of these early approaches are reviewed by Church (2002), Church and Murray (2009) and Murray (2010).

Following the continual development of GIS, it became evident that the links between GIS and location science could progress far beyond the concept of loose coupling described earlier. It can be argued that the two fields are beginning to converge in a number of ways, some of which are analyzed below.

19.4.1 Suitability Analysis and Data Generation

In many practical applications of location science a pre-processing stage is necessary in order to assess all the potential sites for one or more facilities and select those that meet a given set of pre-determined prerequisites for further consideration**.** As noted by Sumathi et al. (2008), apart from determining the set of feasible sites, GIS may also provide a digital data bank for long-term monitoring of these sites, for managing the collection operation and analyzing the routes between different elements of the system. In addition, different data layers, each weighted by a different factor, may easily be combined in order to calculate a suitability score for each possible location and only consider locations whose suitability score exceeds a pre-specified threshold. Several GIS offer interfaces to help determine appropriate weighting factors. For instance, IDRISI's Decision Wizard includes a

module that employs the Analytic Hierarchy Process (AHP) to calculate a land suitability index for each location. However, as emphasized by Church and Murray (2009), special care should be given in handling data of different types (nominal, ordinal, interval or ratio) and thorough analysis of the data transformation functions and the methods for obtaining the weighting factors is required. The wealth of spatial and attribute information available these days and the increasing capabilities of modern GIS allows location modelers to utilize such systems in order to generate data for a variety of location science models. Depending on the scope of the location model in question, the modeler may employ GIS functions to aggregate polygons, generate polygon centroids, derive service zones, calculate distances between different objects and make this data directly available to location science models. Moreover, as noted by Murray (2010), GIS functions may be used to determine more complex spatial relationships such as adjacency, contiguity and/or shape. For example, Murray and Kim (2008) developed a GIS-based procedure to identify cliques of parcels or areal units, namely sets of parcels that are in conflict with each other. These cliques were then used to generate constraints in an integer programming formulation of the Anti Covering Location Problem which regards the positioning of a maximally weighted set of facilities so that no two located facilities are within a specified time or distance measure of each other.

19.4.2 Visualization of Results

Visual representation of large amounts of information is one of the most useful aspects of GIS. Its role in the visualization of location model results has been recognized by several researchers including Densham (1994) and ReVelle and Eiselt (2005). Murray (2010) clearly states that the use of GIS for visualizing the results of location models is far more complicated than a mere depiction of the sites where facilities are located. By exploiting the graphical capabilities of modern GIS, many more aspects of the solution may be represented. For instance, classical GIS tools such as the construction of Voronoi diagrams or spider diagrams have been utilized to represent additional aspects of location model results, namely the coverage of each facility or the allocation of customers to facilities. Other methods such as graduated circles or choropleth maps have also been used to represent other attributes of the solution, e.g., the amount of demand left uncovered after the facilities have been located or the level of cannibalization resulting from the location of several competing facilities. These methods have been employed to good effect in numerous applications including the ones reported by Vijay et al. (2005), Ghose et al. (2005), Dobbins and Jenkins (2011), Suárez-Vega et al. (2011) and Pekin et al. (2013).

The visualization of the results of location science models is vital in most practical applications since it facilitates communication between the various stakeholders involved. It allows analysts and clients to easily experiment with different problem settings, directly compare alternative solutions simultaneously with respect

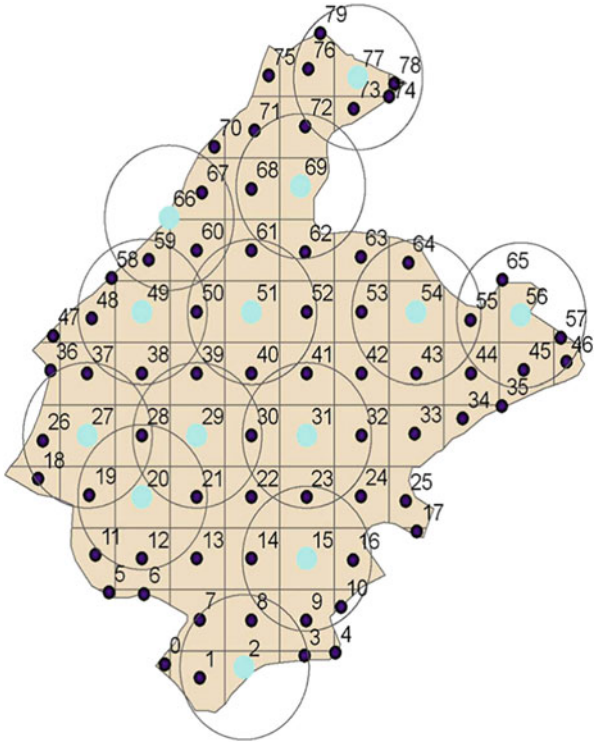


Fig. 19.2 Visualization of solution to a coverage problem

to various criteria and, in the case of dynamic location models, monitor how the solution evolves over time. Finally, an effective visualization through GIS may help identify deficiencies of the solution that would otherwise have been difficult to detect. Typical examples are shown in Murray (2005) and Alexandris and Giannikos (2010) where the Set Covering Location Problem and the Maximal Covering Location Problem respectively are solved with respect to a certain geographical area which is divided into polygonal regions.

As shown in Fig. 19.2, taken from Alexandris and Giannikos (2010), the solution corresponding to 13 available facilities leaves large areas uncovered and also provides coverage to areas outside of the geographical area that must be covered. These obvious deficiencies may lead the decision maker and the analyst to alter the solution or even to modify the underlying model altogether in order to obtain more satisfactory solutions.

19.4.3 Formulation of New Models

The linkages between GIS and location science have progressed far beyond data generation and visualization of results. On a more theoretical note, the data handling

and analytical capabilities of GIS have inspired location analysts to extend classical location science models by taking into account geographical information. A typical example is given by Suárez-Vega et al. (2011) who developed a multicriteria competitive model for locating a hypermarket on a network. They extended traditional Huff-based competitive location models, originally proposed by Huff (1964), by employing GIS procedures to consider geographical aspects such as distance to the main roads, land-use, slope of the terrain and distance to the distribution centers. Another important contribution of GIS concerns the nature of the entities involved in location models. In conventional location models the facilities to be located and the customers are usually represented by points. This is clearly not sufficient in many problem settings since the customers as well as the facilities may be described by general objects other than points, such as lines, polygons or even other irregular shapes. Using the computational geometry techniques implemented within GIS, the adjacency and contiguity between such objects may be determined and their distances to other objects may be calculated. As a result, various models have been developed where the facilities to be located are described by lines, bands (corridors) or polygons. Perhaps the most indicative example is the corridor location model and its variations. This is the problem of identifying one or more paths across a landscape such that some criteria are met. These paths may represent power transmission lines such that their cost is minimized, roads through a region such that visibility of a beautiful landscape is maximized, routes for a military unit that minimize the view or the probability of being detected, etc. Although this problem does not fall exclusively within the domain of GIS, its detailed study requires the capabilities of GIS. Hence, nearly all the approaches to solving this problem employ GIS at some stage. Some of the earlier approaches to the corridor location problem are reviewed by Church (2002) whereas more recent ones can be found in Gonçalves (2010).

A crucial issue in location science, especially when large amounts of data are available, is to determine the level of aggregation or the scale where the elements of a particular problem will be represented. For instance, if population data is available at census block level, should this detailed data be directly used in a location model or should it be aggregated at block group or census tract level? Including all the data may imply significant effort and cost for data collection and require considerable computation time, thus making the problem intractable. On the other hand, aggregating the data reduces the amount of computation work required but introduces errors in the analysis, as originally shown by Goodchild (1979). Shortly afterwards Openswaw and Taylor (1981) remarked that the results of spatial analysis may vary depending on the representation scheme adopted. This effect, known as the Modifiable Areal Unit Problem (MAUP), can be divided into two components, the aggregation level used and the type of unit utilized. Clearly, any model suffering from this effect is problematic and should be used with caution. As far as location models are concerned, Murray (2005) showed that the Set Covering Location Problem is susceptible to MAUP and, then, he developed a GIS-based alternative formulation that can be applied to points, lines, polygons or other objects. Moreover, Tong and Murray (2008) and Alexandris and Giannikos (2010)

observed that the Maximal Covering Location Problem is also prone to MAUP and employed GIS functions to formulate new models that are more robust and are less affected by changes in the scale or units of data. Finally, Kim and Murray (2008) developed a similar model for a specific version of the coverage problem (Backup Coverage Location Problem) which eliminates MAUP in comparison to the original formulation. All these new models are based on the functionality of GIS which provides tools for accurately calculating the portions of the demand areas that are covered by different configurations of the facilities.

19.4.4 Uncertainty and Error Propagation

Uncertainty and error are inherent in many location problems. Demand for a product or service is rarely determined with accuracy and is usually estimated on the basis of historical data. Other parameters, such as transportation costs or distances between facilities are also characterized by uncertainty. The coordinates of the potential facilities themselves or the distribution of the customers may also be recorded incorrectly or inaccurately. A variety of approaches has been proposed in the location science literature to deal with this uncertainty. Most of these approaches are based on the formulation of stochastic models, the performance of sensitivity analysis or the definition of scenario-based problem formulations. A detailed review of such approaches can be found in Snyder (2006).

As far as GIS is concerned, it can be safely said that no map or digital representation of spatial and attribute data is error free. According to Murray and Grubesic (2012), the uncertainty associated with the use of GIS is multifaceted and is related to several aspects including the accuracy, precision, spatial scale and geographic abstraction of the information stored in the GIS. Moreover, this uncertainty and error propagate through the application of GIS functions thus amplifying their effect. Since the early stages of GIS development, it became evident that this uncertainty and error can influence the quality of the analysis and the reliability of the results (Heuvelink 1998). Hence, a broad range of literature has focused on data imperfections and ways to deal with them in GIS. Error propagation has been studied for certain GIS functions such as overlay (see Veregin 1995). Other methods to cater for uncertainty and error include the use of Monte Carlo simulation or multiple analyses conducted on perturbed data in order to obtain more reliable results. A detailed review can be found in Li et al. (2012). It must be noted that although these methods seem to be well known in the GIS community, relatively few applications are encountered in location science problems. This may be due to the fact that a deeper understanding of GIS is required by location analysts in order to fully exploit the capabilities of GIS for dealing with uncertainty and error (see Murray 2010).

A notable example where GIS is used to study uncertainty and error in location science is given by Murray (2003) who considered the planar multi-facility location–allocation problem and used an Avenue script within ArcView to perturb

the optimal facility locations and evaluate the effect on the objective function value. Murray and Grubesic (2012) identified four categories of uncertainty (object geometry, data precision, distance measurement, and proximity interpretation) and proposed improvements of data and/or model quality along each of these dimensions. They also argued that if one improves only one of these dimensions but fails to address the others, the final result may still be problematic. Hence, they concluded that model extensions are necessary to properly address the issues of uncertainty and error in this context. Given the development of GIS based techniques for reducing the effects of uncertainty and error, one would expect many more such extensions to appear in the future.

19.4.5 Problem Solution

When it comes to solving location science problems, GIS may be utilized in several ways depending on the nature of the problems in question. If the number of facilities to be located is limited, as in the location of large infrastructures, then the solution may be obtained by performing a straightforward suitability analysis through GIS in order to determine the sites that meet the selection criteria.

GIS prove to be extremely useful for dealing with problems that can be directly or indirectly solved using certain computational geometry techniques which are standard tools in GIS. A typical example is the heuristic proposed by Suzuki and Okabe (1995) for the continuous p-center problem. This heuristic relies on the generation of a Voronoi diagram corresponding to a set of points at each step of the algorithm. Given that the construction of Voronoi diagrams are standard GIS functions, it seems natural to use GIS to solve this problem. In fact, Wei et al. (2006) implemented this heuristic using a commercial package to locate emergency warning sirens. In the same vein, Matisziw and Murray (2009) addressed the continuous coverage problem and proved that the optimal location lies on the medial axis of the demand area, namely the set of points having more than one closest point to the demand area boundary. They then used GIS to implement a Voronoi-based technique for deriving the medial axis.

When the number of feasible locations is significantly large, then a model is required to determine the sites for the facilities to be located. Combining GIS and some solution routines, either commercial or custom-made, in a loose coupling sense implies a significant exchange of input and output files between the two components and does not really exploit the capabilities of modern GIS. However, following the rapid developments in GIS, several tight coupling possibilities have emerged. In particular, several algorithms for solving location problems are currently available within GIS software packages. The ArcGis Network Analyst toolbox, the TransCad application modules for Territory Management and Site Location Modeling and the Location Intelligence module of MapInfo are examples of packages that offer tools for solving standard location science problems such as the p-median or the Maximal Coverage Location Problem. Each of these

packages employs heuristic procedures to solve a set of specific problems. While this approach may be satisfactory for practitioners facing rather simplified versions of these problems, it may not be sufficient in more realistic applications that involve complexities such as capacitated facilities, additional objectives etc. Moreover, the heuristics available within commercial packages may have not been tested on large data sets. An alternative approach is to use an appropriate programming or scripting language in order to implement more sophisticated solution methods within GIS software packages. For instance, ArcGIS allows customization and access to all of its core objects through a VBA (Visual Basic for Applications) environment. This environment may also be used to link the GIS with dynamic link libraries (DLLs) containing algorithms which can be called upon by the GIS to perform optimizations. In fact, numerous scripts and tools, some of which related to location science, have been prepared by developers for commercial as well as open source GIS and are available in relevant forums. However, they are not always easy to locate amongst the multitude of similar tools available over the internet. Finally, another possibility for linking GIS and location analysis techniques is to invoke both of them through a common programming language such as C++ or Visual Basic. Examples where GIS and location science algorithms are integrated using one of the approaches described above can be found in Johnson (2001), Ribeiro and Antunes (2002), Bender et al. (2002), Liu et al. (2006), Bozkaya et al. (2010), Bruno et al. (2010), García-Palomares et al. (2012) and Xu et al. (2013).

19.5 Using GIS in Location Science Applications

A GIS is recognized as a decision support system based on the integration of spatially referenced data in a problem solving environment (Cowen 1988). This definition is particularly relevant when, in order to appropriately solve a problem, it is necessary to conduct a complex multidimensional analysis, involving a significant set of feasible alternatives and multiple, often conflicting and incommensurable evaluation criteria, which is most often the case in practical applications. Hence, GIS appear to be ideally suited for addressing practical problem situations. A great number of studies have appeared in the literature over the years, reporting applications where GIS have been employed to tackle a wide range of practical FLPs. A review of this literature may be useful for practitioners as well as for researchers since it may identify the main application areas where GIS appear particularly fruitful as an analysis and decision support tool. A first list of GIS application areas is given by Maguire (1991). In this section we focus on the most recent literature on the subject. Although an exhaustive review is practically prohibitive, this literature can be classified into the following broad categories: *land-use suitability analysis, waste management, energy management, transportation* and *private and public sector applications.*

Land-Use Suitability Analysis The objective of land-use suitability analysis is the identification of the most appropriate spatial pattern for future land uses in such a way that a set of requirements, properties and preferences are satisfied (Collins et al. 2001; Malczewski 2004). The use of GIS based approaches in this context can be considered the natural evolution of the hand-drawn overlay techniques used by architects and planners to represent maps where different attributes and characteristics should be shown. This general concept can be applied in a wide variety of situations (i.e. urban and regional planning, environmental impact evaluation, land habitat for animal and plant species, agricultural, ecological and geological applications, public and private site facilities). Given a context where a study area is subdivided into a set of territorial units, it is possible to distinguish between a site selection and a site generation problem. In the first case, given a set of potential feasible sites with known attributes and characteristics, the objective consists in the selection of one or more facilities. This is generally performed by combining the facilities' attributes according to some ranking or rating rules. In the site search problem apart from the location(s) to be selected, it is also necessary to determine the site characteristics (i.e. extension, shape).

Hence, models oriented to land-use suitability analysis may consider very different sectors. A traditional field is represented by the so-called conservation planning, i.e. the activities related to the selection of protected areas based on scientific considerations in order to reduce the risks of habitat fragmentation and, consequently, on the related ecosystem due to the impact of land-use activities. The increasing success in the use of GIS to tackle these problems lies in the fact that most of the criteria for conservation planning are spatial data. In this context, as already indicated by Church (2002), a traditional field of applications of GIS based approaches concerns the forest conservation planning in which the problem mainly consists in the identification of corridors and/or portions of territory in order to extend protected areas and assure continuity (Phua and Minowa 2005; Liu et al. 2014). The same problem can also be tackled by formally assessing the environmental impact of land-use activities such as the location of new infrastructures and by then selecting the minimum impact solution. A wide variety of applications have been developed to consider various aspects and situations related to these conservation issues (Marulli and Mallarach 2005; Liu et al. 2007; do Carmo Giordano and Riedel 2008; Geneletti 2008a, b; Geneletti and van Duren 2008; Silberman and Rees 2010; Sherrouse et al. 2011; Swetnam et al. 2011). The vast majority of these applications and case studies use the multi-layer functionality of GIS to collect data and information (quantitative and qualitative) in order to define criteria-based evaluation for prioritization and selection of potential solutions. Once the criteria have been defined and measured, multi-criteria methods are usually applied to provide the final ranking.

Waste Management It indicates the set of activities (waste reduction, reuse, recycling, composting and disposal) associated with the overall chain of managing solid waste in order to reduce its impact on the environment. The problem of solid waste management has assumed significant dimensions in modern urban centers,

partly due to the acceleration of the phenomenon of urbanization. From the logistic point of view, waste management problems include location and routing aspects. In particular, despite the efforts to reduce waste production at the source, disposal of solid waste at appropriate facilities is still a crucial need. The identification of suitable municipal waste disposal sites is a complex problem either from the technical or from the socio-political point of view. From the technical point of view the process requires environmental, health, economic and engineering considerations to be taken into account. On the other hand, waste disposal sites are a typical example of undesirable facilities, which implies that communities do not accept any feasible solutions on the basis of "not in my backyard" (NIMBY) and "not in anyone's backyard" (NIABY) principles. In this context, it has been argued that the collection of relevant spatial and non spatial information can help planners in order to include in the decision making process the points of view of the various and numerous involved stakeholders. Higgs (2006) underlined the potentiality of integrating multi-criteria approaches with GIS, in order to highlight the opportunities and challenges facing decision makers in their effort to increase the involvement of the public at different stages of the waste management process.

For this reason in the last years a significant proliferation of papers appeared in the literature, showing the ability of GIS to provide crucial support in such complex decisions, in particular in combination with multi-criteria decision making approaches. Sumathi et al. (2008) identified some advantages of applying GIS in the process of identifying appropriate waste disposal sites such as the possibility of determining zones to be excluded according to some screening criteria, performing 'what if' data analysis, investigating different potential scenarios related to population growth and area development, as well as checking the importance of the various influencing factors etc., handling and correlating large amounts of complex geographical data. They used 12 thematic maps and then employed a weighted sum aggregation function to obtain a Composite Suitability Index while the AHP approach was used to calculate relative weights. Integration of GIS with AHP was also proposed by other authors (Guiqin et al. 2009; Sharifi et al. 2009; Sener et al. 2011). A similar approach was used by Gbanie et al. (2013) who built a pair-wise comparison matrix to derive weights using the weight module in IDRISI 15.0. Nas et al. (2010) proposed an alternative approach in which each suitability criterion is represented by a factor map. Each value of a criterion is assigned a different rank, while the maps themselves receive different weights according to the importance of the corresponding criterion. On a different note, Chang et al. (2008) used a fuzzy multicriteria approach for locating waste disposal sites in an urban region. Zamorano et al. (2008) developed a method based on the use of environmental indices calculated through GIS to provide a quantitative assessment of the possible environmental interactions between a waste disposal site and potentially affected environmental components. Finally, in the context of waste management Ghose et al. (2005) proposed a GIS based transportation model for the efficient management of the daily operations for transporting solid wastes.

Energy Management The increasing interest in meeting rising energy demands in a sustainable manner through reducing energy wastes and searching for renewable energy alternatives has stimulated proposals aiming to exploit the functionality of GIS to support the relevant locational decisions. These decisions concern the identification and selection of marginal lands, i.e. lots and areas economically unprofitable due to, for instance, their poor agricultural or residential potential, where biomass production (Niblick et al. 2013), least-cost bioenergy locations (Panichelli and Gnansounou 2008; Kaundinya et al. 2013; Höhn et al. 2014), data center infrastructures (Trigueiros Covas et al. 2013), corridors for electric lines, solar and wind farms (Ramırez-Rosado et al. 2008; Janke 2010; Van Hoesen and Letendre 2010; Molina-Ruiz et al. 2011) or renewable hybrid systems (Aydin et al. 2013) can be located.

Transportation Data availability constitutes a crucial aspect in transportation planning and management. Spatial socio-economic information, historical and current data derived from users' interviews are fundamental to estimate modal-choice, trip generation and distribution among zones of a given study area. Locational decisions in this field concern the optimal positioning of new infrastructures (i.e. roads and highways, parking lots, metro and railway stations, bus lines and stops, intermodal terminals, airports, etc.). Such decisions are traditionally reached through cost-benefit analysis. However the evaluation of cost as well as benefit is often a very complex issue as, in general, many factors need to be taken into account and various effects need to be evaluated. For instance, at regional level, new transport infrastructures may boost local economy by improving productivity and competiveness, while at urban level they may produce significant modifications on the land-use activities, on environmental impact and on the real estate market. For these reasons, GIS represent fundamental tools for combining and synthesizing the multidimensional aspects of the relevant problems. Usually this is performed by calculating, through an analysis of spatial and non spatial data, a set of appropriate composite indicators to describe complex concepts such as the accessibility i.e. the opportunities available to actual and potential users to reach given places (Gutiérrez et al. 2010; Mavoa et al. 2010; Rogalsky 2010; Neutens et al. 2012) or the value of time in the intermodal transport chain (Macharis and Pekin 2009; Pekin et al. 2013). Other examples of transportation location problems in which GIS have been successfully exploited regard, for instance, the coverage of remote communities through "essential air service" (Grubesic et al. 2012), the location of bus stops (Delmelle et al. 2012), of bicycle facilities (Rybarczyk and Wu 2010; García-Palomares et al. 2012) and of hydrogen stations (Kuby et al. 2009).

Private and Public Sector Applications Applications related to GIS based approaches for solving FLPs that involve private or public sector facilities usually concern a wide variety of contexts. Typical applications include the location of emergency services (Liu et al. 2006; Murray and Tong 2009), health care facilities (Cromley and McLafferty 2012), public libraries (Park 2012; Higgs et al. 2013), schools (Teixeira and Antunes 2008; Zolnik et al. 2010), taxi cab stands (Ocalir et al. 2010) and many others. These kinds of problems can generally be defined in

terms of coverage where the main differences among the different approaches lie in the methodology adopted to estimate coverage, by combining and elaborating different data layers.

Special reference should be made to the so-called retail site location problem, consisting in determining the optimal positioning of commercial sites in a given area in order to maximize expected future profits. As this decision involves significant financial resources and risks, the selection process requires a careful analysis of the spatial distribution of the demand (geodemand) and of the competitors (geocompetition). When geographical data regarding population and its attributes, activities, private and public traffic information are available, GIS are the most powerful tool to perform geodemand and geocompetition analysis. The objective is in line with the Maximal Covering Location Problem where it is necessary to define a trade area within which it is assumed that a retailer is able to attract customers and generate sales. Among the various approaches to tackle this problem (see for instance Mendes and Themido 2004; Cheng et al. 2007), Roig-Tierno et al. (2013) proposed a method based on the use of data at the level of single city blocks and the evaluation of geocompetition by defining trade areas of competitors as a function of their facilities' size and then evaluating areas on the basis of the overlap between individual trade areas. Then they finally ranked the set of potential candidate locations by using AHP. Suárez-Vega et al. (2012) used GIS to implement a bi-objective model considering the maximization of the captured demand and the minimization of the cannibalization effect.

19.6 Conclusions

The recent overwhelming advances in Information and Communication Technologies (ICTs) have triggered a profound rethinking of scientific approaches in many fields. As usual, in processes where significant gaps and discontinuity have occurred in the use of consolidated methodologies, passionate and extensive discussions within the relevant scientific communities are generally taking place with researchers debating about the actual and substantial innovations produced by technological change in their respective field. This observation has been also evident in the vast multi and inter-disciplinary community involved in evaluating the impact produced by the development and diffusion of GIS. In particular, in the field of locational analysis, judgments regarding the actual opportunities offered by GIS to effectively solve location problems may be very different. On the one hand, more theoretically oriented researchers tend to downgrade GIS to a mere input-output tool, capable of building sophisticated databases and knowledge bases that external optimization models may use, assigning it a very limited added value as far as the methodological aspects are concerned. On the other hand, researchers and practitioners more interested in discovering opportunities to solve real problems in a more appropriate manner, emphasize the role of GIS as crucial decision support

systems which can be used to analyze new and more challenging problems. It can be argued that a position somewhere between these two extremes better represents the current state of the art and future expectations.

Location problems typically involve a large set of feasible alternatives, multiple and conflicting evaluation criteria as well as the contribution and participation of different actors (decision-makers, stakeholders, interest groups). In order to address all these issues it is necessary, if not vital, for the analyst to have access to accurate and reliable information. This requirement can be even more crucial in all those decision problems such as environmental and land-use planning problems or the location of public services, where the decision making process requires public participation, consensus building and conflict resolution. In such cases appropriate communication tools are required in order to assure that all the information related to certain complex quantitative phenomena is readily available in a user friendly representation. Through the management of multi-layered information, modern GIS are capable of providing a solid and accurate description of such real problems in terms of the variety and richness of information that can be stored and utilized. Consequently, in the near future we expect many more applications concerning GIS-based multi-criteria approaches, consisting in the transformation and combination of spatial data and decision preferences to obtain information for decision making.

Another fundamental issue is represented by the need for tools, both methodological and technological, capable of integrating, in a coherent framework, information of various nature (data, opinions, preferences) that may be also expressed in different and, apparently, incompatible languages. In this context, the widespread use of generalized devices such as smart phones, tablets, or interactive Internet-enabled televisions, may make GIS a reference platform for developing spatial decision support systems able to transform and combine data in such a way that complex problems may be modeled and then appropriately tackled. This aspect may also stress the potential of GIS as a powerful communication tool with interesting implications in institutional, political, social and ethical issues.

The technological evolution will also provide new interesting directions of research as well as opportunities for a whole range of exciting applications. Recent advances in wireless communication technologies have been adding new perspectives to technology integration which is crucial in spatial IT management. More specifically, the increasing popularity of Internet geospatial IT tools such as Google Earth and the massive availability of location-based systems (i.e. Global Positioning System) have been making available a huge amount of accurate spatial information. As a result, Location Based Services (LBS) are expected to proliferate in the near future since the information available within GIS can drive studies and applications that model new problems on the basis of the dynamic availability of customers' positions (people, vehicles, goods). In fact, as argued by Sui (2005), conventional GIS concepts may disappear and GIS functionalities may appear in a pervasive fashion when the idea of ubiquitous computing comes true.

Despite the rapid progress in the technological aspects of GIS, many of the fundamental problems of data modeling, error propagation, uncertainty management and integration with optimization tools are still open and can still represent interesting

directions for future research. More specifically, with the emergence of personalized LBS, more user friendly interfaces as well as faster and more efficient heuristics need to be developed in order to fully exploit the capabilities of GIS in the most appropriate manner.

In addition, at the moment GIS appear to be prevalently a two-dimensional technology driven by the linkage with the geographical map. Even if some approaches have been proposed for handling the third dimension, we are still far from the availability of cost-effective tools whose reliability and efficiency, especially in the function of data acquisition, can be considered comparable to the corresponding 2D versions. Although the representation of the third and, especially, the fourth dimension (time), may represent a serious challenge in terms of database organization and conceptual framework, it will surely offer further interesting opportunities to model and to analyze new location problems.

The extended functionality of modern GIS implies that a significant level of expertise is required in order to utilize their full potential. Hence, the ultimate question for location analysts still remains: is it worth investing the time and effort in GIS in the context of location science? As far as applications are concerned, the answer appears to be obviously affirmative. Furthermore, even in theoretical terms the prospects of combining location science and GIS are constantly improving as new opportunities are presented. We expect that in the years to come the two disciplines will converge even more and the prospects of their interaction will become even brighter.

References

Ahmad S (2011) GIS-based analysis and modeling with empirical and remotely-sensed data on coastline advance and retreat. Electronic theses and dissertations. Paper 446. http://scholar.uwindsor.ca/cgi/viewcontent.cgi?article=1445&context=etd

Alexandris G, Giannikos I (2010) A new model for maximal coverage exploiting GIS capabilities. Eur J Oper Res 202:328–338

Aydin NY, Kentel E, Duzgun HS (2013) GIS-based site selection methodology for hybrid renewable energy systems: a case study from western Turkey. Energy Convers Manag 70:90–106

Bender T, Hennes H, Kalcsics J, Melo MT, Nickel S (2002) Location software and interface with GIS and supply chain management. In: Drezner Z, Hamacher H (eds) Facility location: applications and theory. Springer, Berlin, pp 233–274

Bhat MA, Shah RM, Ahmad B (2011) Cloud computing: a solution to geographical information systems (GIS). Int J Comput Sci Eng 3:594–600

Bozkaya B, Yanik S, Balcisoy S (2010) A GIS-based optimization framework for competitive multi-facility location-routing problem. Netw Spat Econ 10:297–320

Brandeau ML, Chiu SS (1989) An overview of representative problems in location research. Manag Sci 25:645–674

Bruno G, Genovese A, Sgalambro A (2010) An agent-based framework for modeling and solving location problems. TOP 18:81–96

Chang NB, Parvathinathanb G, Breedenc JB (2008) Combining GIS with fuzzy multicriteria decision-making for landfill siting in a fast-growing urban region. J Environ Manag 87:139–153

Cheng T, Adepeju M (2013) Detecting emerging space-time crime patterns by prospective. In: STSS, proceedings of the 12th international conference on geocomputation. http://www.geocomputation.org/2013/papers/77.pdf. Accessed 30 Oct 2013
Cheng EWL, Li H, Yu L (2007) A GIS approach to shopping mall location selection. Build Environ 42:884–892
Chrisman NR (1999) What does "GIS" mean? Trans GIS 3:175–186
Church RL (1999) Location modeling and GIS. In: Longley PA, Goodchild MF, Maguire DJ, Rhind DW (eds) Geographical information systems. Wiley, New York
Church RL (2002) Geographical information systems and location science. Comput Oper Res 29:541–562
Church RL, Garfinkel RS (1978) Locating an obnoxious facility on a network. Transp Sci 2: 107–118
Church RL, Murray AT (2009) Business site selection, location analysis and GIS. Wiley, New York
Church RL, ReVelle C (1974) The maximal covering location problem. Pap Reg Sci Assoc 32: 101–118
Collins MG, Steiner FR, Rushman MJ (2001) Land-use suitability analysis in the United States: historical development and promising technological achievements. Environ Manag 28:611–621
Cooper L (1963) Location–allocation problems. Oper Res 11:311–343
Cowen DJ (1988) GIS versus CAD versus DMBS: what are the differences? Photogramm Eng Remote Sens 54:1551–1555
Cromley EK, McLafferty SL (2012) GIS and public health. The Guilford Press, New York
Current J, Schilling D (1990) Analysis of errors due to demand data aggregation in set covering and maximal covering location problems. Geogr Anal 22:116–126
Current J, Min H, Schilling D (1990) Multiobjective analysis of facility location decisions. Eur J Oper Res 49:295–307
Dasarathy Z, White LJ (1980) A maxmin location problem. Oper Res 32:309–325
de Smith M, Longley P, Goodchild M (2013) Geospatial analysis – a comprehensive guide to principles, techniques and software tools, 4th edn. Winchelsea Press, Winchelsea
Delmelle EM, Li S, Murray AT (2012) Identifying bus stop redundancy: a gis-based spatial optimization approach. Comput Environ Urban 36:445–455
Densham PJ (1994) Integrating GIS and spatial modeling: visual interactive modeling and location selection. Geogr Syst 1:203–219
do Carmo Giordano L, Riedel PS (2008) Multicriteria spatial decision analysis for demarcation of greenway: a case study of the city of Rio Claro, Sao Paulo, Brazil. Landsc Urban Plan 84: 301–311
Dobbins J, Jenkins L (2011) Geographic information systems for estimating coastal maritime risk. Transp Res Board 2222:17–24
Drezner Z, Wesolosky GO (1980) Single facility l_p distance minimax location. SIAM J Algebraic Discrete Methods 1:315–321
Eiselt HA, Laporte G (1995) Objectives in location problems. In: Drezner Z (ed) Facility location: a survey of applications and methods. Springer, Berlin, pp 151–180
Elzinga J, Hearn DW (1972) Geometrical solutions for some minimax location problems. Transp Sci 6:379–394
Erlenkotter D (1978) A dual-based procedure for uncapacitated facility location. Oper Res 26: 992–1009
Farahani RZ, SteadieSeifi M, Asgari R (2010) Multiple criteria location problems. Appl Math Model 34:1689–1709
FAO (2003) Geographic information systems in fisheries management and planning. FAO Fisheries technical paper 449
Francis RL, McGinnis LF, White JA (1983) Locational analysis. Eur J Oper Res 12:220–252
Francis RL, Lowe T, Tamir A (2002) Demand point aggregation for location models. In: Drezner Z, Hamacher H (eds) Facility location: application and theory. Springer, Berlin, pp 207–232
García-Palomares JB, Gutiérrez J, Latorre M (2012) Optimizing the location of stations in bike-sharing programs: a GIS approach. Appl Geogr 35:235–246

Gbanie SP, Tengbe PB, Momoh JS, Medo J, Kabba VTS (2013) Modeling landfill using geographical information system (GIS) and multi-criteria decision analysis (MCDA): case study Bo, southern Sierra Leone. Appl Geogr 36:3–12
Geneletti D (2008a) Impact assessment of proposed ski areas: a GIS approach integrating biological, physical and landscape indicators. Environ Impact Asses Rev 28:116–130
Geneletti D (2008b) Incorporating biodiversity assets in spatial planning: methodological proposal and development of a planning support system. Landsc Urban Plan 84:252–265
Geneletti D, van Duren I (2008) Protected area zoning for conservation and use: a combination of spatial multicriteria and multiobjective evaluation. Landsc Urban Plan 85:97–110
Ghose MK, Dikshit AK, Sharma SK (2005) A GIS based transportation model for solid waste disposal – a case study on Asansol municipality. Waste Manag 26:1287–1293
Goldman AJ (1971) Optimal center location in simple networks. Transp Sci 5:212–221
Gonçalves A (2010) An extension of GIS-based least-cost path modeling to the location of wide paths. Int J Geogr Inf Syst 24:983–996
Goodchild MF (1979) The aggregation problem in location–allocation. Geogr Anal 11:240–255
Goodchild MF (2010) Twenty years of progress: GIScience in 2010. J Spat Inf Sci 1:3–20
Grubesic TH, Matisziw TC, Murray AT (2012) Assessing geographic coverage of the essential air service program. Socio Econ Plan Sci 46:124–135
Guiqin W, Li Q, Guoxue L, Lijun C (2009) Landfill site selection using spatial information technologies and AHP: a case study in Beijing, China. J Environ Manag 90:2414–2421
Gutiérrez J, Condeço-Melhorado A, Martín JC (2010) Using accessibility indicators and GIS to assess spatial spillovers of transport infrastructure investment. J Transp Geogr 18:141–152
Hakimi SL (1964) Optimal locations of switching centers and the absolute centers and medians of a graph. Oper Res 12:450–459
Hakimi SL (1965) Optimal distribution of switching centers in a communication network and some related theoretic graph problems. Oper Res 13:462–475
Hamacher HW, Nickel S (1998) Classification of location models. Locat Sci 6:229–242
Heuvelink GBM (1998) Error propagation in environmental modeling with GIS. Taylor & Francis, London
Higgs G (2006) Integrating multi-criteria techniques with geographical information systems in waste facility location to enhance public participation. Waste Manag Res 24:105–111
Higgs G, Langford M, Fry R (2013) Investigating variations in the provision of digital services in public libraries using network-based GIS models. Libr Inf Sci Res 35:24–32
Höhn J, Lehtonen E, Rasi S, Rintala J (2014) A geographical information system (GIS) based methodology for determination of potential biomasses and sites for biogas plants in southern Finland. Appl Energy 113:1–10
Huff DL (1964) Defining and estimating a trading area. J Mark 28:34–38
IDRISI (2013) http://www.clarklabs.org/products/idrisi.cfm
Janke JR (2010) Multicriteria GIS modeling of wind and solar farms in Colorado. Renew Energy 35:2228–2234
Johnson MP (2001) A spatial decision support system prototype for housing mobility program planning. J Geogr Syst 3:49–67
Kaundinya DP, Balachandra P, Ravindranath NH, Ashok V (2013) A GIS (geographical information system)-based spatial data mining for optimal location and capacity planning of distributed biomass power generation facilities: a case study of Tumkur district, India. Energy 52:77–88
Kim K, Murray AT (2008) Enhancing spatial representation in primary and secondary coverage location modeling. J Reg Sci 48:745–768
Kuby M, Lines L, Schultz R, Xie Z, Kim JG, Lim S (2009) Optimization of hydrogen stations in Florida using flow-refueling location model. Int J Hydrog Energy 34:6045–6064
Li D, Zhang J, Wu H (2012) Spatial data quality and beyond. Int J Geogr Inf Sci 26:2277–2290
Liu N, Huang B, Chandramouli M (2006) Optimal siting of fire stations using GIS and ANT algorithm. J Comput Civ Eng 20:361–369
Liu Y, Lv X, Qin X, Guo H, Yu Y, Wang J, Mao G (2007) An integrated GIS-based analysis system for land-use management of lake areas in urban fringe. Landsc Urban Plan 82:233–246

Liu S, Dong Y, Deng L, Liu Q, Dong S (2014) Forest fragmentation an landscape connectivity change associated with road network extension and city expansion: a case study in the Lancang River valley. Ecol Indic 36:160–168

Macharis C, Pekin E (2009) Assessing policy measures for the stimulation of intermodal transport: a GIS-based policy analysis. J Transp Geogr 17:500–508

Maguire DJ (1991) An overview and definition of GIS. In: Maguire DJ, Goodchild MF, Rhind DW (eds) Geographical information systems: principles and applications, vol 1. Longman, Harlow, pp 9–20

Malczewski J (2004) GIS-based land-use suitability analysis: a critical overview. Prog Plan 62: 3–65

Marsh MT, Schilling DA (1994) Equity measurement in facility location analysis: a review and framework. Eur J Oper Res 74:1–17

Marulli J, Mallarach JM (2005) A GIS methodology for assessing ecological connectivity: application to the Barcelona metropolitan area. Landsc Urban Plan 71:243–262

Matisziw TC, Murray AT (2009) Siting a facility in continuous space to maximize coverage of continuously distributed demand. Socio Econ Plan Sci 43:131–139

Mavoa S, Witten K, McCreanor T, O'Sullivan D (2010) GIS based destination accessibility via public transit and walking in Auckland, New Zealand. J Transp Geogr 20:15–22

Mendes AB, Themido IH (2004) Multi-outlet retail site location assessment. Int Trans Oper Res 11:1–18

Minieka E (1970) The *m*-center problem. SIAM Rev 12:138–141

Molina-Ruiz J, Martínez-Sánchez MJ, Pérez-Sirvent C, Tudela-Serrano ML, García Lorenzo ML (2011) Developing and applying a GIS-assisted approach to evaluate visual impact in wind farms. Renew Energy 36:1125–1132

Murray AT (2003) Site placement uncertainty in location analysis. Comput Environ Urban 27: 205–221

Murray AT (2005) Geography in coverage modeling: exploiting spatial structure to address complementary partial service of areas. Ann Assoc Am Geogr 95:761–772

Murray AT (2010) Advances in location modeling: GIS linkages and contributions. J Geogr Syst 12:335–354

Murray AT, Grubesic TH (2012) Spatial optimization and geographic uncertainty: implications for sex offender management strategies. In: Johnson M (ed) Community-based operations research-decision modeling for local impact and diverse populations. Springer, New York, pp 121–142

Murray AT, Kim H (2008) Efficient identification of geographic restriction conditions in anti-covering location models using GIS. Lett Spat Resour Sci 1:159–169

Murray AT, Tong D (2009) GIS and spatial analysis in the media. Appl Geogr 29:250–259

Nas B, Cay T, Iscan F, Berktay A (2010) Selection of MSW landfill site for Konya, Turkey using GIS and multi-criteria evaluation. Environ Monit Assess 160:491–500

Neutens T, Delafontaine M, Scott DM, De Maeyer P (2012) A GIS-based method to identify spatiotemporal gaps in public service delivery. Appl Geogr 33:253–264

Niblick B, Monnell JD, Zhao X, Landis AE (2013) Using geographical information systems to assess potential biofuel crop production on urban marginal lands. Appl Energy 103:234–242

Ocalir EV, Ercoskun OY, Tur R (2010) An integrated model of GIS and fuzzy logic (FMOTS) for location decisions of taxicab stands. Exp Syst Appl 37:4892–4901

Openswaw S, Taylor PJ (1981) The modifiable areal unit problem. In: Wrigley N, Bennet R (eds) Quantitative geography: a British view. Routledge and Kegan Paul, London, pp 60–69

Panichelli L, Gnansounou E (2008) GIS-based approach for defining bioenergy facilities location: a case study in Northern Spain based on marginal delivery costs and resource competition between facilities. Biomass Bioenergy 32:289–300

Park SJ (2012) Measuring public library accessibility: a case study using GIS. Libr Inf Sci Res 34:13–21

Pekin E, Macharis C, Meers D, Rietveld P (2013) Location analysis model for Belgian intermodal terminals: importance of the value of time in the intermodal transport chain. Comput Ind 64:113–120

Phua MH, Minowa M (2005) A GIS-based multi-criteria decision making approach to forest conservation planning at a landscape scale: a case study in the Kinabalu Area, Sabah, Malaysia. Landsc Urban Plan 71:207–222

Ramırez-Rosado IJ, Garcıa-Garridoa E, Fernandez-Jimenez LA, Zorzano-Santamarıa PJ, Monteiro C, Miranda V (2008) Promotion of new wind farms based on a decision support system. Renew Energy 33:558–566

ReVelle CS, Eiselt HA (2005) Location analysis: a synthesis and survey. Eur J Oper Res 165:1–19

ReVelle CS, Swain RW (1970) Central facilities location. Geogr Anal 2:30–42

ReVelle CS, Eiselt HA, Daskin MS (2008) A bibliography for some categories in discrete location science. Eur J Oper Res 184:817–848

Ribeiro A, Antunes AP (2002) A GIS-based decision-support tool for public facility planning. Environ Plan B 29:553–569

Rogalsky J (2010) The working poor and what GIS reveals about the possibilities of public transit. J Transp Geogr 18:226–237

Roig-Tierno N, Baviera-Puig A, Buitrago-Vera J, Mas-Verdu F (2013) The retail site location decision process using GIS and the analytical hierarchy process. Appl Geogr 40:191–198

Rybarczyk G, Wu C (2010) Bicycle facility planning using GIS and multi-criteria decision analysis. Appl Geogr 30:282–293

Sener S, Sener E, Karagüze R (2011) Solid waste disposal site selection with GIS and AHP methodology: a case study in Senirkent-Uluborlu (Isparta) Basin, Turkey. Environ Monit Assess 173:533–554

Sharifi M, Hadidi M, Vessali E, Mosstafakhani P, Taheri K, Shahoie S, Khodamoradpour M (2009) Integrating multi-criteria decision analysis for a GIS-based hazardous waste landfill sitting in Kurdistan Province, western Iran. Waste Manag 29:2740–2759

Sherrouse BC, Clement JM, Semmens DJ (2011) A GIS application for assessing, mapping, and quantifying the social values of ecosystem services. Appl Geogr 31:748–760

Silberman JA, Rees PW (2010) Reinventing mountain settlements: a GIS model for identifying possible ski towns in the US rocky mountains. Appl Geogr 30:36–49

Snyder L (2006) Facility location under uncertainty: a review. IIE Trans 38:547–564

Suárez-Vega R, Santos-Peñate DR, Dorta-González P, Rodríguez-Díaz M (2011) A multi-criteria GIS based procedure to solve a network competitive location problem. Appl Geogr 31:282–291

Suárez-Vega R, Santos-Peñate DR, Dorta-González P (2012) Location models and GIS tools for retail site location. Appl Geogr 35:12–22

Sui D (2005) Will ubicomp make GIS invisible. Comput Environ Urban 29(1):361–367

Sumathi VR, Natesan U, Sarkar C (2008) GIS-based approach for optimized siting of municipal solid waste landfill. Waste Manag 28:2146–2160

Suzuki A, Okabe A (1995) Using Voronoi diagrams. In: Drezner Z (ed) Facility location: a survey of applications and methods. Springer, New York, pp 103–118

Swetnam RD, Fisher B, Mbilinyi BP, Munishi PKT, Wilcock S, Ricketts T, Mwakalila S, Balmford A, Burgess ND, Marshall AR, Lewis SL (2011) Mapping socio-economic scenarios of land cover change: a GIS method to enable ecosystem service modeling. J Environ Manag 92: 563–574

Teixeira J, Antunes AP (2008) A hierarchical location model for public facility planning. Eur J Oper Res 185:92–104

Tong D, Murray AT (2008) Maximizing coverage of spatial demand for service. Pap Reg Sci 87:479–489

Toregas C, Swain R, ReVelle C, Bergman L (1971) The location of emergency service facilities. Oper Res 19:1363–1373

Trigueiros Covas M, Silva CA, Dias LC (2013) On locating sustainable data centers in Portugal: problem structuring and GIS-based analysis. Sustain Comput Inform Syst 3:27–35

Van Hoesen J, Letendre S (2010) Evaluating potential renewable energy resources in Poultney, Vermont: a GIS-based approach to supporting rural community energy planning. Renew Energy 35:2114–2122

Veregin H (1995) Developing and testing of an error propagation model for GIS overlay operations. Int J Geogr Inf Sci 9:595–619

Vijay R, Gupta A, Kalamdhad AS, Devotta S (2005) Estimation and allocation of solid waste to bin through geographical information systems. Waste Manag Res 23:479–484

Wei H, Murray AT, Xiao N (2006) Solving the continuous space p-center problem: planning application issues. IMA J Manag Math 17:413–425

Xu K, Yi P, Kandukuri Y (2013) Location selection of charging stations for battery electric vehicles. Int J Eng Res Sci Technol 2:1–23

Zamorano M, Molero E, Hurtado A, Grindlay A, Ramos A (2008) Evaluation of a municipal landfill site in Southern Spain with GIS-aided methodology. J Hazard Mater 160:473–481

Zolnik E, Minde J, Gupta DD, Turner S (2010) Supporting planning to co-locate public facilities: a case study from Loudoun County, Virginia. Appl Geogr 30:687–696

Chapter 20
Location Problems in Telecommunications

Bernard Fortz

Abstract Telecommunications is an important area of application in combinatorial optimization. A large class of problems encountered by telecommunications operators are related to location theory. The aim of this chapter is to review recent developments in the application of location models for the design of (wired) telecommunications networks. In particular, we cover the Concentrator Location Problem, the Connected Facility Location Problem, the Regenerator Location Problem and some Ring Location problems.

Keywords Concentrator location • Connected facility location • Network design • Regenerator location • Ring location • Telecommunications

20.1 Introduction

Location problems play a central role in telecommunications network design. In this chapter, we cover a set of problems arising in wired (optical) telecommunications networks. Other location problems arise for wireless networks, such as the location of base stations or location areas planning for mobile users. For a review of these problems (and some problems in wired networks not covered here), we refer to Skorin-Kapov et al. (2006).

The design of a telecommunications network is a very difficult problem. The usual approach is to decompose a problem in three main levels (Balakrishnan et al. 1991):

1. the *long-distance* or *backbone* network that typically connects city pairs through *gateway nodes*;
2. the *inter-office* or *switching center* network within each city, that interconnects *switching centers* in different subdivisions (clusters of customers) and provides access to the gateway(s) node(s);
3. the *local access* network that connects individual subscribers belonging to a cluster to the corresponding switching center.

B. Fortz (✉)
Département d'Informatique, Université libre de Bruxelles (ULB), Brussels, Belgium
e-mail: bernard.fortz@ulb.ac.be

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_20

In the past, high speed optical fiber technologies were used mainly in the long-distance and inter-office networks, while local access networks typically used twisted pair or coaxial cable. However, the current tendency is to bring optical fiber closer and closer to the end-users, which led to the concept of *Fiber-To-The-x (FFTx)*, e.g., Fiber-To-The-Home (FTTH), Fiber-To-The-Building (FFTB), and Fiber-To-The-Cabinet (FTTC). This shift leads to new optimization problems.

At a high level, a telecommunications network can be seen as a set of equipments (computers, routers, etc.) connected by links of different capacities (e.g., fiber optic cables). The capacity of a link often depends of the cable itself (nominal capacity) but also on the equipments installed at its endpoints (network interfaces). For optical cables, for example, the nominal capacity is very high and the real limit is the number of different lightwaves that can be transmitted on the cable. This number of lightwave is in turn determined by the number of interface cards available at the endpoints. When designing telecommunications networks, the decisions to optimize can broadly be categorized as follows.

Equipment location: Due to the hierarchical structure of telecommunications networks described above, and given that a single network often uses different technologies simultaneously, an important question is to locate some particular pieces of equipment dedicated to provide the interface between different technologies or network levels. Such equipments include add-drop multiplexers, concentrators, splitters, regenerators, to name a few. These problems are the main focus of this chapter, as these are typically related to classical location problems.

Link installation: A long term planning question is to determine a set of cables connecting all nodes under some survivability criteria. In this context, the network is seen as a given set of nodes and a set of possible fiber links that have to be placed between these nodes to achieve connectivity and survivability at minimum cost. The long-term horizon considered is such that demand data are not reliable enough, and only topological aspects are considered. For reviews of these problems, see Kerivin and Mahjoub (2005) and Fortz and Labbé (2006).

Network dimensioning and routing decisions: In the mid-term horizon, given a forecast of the demand matrix for this period and the current topology of the network, the problem is to compute how the expected demands will be routed as well as the necessary capacities of the cables. In some models, the addition of new edges is allowed. These problems involve, at the same time, survivable design criteria and routing constraints. At an operational level, the focus is on routing decisions for demands arriving online. For packet-switched networks, the decisions are decentralized and each node takes the decision on the next node to visit in order to reach the packet's final destination. These decisions obey to some protocol rules, and the network operator control on the routing is indirect, possibly only by tweaking the protocol parameters (such as the arc metrics used in shortest-paths routing protocols). An in-depth treatment of these problems can be found in Pióro and Medhi (2004).

Clearly, these different decision making levels are not completely independent and should be integrated as much as possible. However, due to complexity and scalability issues, these levels are often treated separately.

Location problems in the telecommunications context appear mostly for decisions related to the placement of specific equipments into nodes of the network. These problems are closely related to hub location problems (Alumur and Kara 2008). Note that common decisions that appear in most hub location and telecommunication optimization problems are routing decisions for demands between pairs of nodes (while other location problems usually have demands associated to a single node).

In this chapter, we focus on problems that have emerged recently (over the last 5 years) in the literature, that combine network design and equipment location in the context of wired networks (typically fiber optic networks). For surveys of previous work, see, e.g., Skorin-Kapov et al. (2006). To dig further, a unified view on location and network design problems was recently proposed by Contreras and Fernández (2012). The most basic application of equipment location is the Concentrator Location Problem that we study in Sect. 20.2. Using this model as building block, we cover in Sect. 20.3 several variants of the Connected Facility Location Problem, which received much attention recently as operators are trying to bring high-capacity fiber-optic technologies closer to the customers. The Regenerator Location Problem presented in Sect. 20.4 is an example of a problem that emerged only recently as fiber optic cables have almost unlimited capacity, therefore allowing for very sparse designs, but suffer from the degradation of the signal when it travels too long distances. As its name suggests, the problem is concerned with the location of equipments that allow to regenerate the signal to ensure transmission without loss over long distances. Section 20.5 covers problems where some degree of resilience to failures is provided by the usage of rings in the topology of the network. Multi-period and network expansion problems are briefly discussed in Sect. 20.6. The last section concludes the chapter by describing some perspectives for future research on the topic.

20.2 The Concentrator Location Problem

The (capacitated) Concentrator Location Problem is probably the most basic application of equipment placement, and has received much attention in the literature, see, e.g., Pirkul (1987), Boffey (1989), Balakrishnan et al. (1991), and Klincewicz (1998). For a detailed survey of early work on the subject, see Chapter 2 in Yaman (2005). The problem is to determine the number and location of concentrators that are used to aggregate end-user demands before sending them on the backbone network. In addition, the allocation of end-users network nodes to the concentrators has to be determined, without violating the capacities of concentrators.

In this problem, the resulting network has a star-star topology, i.e., the subgraph connecting a given concentrator to its assigned end-users is a star—end-users are

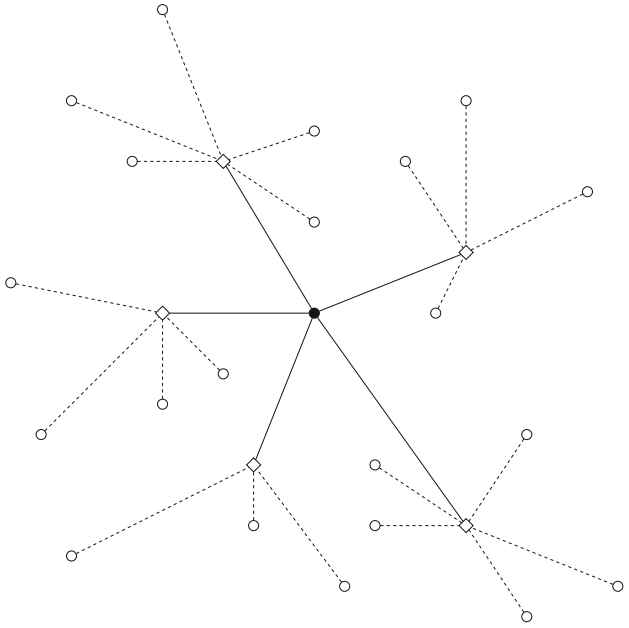


Fig. 20.1 An example of solution to the concentrator location problem

directly connected to the concentrators, and the backbone network connecting the concentrators is also a star—concentrators are directly connected to a central node. Figure 20.1 illustrates such a topology. The concentrators are represented by diamonds, round nodes represent end-user, dashed edges are connections between end-users and concentrators while plain edges form the backbone network.

Let $I = \{1, \ldots, i, \ldots, m\}$ be the set of potential concentrator locations and $J = \{1, \ldots, j, \ldots, n\}$ the set of end-users. The objective is to minimize the sum of the costs c_{ij} incurred by establishing a link between node j and a concentrator at node i, and the sum of costs f_i for installing a concentrator at node i, linked to the central node. Furthermore, we denote by d_j the demand of end-user j and by q_i the capacity of concentrator i.

Using binary variables y_i to indicate if concentrator i is open, and binary variables x_{ij} to indicate if end-user j is assigned to concentrator i, the basic version of the Concentrator Location Problem can be formulated as

$$\text{Minimize} \quad \sum_{i \in I} f_i y_i + \sum_{i \in I} \sum_{j \in J} c_{ij} x_{ij} \tag{20.1}$$

$$\text{subject to} \quad \sum_{i \in I} x_{ij} = 1 \qquad j \in J, \tag{20.2}$$

$$\sum_{j \in J} d_j x_{ij} \leq q_i y_i \qquad i \in I, \tag{20.3}$$

$$x_{ij} \leq y_i \qquad i \in I,\ j \in J, \tag{20.4}$$

$$x_{ij} \in \{0, 1\} \qquad i \in I,\ j \in J, \tag{20.5}$$

$$y_i \in \{0, 1\} \qquad i \in I. \tag{20.6}$$

Constraints (20.2) ensure that each end-user is linked to one of the concentrators. Constraints (20.3) guarantee that no concentrator will serve more end-users than its capacity. Assignment of end-users to open concentrators only is ensured by constraints (20.4). These constraints are redundant in the integer program but improve considerably the linear relaxation of the problem. Finally, constraints (20.5) and (20.6) ensure that the variables are binary.

Note that this problem is also often called the *Capacitated Facility Location Problem with Single-Source Constraints* as it becomes the classical Capacitated Facility Location Problem if the binary requirement on x_{ij} variables is relaxed.

In the context of telecommunications networks, any node is a potential concentrator location, hence $I = J$. In this case, variables y_i are sometimes replaced by x_{ii}.

Many heuristic methods have been proposed to solve that problem, starting with a Lagrangian relaxation algorithm proposed by Pirkul (1987). Holmberg et al. (1999) embedded the Lagrangian relaxation approach in a branch-and-bound framework, leading to a very efficient exact algorithm. More recently, Díaz and Fernández (2002) proposed a branch-and-price algorithm for that problem, and Contreras and Díaz (2008) developed a scatter search approach to provide upper bounds for the optimal solution of the problem.

Ceselli et al. (2009) study different variants of the problem and propose a set-partitioning reformulation that is used in a general branch-and-price framework. Gouveia and Saldanha da Gama (2006) proposed a discretized model for the unit demand case. The model is also used for an extension of the problem where facilities have different possible capacities available (the so-called unit-demand capacitated concentrator location problem with modular interfaces). They also strengthen the model with additional valid inequalities. Correia et al. (2010) also used similar discretization technique for the case of modular link capacities.

More realistic models for telecommunications have end-to-end demands, instead of aggregated demands by end-user as assumed in the model above. The resulting problems become quadratic and can be seen as special cases of the Single Allocation Hub Location Problem. Labbé and Yaman (2006) made a polyhedral analysis of formulations obtained by linearization of the quadratic terms, proposed valid inequalities to strengthen the formulation and solve it by a branch-and-cut algorithm. For the uncapacitated case, Labbé and Yaman (2008) extended their study by adding routing costs. They also did a polyhedral analysis of formulations and proposed a Lagrangian relaxation heuristic. Labbé et al. (2005b) study the variant

in which the backbone network is complete instead of a star, i.e., each pair of concentrators is connected by a direct link.

20.3 The Connected Facility Location Problem

As stated in the introduction of this chapter, telecommunications companies are currently trying to bring rapid and high-capacity fiber-optic technologies closer to the customers (FFTx networks). Outdated copper twisted cable connections are progressively replaced by fiber optic connections. The *Connected Facility Location Problem* (ConFL) aims at optimizing the building cost for networks mixing the two technologies by modeling them as *tree-star* networks: the core network, made of fiber optic connections, has a tree topology and interconnects multiplexers that switch traffic between fiber optic and copper connections. Each multiplexer is the center of a star-network of copper connections to the customers.

20.3.1 Uncapacitated Model

ConFL is a generalization of both the facility location problem and the Steiner tree problem. Formally, an undirected graph $G = (V, E)$ is given, with a set of potential locations for the facilities $I \subseteq V$ and a set of customer nodes $J \subseteq V$. An opening cost $f_i \geq 0$ is incurred for opening facility $i \in I$, each edge $e \in E$ has a cost $c_e \geq 0$, and each customer $j \in J$ has a demand d_j. The edge cost c_e, for an edge e linking a customer j to a facility i, represents the assignment cost for sending the demand of customer j to facility i. We assume the amount of demand d_j is implicitly accounted for in the assignment costs. Nodes in $S := V \setminus J$ are Steiner nodes (i.e., optional nodes that can be used to reduce the cost of the solution but do not necessarily have a facility open). This set includes the set of facilities (i.e., $I \subseteq S$). The cost c_e of an edge between two Steiner nodes represents its installation cost in the Steiner tree. When a facility node is used as pure Steiner node, no opening cost is paid for it. Optionally, a root node $r \in I$ can be given (together with its fixed location) that represents the connection to a higher order (e.g., backbone) network. That root node corresponds to an open facility that is always included in the network.

A solution (F, T) of ConFL is composed of a set of open facilities $F \subseteq I$, such that each customer $j \in J$ is assigned to an open facility $i(j) \in F$ and the open facilities are connected by a Steiner Tree T. An example of such a solution is illustrated in Fig. 20.2. Plain rounded nodes represent the customers, dashed rounded nodes are the Steiner nodes, where no facility is opened, and losanges are the opened facilities. The plain black node in the middle is the root node. The objective is to minimize the sum of assignment, facility opening and Steiner tree costs.

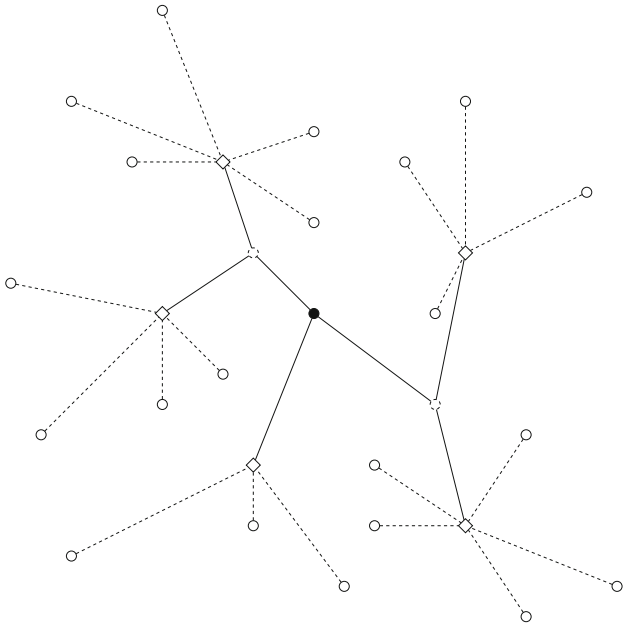


Fig. 20.2 An example of solution to the connected facility location problem

Early work on ConFL concentrated on approximation algorithms, such as the primal-dual procedures proposed by Swamy and Kumar (2004). The currently best-known constant approximation ratio is given by the 4-approximation algorithm of Eisenbrand et al. (2010). Heuristic approaches have been proposed by Ljubić (2007) and Bardossy and Raghavan (2010).

Different MIP models for ConFL were proposed and compared (both theoretically and empirically) by Gollowitzer and Ljubić (2011). As directed formulations for problems with tree topologies usually outperform undirected formulations, the problem is first converted to a directed instance by replacing each edge $e = \{i, j\} \in E$ between two Steiner nodes $i, j \in S$ by two directed arcs (i, j) and (j, i) with costs $c_{ij} = c_{ji} = c_e$, and each edge $e = \{i, j\}$ between a customer $j \in J$ and a facility $i \in I$ by a directed arc (i, j) with cost $c_{ij} = c_e$. If the problem is unrooted, an artificial root r is added to V with cost $f_r = 0$. Arcs (r, i) are added for all facilities $i \in I$ with $c_{ri} = 0$ and the number of arcs emanating from the root r is limited to 1. The resulting set of arcs is denoted by A.

According to the results reported by Gollowitzer and Ljubić (2011), the most effective formulation for solving this problem with a branch-and-cut algorithm is the cut-based formulation, proposed by Ljubić (2007), and described below. We use the following notation: $A_J = \{(i, j) \in A : i \in I,\ j \in J\}$ is the set of arcs connecting customers to facilities, while $A_S = \{(i, j) \in A : i, j \in S\}$ is the set of arcs connecting Steiner nodes. Moreover, for any $W \subset V$, we denote the incoming

and outgoing cuts induced by W as $\delta^-(W) = \{(i,j) \in A : i \notin W,\ j \in W\}$ and $\delta^+(W) = \{(i,j) \in A : i \in W,\ j \notin W\}$.

To formulate the problem, binary variables y_i are considered to indicate if facility i is open, as well as binary variables x_{ij} to indicate if arc (i,j) is used, as a Steiner tree arc if $(i,j) \in A_S$, or to assign customer j to facility i if $(i,j) \in A_J$. The cut-based model of Ljubić (2007) can then be written as:

$$\text{Minimize} \quad \sum_{i \in I} f_i y_i + \sum_{(i,j) \in A} c_{ij} x_{ij} \tag{20.7}$$

$$\text{subject to} \quad \sum_{(j,k) \in \delta^-(W)} x_{jk} \geq y_i \qquad W \subseteq S \setminus \{r\},\ i \in W \cap F \neq \emptyset, \tag{20.8}$$

$$\sum_{i:(i,j) \in A_J} x_{ij} = 1 \qquad j \in J, \tag{20.9}$$

$$x_{ij} \leq y_i \qquad (i,j) \in A_J, \tag{20.10}$$

$$y_r = 1 \tag{20.11}$$

$$x_{ij} \in \{0,1\} \qquad (i,j) \in A, \tag{20.12}$$

$$y_i \in \{0,1\} \qquad i \in I. \tag{20.13}$$

Cut constraints (20.8) impose that there is an arc entering any subset containing an open facility but not containing the root. Together, these constraints ensure that a path exists from the root to any open facility. Constraints (20.9) ensure that each customer is linked to one of the facilities, while constraints (20.10) force to assign customers to open facilities only. The root is always open by constraint (20.11). Finally, constraints (20.12) and (20.13) ensure that the variables are binary.

20.3.2 The Capacitated Connected Facility Location Problem

Recently, Gollowitzer et al. (2013) extended ConFL by considering that each facility $i \in I$ has a fixed capacity q_i, and the sum of customers' demands assigned to facility i cannot exceed such value. Moreover, each facility has a demand b_i that must be sent from the root. Each arc $(i,j) \in A_S$ has a capacity u_{ij}, and the flows routed in the Steiner tree to satisfy the demands of the facilities cannot exceed these capacities.

Note that this problem is a generalization of both ConFL and the Concentrator Location Problem. One way to formulate the problem is to explicitly introduce

continuous variables g_{ij} for $(i, j) \in A_S$, representing the amount of flow routed on arc (i, j) to satisfy the demands of the facilities.

$$\text{Minimize} \quad \sum_{i \in I} f_i y_i + \sum_{(i,j) \in A} c_{ij} x_{ij} \tag{20.14}$$

$$\text{subject to} \quad \sum_{j:(j,i) \in A_S} g_{ji} - \sum_{j:(i,j) \in A_S} g_{ij}$$

$$= \begin{cases} b_i y_i & \text{if } i \in I, \\ -\sum_{k \in I} b_k y_k & \text{if } i = r, \\ 0 & \text{otherwise,} \end{cases} \qquad i \in S, \tag{20.15}$$

$$0 \le g_{ij} \le u_{ij} x_{ij} \qquad (i, j) \in A_S, \tag{20.16}$$

$$\sum_{i:(i,j) \in A_J} x_{ij} = 1 \qquad j \in J, \tag{20.17}$$

$$\sum_{j:(i,j) \in A_J} d_j x_{ij} \le q_i y_i \qquad i \in I, \tag{20.18}$$

$$x_{ij} \le y_i \qquad (i, j) \in A_J, \tag{20.19}$$

$$x_{ij} \in \{0, 1\} \qquad (i, j) \in A, \tag{20.20}$$

$$y_i \in \{0, 1\} \qquad i \in I. \tag{20.21}$$

Constraints (20.17)–(20.21) are similar to the Concentrator Location Problem (see Sect. 20.2). Constraints (20.15) are flow conservation constraints that ensure the demands of open facilities are routed from the root node, while constraints (20.16) make sure only arcs belonging to the Steiner tree are used for satisfying the demands of the facilities, and that flows do not exceed arc capacities.

Although this formulation is compact, it is attractive from a computational point of view to project out flow variables and replace constraints (20.15) and (20.16) by the capacitated cut set inequalities (Ljubić et al. 2012):

$$\sum_{(i,j) \in \delta^-(W)} u_{ij} x_{ij} \ge \sum_{k \in F \cap W} d_k y_k \qquad W \subseteq S \setminus \{r\}.$$

These inequalities can easily be strengthened as follows:

$$\sum_{(i,j) \in \delta^-(W)} \min \left(u_{ij}, \sum_{k \in F \cap W} d_k \right) x_{ij} \ge \sum_{k \in F \cap W} d_k y_k \qquad W \subseteq S \setminus \{r\}.$$

The model can be further strengthened by adding constraints (20.8) and cut set inequalities that ensure connectivity between the root and every customer:

$$\sum_{(i,j)\in\delta^-(W)} x_{ij} \geq 1 \qquad W \subseteq V \setminus \{r\},\ W \cap J \neq \emptyset.$$

Gollowitzer et al. (2013) also proposed additional valid inequalities based on the combination of known inequalities from the literature on the facility location and network design. They study the separation problems associated to all these inequalities and show that the combination of all valid inequalities in a branch-and-cut algorithm provides an effective algorithm for solving large size, realistic instances.

20.3.3 Other Variants of the Connected Facility Location Problem

The Connected Facility Location problem is still getting much attention, and many extensions have been studied. Bardossy and Raghavan (2013) proposed a robust version of the problem based on the framework introduced by Bertsimas and Sim (2003). In particular, they proposed a heuristic based on the dual-ascent based local search for the basic ConFL problem proposed in the latter paper.

Leitner et al. (2013) proposed a model and a branch-and-cut algorithm for the Connected Facility Location with Two Architectures problem. This is an extension of the ConFL problem for networks that mix two architectures in a combined deployment (e.g., FFTH and FFTC/FFTB). Two types of facilities (one for each architecture) exist in the network. Central offices in the network are nodes where switching between the two architectures is possible. Each open facility must be connected by a path to an open central office. In addition, a certain fraction of customers (determined according to minimum coverage rates) must be served by each architecture.

Related to ConFL, Contreras et al. (2010) studied the Tree of Hubs Location Problem where exactly p hubs (facilities) must be opened, connected by a spanning tree. The major differences with ConFL lie in the fixed number of facilities to open, and the cost structure that considers routing costs explicitly. Moreover, there are no Steiner nodes in this problem. A four index formulation was also proposed in Contreras et al. (2009), leading to better lower bounds. However, this improvement comes at the price of a considerable increase in the computational time used to solve the linear relaxation of the problem. The authors therefore suggested a Lagrangian relaxation method leading to an efficient decomposition method.

20.4 The Regenerator Location Problem

Fiber optic cables used today provide large capacity, and technologies like Wave Division Multiplexing (WDM) are frequently used to increase it further by using multiple wavelengths to transmit different signals on the same fiber. In practice, this means that an optical network can transport almost unlimited amounts of bandwidth. However, optical networks face a fundamental restriction related to the geographical extent of transmission: an optical signal becomes weaker as it gets farther from its source, and therefore the distance over which an optical signal can be sent without loss or errors is limited. To overcome this limitation, regenerating can be done with some specific equipments (called regenerators) to allow the signal to be sent farther. These regenerators are located in nodes of the network.

Network design models usually ignore the distance aspects and do not deal with the placement of the regenerators. Since planners typically work in a hierarchical fashion, Chen et al. (2010) suggested to address the *Regenerator Location Problem* (RLP) at the start of the network design phase, to ensure that regenerators are placed at nodes of the network so that all nodes of the network may communicate without worry of losses due to distances.

The problem can be formally defined as follows: given an undirected network $G = (V, E)$, with a length d_e associated to each edge $e \in E$, and a maximum distance $d_{\max}$ that the signal can travel without being regenerated, find a minimum cardinality subset of nodes L such that for every pair of nodes in V, there exists a path between the nodes that contains no subpath of length $\geq d_{\max}$ without at least an internal node in L.

Chen et al. (2010) showed that the problem is NP-hard and suggested three constructive heuristics and an improvement procedure. Furthermore, they showed that RLP is equivalent to a Steiner arborescence problem with a unit degree constraint. The transformation can be summarized as follow: a new directed graph $H = (N, A)$ is created with two copies $i_1 \in N_1$ and $i_2 \in N_2$ for each node $i \in V$, and a dummy root node r. The set of nodes in H is thus $N = N_1 \cup N_2 \cup \{r\}$. The set of arcs is defined as $A = A_1 \cup A_2 \cup A_r$ where

- for each $i \in V$, there is an arc $(i_1, i_2) \in A_1$ with cost $c_{i_1 i_2} = 1$;
- there is an arc $(r, i_1) \in A_r$ from the root to each node $i_1 \in N_1$ with cost $c_{r i_1} = 0$;
- A_2 is constructed by first applying the all pairs shortest path algorithm to graph G. For edge $\{i, j\} \in E$, if the shortest path between i and j has a length $\leq d_{\max}$, then there are two arcs $(i_2, j_1), (j_2, i_1) \in A_2$ with costs $c_{i_2 j_1} = c_{j_2 i_1} = 0$. These correspond to nodes that can communicate directly without adding any concentrator.

With this transformation, a Steiner arborescence, rooted at r, with unit degree at the root node and that spans N_1, has a cost equal to the number of internal terminal nodes in the arborescence. It can be proved that these internal terminal nodes correspond to a feasible set of regenerators for RLP, and that conversely, a feasible Steiner arborescence of cost $|L|$ can be built from a solution L of RLP.

Hence the two problems are equivalent, and Chen et al. (2010) solved the RLP indirectly with a branch-and-cut algorithm for the unit degree Steiner Arborescence problem.

Introducing binary variables y_i to indicate if a node $i \in N_2$ is used in the arborescence, and binary variables x_{ij} to indicate if arc $(i,j) \in A$ is used, the problem can be formulated as:

$$\text{Minimize} \quad \sum_{(i,j)\in A} c_{ij}x_{ij} \tag{20.22}$$

$$\text{subject to} \quad \sum_{i:(i,j)\in\delta^-(\{j\})} x_{ij} = 1 \quad j \in N_1, \tag{20.23}$$

$$\sum_{i:(i,j)\in\delta^-(\{j\})} x_{ij} = y_j \quad j \in N_2, \tag{20.24}$$

$$\sum_{(i,j)\in\delta^-(W)} x_{ij} \geq 1 \quad W \subseteq N,\ W \cap N_1 \neq \emptyset, \tag{20.25}$$

$$\sum_{(i,j)\in\delta^-(W)} x_{ij} \geq y_k \quad W \subseteq N,\ k \in W \cap N_2, \tag{20.26}$$

$$\sum_{j:(r,j)\in\delta^+(\{r\})} x_{rj} = 1 \tag{20.27}$$

$$x_{ij} \in \{0,1\} \qquad (i,j) \in A, \tag{20.28}$$

$$y_i \in \{0,1\} \qquad i \in N_2. \tag{20.29}$$

Constraints (20.23) and (20.24) are degree constraints imposing that each node used in the arborescence has exactly one incoming arc, cut constraints (20.25) and (20.26) ensure connectivity, while (20.27) is the degree constraint on the root.

A weighted version of the problem was also discussed by Chen et al. (2010), where a cost w_i is associated with the placement of a regenerator in node $i \in V$, depending on the location in the network. The model above can again be used by setting the cost of arc (i_1, i_2) to w_i instead of one.

Computational results reported show the effectiveness of the branch-and-cut algorithm for small to medium size instances. For large scale instances, heuristics appear to be the only viable approach as the lower bounds provided by the branch-and-cut algorithm are really weak. Very recently, Duarte et al. (2014) proposed randomized heuristics that outperform the constructive heuristics of Chen et al. (2010).

20.5 Ring Location Problems

All the problems studied so far in this chapter impose simple connectivity in the network, i.e. only a single path is required between nodes that have to communicate with each other. However, this is clearly not sufficient to build a network resilient to failures, since a single link (or node) failure would disconnect the network. A network is said to be *survivable* if traffic interrupted by the failure of some of its elements can be rerouted via spare or excess capacity specifically placed in the network for that purpose.

It is generally considered that failures affecting more than one element at a time are extremely improbable. With the advent of SDH/SONET networks, a protection technique, known as *Self-Healing Rings*, was introduced which maintains very fast reconfiguration times while achieving low redundant capacity requirements. Demand nodes are grouped together forming a closed loop or cycle in the network. Within such a ring architecture, there always exist two link- and node-disjoint paths connecting any pair of nodes belonging to the ring. All traffic flowing through a ring is therefore protected against any single failure (as well as some multiple failures) of the links or nodes forming it by providing enough spare capacity on the alternate path of each demand.

Survivability is particularly important in the backbone network. A natural extension to the concentrator location problem is then to ensure that concentrators are interconnected through a ring structure, while customers are connected to concentrators by point-to-point links, resulting in a star topology. The resulting problem, called the *Ring Star Problem* (RSP) was first introduced and studied by Labbé et al. (2004). They model the problem as an integer program, propose several classes of valid inequalities and solve the problem with a branch-and-cut algorithm. On the heuristic side, a hybrid metaheuristic combining General Variable Neighborhood Search with a Greedy Randomized Adaptive Search Procedure was proposed by Dias et al. (2006).

In order to be consistent with the notation used throughout the chapter, we use here slightly modified notation and formulation from the ones in Labbé et al. (2004).

Formally, let $G = (V, E)$ be an undirected graph, and $r \in V$ be a given root node in that graph. An assignment cost c_{ij} is incurred for establishing a direct link between node j and concentrator node i on the cycle ($i, j \in V$), and a ring cost d_{ij} is paid for using edge $\{i, j\} \in E$ in the ring connecting the nodes chosen as concentrators. A solution of the problem is defined as

- a subset $L \subseteq V$ of nodes opened as concentrators, with $r \in L$;
- an assignment of each node in V to an open concentrator (a concentrator is assigned to itself);
- a set of edges defining a cycle going through all the open concentrators.

The objective is to minimize the sum of assignment and ring costs. Such a solution is illustrated in Fig. 20.3.

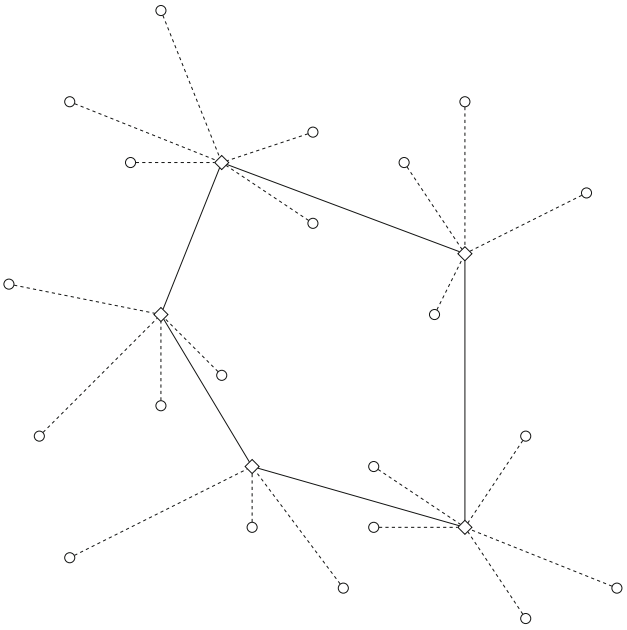


Fig. 20.3 An example of solution to the ring star problem

Using binary variables x_{ij} to indicate if node j is assigned to concentrator i, $i, j \in V$, with $x_{ii} = 1$ indicating that i is selected as concentrator, and variables z_{ij} to indicate if edge $\{i, j\} \in E$ is used in the cycle, the problem can be formulated as:

$$\text{Minimize} \sum_{i \in V} \sum_{j \in V} c_{ij} x_{ij} + \sum_{\{i,j\} \in E} d_{ij} z_{ij} \tag{20.30}$$

$$\text{subject to} \quad \sum_{i \in V} x_{ij} = 1 \qquad j \in V, \tag{20.31}$$

$$\sum_{j:\{i,j\} \in \delta(\{i\})} z_{ij} = 2x_{ii} \qquad i \in V, \tag{20.32}$$

$$\sum_{\{j,k\} \in \delta(W)} z_{jk} \geq 2 \sum_{i \in W} x_{ii} \qquad W \subset V \setminus \{r\},\ i \in W, \tag{20.33}$$

$$x_{rr} = 1 \tag{20.34}$$

$$x_{ij} \in \{0, 1\} \qquad i, j \in V, \tag{20.35}$$

$$z_{ij} \in \{0, 1\} \qquad \{i, j\} \in E, \tag{20.36}$$

where, for any $W \subset V$, $\delta(W) = \{\{i, j\} \in E : i \notin W,\ j \in W\}$ denotes the cut induced by W. Constraints (20.31) ensure that each node is linked to one of the

open concentrators. Constraints (20.32) impose a degree 2 for each concentrator in the cycle, and cut constraints (20.33) make sure the cycle is connected (i.e. disjoint cycles are avoided). Finally, constraint (20.34) imposes to open the root, and constraints (20.35) and (20.36) ensure variables are binary.

A variant of the *RSP* is the *Median Cycle Problem* (MCP), studied by Labbé et al. (2005a). In this case, the objective function only contains ring costs, i.e.

$$\text{Minimize} \sum_{\{i,j\}\in E} d_{ij} z_{ij},$$

and the problem incorporates the assignment costs as a budget constraint

$$\sum_{i\in V} \sum_{j\in V} c_{ij} x_{ij} \leq B,$$

for a given maximal budget B.

For larger scale networks, instead of connecting customers directly to concentrators, additional resilience to failures can be obtained by interconnecting customers assigned to the same concentrator through a self-healing ring connected to the backbone ring of concentrators (sometimes called the federal ring in this context). Goldschmidt et al. (2003) study a basic version of this problem, called the *SONET ring assignment problem* (SRAP). The problem can be described as a node-partitioning problem consisting of assigning nodes to local rings and interconnecting the local rings by a federal ring. The goal is to minimize the number of rings needed, while ensuring that the sum of demands between nodes in the same ring do not exceed the capacity of the ring, and also that the sum of inter-ring demands does not exceed the capacity of the federal ring. They report results with an integer programming approach as well as several heuristics for solving the problem. Note, however, that they do not consider the physical topology of the rings.

As reported by Goldschmidt et al. (2003), another drawback of this kind of designs is that many instances are infeasible. Recently, Carroll et al. (2013) studied a generalization of both the SRAP and the RSP, called the *Ring Spur Assignment Problem* (RSAP). In this problem, the objective is to design a set of bounded disjoint local rings that are interconnected by a federal ring, like in the SRAP. The topology of the rings must also be determined. Since no SRAP solution exist in some real world instances, locations that have insufficient spare capacity or no possible physical route due to limitations of geography can be connected to local rings by spurs off the local rings. Spur nodes must be connected to a local ring by a single edge, like in the RSP. Carroll et al. (2013) proposed a formulation for the problem that combines blocks of constraints similar to the formulation above for the RSP for each local ring, plus additional constraints linking the local rings together and defining the federal ring. We do not reproduce the formulation here due to its complexity and large size. Carroll et al. (2013) also proposed some valid inequalities and solved the problem with a branch-and-cut algorithm. To the best

of our knowledge, no heuristic algorithm has been proposed to tackle large-size instances of the problem.

20.6 Network Expansion and Multi-Period Problems

Models covered so far in this chapter consider telecommunications networks that are built from scratch. This static (single-step) setting is not realistic in all situations, and network operators are sometimes faced to the upgrading of an existing network. As stated in the introduction of this chapter, a network is often composed of a backbone network, for the transfer of large volumes of data, and local access networks that connect terminals to an access node of the backbone network. Network capacity expansion problems for backbone networks have been studied since the pioneering work of Balakrishnan et al. (1991).

For local access networks, a basic model, in which growing demand can be satisfied by expanding cable capacities and/or installing concentrators in the network, was introduced by Balakrishnan et al. (1995). They showed that the problem is NP-hard and proposed a Lagrangian-based decomposition heuristic to solve it. Flippo et al. (2000) later showed that the problem is weakly NP-hard and presented a pseudo-polynomial dynamic programming algorithm. A multi-period expansion problem for the particular case of a local access network that has a tree topology was solved heuristically by Gendreau et al. (2006). Gourdin and Klopfenstein (2008) studied a multi-period capacitated problem with modular concentrators and link capacities. They proposed an integer linear model, and after a polyhedral analysis of the problem, presented some facet-defining inequalities.

20.7 Conclusions

This chapter described some applications of location problems in telecommunications. Some of these problems have been studied only recently, and advances in understanding their structure to provide better exact algorithms are still necessary. Furthermore, these problems have been mostly studied in their uncapacitated versions. Since capacitated versions are usually much harder to solve, it is expected that they will attract more interest in the near future. Additionally, demands in the telecommunications industry fluctuate a lot, and the development of robust optimization approaches has emerged as a hot topic. The development of robust counterparts for the problems presented in this chapter is therefore another important trend for future research. If probability distributions can be associated to demand and/or failure scenarios, stochastic programming approaches might be used, although the literature on the subject is currently very limited. Finally, advances in heuristics are still expected for some of these problems, in order to tackle very large scale instances.

Acknowledgements The work of Bernard Fortz is supported by the Interuniversity Attraction Poles Program of the Belgian Science Policy Office.

References

Alumur S, Kara BY (2008) Network hub location problems: the state of the art. Eur J Oper Res 190:1–21

Balakrishnan A, Magnanti T, Wong R (1995) A decomposition algorithm for local access telecommunications network expansion planning. Oper Res 43:58–76

Balakrishnan A, Magnanti T, Shulman A, Wong R (1991) Models for planning capacity expansion in local access telecommunication networks. Ann Oper Res 33:239–284

Bardossy M, Raghavan S (2010) Dual-based local search for the connected facility location and related problems. INFORMS J Comput 22:584–602

Bardossy MG, Raghavan S (2013) Robust optimization for the connected facility location problem. Electron Notes Discrete Math 44:149–154

Bertsimas D, Sim M (2003) Robust discrete optimization and network flows. Math Program 98:49–71

Boffey TB (1989) Location problems arising in computer networks. J Oper Res Soc 40:347–354

Carroll P, Fortz B, Labbé M, McGarraghy S (2013) A branch-and-cut algorithm for the ring spur assignment problem. Networks 61:89–103

Ceselli A, Liberatore F, Righini G (2009) A computational evaluation of a general branch-and-price framework for capacitated network location problems. Ann Oper Res 167:209–251

Chen S, Ljubić I, Raghavan S (2010) The regenerator location problem. Networks 55:205–220

Contreras I, Díaz J (2008) Scatter search for the single source capacitated facility location problem. Ann Oper Res 157:73–89

Contreras I, Fernández E (2012) General network design: a unified view of combined location and network design problems. Eur J Oper Res 219:680–697

Contreras I, Fernández E, Marín A (2009) Tight bounds from a path based formulation for the tree of hub location problem. Comput Oper Res 36(12):3117–3127

Contreras I, Fernández E, Marín A (2010) The tree of hubs location problem. Eur J Oper Res 202(2):390–400

Correia I, Gouveia L, Saldanha da Gama F (2010) Discretized formulations for capacitated location problems with modular distribution costs. Eur J Oper Res 204(2):237–244

Dias TCS, Sousa Filho GF, Macambira EM, Anjos L, Cabral L, Fampa MHC (2006) An efficient heuristic for the ring star problem. In: Àlvarez C, Serna M (eds) Experimental algorithms. Lecture notes in computer science, vol 4007. Springer, Berlin/Heidelberg, pp 24–35

Díaz J, Fernández E (2002) A branch-and-price algorithm for the single-source capacitated plant location. J Oper Res Soc 53:728–470

Duarte A, Martí R, Resende M, Silva R (2014) Improved heuristics for the regenerator location problem. Int Trans Oper Res 21:541–558

Eisenbrand F, Grandoni F, Rothvoß T, Schäfer G (2010) Connected facility location via random facility sampling and core detouring. J Comput Syst Sci 76:709–726

Flippo OE, Kolen AW, Koster AM, van de Leensel RL (2000) A dynamic programming algorithm for the local access telecommunication network expansion problem. Eur J Oper Res 127:189–202

Fortz B, Labbé M (2006) Design of survivable networks. In: Resende M, Pardalos P (eds) Handbook of optimization in telecommunications, Chap. 15. Springer, New York, pp 367–389

Gendreau M, Potvin JY, Smires A, Soriano P (2006) Multi-period capacity expansion for a local access telecommunications network. Eur J Oper Res 172(3):1051–1066

Goldschmidt O, Laugier A, Olinick EV (2003) SONET/SDH ring assignment with capacity constraints. Discret Appl Math 129:99–128

Gollowitzer S, Ljubić I (2011) MIP models for connected facility location: a theoretical and computational study. Comput Oper Res 38:435–449
Gollowitzer S, Gendron B, Ljubić I (2013) A cutting plane algorithm for the capacitated connected facility location problem. Comput Optim Appl 55:647–674
Gourdin É, Klopfenstein O (2008) Multi-period capacitated location with modular equipments. Comput Oper Res 35:661–682
Gouveia L, Saldanha da Gama F (2006) On the capacitated concentrator location problem: a reformulation by discretization. Comput Oper Res 33:1242–1258
Holmberg K, Rönnqvist M, Yuan D (1999) An exact algorithm for the capacitated facility location problems with single sourcing. Eur J Oper Res 113:544–559
Kerivin H, Mahjoub A (2005) Design of survivable networks: a survey. Networks 46:1–21
Klincewicz JG (1998) Hub location in backbone/tributary network design: a review. Locat Sci 6:307–335
Labbé M, Yaman H (2006) Polyhedral analysis for concentrator location problems. Comput Optim Appl 34:377–407
Labbé M, Yaman H (2008) Solving the hub location problem in a star-star network. Networks 51:19–33
Labbé M, Laporte G, Rodríguez-Martín I, Salazar-González JJ (2004) The ring star problem: polyhedral analysis and exact algorithm. Networks 43:177–189
Labbé M, Laporte G, Rodríguez-Martín I, Salazar-González JJ (2005a) Locating median cycles in networks. Eur J Oper Res 160:457–470
Labbé M, Yaman H, Gourdin É (2005b) A branch and cut algorithm for hub location problems with single assignment. Math Program 102:371–405
Leitner M, Ljubić I, Sinnl M, Werner A (2013) On the two-architecture connected facility location problem. Electron Notes Discrete Math 41:359–366
Ljubić I (2007) A hybrid vns for connected facility location. Lect Notes Comput Sci 4771:157–169
Ljubić I, Putz P, Salazar-González JJ (2012) Exact approaches to the single-source network loading problem. Networks 59:89–106
Pióro M, Medhi D (2004) Routing, flow, and capacity design in communication and computer networks. In: The Morgan Kaufmann series in networking. Elsevier, San Francisco
Pirkul H (1987) Efficient algorithms for the capacitated concentrator location problem. Comput Oper Res 14:197–208
Skorin-Kapov D, Skorin-Kapov J, Boljunčić V (2006) Location problems in telecommunications. In: Resende M, Pardalos P (eds) Handbook of optimization in telecommunications. Springer, New York, pp 517–544
Swamy C, Kumar A (2004) Primal-dual algorithms for connected facility location problems. Algorithmica 40:245–269
Yaman H (2005) Concentrator location in telecommunications networks. In: Combinatorial optimization, vol 16. Springer, New York

Chapter 21
Location Problems in Healthcare

Evrim Didem Güneş and Stefan Nickel

Abstract In this chapter, we discuss facility location problems arising in the context of healthcare. We concentrate on three main areas: the most classical one is healthcare facility location which is closely related to public facility location. Secondly, we look at ambulance planning which includes ambulance location and relocation problems. In the last part, we give an overview of hospital layout problems. For all three parts, we state some important models and give an overview of relevant literature as well as current research directions. A comprehensive reference list is included at the end of the chapter.

Keywords Ambulance location • Healthcare facility location • Layout problems • Public facility location

21.1 Introduction

The ageing society together with a high cost pressure on the healthcare sector brings methods from operations research in a quite prominent place. From a perspective of facility location, healthcare applications bring together different models from location theory and moreover, they give raise to new models as we will see in this chapter.

One of the most discussed topics in healthcare is the equal access to health services and a high level of health protection at the same time which is a universal and ageless problem. This leads to the first topic that we deal with in this chapter: Sect. 21.2 is devoted to healthcare facility location; we review the literature and present some classical models in that area. The reader needs some basic knowledge on discrete facility location problems as discussed in Chaps. 2–5 of Part I of this book. Another crucial issue in healthcare is the time interval between an emergency

E.D. Güneş (✉)
Koç University, Istanbul, Turkey
e-mail: egunes@ku.edu.tr

S. Nickel
Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
e-mail: stefan.nickel@kit.edu

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_21

call and the delivery of the patient to an appropriate health service provider. We devote Sect. 21.3 to ambulance location problems which integrate aspects from covering location models, multi-period location models, and location problems under uncertainty. In these problems, the influence of local law regulation on constraints and objective functions is quite remarkable as well. The third and last topic we are dealing with in this chapter concerns layout problems in hospitals. As a result of a good layout, hospitals prepare themselves for changes in the structure of patient groups and the mix of medical cases as well as for a trend from surgery-centered care to chronic disease care. In Sect. 21.4, basic models are presented and modern trends are discussed, such as the inclusion of multiple floors, multiple objectives or uncertainty. At the end of this chapter, the reader will find some conclusions and a comprehensive list of references.

21.2 Healthcare Facility Location

In this section, we focus on applications of discrete network location problems to health facilities. Such facilities involve community health clinics, primary care centers, public and private hospitals, or specialized clinics. The problems are therefore closely related to public facility location. We do not discuss continuous location models. In the literature, there are only a few papers applying such models; see, e.g., Dokmeci (1977, 1979).

Location of healthcare facilities can be a critical decision for developing countries since they have scarce resources and the majority of their population living in rural areas. The low population density in these regions makes the provision of health services a challenge. Within this context, location-allocation models can therefore be successfully applied for the design of health facility networks. One of the earliest applications is due to Gould and Leinbach (1966) who considered locating hospitals and determining their capacities in Western Guatemala. For an extensive review of such applications, see Rahman and Smith (2000); for a review on healthcare facility location problems, see Daskin and Dean (2004).

In the following, we give an overview of health facility location applications by first discussing the relevant objective functions and then presenting important aspects of these problems with examples from the literature.

21.2.1 Objective Functions in Healthcare Facility Location

Healthcare facility location problems are inherently multi-objective since there are different stakeholders and the facilities are predominantly public. The decisions affect health consumers and healthcare providers as well as the public community. These three sectors can have different priorities and utility functions. For example, consumers are influenced by the travel cost and time, quality of service, comfort and

convenience of the facility, waiting time at the facility, and the cost of service. On the other hand, providers are influenced by setup and operating costs, travel costs for the staff, and availability of supporting facilities (Calvo and Marks 1973). From the community perspective, equity in access among different districts is an important issue. Moreover, workload equity can be a concern for healthcare staff. Notice also that some of these factors are very difficult to quantify and measure. Consequently, the literature focuses on a few of these criteria. Relevant objectives most commonly applied in the healthcare facility location literature are the following:

- *Minimize access cost for health consumers.* This cost type can be defined as travel costs, distance, or travel time from a population district to a health facility, weighted by the population size of that district. When this is the only objective, the standard p-median formulation is commonly used for deciding where to locate a set of health facilities. The following function may represent access cost:

$$\sum_{i \in I} \sum_{j \in J} d_j c_{ij} x_{ij}, \tag{21.1}$$

  where I is the set of potential locations for the facilities, J is the set of populations or districts to serve, d_j corresponds to the population size in district $j \in J$, c_{ij} represents the distance between location $i \in I$ and district $j \in J$, and x_{ij} is a binary decision variable that is equal to 1 if the population in district j is served from a facility at location i and 0 otherwise.
- *Maximize population with access to a health facility, or maximize covered demand.* A cover type objective assumes that a population in a district is covered (has access) only if it can be assigned to a facility within a pre-determined maximum distance, and aims to maximize the covered population. Such a type of objective is appropriate to locate emergency medical services or primary care centers for under-served populations. When the objective is to minimize the total access cost, some districts can have very high access costs. Cover type objectives overcome such undesirable solutions.

  Some health services, such as preventive care, are not perceived as essential by the consumers. However, providing these services is an important public health goal. Therefore, maximizing the utilization of health facilities is another cover related objective that was first defined by Calvo and Marks (1973). There are several socio-economical factors that affect health service utilization, such as income, age, insurance coverage of the population, and convenience and proximity of the facilities (Institute of Medicine 1993). Location models are best suited to account for the "proximity of the facilities" among these factors. Zhang et al. (2009) introduced the concept of "participation" which they measure using a decreasing function of travel time plus waiting time. In that paper, the goal was to maximize participation as opposed to coverage. Güneş et al. (2014) defined participation as a decreasing function of distance, and solved models aiming at maximizing coverage and participation for a primary care network design problem.

A simple participation function can be defined as follows: $\sigma_{ij} = 1 - c_{ij}/c_{\max}$ if c_{ij} is less than or equal to $c_{\max}$ and $\sigma_{ij} = 0$ otherwise, where $c_{\max}$ is the predetermined maximum distance between a facility $i \in I$ and a district $j \in J$ that can be covered by that facility. The total weighted participation function is the following:

$$\sum_{i \in I} \sum_{j \in J} d_j \sigma_{ij} x_{ij}. \tag{21.2}$$

- *Maximize equity in access.* There is an increasing interest in incorporating equity in healthcare facility location applications. Nevertheless, there is no agreement on how to define equity, and various definitions have been used in the literature. For a review of these definitions, see Marsh and Schilling (1994). Commonly used equity objectives are: minimize the maximum distance that patients must travel (Mitropoulos et al. 2006; Güneş et al. 2014), minimize deviations from a standard distance (Smith et al. 2009, 2013), minimize differences of utilization from a national norm (Oliveira and Bevan 2006), or minimize standard deviation of the distribution of the allocated populations to healthcare facilities (Güneş et al. 2014).

All of these objectives are important, and it may be difficult to choose one in realistic applications. Therefore, multi-criteria models have gained popularity in recent years. We note that the equity criterion is commonly considered in combination with the efficiency (access) criterion since the equity objectives alone can produce undesirable solutions (Smith et al. 2013). The reader can refer to Mayhew and Leonardi (1982), Cho (1998), Mitropoulos et al. (2006), and Smith et al. (2009, 2013) for examples on applications with bi-criteria equity-efficiency objectives. Stummer et al. (2004) developed a multi-objective model to determine the size and location of departments in facilities within a given network of hospitals. The objectives considered are: minimize total access cost for patients, minimize total cost of the network, minimize number of patients rejected due to low capacity, and minimize total number of changes required in the network. Güneş et al. (2014) considered the objectives of minimizing access cost for patients, maximizing coverage, maximizing participation, and maximizing equity among physicians.

A common solution approach in multi-criteria problems is to construct efficient solution sets to inform decision makers (cf. Stummer et al. 2004; Smith et al. 2013; Güneş et al. 2014). In bi-criteria problems, the efficient frontier can be found by solving the problem with one of the objectives and then including the obtained result for the objective value as a constraint while solving for the second objective (cf. Ehrgott 2005; Smith et al. 2013). Another approach, which is not restricted to the bi-criteria case, is to include all criteria in the objective function with different weights. For example, Bruni et al. (2006) modeled the location of transplant centers considering distance, waiting list, and maximum waiting list (as a proxy for equity) with different weights in the objective.

21.2.2 An Overview of Healthcare Facility Location Models

The classical p-median problem seeks for the optimal location of p facilities to minimize a demand-weighted cost of access (or equivalently distance, or time) for the population residing at the nodes of the network (see Chap. 2 for a detailed discussion of this problem). Therefore, the problem that consists of deciding where to locate a set of primary care facilities, such as community clinics or family centers, or hospitals, is often casted as a p-median problem. Assuming, as before, that I denotes the set of potential locations for the facilities and J the set of districts or populations to serve, the basic formulation is as follows:

$$\text{minimize} \quad \sum_{i \in I} \sum_{j \in J} d_j c_{ij} x_{ij} \tag{21.3}$$

$$\text{subject to} \quad \sum_{i \in I} y_i = p \tag{21.4}$$

$$\sum_{i \in I} x_{ij} = 1 \qquad \forall j \in J \tag{21.5}$$

$$x_{ij} \leq y_i \qquad \forall i \in I, j \in J \tag{21.6}$$

$$y_i \in \{0, 1\} \qquad \forall i \in I \tag{21.7}$$

$$x_{ij} \in \{0, 1\} \qquad \forall i \in I, j \in J, \tag{21.8}$$

where d_j is the population in district j, c_{ij} is the distance between location i and district j, x_{ij} is a binary variable equal to 1 if the population in district j is served from the facility at location i and 0 otherwise, y_i is a binary variable equal to 1 if a facility is opened at location i and 0 otherwise, and p is the total number of facilities to open.

The formulation assumes an unlimited capacity for each facility which is rarely the case in practice. Therefore, most practical applications use a capacitated formulation by adding the following constraint:

$$\sum_{j \in J} d_j x_{ij} \leq q_i \qquad \forall i \in I, \tag{21.9}$$

where q_i is the exogenous capacity of the facility at node i. In some situations, facility capacities can also be decision variables. This can be modeled by incorporating the cost associated with building capacity in the objective function.

21.2.2.1 Modeling Capacity

Explicit modeling of capacity decisions is facilitated by a resource-based view of facilities. For example, the capacity of a health center is determined by the number of physicians assigned to that clinic. Similarly, the number of beds is a significant determinant for hospital capacity. Many healthcare facility location models consider the amount of resources in facilities also as decision variables. For example, Güneş and Yaman (2009) modeled the resource re-allocation problem for a hospital network with beds as resources. Oliveira and Bevan (2006), Griffin et al. (2008), Zhang et al. (2009, 2010) and Güneş et al. (2014) modeled the staff in each facility as a decision variable. In addition, these models can incorporate the decision about the services to offer in each facility (cf. Oliveira and Bevan 2006; Griffin et al. 2008). With R denoting the set of resource types and S the set of service types, such a model can be built by defining resource sets $R_s \subseteq R$ required to serve demand for service $s \in S$. To this end, let κ_{sr} be the amount of resource r that is utilized to serve a patient requiring service s. Then, the decisions concerning the capacity (number of patients that can be served) for service s in location i, q_{is}, and the amount of resource r in location i, w_{ri}, are modeled by the following constraints:

$$\sum_{s \in S: r \in R_s} \kappa_{sr} q_{is} \leq w_{ri} \quad \forall i \in I, r \in R \tag{21.10}$$

$$\sum_{j \in J} d_j x_{ijs} \leq q_{is} \quad \forall i \in I, s \in S \tag{21.11}$$

$$\sum_{i \in I} x_{ijs} = 1 \quad \forall j \in J, s \in S, \tag{21.12}$$

where x_{ijs} is a binary variable defining the assignment of patients for service s from district j to the facility at location i.

In some cases, there may be restrictions on the minimum number of patients assigned to a facility. In general, such restrictions are motivated by economies of scale arguments. For healthcare services, there may also be regulations on minimum number of patients assigned to a physician because for some specialties (such as mammography interpretation or surgery), regular practice is important to maintain high service quality. See Verter and Lapierre (2002), Güneş and Yaman (2009), Mestre et al. (2012), Güneş et al. (2014) for examples on how to incorporate such type of constraints.

21.2.2.2 Assumptions on Allocation

The classical p-median formulation assumes that when $x_{ijs} = 1$, all the population in district j is served from the facility at location i for service s. This single assignment assumption may be appropriate when it is desired to provide the same

service for all patients in a location. However, in case of capacity constrained systems, this may not be a reasonable assumption since the capacity of a facility may not be sufficient to serve large population centers. In that case, multiple assignment can be modeled by redefining the variable x_{ijs} as the number of patients from district j assigned to location i for service s, and by changing the assignment constraint (21.12) as follows:

$$\sum_{i \in I} x_{ijs} = d_j \qquad \forall j \in J, s \in S. \tag{21.13}$$

Notice that these models do not account for preferences of patients in different locations, while healthcare facilities are utilized by consumers who may have discretion on which one to patronize. A common approach to incorporate these preferences is to use *closest assignment constraints* in order to ensure that each population will patronize its assigned facility, assuming that the closest facility is the most preferred one (cf. Verter and Lapierre 2002). The following set of constraints can be added to model (21.3)–(21.8) (see, e.g., Canovas et al. 2007; Güneş et al. 2014):

$$\sum_{k \in I : c_{kj} > c_{ij}} x_{kj} + y_i \leq 1 \qquad \forall i \in I, j \in J. \tag{21.14}$$

These constraints ensure that for a given zone $j \in J$, if a facility at location $i \in I$ is open, then j is not assigned to any facility whose distance to j is more than the distance between j and i. For other examples of closest assignment constraints in a healthcare context see Verter and Lapierre (2002), Smith et al. (2009, 2013).

21.2.2.3 Assumptions on Demand and Patient Choice

The problem of locating healthcare facilities is characterized by various complexities due to the central presence of the human element in the system. Consequently, the demand for health services is uncertain and its estimation is not trivial since there are various relevant factors influencing it, such as disease prevalence, insurance coverage, demographics, and accessibility of the facilities. Therefore, there is a need for a better understanding of the patient behavior and preferences, and for incorporating them in location models.

Parker and Srinivasan (1976) were the first authors incorporating consumer preferences. Their model was built for expanding a rural primary care facility network. They estimated the benefit of a patient when getting service from a facility as a function of several attributes, such as distance, waiting time, time to get an appointment, and the type of facility. In that paper, the total benefit was maximized using an iterative procedure which finds the equilibrium allocation. Some recent papers investigate models that include demand estimation. For example, Griffin et al. (2008) embedded statistical estimation of demand for community health clinics.

Cardoso et al. (2012) proposed a simulation model based on a short term decision tree and long term Markov model in order to predict annual demand for long term care services over the next few years.

Location-allocation models are commonly used for healthcare facility planning. In some applications, the assumption that some patients will patronize the designated facility may be realistic. This may be forced by regulations dictating that patients must be served from the facilities they are assigned to. However, in many health service systems, patients have free choice of where to get service from. If this is the case, then a *user-choice model* defining patient behavior should be considered. One approach is to assume that patients patronize each facility with a certain probability that depends on its location as well as on other relevant factors. For example, Oliveira and Bevan (2006) used a gravity model to define the probability that patients in some district or region choose some hospital.

An alternative approach is to assume that patients patronize their first choice given by an optimization model. It is common to assume that patients patronize the closest facility, i.e., to use the closest assignment constraints in (21.14). However, although the distance to a facility is very important, it is not the only factor influencing the choice of users. In fact, the waiting time at a facility is another important factor that can be considered. Capturing congestion and its effects on patient preferences is an interesting aspect to improve realism in healthcare facility location models. In this case, the number of people using a facility determines the waiting time at the facility. Since waiting time, in turn, affects the number of people using the facility, models should incorporate equilibrium constraints. In the equilibrium, allocation should ensure that patients are assigned to their best choice and do not want to switch facilities. One such example was proposed by Chao et al. (2003) where resource allocation decisions for a public hospital network are made in order to minimize the waiting time at the facilities. The resulting allocation is incentive compatible, i.e., it is also optimal from the perspective of the patients. Zhang et al. (2009) modeled the location of preventive healthcare facilities where patients choose the facility with minimum total service time. The latter is defined as the sum of travel time and waiting time at the facility. In turn, the waiting time at a facility can be modeled using steady-state equations found in queuing theory. The resulting formulation proposed by Zhang et al. (2009) is highly nonlinear and a heuristic approach was suggested in that paper. Zhang et al. (2010) proposed a bi-level model with equilibrium constraints for a preventive healthcare facility network design problem. The solution approach uses a gradient projection method and a tabu search heuristic.

21.2.2.4 Assumptions on Facility Types and Patient Flows: Hierarchical Models

In most countries, healthcare systems are organized in hierarchical structures. There are different types of facilities, such as physicians' offices, community health centers, specialty clinics, and general hospitals. Notice that there is a hierarchy

in the services offered by these facilities. For instance, a hospital can usually provide all the services offered in a clinic. Moreover, some health systems require a referral from a general practitioner before a patient can ask for service at a hospital. Hierarchical location models can incorporate such characteristics. Şahin and Süral (2007) provided a comprehensive review on hierarchical systems with a discussion of modeling approaches and applications.

Hierarchical systems are commonly classified as successively inclusive or exclusive: in a *successively inclusive hierarchy*, a facility at some level provides all the services offered by lower level facilities (Calvo and Marks 1973; Narula 1984). This is a typical structure for healthcare facilities. Conversely, a *successively exclusive hierarchy* implies that facilities at each level offer a service that is unique to that level (Tien et al. 1983). This is the case for specialized service facilities. We now assume that $I = J = \{1, \ldots, n\}$, i.e., in each district $j \in J$ there is exactly one potential location $i \in I$ for a facility. The formulation provided by Calvo and Marks (1973) for a successively inclusive hierarchy is an assignment based p-median type model with an objective function that quantifies the total distance traveled:

$$\text{minimize} \quad \sum_{i \in I} \sum_{j \in J} c_{ij} \sum_{s \in S} d_{js} x_{ijs} \tag{21.15}$$

$$\text{subject to} \quad \sum_{i \in I} x_{ijs} = 1 \qquad \forall j \in J, s \in S \tag{21.16}$$

$$x_{iis} \geq x_{ijs} \qquad \forall i \in I, j \in J, s \in S \tag{21.17}$$

$$\sum_{i \in I} x_{iik} = \sum_{s \in S : s \geq k} p_s \qquad \forall k \in S \tag{21.18}$$

$$x_{ijs} \in \{0, 1\} \qquad \forall i \in I, j \in J, s \in S. \tag{21.19}$$

where c_{ij} is the distance between location i and district j, x_{ijs} is a binary variable equal to 1 if individuals residing in district j that require service type s are assigned to location i and 0 otherwise, d_{js} is the number of individuals residing in district j and requiring service type s, and p_s is the number of facilities offering type s services to be located. Constraints (21.16) ensure that all districts are assigned to a facility for all services. Constraints (21.17) ensure that assignments are done to open facilities only, and constraints (21.18) specify the possible number of self-assignments (i.e., the assignment of the groups of individuals residing at a location to the facility at that location). Finally, constraints (21.19) are the variable domain constraints.

Narula and Ogbu (1979) developed a two-level hierarchical model with an approach based on network flows where p_1 health centers (level $s = 1$) and p_2 hospitals (level $s = 2$) are to be located among the population centers, and a proportion of patients, θ, at health centers are transferred to hospitals. In each location, at most one facility type can be located. y_{is} is a binary variable equal to 1 if a facility of service type s is located in location i. x_{ij}^{0s} is the number of patients

from district j allocated to a facility of type s located at i; x_{ij}^{12} is the number of patients that are transferred from a health center in location i to a hospital in location j. Finally, q_s is the exogenous capacity of a facility with service type s, c_{ij} is the minimum distance between locations i and j, and d_j is the number of individuals of population j. Then a mixed-integer programming formulation to minimize total distance traveled is as follows:

$$\text{minimize} \quad \sum_{i\in I}\sum_{j\in J} c_{ij}(x_{ij}^{01} + x_{ij}^{02} + x_{ij}^{12}) \tag{21.20}$$

$$\text{subject to} \quad \sum_{i\in I}(x_{ij}^{01} + x_{ij}^{02}) = d_j \qquad \forall j \in J \tag{21.21}$$

$$\sum_{i\in I} x_{ij}^{12} = \theta \sum_{i\in I} x_{ij}^{01} \qquad \forall j \in J \tag{21.22}$$

$$\sum_{j\in J} x_{ij}^{01} \leq q_1 y_{i1} \qquad \forall i \in I \tag{21.23}$$

$$\sum_{j\in J}(x_{ij}^{02} + x_{ij}^{12}) \leq q_2 y_{i2} \qquad \forall i \in I \tag{21.24}$$

$$\sum_{i\in I} y_{is} = p_s \qquad s \in S \tag{21.25}$$

$$y_{i1} + y_{i2} \leq 1 \qquad \forall i \in I \tag{21.26}$$

$$0 \leq x_{ij}^{01} \leq d_j \qquad \forall i \in I, j \in J \tag{21.27}$$

$$0 \leq x_{ij}^{02} \leq d_j \qquad \forall i \in I, j \in J \tag{21.28}$$

$$0 \leq x_{ij}^{12} \leq \theta q_1 \qquad \forall i \in I, j \in J \tag{21.29}$$

$$y_{is} \in \{0, 1\} \qquad \forall i \in I, s \in S. \tag{21.30}$$

Narula and Ogbu (1979) proposed heuristic procedures for tackling this model.

Some examples of hierarchical facility location models include Hodgson (1988) for primary care facilities, Smith et al. (2009, 2013) for community health facilities, and Mestre et al. (2012) for regional and central hospitals. Typically, these models can be solved by commercial solvers. Galvão et al. (2002) applied a three-level hierarchical model for the delivery of perinatal care in the municipality of Rio de Janeiro with service referrals, and Galvão et al. (2006) extended this model to include capacitated facilities. The increased complexity of the models motivated the use of Lagrangian relaxation based procedures.

21.2.2.5 Modeling Dynamic Aspects of Location Decisions

A majority of health facility location applications discussed in this section assume a static environment: demand is known and fixed, and facilities are static. These assumptions may be realistic for short term planning problems. However, facility location decisions are often made at a strategic level with a long term impact. Therefore, if changes in the demand or in other relevant parameters are expected in the long term, then multi-period models may be more appropriate. For instance, we may observe seasonal effects in demand because of nomadic population groups or because of tourism. Ndiaye and Alfares (2008) developed a multi-period integer programming model to minimize the total cost for locating primary health centers where the populations to be served occupy different locations in different seasons. Benneyan et al. (2012) considered a multi-period model for the location of specialty care clinics for veteran administration to minimize the total cost subject to access constraints where the demand changes over time. Harper et al. (2005) developed a discrete event geographical simulation model incorporating changes over time in many aspects of the system, such as demand, services offered, and facilities opened. Such changes can be used for a scenario analysis in the environment of simulation models.

Mobile healthcare facilities are commonly used in rural areas to improve access. Hodgson et al. (1998) developed an integer programming formulation for the problem of covering tour planning for the mobile healthcare facilities in Ghana. The objective is to minimize the total travel time of the facility while serving all population centers within a range of the feasible stops. Note that this problem is different from ambulance location problems since mobile facilities here serve for primary care needs as opposed to emergency care situations.

21.3 Ambulance Location

A usual goal of ambulance location problems is to find locations for ambulances (or ambulance stations) minimizing the number of ambulances (or ambulance stations) needed while fulfilling a certain level of demand. Another possibility is to maximize the coverage having a fixed number of ambulances (or ambulance stations) available. The main aspect of the corresponding coverage models is that the demand points must be reachable from the determined locations within a given time interval. Concerning ambulance planning, a large variety of literature exists. Reviews can be found in Marianov and ReVelle (1995), Owen and Daskin (1998), Brotcorne et al. (2003), Galvão et al. (2005), and Li et al. (2011).

In general, ambulance planning can be done at three different levels: strategic, tactical and operational level. At the strategic level, decisions concerning the locations of ambulance stations are made. These decisions often have a long term effect and last for several decades. The number of ambulances per station and the movable locations are determined at the tactical level. The operational level includes

the dispatching of ambulances, i.e., the allocation and reallocation to emergencies and stations. In the next two sections, exemplary models for the planning problems at the three levels are presented. Section 21.3.1 looks at strategic and tactical models, while Sect. 21.3.2 concentrates on operational aspects.

21.3.1 The Strategic and Tactical Level: Finding Ambulance Base Locations and Assigning Ambulances

One possibility for determining ambulance base locations is to use the location set covering model (LSCM) that has been first introduced by Toregas et al. (1971). The objective is to find the minimum number of ambulance bases needed to cover all demand points.

For the LSCM, a set J of demand nodes is given, and these nodes are also the potential locations for the ambulances. Moreover, as usually done in covering problems in ambulance planning, a maximum response time T is defined. Therefore, a node i can cover an emergency in node j if and only if the driving time t_{ij} between the two nodes is less than or equal to T. The set of all the nodes i that fulfill this condition is denoted by $J_j = \{i \in J \mid t_{ij} \leq T\}$, $\forall j \in J$. For each node $j \in J$, a binary decision variable x_j is considered, equal to 1 if an ambulance is located at site j and 0 otherwise. The objective function represents the number of ambulances, which is to be minimized. The constraints ensure that each demand node can be served within the given response time by at least one ambulance. The LSCM therefore looks as follows:

$$\text{minimize} \quad \sum_{j \in J} x_j \tag{21.31}$$

$$\text{subject to} \quad \sum_{i \in J_j} x_i \geq 1 \quad \forall j \in J \tag{21.32}$$

$$x_j \in \{0, 1\} \quad \forall j \in J. \tag{21.33}$$

21.3.1.1 A Double Coverage Model

The model by Toregas et al. (1971) only assures that all demand points can be reached within a given time interval, but it does not consider the possibility of covering demands from multiple nodes. Therefore, Gendreau et al. (1997) presented a so-called double standard model (DSM) that includes what is referred to as double coverage for the demand points. Compared to LSCM, DSM includes several additional features. First, the number of ambulances to be located is fixed and equal to p. Second, for demand and potential ambulance locations, two node sets I and J are considered, which may be distinct. Third, for each node $i \in I$, up to p_i

ambulances can be placed. Additionally, instead of a single maximum response time, two values, t_1 and t_2, are considered with $t_2 \geq t_1$. Note that t_2 is equivalent to T since all demand must be covered by an ambulance located within time t_2. Finally, a proportion α is defined for which the demand must also be fulfilled within t_1 time units by some of the ambulances (which can be the same ambulances or different ones). Consider now a complete graph whose nodes correspond to the elements in $I \cup J$, and whose edges $\{i, j\}$ with $i \in I$ and $j \in J$ are weighted with the travel time t_{ij} between these two nodes. Further, let d_j denote the demand at node $j \in J$, and define the following two coefficients for $i \in I$ and $j \in J$:

$$\gamma_{ij}^1 = \begin{cases} 1 & \text{if } t_{ij} \leq t_1 \\ 0 & \text{otherwise} \end{cases} \qquad (j \text{ is covered by location } i \text{ within time } t_1) \tag{21.34}$$

and

$$\gamma_{ij}^2 = \begin{cases} 1 & \text{if } t_{ij} \leq t_2 \\ 0 & \text{otherwise} \end{cases} \qquad (j \text{ is covered by location } i \text{ within time } t_2) \tag{21.35}$$

Two sets of decision variables can be considered: y_i denotes the (integer) number of ambulances to locate at $i \in I$ (bounded by p_i), and x_{jk} is a binary variable equal to 1 if j is covered at least k times within t_1 for $k \in \{1, 2\}$ and 0 otherwise. The double standard model (DSM) proposed by Gendreau et al. (1997) is the following:

$$\text{maximize} \quad \sum_{j \in J} d_j x_{j2} \tag{21.36}$$

$$\text{subject to} \quad \sum_{i \in I} \gamma_{ij}^2 y_i \geq 1 \qquad \forall j \in J \tag{21.37}$$

$$\sum_{j \in J} d_j x_{j1} \geq \alpha \sum_{j \in J} d_j \tag{21.38}$$

$$\sum_{i \in I} \gamma_{ij}^1 y_i \geq x_{j1} + x_{j2} \qquad \forall j \in J \tag{21.39}$$

$$x_{j2} \leq x_{j1} \qquad \forall j \in J \tag{21.40}$$

$$\sum_{i \in I} y_i = p \tag{21.41}$$

$$y_i \leq p_i \qquad \forall i \in I \tag{21.42}$$

$$x_{j1}, x_{j2} \in \{0, 1\} \qquad \forall j \in J \tag{21.43}$$

$$y_i \in \mathbb{Z}_0^+ \qquad \forall i \in I. \tag{21.44}$$

The objective function (21.36) maximizes the amount of demand that is covered twice within t_1. Each node must be covered at least once within time t_2 as assured by constraints (21.37). Constraint (21.38) states that a proportion α of the demand must be covered within t_1. A location can only be covered twice within t_1 if it is covered once, as expressed by constraints (21.39) and (21.40). Exactly p ambulances must be located in total (21.41) and only p_i can be located at node i (21.42). Constraints (21.43) and (21.44) define the domains of the decision variables. The model (21.36)–(21.44) has been tackled in Gendreau et al. (1997) by a tabu search heuristic.

21.3.1.2 Considering Ambulance Utilization

In practice, ambulances are not always available when they are needed. Therefore, the strategic and tactical level planning has to take some aggregated data from the operational level into account (if possible): the utilization of ambulances. For this situation, the expected coverage of a region can be determined. When the number of ambulances to be placed is fixed and the expected coverage is to be maximized, the problem can be formulated as the maximum expected location covering problem (MEXCLP) proposed by Daskin (1983).

The set of demand nodes is denoted by J, and each node has a demand d_j. I is the set of possible ambulance locations, and the maximum number of ambulances that can be located is bounded by p. In the original model, we have $I = J = \{1, \ldots, n\}$. The probability that an ambulance is occupied is defined by P and P^k is the probability that k ambulances are busy at the same time. If node $j \in J$ is covered by k ambulances, $E_k^j = d_j(1 - P^k)$ gives the corresponding expected covered demand and $E_k^j - E_{k-1}^j = d_j(1 - P)P^{k-1}$ is the marginal contribution of the kth ambulance to this expected value. A decision variable y_i is considered representing the number of ambulances to locate at node i. Moreover, we use set $K = \{1, \ldots, n\}$ in order to refer to the number of times that a node is covered by an ambulance. A decision variable x_{jk} is equal to 1 if node j is covered k times and 0 otherwise. In addition, γ_{ij} is a binary parameter with:

$$\gamma_{ij} = \begin{cases} 1 & \text{if } t_{ij} \leq T \\ 0 & \text{otherwise} \end{cases} \qquad \text{(an ambulance at } i \text{ covers demands at } j\text{)} \tag{21.45}$$

Here, t_{ij} states the driving time from node i to node j and T expresses the maximal allowed driving time. The MEXCLP can be written as follows:

$$\text{maximize} \quad \sum_{k \in K} \sum_{j \in J} d_j(1 - P)P^{k-1}x_{jk} \tag{21.46}$$

$$\text{subject to} \quad \sum_{k \in K} x_{jk} \leq \sum_{i \in I} \gamma_{ij} y_i \qquad \forall j \in J \tag{21.47}$$

$$\sum_{i \in I} y_i \leq p \tag{21.48}$$

$$y_i \in \{0, 1, \ldots, p\} \qquad \forall i \in I \tag{21.49}$$

$$x_{jk} \in \{0, 1\} \qquad \forall j \in J, k \in K. \tag{21.50}$$

The objective function (21.46) maximizes the expected demand that is covered. Note that this expression adds the expected coverage over all possible numbers of ambulances. Constraints (21.47) ensure that the number of ambulances used to cover j is bounded by the number of ambulances located not farther away than time T from j. Constraints (21.48) impose that in total at most p ambulances are located. Constraints (21.49) and (21.50) are the variable domain constraints. A heuristic for the problem has also been devised in Daskin (1983).

21.3.1.3 Further Reading

In addition to the models presented in the previous sections, several more can be found in the literature. Chapman and White (1974) proposed the first probabilistic approach by considering a probabilistic set covering model in which servers are not always available. Nowadays, different kinds of probabilistic approaches can be found for ambulance location planning. They use, for example, reliability constraints and busy fractions for servers. The same probabilistic approach is used in the maximal cover location problem investigated by ReVelle and Hogan (1988). The maximum availability location problem by ReVelle and Hogan (1989) is also worth mentioning. Overall, we can identify two main approaches for including stochasticity into the ambulance location problem, namely hypercube queuing models and stochastic programming. Larson (1974) introduced the first hypercube queuing model which represents a general planning approach where a set of states is considered as well as the transition probabilities between them. Based on that, different variations can be found, such as in Geroliminis et al. (2009), Iannoni and Morabito (2007), Iannoni et al. (2011), Silva and Serra (2008), and Takeda et al. (2007). Stochastic programming approaches have also been proposed as it is the case with the works by Beraldi et al. (2004), Beraldi and Bruni (2009), and Noyan (2010).

21.3.2 The Operational Level: Ambulance Relocation

At an operational decision level, decisions usually concern the allocation of ambulances to emergencies and the reassignment of ambulances to bases after having finished a service. In addition, relocations of ambulances during some time period (e.g., 1 day) are possible, and they can either be predefined or dynamically

determined throughout the period. A review on relocation models can be found in Brotcorne et al. (2003).

Relocation approaches proposed so far are based on Markov chain models (Alanis et al. 2013) or on approximate dynamic programming (Maxwell et al. 2009, 2013; Schmid 2012). Gendreau et al. (2001) use a parallel tabu search heuristic for solving the dynamic relocation problem. Further approaches were presented by Rajagopalan et al. (2008) and Schmid and Doerner (2010).

Because of real-time requirements encountered in practical settings, literature on ambulance relocation focuses mainly on heuristic solution methods. One such heuristic was proposed by Andersson and Värbrand (2007). The main idea was to include a so-called preparedness of ambulances. For this purpose, the area to serve is divided into a number of zones. Denote by I the set of ambulances and by J the set of zones which have a demand for ambulances. A weight d_j is assigned to each zone j which states the demand for ambulances in the zone. p_j is the (exogenous) number of ambulances that contribute to the preparedness in zone j and τ_{ij} represents the driving time from ambulance location i to zone j. Moreover, let t_j^l denote the travel time of the lth closest ambulance to zone j and let x be the vector form of the decision variables x_{ij} which are equal to 1 if ambulance i is relocated to zone j and 0 otherwise. Clearly, $t_j^l(x)$ is a function of the x-variables since the travel time depends on where the ambulances are located currently as decided by the values in x. In addition, let γ^l be the (user-defined) contribution factor of the lth closest ambulance and let the following two properties be fulfilled:

$$t_j^1 \leq t_j^2 \leq \ldots \leq t_j^{p_j}, \tag{21.51}$$

$$\gamma^1 > \gamma^2 > \ldots > \gamma^{p_j}. \tag{21.52}$$

The preparedness in zone j is then defined as

$$q_j = \frac{1}{d_j} \sum_{l=1}^{p_j} \frac{\gamma^l}{t_j^l}. \tag{21.53}$$

Andersson and Värbrand (2007) proposed a tree search algorithm that tackles the following relocation model in order to minimize the maximum travel time for the ambulances (21.54) while restricting possible relocations:

$$\text{minimize} \quad z \tag{21.54}$$

$$\text{subject to} \quad z \geq \sum_{j \in J_i} \tau_{ij} x_{ij} \qquad \forall i \in I \tag{21.55}$$

$$\frac{1}{c_j} \sum_{l=1}^{p_j} \frac{\gamma^l}{t_j^l(x)} \geq q_{min} \quad \forall j \in J \tag{21.56}$$

$$\sum_{j \in J_i} x_{ij} \leq 1 \qquad \forall i \in I \tag{21.57}$$

$$\sum_{i \in I} \sum_{j \in J} x_{ij} \leq p \tag{21.58}$$

$$x_{ij} \in \{0, 1\} \qquad \forall i \in I, j \in J. \tag{21.59}$$

In this formulation, J_i is the set of zones that can be reached by ambulance i within a given time frame. The objective function (21.54) in conjunction with constraints (21.55) (which ensure that z must not be smaller than any of the driving times τ_{ij}) represents the minimum time it will take so that the preparedness in each zone is at least q_{min} as prescribed in constraints (21.56). Hence, the left side of these constraints can be interpreted as the preparedness for zone j that must be greater or equal to a minimum value q_{min}. Constraints (21.57) assure that each ambulance can only be relocated to at most one zone in J_i. Constraints (21.58) guarantee that at most p ambulances are relocated in total. Finally, constraints (21.59) are the variable domain constraints.

21.4 Hospital Layout Planning

A special class of location problems are layout planning problems which aim at minimizing in-house travel distances or costs associated with the positions of organizational units (OUs) inside a building. This class of problems mainly originates from industrial applications for layout planning of public buildings.

Layout planning problems in healthcare were first introduced by Elshafei (1977). He modeled a hospital layout problem as a quadratic assignment problem (QAP) and developed heuristics to solve it. In the framework for hospital planning and control, the hospital layout planning problem is classified as a resource capacity planning problem on a strategic level (Hans et al. 2011). Although it is a long term decision, the spatial organization within hospitals directly influences the quality and efficiency of healthcare and secondary services of the daily routine (Choudhary et al. 2010; Hignett and Lu 2010) as well as patient satisfaction (Chaudhury et al. 2005). The challenge lies in developing a holistic approach in order to combine the architectural and legal aspects with logistics, i.e., patient, personnel, and material flows inside the future hospital building.

In the next section, the quadratic assignment problem (QAP) is presented. Section 21.4.2 details a mixed-integer programming (MIP) formulation. Thereafter, in Sect. 21.4.3, suggestions for further reading are provided in order to show some extensions of the presented QAP and MIP models with respect to the underlying assumptions.

21.4.1 The Quadratic Assignment Problem

The well-known QAP (see Chap. 13) as introduced by Koopmans and Beckmann (1957) has been first applied to hospital layout planning by Elshafei (1977) who developed heuristics to solve large instances of the problem since it is NP-hard. A solution of the QAP determines the assignment of each OU $j \in J$ to a predefined location (e.g., a room) $i \in I$ inside a building. It is assumed that each OU can be assigned to each location. The solution of a QAP instance is an assignment of $|J|$ OUs to $|I|$ locations.

Denote by f_{jk} the flow between each pair of OUs $j, k \in J$. The distance between each pair of locations $h, i \in I$ is given by d_{hi}. A binary decision variable x_{ij} can be considered, indicating whether OU j is assigned to location i ($x_{ij} = 1$) or not ($x_{ij} = 0$). Moreover, we now assume that $I = J = \{1, \ldots, n\}$ in order to obtain a mathematical formulation of the QAP as follows:

$$\text{minimize} \quad \sum_{h \in I} \sum_{i \in I} \sum_{j \in J} \sum_{k \in J} f_{jk} d_{hi} x_{hj} x_{ik} \tag{21.60}$$

$$\text{subject to} \quad \sum_{i \in I} x_{ij} = 1 \qquad \forall j \in J \tag{21.61}$$

$$\sum_{j \in J} x_{ij} = 1 \qquad \forall i \in I \tag{21.62}$$

$$x_{ij} \in \{0, 1\} \qquad \forall i \in I, j \in J. \tag{21.63}$$

The objective function (21.60) minimizes the sum of all travel distances. Constraints (21.61) ensure that each OU is assigned to exactly one room whereas constraints (21.62) guarantee that each room is only occupied by one OU. Constraints (21.63) define the domain of the decision variables.

In this basic formulation of the QAP, the area and shapes of the OUs and locations are not regarded explicitly. This means that each OU fits to each location. This is a very strong assumption which is not realistic in many applications, such as hospital layout planning, since the dimensions (area, length, width) of the OUs to be assigned can vary in a wide range. In the next section, a MIP formulation is presented which overcomes this drawback.

21.4.2 A Mixed-Integer Program

In contrast to the discrete layout representation by the QAP formulation, the MIP formulation presented next allows for a continuous representation of the layout. Thus, the length and width of each OU can be modeled explicitly as decision variables considering the defined area of the OU. Furthermore, the location of each OU can be chosen in a more flexible way within a given floor area, i.e., not only by

predefined locations as in the QAP model. Again, the objective is to minimize the total travel distance. The model presented here goes back to Montreuil (1991) and has been linearized and explained in detail by Tompkins et al. (2010).

The following parameters are given: B_a and B_b represent the length and width of the building, respectively. The lower and upper limits on the length and width of OU j are given by L_j^l, L_j^u, W_j^l and W_j^u, respectively. P_j^l and P_j^u are lower and upper limits on the perimeter of OU j, respectively. M represents a sufficiently large number (Big M). Again, f_{jk} is the flow between two OUs j and k. Furthermore, the following decision variables are defined: α_j and β_j are the x- and y-coordinates of the centroid of OU j. The x-coordinates of the left and right sides of OU j are defined by a'_j and a''_j, respectively. The y-coordinates of the bottom and top of OU j are represented by b'_j and b''_j, respectively. Furthermore, the binary variables z_{jk}^a (z_{jk}^b) are considered which are equal to 1 if OU j is strictly to the right (top) of OU k and 0 otherwise. The layout problem can be formulated as follows:

$$\text{minimize} \quad \sum_{j\in J}\sum_{k\in J} f_{jk}\left(\alpha_{jk}^+ + \alpha_{jk}^- + \beta_{jk}^+ + \beta_{jk}^-\right) \tag{21.64}$$

$$\text{subject to} \quad \alpha_j - \alpha_k = \alpha_{jk}^+ - \alpha_{jk}^- \qquad \forall j,k \in J, j \neq k \tag{21.65}$$

$$\beta_j - \beta_k = \beta_{jk}^+ - \beta_{jk}^- \qquad \forall j,k \in J, j \neq k \tag{21.66}$$

$$L_j^l \leq \left(a''_j - a'_j\right) \leq L_j^u \qquad \forall j \in J \tag{21.67}$$

$$W_j^l \leq \left(b''_j - b'_j\right) \leq W_j^u \qquad \forall j \in J \tag{21.68}$$

$$P_j^l \leq 2\left(a''_j - a'_j + b''_j - b'_j\right) \leq P_j^u \qquad \forall j \in J \tag{21.69}$$

$$0 \leq a'_j \leq a''_j \leq B_a \qquad \forall j \in J \tag{21.70}$$

$$0 \leq b'_j \leq b''_j \leq B_b \qquad \forall j \in J \tag{21.71}$$

$$\alpha_j = 0.5\left(a'_j + a''_j\right) \qquad \forall j \in J \tag{21.72}$$

$$\beta_j = 0.5\left(b'_j + b''_j\right) \qquad \forall j \in J \tag{21.73}$$

$$a''_k \leq a'_j + M\left(1 - z_{jk}^a\right) \qquad \forall j,k \in J, j \neq k \tag{21.74}$$

$$b''_k \leq b'_j + M\left(1 - z_{jk}^b\right) \qquad \forall j,k \in J, j \neq k \tag{21.75}$$

$$z_{jk}^a + z_{kj}^a + z_{jk}^b + z_{kj}^b \geq 1 \qquad \forall j,k \in J, j < k \tag{21.76}$$

$$\alpha_j, \beta_j, a'_j, a''_j, b'_j, b''_j \geq 0 \qquad \forall j \in J \tag{21.77}$$

$$\alpha_{jk}^+, \alpha_{jk}^-, \beta_{jk}^+, \beta_{jk}^- \geq 0 \qquad \forall j,k \in J, j \neq k \tag{21.78}$$

$$z_{jk}^a, z_{jk}^b \in \{0,1\} \qquad \forall j,k \in J, j \neq k. \tag{21.79}$$

The objective function (21.64) minimizes the sum of the rectilinear distances of all the flows between the centroids of the OUs. Constraints (21.65) and (21.66) are needed in order to linearize the model given by Montreuil (1991): in order to get a linear objective function, the auxiliary decision variables $\alpha_{jk}^{+}, \alpha_{jk}^{-}, \beta_{jk}^{+}$ and β_{jk}^{-} have to be introduced such that with (21.65) and (21.66), we have $\left|\alpha_j - \alpha_k\right| = \alpha_{jk}^{+} + \alpha_{jk}^{-}$ and $\left|\beta_j - \beta_k\right| = \beta_{jk}^{+} + \beta_{jk}^{-}$. Constraints (21.67), (21.68) and (21.69) control the lower and upper limits of the length, width and perimeter of the OUs, respectively. The correct definition of the sides of the OUs as well as their location inside the building is ensured by constraints (21.70) and (21.71). The centroid of each OU is defined by constraints (21.72) and (21.73). The non-overlapping requirements for the OUs are formulated by constraints (21.74)–(21.76). The domains of the decision variables are given in constraints (21.77)–(21.79). We finally remark that the model has been first used by Montreuil (1991) in order to devise a comprehensive modeling framework which aims at integrating layout design and material flow network design in material handling and logistics systems.

21.4.3 Further Reading

In this section, some possible extensions to the two models discussed in Sects. 21.4.1 and 21.4.2 are presented. Important characteristics which were not considered above, but which are also of importance for hospital layout planning problems comprise the consideration of multiple periods, multiple floors, multiple objectives as well as uncertainty in patient, personnel, and material flows. Overall, there are very few publications considering the application of layout planning problems in hospitals from a mathematical perspective. General surveys on layout planning have been conducted, among others, by Drira et al. (2007) and Singh and Sharma (2006). Textbooks on facility layout planning and design are given by Tompkins et al. (2010) and Heragu (2008).

A general review on dynamic layout problems which takes into account multiple periods and, thus, changing process flows, is given by Balakrishnan and Cheng (1998). A very recent approach for a multi-period ward layout planning problem for hospitals has been presented by Arnolds and Nickel (2013b).

Since hospital buildings usually have more than one floor, another extension comprises multiple floors. In this respect, the planning of elevators such as their location, number, capacity and control is a quite new and challenging field that has been addressed for example by Matsuzaki et al. (1999), Goetschalckx and Irohara (2007a,b), and Krishnan et al. (2009). Further modeling and solution approaches for multi-floor layout problems can be found in Bozer et al. (1994), Patsiatzis and Papageorgiou (2002), and Meller and Bozer (1997).

In the last years, a number of papers has been published with respect to multiple objectives (Chen and Sha 1999, 2005; Sha and Chen 2001; Tenfelde-Podehl 2002; Aiello et al. 2006; Chen and Rogers 2009a,b; Bashiri and Dehghan 2010). This is

a very important issue for hospital layout planning problems since, for example, travel distances or times of patients, personnel and materials somehow have to be regarded and balanced.

One additional aspect worth discussing is the uncertainty that can impact data. For example, future patient figures for certain diseases are unknown. Accordingly, processes, i.e., the flow of patients, personnel, and materials, depend on outcomes and reconvalescence and, thus, are not deterministic. This uncertainty should be reflected in the design process. Some works taking into account different sources of uncertainty in general layout planning problems include Liu et al. (2006), Norman and Smith (2006), Kulturel-Konak (2007), and Tavakkoli-Moghaddam et al. (2007). Another approach has been developed by Arnolds and Nickel (2013a) who apply a simulation-optimization approach in order to take into account the uncertainty in patient, personnel, and material flows: while solving a mathematical layout model results in optimal solutions under deterministic data, discrete-event simulation scenarios help to create a robust layout which will show high performance even when patient, personnel, and material flows are uncertain.

21.5 Conclusions

In this chapter, we have seen that mathematical models of facility location can be applied to the healthcare sector at all planning levels. Considering the challenge of an ageing population on the one hand and the increased significance of an efficient resource management in the medical sector on the other hand, the topic will be receiving even more attention over the next decades. Future research directions could integrate planning problems at different levels with the goal of developing advanced planning instruments focused on healthcare applications. Likewise, advancements in solution methods for current problems as discussed in this chapter, as well as the identification of future problems along with the development of corresponding solution methodologies represent interesting challenges for future research on location problems in healthcare.

Acknowledgements The second author would like to thank Ines Arnolds and Melanie Reuter for their support in preparing this text.

References

Aiello G, Enea M, Galante G (2006) A multi-objective approach to facility layout problem by genetic search algorithm and electre method. Robot Cim-Int Manuf 22:447–455

Alanis R, Ingolfsson A, Kolfal B (2013) A markov chain model for an EMS system with repositioning. Prod Oper Manag 22:216–231

Andersson T, Värbrand P (2007) Decision support tools for ambulance dispatch and relocation. J Oper Res Soc 58:195–201

Arnolds IV, Nickel S (2013a) An iterative simulation-optimization approach for hospital layout planning. Tech. Rep. Institute of Operations Research, Discrete Optimization and Logistics, Karlsruhe Institute of Technology

Arnolds IV, Nickel S (2013b) Multi-period layout planning for hospital wards. Socio Econ Plan Sci 47:220–237

Balakrishnan J, Cheng CH (1998) Dynamic layout algorithms: a state-of-the-art survey. Omega 26:507–521

Bashiri M, Dehghan E (2010) Optimizing a multiple criteria dynamic layout problem using a simultaneous data envelopment analysis modeling. Int J Comput Sci Eng 2:28–35

Benneyan JC, Musdal H, Ceyhan ME, Shiner B, Watts BV (2012) Specialty care single and multi-period location-allocation models within the veterans health administration. Socio Econ Plan Sci 46:136–148

Beraldi P, Bruni ME (2009) A probabilistic model applied to emergency service vehicle location. Eur J Oper Res 196:323–331

Beraldi P, Bruni ME, Conforti D (2004) Designing robust emergency medical service via stochastic programming. Eur J Oper Res 158:183–193

Bozer YA, Meller RD, Erlebacher SJ (1994) An improvement-type layout algorithm for single and multiple-floor facilities. Manag Sci 40:918–932

Brotcorne L, Laporte G, Semet F (2003) Ambulance location and relocation models. Eur J Oper Res 147:451–463

Bruni ME, Conforti D, Sicilia N, Trotta S (2006) A new organ transplantation location-allocation policy: a case study of Italy. Health Care Manage Sci 9:125–142

Calvo AB, Marks DH (1973) Location of health care facilities: an analytical approach. Socio Econ Plan Sci 7:407–422

Canovas L, García S, Labbé M, Marín A (2007) A strengthened formulation for the simple plant location problem with order. Oper Res Lett 35:141–150

Cardoso T, Oliveira MD, Barbosa-Póvoa A, Nickel S (2012) Modeling the demand for long-term care services under uncertain information. Health Care Manage Sci 15:385–412

Chao X, Liu L, Zheng S (2003) Resource allocation in multisite service systems with intersite customer flows. Manag Sci 49:1739–1752

Chapman SC, White JA (1974) Probabilistic formulations of emergency service facilities location problems. Paper presented at the 1974 ORSA/TIMS Conference, San Juan, Puerto Rico

Chaudhury H, Mahmood A, Valente M (2005) Advantages and disadvantages of single- versus multiple-occupancy rooms in acute care environments: a review and analysis of the literature. Environ Behav 37:760–86

Chen GY, Rogers KJ (2009a) Managing dynamic facility layout with multiple objectives. In: Proceedings of PICMET 2009 – Portland International Center for Management of Engineering and Technology, Portland, pp 1175–1184

Chen GY, Rogers KJ (2009b) Proposition of two multiple criteria models applied to dynamic multi-objective facility layout problem based on ant colony optimization. In: Proceedings of IEEEM 2009 – international conference on industrial engineering and engineering management, Hong Kong, pp 1553–1557

Chen CW, Sha DY (1999) A design approach to the multi-objective facility layout problem. Int J Prod Res 37:1175–1196

Chen CW, Sha DY (2005) Heuristic approach for solving the multi-objective facility layout problem. Int J Prod Res 43:4493–4507

Cho C (1998) An equity-efficiency trade-off model for the optimum location of medical care facilities. Socio Econ Plan Sci 32:99–112

Choudhary R, Bafna S, Heo Y, Hendrich A, Chow M (2010) A predictive model for computing the influence of space layouts on nurses' movement in hospital units. J Build Perform Simul 3:171–184

Daskin MS (1983) A maximum expected covering location model: formulation, properties and heuristic solution. Transp Sci 17:48–70

Daskin MS, Dean LK (2004) Location of health care facilities. In: Brandeau M, Sainfort F, Pierskalla W (eds) Location of health care facilities, in operations research and health care: a handbook of methods and applications. Kluwer, New York, pp 43–76
Dokmeci VF (1977) A quantitative model to plan regional health facility systems. Manag Sci 24:411–419
Dokmeci VF (1979) A multiobjective model for regional planning of health facilities. Environ Plan A 11:517–525
Drira A, Pierreval H, Hajri-Gabouj S (2007) Facility layout problems: a survey. Annu Rev Control 31:255–267
Ehrgott M (2005) Multicriteria optimization, 2nd edn. Springer, Berlin/Heidelberg
Elshafei AN (1977) Hospital layout as a quadratic assignment problem. Oper Res Q 28:167–179
Galvão RD, Espejo LGA, Boffey B (2002) A hierarchical model for the location of perinatal facilities in the municipality of Rio de Janeiro. Eur J Oper Res 138:495–517
Galvão RD, Chiyoshi FY, Morabito R (2005) Towards unified formulations and extensions of two classical probabilistic location models. Comput Oper Res 32:15–33
Galvão RD, Espejo LGA, Boffey B, Yates D (2006) Load balancing and capacity constraints in a hierarchical location model. Eur J Oper Res 172:631–646
Gendreau M, Laporte G, Semet F (1997) Solving an ambulance location model by tabu search. Locat Sci 5:75–88
Gendreau M, Laporte G, Semet F (2001) A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. Parallel Comput 27:1641–1653
Geroliminis N, Karlaftis MG, Skabardonis A (2009) A spatial queuing model for the emergency vehicle districting and location problem. Transp Res B Methodol 43:798–811
Goetschalckx M, Irohara T (2007a) Efficient formulations for the multi-floor facility layout problem with elevators. Tech. Rep. School of Industrial and Systems Engineering Research
Goetschalckx M, Irohara T (2007b) Formulations and optimal solution algorithms for the multi-floor facility layout problem with elevators. In: Proceedings of IIE annual conference and Expo 2007 – industrial engineering's critical role in a flat world, Nashville, pp 1446–1452
Gould PR, Leinbach TR (1966) An approach to the geographic assignment of hospital services. Tijdschrift voor Economische en Sociale Geografie 57:203–206
Griffin PM, Scherrer CR, Swann JL (2008) Optimization of community health center locations and service offerings with statistical need estimation. IIE Trans 40:880–892
Güneş ED, Yaman H (2009) Health network mergers and hospital re-planning. J Oper Res Soc 61:275–283
Güneş ED, Yaman H, Cekyay B, Verter V (2014) Matching patient and physician preferences in designing a primary care facility network. J Oper Res Soc 65:483–496
Hans EW, van Houdenhoven M, Hulshof PJH (2011) A framework for health care planning and control. In: Hall R (ed) Handbook of health care systems scheduling. Springer international series in operations research & management science, vol 168, Chap 12. Springer, Berlin, pp 303–320
Harper PR, Shahani AK, Gallagher JE, Bowie C (2005) Planning health services with explicit geographical considerations: a stochastic location-allocation approach. Omega 33:141–152
Heragu SS (2008) Facilities design, 3rd edn. CRC Press, Boca Raton
Hignett S, Lu J (2010) Space to care and treat safely in acute hospitals: recommendations from 1866 to 2008. Appl Ergon 41:666–673
Hodgson MJ (1988) An hierarchical location-allocation model for primary health care delivery in a developing area. Soc Sci Med 26:153–161
Hodgson MJ, Laporte G, Semet F (1998) A covering tour model for planning mobile health care facilities in Suhum District, Ghana. J Reg Sci 38:621–638
Iannoni AP, Morabito R (2007) A multiple dispatch and partial backup hypercube queuing model to analyze emergency medical systems on highways. Transp Res E Log 43:755–771
Iannoni AP, Morabito R, Saydam C (2011) Optimizing large-scale emergency medical system operations on highways using the hypercube queuing model. Socio Econ Plan Sci 45:105–117

Institute of Medicine (1993) Access to health care in America. National Academy Press, Washington
Koopmans T, Beckmann M (1957) Assignment problems and the location of economic activities. Econometrica 25:53–76
Krishnan KK, Jaafari AA, Abolhasanpour M, Hojabri H (2009) A mixed integer programming formulation for multifloor layout. Afr J Bus Manag 3:616–620
Kulturel-Konak S (2007) Approaches to uncertainties in facility layout problems: perspectives at the beginning of the 21st century. J Intell Manuf 18:273–284
Larson RC (1974) A hypercube queuing model for facility location and redistricting in urban emergency services. Comput Oper Res 1:67–95
Li X, Zhao Z, Zhu X, Wyatt T (2011) Covering models and optimization techniques for emergency response facility location and planning: a review. Math Method Oper Res 74:281–310
Liu F, Dong M, Hou F, Chen F (2006) Facility layout optimization with stochastic logistic flows. In: Proceedings of SOLI 2006 - IEEE international conference on service operations and logistics, and informatics, Shanghai, pp 534–539
Marianov V, ReVelle C (1995) Siting emergency services. In: Drezner Z (ed) Facility location: a survey of applications and methods. Springer, New York, pp 199–223
Marsh MT, Schilling DA (1994) Equity measurement in facility location analysis: a review and framework. Eur J Oper Res 74:1–17
Matsuzaki K, Irohara T, Yoshimoto K (1999) Heuristic algorithm to solve the multi-floor layout problem with the consideration of elevator utilization. Comput Ind Eng 36:487–502
Maxwell MS, Henderson SG, Topaloglu H (2009) Ambulance redeployment: an approximate dynamic programming approach. In: Proceedings of WSC 2009 - winter simulation conference 2009, Austin, pp 1850–1860
Maxwell MS, Henderson SG, Topaloglu H (2013) Tuning approximate dynamic programming policies for ambulance redeployment via direct search. Stoch Syst 3:322–361
Mayhew LD, Leonardi G (1982) Equity, efficiency, and accessibility in urban and regional health-care systems. Environ Plan A 14:1479–1507
Meller RD, Bozer YA (1997) Alternative approaches to solve the multi-floor facility layout problem. J Manuf Syst 16:192–203
Mestre AM, Oliveira MD, Barbosa-Póvoa A (2012) Organizing hospitals into networks: a hierarchical and multiservice model to define location, supply and referrals in planned hospital systems. OR Spectrum 34:319–348
Mitropoulos P, Mitropoulos I, Giannikos I, Sissouras A (2006) A biobjective model for the locational planning of hospitals and health centers. Health Care Manage Sci 9:171–179
Montreuil (1991) A modelling framework for integrating layout design and flow network design. In: Graves RJ, McGinnis LF, Wilhelm MR, Ward RE (eds) Material handling 1990. Progress in material handling and logistics, vol 2. Springer, Berlin/Heidelberg, pp 95–115
Narula SC (1984) Hierarchical location-allocation problems: a classification scheme. Eur J Oper Res 15:93–99
Narula SC, Ogbu UI (1979) An hierarchal location-allocation problem. Omega 7:137–143
Ndiaye M, Alfares H (2008) Modeling health care facility location for moving population groups. Comput Oper Res 35:2154–2161
Norman BA, Smith AE (2006) A continuous approach to considering uncertainty in facility design. Comput Oper Res 33:1760–1775
Noyan N (2010) Alternate risk measures for emergency medical service system design. Ann Oper Res 181:559–589
Oliveira MD, Bevan G (2006) Modelling the redistribution of hospital supply to achieve equity taking account of patient's behaviour. Health Care Manage Sci 9:19–30
Owen SH, Daskin MS (1998) Strategic facility location: a review. Eur J Oper Res 111:423–447
Parker BR, Srinivasan V (1976) A consumer preference approach to the planning of rural primary health-care facilities. Oper Res 24:991–1025
Patsiatzis DI, Papageorgiou LG (2002) Optimal multi-floor process plant layout. Comput Chem Eng 26:575–583

Rahman S, Smith D (2000) Use of location-allocation models in health service development planning in developing nations. Eur J Oper Res 123:437–452
Rajagopalan HK, Saydam C, Xiao J (2008) A multiperiod set covering location model for dynamic redeployment of ambulances. Comput Oper Res 35:814–826
ReVelle C, Hogan K (1988) A reliability-constrained siting model with local estimates of busy fractions. Environ Plan B 15:143–152
ReVelle C, Hogan K (1989) The maximum availability location problem. Transp Sci 23:192–200
Şahin G, Süral H (2007) A review of hierarchical facility location models. Comput Oper Res 34:2310–2331
Schmid V (2012) Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. Eur J Oper Res 219:611–621
Schmid V, Doerner KF (2010) Ambulance location and relocation problems with time-dependent travel times. Eur J Oper Res 207:1293–1303
Sha DY, Chen CW (2001) A new approach to the multiple objective facility layout problem. Integr Manuf Syst 12:59–66
Silva F, Serra D (2008) Locating emergency services with different priorities: the priority queuing covering location problem. J Oper Res Soc 59:1229–1238
Singh SP, Sharma RRK (2006) A review of different approaches to the facility layout problems. Int J Adv Manuf Technol 30:425–433
Smith HK, Harper PR, Potts CN, Thyle A (2009) Planning sustainable community health schemes in rural areas of developing countries. Eur J Oper Res 193:768–777
Smith HK, Harper PR, Potts CN (2013) Bicriteria efficiency/equity hierarchical location models for public service application. J Oper Res Soc 64:500–512
Stummer C, Doerner K, Focke A, Heidenberger K (2004) Determining location and size of medical departments in a hospital network: a multiobjective decision support approach. Health Care Manage Sci 7:63–71
Takeda RA, Widmer JA, Morabito R (2007) Analysis of ambulance decentralization in an urban emergency medical service using the hypercube queueing model. Comput Oper Res 34:727–741
Tavakkoli-Moghaddam R, Javadian N, Javadi B, Safaei N (2007) Design of a facility layout problem in cellular manufacturing systems with stochastic demands. Appl Math Comput 184:721–728
Tenfelde-Podehl D (2002) Facilities layout problems: polyhedral structure, multiple objectives and robustness. Ph.D. thesis, Universität Kaiserslautern
Tien JM, El-Tell K, Simons GR (1983) Improved formulations of the hierarchical health facility location-allocation problem. IEEE Trans Syst Man Cybern 13:1128–1132
Tompkins JA, White JA, Bozer YA, Tanchoco JMA (2010) Facilities planning, 4th edn. Wiley, Hoboken
Toregas C, Swain R, ReVelle C, Bergman L (1971) The location of emergency service facilities. Oper Res 19:1363–1373
Verter V, Lapierre SD (2002) Location of preventive health care facilities. Ann Oper Res 110:123–132
Zhang Y, Berman O, Verter V (2009) Incorporating congestion in preventive healthcare facility network design. Eur J Oper Res 198:922–935
Zhang Y, Berman O, Marcotte P, Verter V (2010) A bilevel model for preventive healthcare facility network design with congestion. IIE Trans 42:865–880

Chapter 22
The Design of Rapid Transit Networks

Gilbert Laporte and Juan A. Mesa

Abstract Metro and other rapid transit systems increase the mobility of urban populations while decreasing congestion and pollution. There are now 191 cities with a metro system in the world, 49 of which were inaugurated in the twenty-first century. The design of a rapid transit system is a hard problem involving several players, multiple objectives, sizeable costs and a high level of uncertainty. Operational research techniques cannot fully solve the problem, but they can generate alternative solutions among which the decision makers can choose, and be employed to solve some specific subproblems. The scientific literature on rapid transit location planning has grown at a fast rate over the past 20 years. In this chapter an account of some of the most important results are provided. First the main objectives and indices used in the assessment of rapid transit systems are described. Then the main models and algorithms used to design such systems are reviewed. The case of a single alignment and of a full network are treated separately. Then follows a section on the location of stations on an already existing network.

Keywords Location • Metro • Network design • Rapid transit • Stations

22.1 Introduction

Due to the increasing population and the spread of urbanized zones, many cities and metropolitan areas around the world are planning, constructing or extending their transit systems. Among these, metro systems are the most efficient because they consume less energy and are able to transport more passengers per surface unit than any other form of public transport. Metro systems help decrease private car traffic, therefore reducing congestion and pollution. The term metro is sometimes used synonymously with rapid transit but the latter has a wider acceptation. In the

G. Laporte (✉)
HEC Montréal, Montréal, QC, Canada
e-mail: gilbert.laporte@cirrelt.ca

J.A. Mesa
Department of Applied Mathematics II, University of Seville, Sevilla, Spain
e-mail: jmesa@us.es

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_22

technical literature, rapid transit usually covers not only metro, but also commuter train, light metro, light rail, monorail and others urban mass rapid public transport systems. A metro system is independent from other traffic, even though some light metros or German *stadtbahn* are underground in city centers, but at grade with preference level crossings in suburban areas. According with the World Metro Database (Rhode 2014), 49 out of 191 cities with a metro system have inaugurated it in the twenty-first century. The latter figure can be compared with that of 1991 (Gendreau et al. 1995), when there existed fewer than 90 metro systems. Even though the list is not exhaustive, in July 2014 Wikipedia reported that 36 such systems were under construction (Wikipedia 2014). Bus rapid transit (BRT) systems are sometimes considered as rapid transit systems. They share several characteristics with those using rails but they exhibit several differences, such as slower vehicles, level crossings, and less capacity. They are usually treated separately in planning processes and in academic research.

In practice, rapid transit planning is a very complex task involving agents with different backgrounds and loyalties (politicians, urban planners, transit agencies, engineers, construction companies, citizen groups, etc.). These players may therefore have different and sometimes conflicting goals. The planning process usually starts by analyzing the area under consideration and the main travel patterns. Then, based on travel patterns codified by origin-destination flow matrices, some broad traffic corridors are identified and combined, giving rise to several network scenarios which can be evaluated from different points of view, often using finite multi-criteria analysis. Since, the problem is inherently strategic, this process usually takes a long time.

Rapid transit planning can be broadly classified depending on whether the network is to be constructed from scratch or whether it is to be extended by adding new lines or extending some existing ones. Rail rapid transit planning lies within the broader field of rail network planning. The sequential process of rail planning is based on the knowledge of the travel patterns and starts with network design. Line planning, timetabling and resource scheduling are the subsequent stages in this process. Other related important issues are reliability, robustness, timetabling information, shunting, platforming, etc. However, due to their special characteristics rapid transit planning deserves a particular study. Usually, the tracks of metro lines are not interconnected. There are exceptions to this rule, for example the cases where there is a common trunk for several lines (Los Angeles, Brussels and Bilbao metros), or the case of a line working as a set of lines but most of the lines work independently. This is the case of the London Underground Northern line with three northern termini and two different routes in the city center, see Fig. 22.1. Some lines in commuter systems also share the railway system in the city centre. This implies that network design and line planning (except frequency setting) are considered together in the modeling process. A second specific characteristic of metros is that they carry a large number of passengers traveling over short distances compared with medium and long distance railways. This implies that headways are very short (with the new telecommunication technologies, in some cases these are reduced to one minute and a half). Another distinguishing feature is the importance of mode

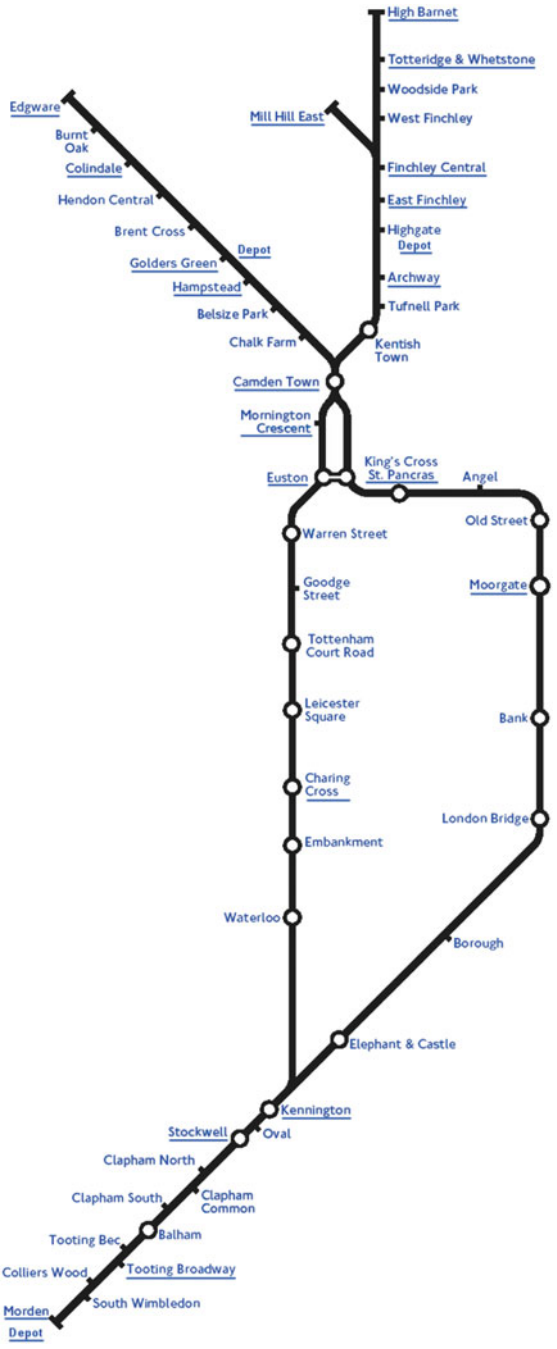


Fig. 22.1 The Northern Line, London underground

selection due to the fact that in most metropolitan areas where such systems are planned, several competing modes of transportation (bus, private car) are available.

Rapid transit network design is made up of two intertwined problems: the determination of alignments and the location of stations. There are other related location problems such as those of locating park-and-ride facilities and depots, but usually their corresponding feasible sets are limited to very few possibilities and thus do not give rise to interesting location problems. The location of stations is a typical attractive facility location problem for which several criteria can be applied depending on the goals of the decision maker. However, a station located in a high density area could be non-efficient because of the direction of the line to which it belongs. For example, if the line goes north-south but the people located close to the station work east of the station, this station will not be useful for their working trips. Therefore, it is crucial to concentrate on the location of the alignments and not only on that of the stations. Since the facility to be located is a network, and therefore very large with respect to its environment, the problem under consideration is an extensive or multi-dimensional facility location problem (Mesa and Boffey 1996).

Our aim is to review some of the main aspects of rapid transit location. For the sake of readability, we have avoided the use of lengthy formulations and formulas as much as possible. These can be found in the original sources. We will first describe in Sect. 22.2 the main indicators used to assess the quality of a rapid transit network. Models and algorithms used for rapid transit network design will be described in Sect. 22.3. In Sect. 22.4 we focus on the location of stations. Conclusions follow in Sect. 22.5.

22.2 Objectives and Network Assessment

The main objective of a collective transit system is to improve the population mobility. Since rapid transit systems usually have a high capacity, they extensively reduce traffic congestion, airborne pollution and energy consumption, thus providing sustainable mobility. Moreover, these systems are among the quickest collective mode of ground transportation, and therefore they usually provide the shortest travel times. Another important feature is their structuring influence on cities since they provide the backbone for the development of residential, business and commercial areas. Rapid transit systems require high-level investments, both for construction and maintenance. The initial investment is related to the construction of tunnels, elevated or at grade right-of-ways, communication systems and the purchase of rolling stocks. Operating cost include fixed and variable costs on a daily basis.

The agents interested in the planning processes can be broadly classified into three groups: the society in general, which is represented by transportation agencies and government sections, the potential riders, and the company offering the service. The first group is mainly interested in global advantages such as those mentioned above and they are therefore concerned with the population to be served by the system. A measure frequently used at the planning stage is the population covered

by the system, often defined as the population living within a certain distance threshold from stations. This limit has been fixed to 400 m or 5 min walk in dense areas (Vuchic 2005), but it can grow to 1 km in less populated regions. Moreover, the catchment areas of stations are not always limited to pedestrian traffic but also to combined modes (Mesa and Ortega 2001). However, ridership is not only a function of the distance to the line, but also of the design of the network (Gendreau et al. 1995). A better measure is the predicted trip coverage which can be measured by origin-destination surveys, coupled with traffic equilibrium models. Potential users are mainly interested in reducing their travel time. A secondary objective of the passengers is to transfer between lines as little as possible. Of course this can be included into a more general and difficult to measure concept of comfort or generalized cost. Finally, the third group, that of construction and operating companies, is mainly concerned with fixed and variable construction and operating costs and revenues.

An existing rapid transit network can be evaluated by means of network measures and indicators, but the same measures can also be used to evaluate potential networks, in particular those resulting of the process of combining corridors. To this end graph theory is a useful tool. Furthermore, these measures can be used as objective functions or as constraints in mathematical programming models. Musso and Vuchic (1988) have developed some network topology indicators such as circle availability, network complexity and connectivity. They have also considered service measures and utilization indicators. Laporte et al. (1997) have also measured the efficiency of rapid transit networks via the passengers/network and passengers/plane measures. For example, these authors have shown that in a circular city, triangle and cartwheel designs are preferable to star designs (Fig. 22.2) in terms of connectivity and travel directness.

Gattuso and Miriello (2005) provide a comparative analysis of 13 existing metro networks with respect to 10 indicators. Other indicators such as regularity, service availability, punctuality and reliability can be found in UITP (2011). Nowadays, the values of some of these indicators are often presented in the technical reports of operating companies. Whereas most of the early research on indicators and measures concerns the description and efficiency of the networks with respect to different topological indicators, in recent years we have witnessed the emergence of

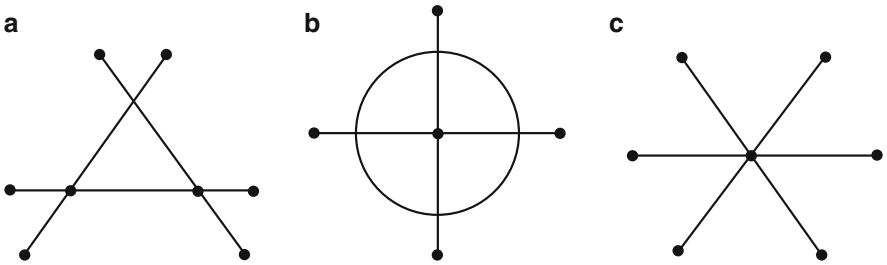


Fig. 22.2 Three basic metro designs. (**a**) Triangle. (**b**) Cartwheel. (**c**) Star

new indices based on the assessment of transportation networks from the angle of complex network theory and robustness. In accordance with the glossary of the IEEE (1990), robustness can be defined as the degree to which a system or component can function correctly in the presence of invalid inputs or stressful environmental conditions. In the case of rapid transit networks planning, future ridership is an uncertainty input variable which also depends on the travel times of alternative transportation modes.

Another issue affecting robustness lies in the disturbances of normal operations. The paper of De-Los-Santos et al. (2012) considers robustness from the angle of passengers in the presence of disruptions. The auxiliary function applied to define robustness measures is the total transit time of passengers. Two cases are considered. In the first case, passengers affected by the disruption have to wait for the failure to be repaired or have to take an alternative route in the same network. In the second case, the operator provides a bus-bridge service. An example for the Madrid commuter system illustrates the applicability of the robustness indices developed by the authors.

Over the past 15 years there has been an increased research interest in the structural properties of the networks representing complex systems, which is interesting for understanding the functioning of these systems. One of the most cited examples in the scientific literature is that of transportation networks and, in particular, metro networks. The concept of small-world phenomenon comes from sociology. The corresponding networks are an intermediate class between regular networks (with equal-degree nodes) and random networks (edge-generated by a given probability). Small-world networks are highly clustered, like regular networks, but they have a low average shortest path length between pairs of nodes (Watts and Strogatz 1998). Let $G = (V, E)$ be a graph and let $d_{ij}, v_i, v_j \in V$ be the topological distance between v_i and v_j (the minimum number of edges in a path between v_i and v_j). Then the characteristic path length L and the clustering coefficient C are defined as

$$L = \frac{1}{|V|(|V|-1)} \sum_{i \neq j} d_{ij}, \quad \text{and} \quad C = \frac{1}{|V|} \sum_{v_i \in V} C_i,$$

where C_i is the number of edges in $G_i = (V_i, E_i)$, the subgraph of the neighbors of v_i, divided by the maximum possible number $|V_i|(|V_i|-1)/2$.

In order to adapt these concepts to metric networks and to overcome some indetermination, the average length of shortest paths and clustering coefficients were substituted by global and local efficiency (Latora and Marchiori 2001):

$$E_{glob}(G) = \frac{2}{|V|(|V|-1)} \sum_{i<j} \frac{1}{d_{ij}}, \quad \text{and} \quad E_{loc}(G) = \frac{1}{|V|} \sum_{v_i \in V} E_{glob}(G_i),$$

where G_i is the subgraph of neighbors of v_i.

In small-world networks it is easy to travel both at the local and at the global levels. Since such networks are tolerant against disruptions, they are robust. However, metro networks have been shown not to be robust at the local level. Nevertheless, networks of direct connections, where there exists an edge between all pairs of stations for which passengers do not need to transfer to other line, may be seen as small world networks (Sen et al. 2002; Seaton and Hackett 2004). Other papers dealing with efficiency, robustness, vulnerability and small-world phenomenon of metro networks are those of Latora and Marchiori (2002), Criado el al. (2007), Derrible and Kennedy (2010), Barbadillo and Saldaña (2011) and Zhang et al. (2013). The paper by Roth et al. (2012) also deserves a mention. These authors consider the dynamics of the largest metro networks and prove that they converge to a unique network shape.

A new approach to the study of the connectivity of metro networks and thus their robustness is grounded in the concept of hypergraphs and their associated linear graphs. Given a collective transportation network made up of a set of lines $\{L_1, \ldots, L_l\}$, where $L_i = \{s_1^i, \ldots, s_{l_i}^i\}$ is the set of stations of line L_i, the associated hypergraph is the pair $H = (V(H), E(H))$, where $V(H)$ is the set of all stations, and the hyperedge set $E(H) = \{L_1, \ldots, L_l\}$ consists of the station sets of the lines. The associated linear graph is $L(H) = (\{L_1, \ldots, L_l\}, E(L(H)))$, where the edge set $E(L(H))$ is the set representing the transfer stations. In Barrena et al. (2013) the indices defined above are extended to collective transportation networks in order to allow them to extract information on the easiness of transfer and to compare different metro networks from this viewpoint. In that paper, the notions of clustering, characteristic path length, local efficiency and global efficiency are extended to hypergraphs and are applied to the comparison of several metro networks.

22.3 Location of Rapid Transit Networks: Models and Algorithms

Construction projects for rapid transit networks can be classified into three groups: those in which a single line is planned from scratch (Metro de Granada 2013), those in which several lines are planned from scratch and simultaneously (for example, Sociedad del Metro de Sevilla 2001), and those in which an existing network is to be extended, which corresponds to a conditional network design problem (Metro de Lisboa 2014).

22.3.1 Location of a Single Alignment

The problem of locating an alignment for a rapid transit system lies within the area of location of dimensional structures either in a discrete or in a continuous space

(Mesa and Boffey 1996; Díaz et al. 2004), more precisely that of locating paths and networks. Cast in the framework of graph theory, the problem is to select a path between two nodes (which could be fixed a priori) and some of the intermediate nodes to be stations, in order to optimize an objective function subject to certain constraints. In the continuous setting, the problem is that of selecting a straight-line, a broken line (polygonal segment) or a curved segment and some points on it. If the rapid transit line is planned to be at grade, it is almost always necessary to work with a discrete setting, but if the network is to be constructed underground, then a mixed network-continuous space fits better. Here we consider the problem of locating a path and the points on it, leaving the case of locating the stations on a given alignment to Sect. 22.4. Therefore, the decision variables of the problems considered in this section are those of the coordinates of the stations and of the links connecting adjacent stations.

In order to realistically model the problem of locating an alignment, it is necessary to consider several features in addition to those encountered in covering-path problems (Current et al. 1985). These include interstation spacing constraints, competition or intermodality with other means of transportation, demand allocated to pairs of points instead of single points, etc. The early paper of Gendreau et al. (1995) proposes a simple algorithmic approach to the problem of locating a transit line, but without any computational implementation. To our knowledge, Dufourd et al. (1996) provide the first real attempt to solve the problem of locating a transit line taking into account maximum and minimum station interspacing and the number of allowed stations to be located. In that paper, the objective is to maximize the population covered by the stations. This is computed by using several levels of catchment with the use of the Manhattan or ℓ_1 metric. The authors solve the problem by means of tabu search. The paper by Bruno et al. (1998) incorporates the more realistic criterion of maximizing trip coverage, as opposed to population coverage. In order to introduce real-world features into their model, the authors consider a private mode of transportation competing with the bimodal pedestrian-public transit mode. Each mode uses its respective network and the demand is assigned to the mode with the least travel cost. The problem consists of computing non-dominated solutions with respect to cost and trip coverage objectives. Bruno et al. (2002) consider the same model as in Dufourd et al. (1996), except for the use of the ℓ_2 metric instead of the ℓ_1 metric for interstation distances. They develop a heuristic consisting of two phases: the construction of the path and the iterative improvement of it. This heuristic is shown to produce better solutions in less time than the tabu search approach of Dufourd et al. (1996).

A similar approach was used in Laporte et al. (2005) to solve the more complex problem of maximizing trip coverage in the presence of an alternative mode of transportation. Instead of considering a binary variable to decide to which mode the demand pair should be allocated, the authors use a continuous variable representing the distribution of the demand between each mode, according to a logit function which depends on the difference between travel times (or costs) of both modes.

Finally, in order to avoid possible damage to historical building a modified anticenter path location problem is used in Laporte et al. (2009) to design a metro line as far away as possible from some patrimonial buildings to be protected. The problem is solved with the help of a Voronoi diagram constructed around the protected sites.

22.3.2 Rapid Transit Network Design

We now consider the problem of locating a rapid transit network from scratch, as well as the problem of extending an already located network. The first attempt at modeling and solving the general rapid transit network design problem is presented in the paper by Laporte et al. (2007), which provides a computationally tractable approach consisting of three stages. The first is the selection of key stations, which are the main attraction points: railway or bus stations and airports, hospitals, university campuses, large stores and commercial centers and densely populated areas far away from the central area of the city, etc. The second stage is to connect the key stations to form a core network. Finally, the intermediate stations are located on the alignment resulting from the second stage. In the same paper, a linear integer programming model aiming at maximizing the trip coverage is used in order to solve the core network design problem in presence of an alternative mode of transportation. Later, Marín (2007) relaxed some restrictions on the lines. In his model the number of lines and the extremes of them are not fixed.

With the aim of modeling the user's behavior, Marín and García-Ródenas (2009) introduced a logit function in order to distribute the travelers between the rapid transit and private modes. In order to maintain the linear character of the program, they consider a piecewise linear interpolation of the logit function. In the paper of Escudero and Muñoz (2009) the problem is decomposed into two stages. The first one consists of determining the infrastructure network, and the second one determines the lines.

A methodological contribution to modeling and solving the transit network design problem can be found in Gutiérrez-Jarpa et al. (2013). These authors take into account the fact that the rapid transit networks are composed of line segments which often have to be constructed within broad corridors. These segments are later assembled into lines. The authors apply two criteria: minimizing construction cost and maximizing origin-destination traffic capture and computed Pareto-optimal solutions.

A multi-period capacity expansion problem was studied in Marín and Jaramillo (2008). In this paper the lines to be opened in each period are determined by taking into account an objective function which is a combination of community, passenger and operator oriented objectives. Since the general problem cannot be solved exactly, a heuristic procedure is designed to solve it.

Other approaches to solve the mathematical programming model for rapid transit network design problems are based on Benders decomposition (Marín and

Jaramillo 2009), genetic algorithms (Wang and Lin 2010) and simulated annealing (Kemanshani et al. 2010). Line configuration with assignment of passengers is studied in Guan et al. (2006). Finally, a recent line of research deals with network robustness aspects. Several ways of treating robustness have been studied: through the application of game theory (Laporte et al. 2010), by providing alternative routes to be used in case of a disruption (Laporte et al. 2011), through the concept of recoverable robustness (Cadarso and Marín 2012), and by the application of a GRASP to infrastructure railway network design problem (García-Archilla et al. 2013).

22.4 Location of Stations

The problem of locating stations is different in the case of locating a network from scratch than in the case of extending an already existing network. In the first case, several locations attract large volume of passengers and are obvious candidates for stations. The remaining stations must then be located with the help of analytical tools. Assuming that the alignments of the network are given, the problem of efficiently locating the stations arises. The first objective for the community and one of the most important ones for the operating company is to attract as many travelers as possible. To this end, in technical projects the population living in a circle centered at each station is used as an approximation. However, since walking distances are not Euclidean, this is a rough measure for the station attractiveness. In their paper, Laporte et al. (2002) use census tracts coupled with information on population density to estimate the actual walking distances. Different levels of attraction are applied in order to obtain a better estimation of the population covered (see Fig. 22.3). For each given location of the stations in a corridor, line coverage is subsequently defined. In that paper, given a discrete set of potential sites for stations, optimal locations are obtained by maximizing the line coverage with the help of an ad hoc defined acyclic graph and a longest-path algorithm.

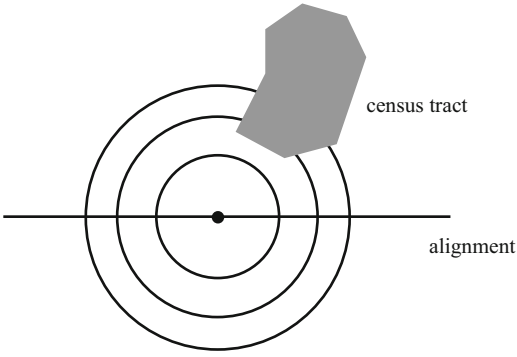


Fig. 22.3 Concentric catchment areas around a station intersecting with a census tract

However, the estimation of future ridership cannot only be based on line coverage since it depends not only on the location of the stations of the line, but also on the overall location of the network. In their paper, De Cea et al. (1986) use origin-destination pairs for computing the total population affected by an improvement of a transportation network. In Laporte et al. (2005), trip coverage is analytically defined and used to compute the network coverage as a good estimate of future ridership. The objective of minimizing the total travel time of passengers was introduced in Vuchic and Newell (1968). These authors considered the case of a population concentrated in a specified area and commuting to a central point. Their aim was to determine an optimal interstation spacing, while taking access time, kinematics of trains, dwell times and intermodal transfer times into account.

There exist a number of papers dealing with the location of new stations on general railway lines. Here we will highlight some of them. Hamacher et al. (2001) studied a problem in which the objective is to maximize the saving in passenger travel time when introducing new stations. Schöbel (2005) considered the maximization of coverage and the minimization of the number of new stations as bicriteria problems. Gross et al. (2009) presented two models combining the number of stations and the distances to them. In the first one, the objective is to minimize the number of new stations assuming that the demand is covered within a predefined distance. The second problem is NP-hard and consists of minimizing the sum of distances from the demand points to the closest (old or new) station under the constraint that the number of new stations is bounded above. They have considered two environments for each problem (a planar space with an ℓ_1 metric, and a network) thus giving rise to four cases. For each case, they have identified a polynomial complexity dominating set for the new stations. Körner et al. (2012) have dealt with the problem of locating two new facilities in a mixed planar-network space so that the number of trips between each pair of demand points is maximized. In this paper it is assumed that an alternative mode of transportation exists. The authors have analyzed the cases of segments and tree-networks and have also designed polynomial time algorithms. For the case of more than two facilities to be located on a segment, the big-cube-small-cube method has been shown to be efficient. In a very recent paper by Carrizosa et al. (2013), the kinematics of the trains are taken into account in order to minimize the total travel time when a given number of new stops are located, as well as the total travel time of traversing all edges, subject to the coverage of all demand points.

22.5 Conclusions

The design of rapid transit systems is a complex process which involves the participation of many players. These projects are fraught with high costs and uncertainty. Formulating models and designing algorithms for such problems is difficult since the objectives and constraints are not as well defined as in many operational research problems. Analytical techniques can be employed to assist

decision making or to solve some specific subproblems, but human judgment and intervention remain critical in the planning process. Over the past 20 years we have witnessed a number of important methodological advances in the area of rapid transit location planning. Several quality indices have been developed and mathematical models of increasing realism have been proposed, some of which can be solved directly by off-the-shelf solvers or by powerful heuristics. We expect to see in the near future models and algorithms capable of integrating operational and tactical considerations when solving the problem at the strategic planning level.

Acknowledgements This work was partially supported by the Canadian Natural Sciences and Engineering Research Council under grant 39682-10, by the Ministerio de Economía y Competitividad (Spain)/FEDER under projects MTM2009-14243 and MTM2012-37040, and by Junta de Andalucía (Spain)/FEDER under excellence projects P09-TEP-5022 and FQM-5849.

References

Barbadillo J, Saldaña J (2011) Navigation in large subway networks: an informational approach. Physica A 390:374–386

Barrena E, De-Los-Santos A, Mesa JA, Perea F (2013) Analyzing connectivity in collective transportation line networks by means of hypergraphs. Eur Phys J ST 215:93–108

Bruno G, Ghiani G, Improta G (1998) A multi-modal approach to the location of a rapid transit line. Eur J Oper Res 104:321–332

Bruno G, Gendreau M, Laporte G (2002) A heuristic for the location of a rapid transit line. Comput Oper Res 29:1–12

Cadarso L, Marín A (2012) Recoverable robustness in rapid transit network design. In: 15th meeting of the Euro Working Group on transportation, September 2012, Paris, pp 1–10

Carrizosa E, Harbering J, Schöbel A (2013) The stop location problem with realistic traveling time. In: Frigioni D, Stiller S (eds) 13th workshop on algorithmic approaches for transportation modeling, optimization and systems (ATMOS'13), OASICS Schloss Dagstuhl, Germany, pp 80–93

Criado R, Hernández-Bermejo B, Romance M (2007) Efficiency, vulnerability and cost: an overview with applications to subway networks worldwide. Int J Bifurcation Chaos 17:2289–2301

Current JR, ReVelle CS, Cohon J (1985) The maximum covering/shortest path problems: a multiobjective network design and routing formulation. Eur J Oper Res 21:189–199

De Cea J, Ortúzar JD, Willumsen LG (1986) Evaluating marginal improvements to a transport network: an application to the Santiago underground. Transportation 13:211–233

De-Los-Santos A, Laporte G, Mesa JA, Perea F (2012) Evaluating passenger robustness in a rail transit network. Transp Res C Emerg Technol 20:34–46

Derrible S, Kennedy C (2010) The complexity and robustness of metro networks. Physica A 389:3678–3691

Díaz JM, Mesa JA, Schöbel A (2004) Continuous location of dimensional structures. Eur J Oper Res 152:22–44

Dufourd H, Gendreau M, Laporte G (1996) Locating a transit line using tabu search. Locat Sci 4:1–19

Escudero LF, Muñoz S (2009) An approach for solving a modification of the extended rapid transit network design problem. TOP 17:320–334

García-Archilla B, Lozano AJ, Mesa JA, Perea F (2013) GRASP algorithms for the robust railway network design problem. J Heuristics 19:399–422

Gattuso D, Miriello E (2005) Compared analysis of metro network supported by graph theory. Netw Spat Econ 5:395–414
Gendreau M, Laporte G, Mesa JA (1995) Locating rapid transit lines. J Adv Transp 29:145–162
Gross DRP, Hamacher HW, Horn S, Schöbel A (2009) Stop location design in public transportation networks: covering and accessibility objectives. TOP 17:335–346
Guan JF, Yang H, Wirasinghe SC (2006) Simultaneous optimization of transit line configuration and passenger line assignment. Transp Res B Methodol 40:885–902
Gutiérrez-Jarpa G, Obreque C, Laporte G, Marianov V (2013) Rapid transit network design for optimal cost and origin-destination demand capture. Comput Oper Res 40:3000–3009
Hamacher HW, Liebers A, Schöbel A, Wagner D, Wagner F (2001) Locating new stops in a railway network. ENTCS 50:1–11
Institute of Electrical and Electronics Engineers (1990) IEEE Standard Computer Dictionary: a compilation of IEEE standard computer glossaries
Kermanshani S, Shafahi Y, Mollanejad M, Zangui M (2010) Rapid transit network design using simulated annealing. In: 12th WCTR, pp 1–15
Körner M-C, Mesa JA, Perea F, Schöbel A, Scholz D (2012) A maximum trip covering location problem with an alternative mode transportation on tree networks and segments. TOP 22:227–253
Laporte G, Mesa JA, Ortega FA (1997) Assessing the efficiency of rapid transit configurations. TOP 5:95–104
Laporte G, Mesa JA, Ortega FA (2002) Locating stations on rapid transit lines. Comput Oper Res 29:741–759
Laporte G, Mesa JA, Ortega FA, Sevillano I (2005) Maximizing trip coverage in the location of a single rapid transit alignment. Ann Oper Res 136:49–63
Laporte G, Marín A, Mesa JA, Ortega, FA (2007) An integrated methodology for the rapid transit network design problem. In: Geraets F, Kroon L, Schöbel A, Wagner D, Zaroliagis CD (eds) Algorithmic methods for railway optimization (Proceedings of ATMOS 2004). Lecture notes in Computer Science, vol 4359. Springer, Berlin/Heidelberg, pp 187–199
Laporte G, Mesa JA, Ortega FA, Pozo MA (2009) Locating a metro line in a historical city centre: application to Sevilla. J Oper Res Soc 60:1462–1466
Laporte G, Mesa JA, Perea F (2010) A game theoretic framework for the robust railway transit network design problem. Transp Res C Methodol 44:447–459
Laporte G, Marín A, Mesa JA, Perea F (2011) Designing robust rapid transit networks with alternative routes. J Adv Trans 45:54–65
Latora V, Marchiori M (2001) Efficient behavior of small-world networks. Phys Rev Lett 87:1987011–1987014
Latora V, Marchiori M (2002) Is the Boston subway a small-world network? Physica A 314:109–113
Marín A (2007) An extension to rapid transit design problem. TOP 15:231–241
Marín A, García-Ródenas R (2009) Location of infrastructure in urban railway networks. Comput Oper Res 36:1461–1477
Marín A, Jaramillo P (2008) Urban rapid transit network capacity expansion. Eur J Oper Res 191:45–60
Marín A, Jaramillo P (2009) Urban rapid transit network design: accelerated Benders decomposition. Ann Oper Res 169:35–53
Mesa JA, Boffey B (1996) A review of extensive facility on networks. Eur J Oper Res 95:592–603
Mesa JA, Ortega FA (2001) Park-and-ride station catchment areas in metropolitan rapid transit systems. In: Pursula M, Nittmäki J (eds) Mathematical methods on optimization in transportation systems. Kluwer, Dordrecht, pp 81–93
Metro de Granada (2013) http://www.urbanrail.net/eu/es/granada/granada.htm. Accessed 11 Nov 2013
Metro de Lisboa (2014) http://www.metrolisboa.pt/obras/projectos-de-expansao/. Accessed 27 July 2014

Musso A, Vuchic VR (1988) Characteristics of metro network and methodology for their evaluation. Transp Res Rec 1162:22–33

Rhode M (2014) World Metro Database. http://www.mic-ro.com/metro/table.html. Accessed 30 July 2014

Roth C, Kang SM, Batty M, Barthelemy M (2012) A long-time limit for world subway networks. J R Soc Interface 9:2540–2550

Schöbel A (2005) Locating stops along bus or railway lines-a bicriteria problem. Ann Oper Res 136:211–227

Seaton KA, Hackett LM (2004) Stations, trains and small-world networks. Physica A 339:635–644

Sen P, Dasgupta S, Chatterjee A, Sreeran PA, Mukherjee G, Manna SS (2002) Small-world properties of the Indian railway network. arXiv:cond-math/0208535v2 [cond-mat.soft] 31 Dec 2002

Sociedad del Metro de Sevilla S.A. (2001) Proyecto general básico de la red de metro de Sevilla y programación de fases (in Spanish), UTE Iberinsa and Ghesa

UITP (International Association of Public Transports) (2011) Metro service performance indicators. http://www.uitp.org/publications/corebriefs.cfm

Vuchic VR (2005) Urban transit operations, planning and economics. Wiley, Hoboken

Vuchic VR, Newell GF (1968) Rapid transit interstation spacings for minimum travel time. Transp Sci 2:303–339

Wang J-Y, Lin C-M (2010) Mass transit route network design using genetic algorithm. J Chin Inst Eng 33:301–315

Watts DJ, Strogatz SH (1998) The dynamics of 'small-world' networks. Nature 393:440–442

Wikipedia (2014) http://eu.wikipedia.org/wiki/list-of-metro-systems. Accessed 30 July 2014

Zhang J, Zhao M, Liu H, Xu X (2013) Networked characteristics of the urban rail transit networks. Physica A 392:1538–1546

Chapter 23
Districting Problems

Jörg Kalcsics

Abstract Districting is the problem of grouping small geographic areas, called basic units, into larger geographic clusters, called districts, such that the latter are balanced, contiguous, and compact. Balance describes the desire for districts of equitable size, for example with respect to workload, sales potential, or number of eligible voters. A district is said to be geographically compact if it is somewhat round-shaped and undistorted. Typical examples for basic units are customers, streets, or zip code areas. Districting problems are motivated by quite different applications ranging from political districting over the design of districts for schools, social facilities, waste collection, or winter services, to sales and service territory design. Despite the considerable number of publications on districting problems, there is no consensus on which criteria are eligible and important and, moreover, on how to measure them appropriately. Thus, one aim of this chapter is to give a broad overview of typical criteria and restrictions that can be found in various districting applications as well as ways and means to quantify and model these criteria. In addition, an overview of the different areas of application for districting problems is given and the various solution approaches for districting problems that have been used are reviewed.

Keywords Districting criteria • Political districting • Sales territory design • Service districting

23.1 Introduction

Most problems discussed in this book focus on the location of facilities: where to locate, how many to locate, when to locate, which type to locate, etc. However, although the driving force is the location of facilities, equally important is the second aspect of location problems that is usually not mentioned explicitly: the allocation of customers to facilities. Even if this task is trivial in many classical location problems like the p-median or the p-center problem (see Chaps. 2 and 4), only after deciding

J. Kalcsics (✉)
Institute of Operations Research, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
e-mail: kalcsics@kit.edu

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_23

about allocations can we evaluate a given facility configuration and, thus, try to find the optimal one. Hence, the allocations have a fundamental impact on the location of facilities and different rules of allocation will result in different evaluations of the same facility configuration. The aim of districting problems is now the other way around: we first find allocations—or, more generally, determine which customers should be served together—and then, if necessary, we find locations for the facilities serving the customers.

In general, districting is the problem of grouping small geographic areas, called basic units, into larger geographic clusters, called districts, in a way that the latter are acceptable according to relevant planning criteria. Typical examples for basic units are customers, streets, or zip code areas. Depending on the practical context, districting is also called territory design, territory alignment, zone design, or sector design. Three important criteria are balance, contiguity, and compactness. Balance describes the desire for districts of equitable size with respect to some performance measure for the districts. Depending on the context, this criterion can either be economically motivated, for example, equal sales potentials, workload, or number of customers, or have a demographic background, for example, the same number of inhabitants or eligible voters. A district is called contiguous if it is possible to travel between the basic units of the district without having to leave the district. Finally, a district is said to be geographically compact if it is somewhat round-shaped, undistorted, and without holes. Contiguous and compact districts usually reduce the travel time of the person responsible for servicing the district. Unfortunately, a rigid and concise mathematical definition of contiguity and compactness is often difficult and strongly depends on the available data. In addition, for each district often the location of a "facility" is either given or should be sought. This facility can be a branch office, a depot, or the home address of a sales person. Figure 23.1 shows an example of a districting plan for streets and for zip code areas.

Districting problems are motivated by quite different applications ranging from political districting over the design of districts for schools, social facilities, waste collection, or winter services, to sales and service territory design. Looking at the literature, it is striking that only few authors consider the districting problem independently from a practical background. Therefore, the aim of this chapter is to

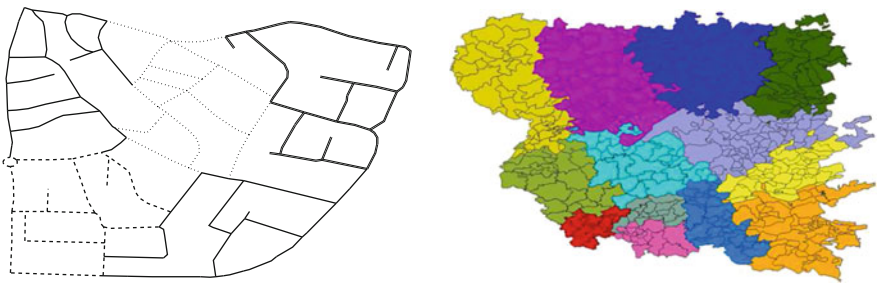


Fig. 23.1 An example of a districting plan for streets and for zip-code areas

give a broad overview of typical criteria and restrictions that can be found in the various districting applications as well as ways and means to quantify and model these criteria. As most districting applications have a strong spatial component, it is natural to integrate the algorithms into a Geographic Information System (GIS). In a modern GIS, users can access and utilize the rich variety of maps, spatial databases, and geographical objects available to appropriately mark out the problem and display the solutions, see also Chap. 19.

The rest of the chapter is organized as follows. The next section reviews the broad range of districting applications and identifies and motivates the different planning restrictions. In Sect. 23.3, basic notations are introduced. The next section discusses the most common criteria found in districting applications and discusses possible approaches to quantify these criteria and to incorporate them into districting models. Finally, Sect. 23.5 presents an overview of the different solution techniques for solving districting problems.

23.2 Applications

There are four major areas of application for districting problems: political districting, sales territory design, service districting, and distribution districting, and this section provides a comprehensive but non-exhaustive overview. But before we start, we mention a first "application" in the context of facility location that derives from the problem of aggregating demand points for location problems with the aim of reducing the complexity of the problem. Simchi-Levi et al. (2003) formulate the following guidelines (among others): aggregate demand points for 150–200 zones, make sure each zone has an approximately equal amount of total demand, and place aggregated points at the center of the zone. These guidelines read as a classical districting problem.

23.2.1 Political Districting

Political districting is the problem of dividing a governmental area, such as a city or a state, into constituencies from which political candidates are elected. Basic units typically correspond to census tracts, which are given as polygons, and the districts to the electoral constituencies. In general, the process of redistricting has to be periodically undertaken to account for population shifts. The length of these periods varies from country to country, e.g., in New Zealand every 5 years, in Canada and the U.S. every decade (after each census). In the past, political districting has often been flawed by manipulation aiming to favor some particular party or to discriminate against social or ethnic minorities. Since the responsibility for approving state and local districting plans usually falls to elected representatives, plans are likely to be shaped implicitly, if not overly, by political considerations, e.g., to keep them in

power. A famous case arose in Massachusetts in the early nineteenth century when the state legislature proposed a salamander-shaped electoral district in order to gain electoral advantage. The governor of the state at that time was Elbridge Gerry, and this practice became known as gerrymandering. See Lewyn (1993) for an interesting description of gerrymandering cases.

To avoid political interference, many states have set up a neutral commission to determine political boundaries satisfying a number of legislative and common sense criteria. Depending on the country or jurisdiction involved, these criteria may be enforced by legislative directive, judicial mandate, or historical precedent. However, there is no consensus in political science, law, or geography on which criteria are legitimate for the districting process, i.e., satisfy the neutrality condition. Moreover, it is often unclear how they should be measured (Williams 1995). One important issue at stake is population equality. To respect the principle of "one man-one vote", i.e., every vote has the same power, all districts should contain approximately the same number of voters, i.e., be balanced. In the U.S., population equality has been deemed by the courts to be very important, and as a result, the total deviation of congressional districts from perfect balance was less than 1 % after the last census in 2000 (Webster 2013). In other countries, the allowed deviations are usually higher (Handley and Grofmann 2008). Two other important criteria always being mentioned are contiguity and compactness which both aim at preventing gerrymandering. While contiguity is generally undisputed and easy to verify, this is not the case for compactness. There is a broad discussion on how to quantify this criterion adequately (Horn et al. 1993), and whether it is relevant in the first place because an algorithm will never gerrymander on purpose as long as it is does not use political data (Garfinkel and Nemhauser 1970). Moreover, if an adequate minority representation is sought for, this may sometimes only be achieved through non-compact districts (Dixon 1968). Other—often disputed—criteria are the conformity to administrative boundaries, e.g., cities or counties, the preservation of communities of interest, socio-economic homogeneity or a fair representation of minority voters across the districts, the similarity with the previous electoral districts, or the consideration of topological obstacles, like mountain ranges, lakes, or rivers (cf. George et al. 1997; Parker 1990; Bozkaya et al. 2011). An excellent review on typical criteria for political districting and their eligibility is given in Webster (2013).

When discussing automated procedures in the literature, it is always noted that they are non-partisan and neutral as long as they do not use political data and, hence, prevent gerrymandering. However, even if the computer does not gerrymander on purpose, it may still do it accidentally, precisely because no political data is taken into account. Therefore, Puppe and Tasnádi (2008) recently introduced the notion of an (ex post) unbiased districting plan. In such a plan the number of districts won by each party respects the relative strength of the party in the population as close as possible. They focus on game theoretical aspects of the problem; see also Nagel (1965). However, one has to do a careful weighing up to avoid forthright politically biased criteria that lead, in spirit, to gerrymandering.

23.2.2 Sales Territory Design

The important but expensive task of designing sales territories is common to all companies that operate a sales force and need to subdivide the market area into regions of responsibility that are each attended to by one or more sales representatives. According to Zoltners and Sinha (2005), approximately every tenth full-time employee in the U.S. is working as a field and retail sales person and the expenditure for them is more than three trillion dollars every year. Territories with low sales potential, intense competition, or too many small accounts lead to low morale, poor performance, a high turnover rate, and an inability to assess the productivity of individual territories. Therefore, well-planned decisions enable an efficient market penetration and lead to decreased costs and improved customer service and sales. Zoltners and Sinha (2005) "guestimate" that a good territory alignment can increase sales by 2–7 % compared to an average alignment. In the related literature, districts are predominantly called territories and districting is termed territory alignment or territory design.

In the classical problem, the task is to assign a given set of (prospective) customer accounts, each with a fixed market potential, to the individual members of the sales force such that each customer has a unique representative and each sales person faces equitable workload and travel time and has an equal income opportunity in terms of incentive pay (Zoltners and Sinha 2005). Thus, basic units correspond to accounts and are usually given as points. Concerning the travel time, if a sales person visits each customer every day, then the travel time is proportional to the length of a TSP tour. However, the workload of districts is usually balanced over 2–4 weeks and some customers may have to be visited only once during this time whereas others require weekly service. Moreover, customers may have time windows, tours may include overnight stays, and so on, which makes the actual computation of the travel times almost impossible. Hence, in most cases one has to rely on estimates. Typically, a sales person is exclusively responsible for all customers within a specific geographic region. However, in large companies sometimes a sales person is only responsible for a certain product segment or accounts of a particular size within his region. In such cases, sales territories may overlap. For practical examples of sales territory design see Fleischmann and Paraschis (1988), Zoltners and Sinha (2005), López-Pérez and Ríos-Mercado (2013).

Three classical sales districting criteria are again balance, contiguity, and compactness. In contrast to political districting, typically more than one performance measure has to be balanced, for example workload and sales potential. A district with comparatively many small accounts or customers with low sales potential will yield lower sales and, hence, lower incentives for the responsible sales person than a district with an equitable workload but only high potential accounts. This disparity will lead to discontent among the sales persons and, in the long run, lower sales for the company. Having said that, only few authors consider more than one balancing criterion: Deckro (1977), Zoltners and Sinha (1983), Ríos-Mercado and Fernández (2009). Contiguous districts are desired to obtain clearly defined geographic areas

of responsibility and, together with compactness, to reduce the unproductive travel time of the sales force. Unfortunately, as basic units are typically points, it is not clear how to assess contiguity. Moreover, it is important to point out that the desire for compact districts is born out of necessity because the actual travel times are usually impossible to determine efficiently. The hope is that geographically compact and contiguous districts result in smaller travel times on a day-to-day basis than non-compact and/or non-contiguous districts.

As the main goal of most companies is to maximize profit, several authors relax the assumption that the sales potential of customers is fixed. Instead, they propose an integration of time-effort allocation and territory design methods to increase profit while maintaining the equitable workload criterion (cf. Lodish 1975; Glaze and Weinberg 1979; Zoltners and Sinha 1983). These models not only assign customers to sales people but also determine how much time should be invested in the customer. Some authors even object that equity is not the primary goal for most companies. Instead, the aim should be to maximize profits, regardless of any balancing aspect (Skiera and Albers 1994; Drexl and Haase 1999). However, in practice sales persons are typically reluctant to implement such detailed call plans resulting from pure profit maximizing approaches (Zoltners and Sinha 2005). Moreover, designing territories is a mid- or even long-term decision whereas time-effort allocation is an operational problem that is influenced by weather (espc. in the beverage industry), sales promotions, etc. Thus, these two problems should be addressed separately.

Often, the number of districts to be designed is predetermined by the designated sales force size (Fleischmann and Paraschis 1988). If the size is not self-evident, methods based on the total workload involved in covering the entire market compared to the available time per sales person can be used. Another possibility is to follow the "decreasing returns" principle and add sales persons to the sales force as long as the expected increase in profit exceeds the expected increase in costs (Howick and Pidd 1990; Zoltners and Sinha 2005).

As sales persons have to visit their territories regularly, their home-base, e.g., office or residence, is an important factor to be considered in the alignment process. However, there is no consensus as to whether predetermined locations should be kept or be subject to the planning process. On the one hand, most sales persons have strong preferences for home-base cities. Hence, such locations should be respected or determined prior to the alignment to socialize them with the sales management (Zoltners and Sinha 2005). On the other hand, addresses and sales personnel frequently change and the management often does not want sales persons residences to overly influence the definition of territories (Fleischmann and Paraschis 1988).

23.2.3 Service Districting

The problem of designing service districts appears in various contexts. One area of applications focuses on social facilities, like hospitals or public utilities. Sometimes

districts are needed to define for each inhabitant which facility he should visit to obtain service, for example for preventive medical examinations, or to determine areas of responsibility of home-care visits by healthcare personnel, like nurses or physiotherapists. The goal is to determine contiguous districts that have a good accessability with respect to public transportation and have an equitable workload based on service and travel time or account for a high capacity utilization of the social facility (cf. Minciardi et al. 1981; Blais et al. 2003; Benzarti et al. 2013).

A second field of applications deals with providing service to streets. A classical problem concerns the design of districts for postal or leaflet delivery. Instead of considering each household separately, districts are composed of whole streets. Thus, basic units correspond to streets and each basic unit typically has two attributes: the times required to traverse the street with and without providing service. The task is to partition the streets into a given number of districts such that the required delivery time is approximately the same for all districts and does not exceed the working time restriction of the deliverer. The delivery time is proportional to the length of a Chinese postman tour through the district, which can be computed efficiently. Moreover, the delivery districts should be contiguous, incur little deadheading, and should not overlap, i.e., be geographically compact (Bodin and Levy 1991; Butsch et al. 2014). A common characteristic of these applications is that the deliverer either walks through his district on foot or goes by bike so that one-way streets are no hindrance. If a street is too wide or has too much traffic to serve it in a zig-zag pattern, then each side of the street is modeled as a separate basic unit. A similar problem arises in the context of meter reading in power distribution networks (Silva de Assis et al. 2014). Closely related are districting problems for solid waste disposal, salt spreading, and winter gritting (Hanafi et al. 1999; Muyldermans et al. 2002; Lin and Kao 2008). The criteria are almost identical to postal delivery. The only differences are that vehicles typically have to respect one-way streets and have difficulties making U-turns, and that their tours have to include a depot, e.g., to drop off waste or refill salt. All these aspects make the computation of the travel times more difficult. Other applications deal with the design of patrol districts for police cars and primary response areas for ambulances, where the districts additionally should have an average response time and/or incident arrival rate below a given threshold (Baker et al. 1989; D'Amico et al. 2002; Xu and Yum 2010).

Other applications deal with the problem of assigning residential areas to schools (Ferland and Guénette 1990; Schoepfle and Church 1991). Criteria to be taken into account are capacity limitations and an equal utilization of the schools, maximal or average travel distances for students, good accessability, and ethnic balance. Another aspect is to decide which students should walk to school and which should take the school bus. Districting problems also occur in electric power networks. According to Bergey et al. (2003), the World Bank regularly faces the challenge of helping developing countries to move from state owned, monopolistic electric utilities to a more competitive environment with multiple electricity service providers. At that, they face the task of partitioning the physical power grid into economically viable districts (distribution companies). The main aim is to determine non-overlapping and contiguous districts with approximately

equal revenue potential (to foster competition) which are compact over a geographic region (to be easier to manage and more economical to maintain).

23.2.4 Distribution Districting

Another important field of applications is the design of pickup and delivery districts in logistics. Typically, such problems are modeled and solved as vehicle routing problems. However, if there exists considerable uncertainty in the demand of customers, several authors propose a two-phase approach that first builds the pickup and delivery districts and then does the routing on a day-to-day basis. This conforms with the well-known "cluster first–route second" paradigm for vehicle routing problems. Hence, basic units correspond to potential customers, given as points, and the task is to partition the set of customers into districts, one for each driver, such that the districts satisfy certain planning criteria. A first advantage of these fixed customer assignments is that the driver becomes familiar with his district. This, in turn, increases the driver's performance since he becomes quicker at finding customer addresses, localizing offices within buildings as well as organizing his routes (Zhong et al. 2007). A second advantage is that customers become familiar with their drivers, which increases customer satisfaction (Jarrah and Bard 2012). These advantages however have to be carefully weighed against flexible customer assignments on a daily basis which enable the planner to maximize the driver utilization and minimize the routing costs (Zhong et al. 2007).

Concerning the criteria for the districting process, districts should be contiguous and compact, and the workload should either be balanced or at least not exceed a given upper bound, e.g., the driver working time. The workload includes the service time at the customers and typically also an estimate of the average travel time within the district and to a centralized depot (Galvão et al. 2006; Haugland et al. 2007; Zhong et al. 2007; Jarrah and Bard 2012; Lei et al. 2012).

A final application concerns the establishment of a distribution center which involves a considerable level of risk due to its enormous start-up investment and volatile customer demand patterns. One way of reducing this risk is to avoid both overcrowding and, especially, underutilization of centers by balancing the allocation of customers to them (Zhou et al. 2002).

23.3 Notations

This section introduces notations for the main components of districting problems.

23.3.1 Basic Units

A districting problem comprises a set $J = \{1, \ldots, n\}$ of *basic units*, sometimes called sales coverage units, basic areas, or geographical units. Each basic unit represents a geometric object in the plane: a point, e.g., a geo-coded address, a line segment, e.g., a street, or a polygonal area, e.g., a zip code area, county, or predefined company trading area. The distance between two basic units $i, j \in J$ is denoted as $d_{ij} = d(i, j)$. Typical examples for d_{ij} are Euclidean (cf. Fleischmann and Paraschis 1988) or road distances (cf. Ríos-Mercado and Salazar-Acosta 2011). The latter have the advantage that they can properly reflect obstacles like rivers or mountain ranges. For non-point objects, distances are either computed between representative points, e.g., the midpoint of a street or the centroid of a polygon, or as the surface-to-surface distance.

Moreover, one or more quantifiable attributes, called *activity measures*, are associated with each basic unit. Typical examples are service times, estimated sales potential, or number of voters. Sometimes, they also include an estimate of the travel time for visiting the basic unit (Jarrah and Bard 2012). The activity measures are all assumed to be deterministic. Let w_j^q denote the q-th activity measure of basic unit $j \in J$, $1 \leq q \leq Q$, where Q is the number of different attributes to be considered. If $Q = 1$, the superscript is usually omitted.

If explicit neighborhood information is given for the basic units, then $G = (V, E)$ denotes the *neighborhood* or *contiguity graph* where $v_j \in V$ corresponds to $j \in J$ and $\{v_i, v_j\} \in E$ iff basic units i and j are neighboring. The length of edge $\{v_i, v_j\}$ is d_{ij}. Finally, $N(j) \subseteq V$ denotes the set of basic units adjacent to $v_j \in V$.

23.3.2 Districts

A *district* D_k, $1 \leq k \leq p$, is a subset of basic units, where p is the total number of districts. The number of districts can either be fixed in advance, e.g., the number of political districts to create or the number of available nurses for elderly care, or be subject to planning, e.g., the minimal number of salespersons required to service all customers or the minimal number of patrol cars to ensure a certain response time. The q-th activity measure of a district is the sum of the activity measures of its basic units, i.e., $w^q(D_k) = \sum_{j \in D_k} w_j^q$. For $Q = 1$, $w^1(D_k)$ is simply called the *size* of the district. Note that sometimes the size also includes an estimate of the (expected) travel time. However, as travel times are represented through the compactness criterion, we refrain from including them and just mention when this may change things.

In some applications the location c_k of a facility is associated with each district D_k. This may be some predefined site, e.g., a hospital providing preventive medical care, or be an outcome of the districting process, e.g., the optimal location of a sales

office. In districting, this location is called the *center* of the district. One has to be aware of the ambiguity with the notion of a center in location theory, which is something different, see Chap. 4. Typically, the center coincides with a basic unit, i.e., $c_k \in J$. A predetermined set of centers is denoted by J_c.

Finally, a *districting plan* $\mathscr{D}$ is defined as a set of p districts $\mathscr{D} = \{D_1, \ldots, D_p\}$.

23.3.3 Problem Formulation

The districting problem can now informally be described as follows: Partition all basic units J into a number of p districts that satisfy the planning criteria of balance, compactness, and contiguity and, if required, locate a center within each district. Unfortunately, in contrast to many other optimization problems, there does not exist *the* mathematical model for districting problems. This is mainly due to the considerable ambiguity on how to quantify the different planning criteria and in the motivation and relevance of some of them.

23.4 Districting Criteria

This section presents an overview over typical criteria employed in districting problems and various ways and means to quantify them. In the following, a measure for a criterion applied to a single district (the whole districting plan) is termed a local (global) measure. Moreover, if not explicitly stated otherwise, let $Q = 1$.

23.4.1 Complete and Exclusive Assignment

In most cases, each basic unit is assigned to exactly one district, i.e., the districts define a partition of the set J of basic units:

$$D_1 \cup \cdots \cup D_p = J \quad \text{and} \quad D_l \cap D_k = \emptyset, \quad 1 \leq l, k \leq p, l \neq k\,.$$

The requirement of exclusive assignment is sometimes also termed *integrity*. For political districting, these criteria are obvious. In sales territory design, unique allocations result in transparent responsibilities for the sales force avoiding contentions and allowing the establishment of long-term customer relations.

23.4.2 Balance

This criterion is one of the trademarks of districting problems. It expresses the desire for districts of equitable size with respect to the activity measure(s). In political districting, this criterion is employed to ensure the "one man–one vote" principle, and in sales territory design to avoid districting plans with large discrepancies in terms of workload, sales potential, or travel time.

Due to the discrete structure of the problem and the integrity assumption, perfectly balanced districts can generally not be accomplished. There exist different approaches in the literature to quantify imbalance and to incorporate the criterion into the districting process. The most common local measure is based on the relative deviation of the district size $w(D_k)$ from the mean district size $\mu = w(J)/p$:

$$bal(D_k) \;=\; \left| \frac{w(D_k) - \mu}{\mu} \right|, \quad 1 \le k \le p$$

(cf. Forman and Yue 2003; Ríos-Mercado and Fernández 2009; Silva de Assis et al. 2014). The larger this deviation is, the worse is the balance. A district D_k is perfectly balanced, if $bal(D_k) = 0$. If the district sizes also contain a solution dependent performance measure, like travel times, then this affects μ and the balance of one and the same district may be different in different districting plans. Another approach concedes a priori a certain relative deviation $\alpha > 0$ from perfect balance and only measures the imbalance exceeding this threshold (Bodin and Levy 1991; Bozkaya et al. 2011)

$$bal(D_k) \;=\; \frac{1}{\mu} \max\{w(D_k) - (1+\alpha)\mu,\; (1-\alpha)\mu - w(D_k),\; 0\}\,,$$

i.e., the district is balanced if its size is between this lower and upper bound. Instead of determining the bounds based on the mean district size, they are sometimes directly motivated by the application, e.g., the working time restrictions of the mailman or the sales potential required to ensure a decent living for the sales person.

Using these local measures, the global balance of a districting plan is then typically computed as the maximal balance of a district

$$bal^{max}(\mathscr{D}) \;=\; \max_{k=1,\ldots,p} bal(D_k)\,.$$

Less common are the sum over all districts (Bozkaya et al. 2003; Bodin and Levy 1991) or a convex combination of both (Butsch et al. 2014):

$$bal^{sum}(\mathscr{D}) \;=\; \sum_{k=1}^{p} bal(D_k) \quad \text{and} \quad bal^{cv}(\mathscr{D}) \;=\; \lambda\, bal^{sum}(\mathscr{D}) + (1-\lambda)\, bal^{max}(\mathscr{D})\,,$$

with $\lambda \in (0,1)$. The convex combination alleviates some of the weaknesses of bal^{sum} and bal^{max}. The latter does not take into account the balance of all districts and sometimes yields rather poor solutions on average whereas the former allows a few highly unbalanced districts to be compensated by some well-balanced districts. A different global approach computes the range of district sizes (Tavares-Pereira et al. 2007)

$$bal^{rng}(\mathscr{D}) = \max_{k=1,\dots,p} w(D_k) - \min_{k=1,\dots,p} w(D_k)\,.$$

23.4.2.1 Mathematical Modelling

In districting models, there is no clear trend on whether to treat balance as a hard constraint (Hess et al. 1965; Fleischmann and Paraschis 1988; Zoltners and Sinha 2005) or to include it in the objective function (Blais et al. 2003; Ricca and Simeone 2008; Silva de Assis et al. 2014). In the former case, the size of each district is required to lie between a given lower and upper bound. Some authors even do both (Bergey et al. 2003; Salazar-Aguilar et al. 2013b). All of the above measures easily give rise to linear expressions.

23.4.3 Contiguity

Almost all districting approaches require districts to be contiguous. In political districting, this criterion should prevent gerrymandering. For the other types of applications, contiguous districts reduce the day-to-day travel distances for sales persons, delivery vans, snow ploughs, mailmen, etc. Unfortunately, a rigid and concise mathematical formulation of contiguity is difficult for basic units representing points.

23.4.3.1 Graph-Based Measures

If basic units are lines or polygons, it is easy to derive explicit neighborhood information. For example, two zip-code areas are neighboring if they share a common border, or two streets if they meet in a crossroad. In the former case, sometimes an additional requirement is the existence of a direct road connection between the two basic units. In general, two basic units are called *neighboring*, if their geometric representations have a nonempty intersection. This information is stored in the neighborhood graph $G = (V, E)$, and a district is contiguous if the basic units of the district induce a connected subgraph in G.

If basic units are represented by points, e.g., customer addresses, it is not clear how to assess contiguity. Over the years, different surrogate definitions for

contiguity have been proposed. One approach is based on proximity graphs to estimate the adjacency of points. One such graph is the Gabriel graph, in which two nodes v_i and v_j are connected by an edge if and only if the disc with antipodal points v_i and v_j does not contain any other node in its interior (Gross and Yellen 2003). A second approach to construct a contiguity graph is based on the Voronoi diagram (Lei et al. 2012). Two basic units are defined to be adjacent, iff their Voronoi cells have a common link within the smallest axis-parallel rectangle enclosing all basic units (for a definition of Voronoi diagrams and cells see Aurenhammar et al. 2013). A third construction of the proximity graph is to start with a complete graph and then sequentially go over all edges and delete for two intersecting edges in the planar representation of the graph the longer or more costly one (Haugland et al. 2007). All three graphs are planar. Moreover, by definition the Gabriel graph is a subset of the Voronoi-based graph.

Example 23.1 An example for these three proximity graphs for a point set with 26 basic units is depicted in Fig. 23.2. The Gabriel graph defines the most strict neighborhood relation. The graphs obtained by Lei et al. (2012) and Haugland et al. (2007) are fairly similar. The main difference is that the latter typically establishes more adjacencies along the boundary of the convex hull of the point set. Just by looking at the graphs it is difficult to decide which one is more suitable.

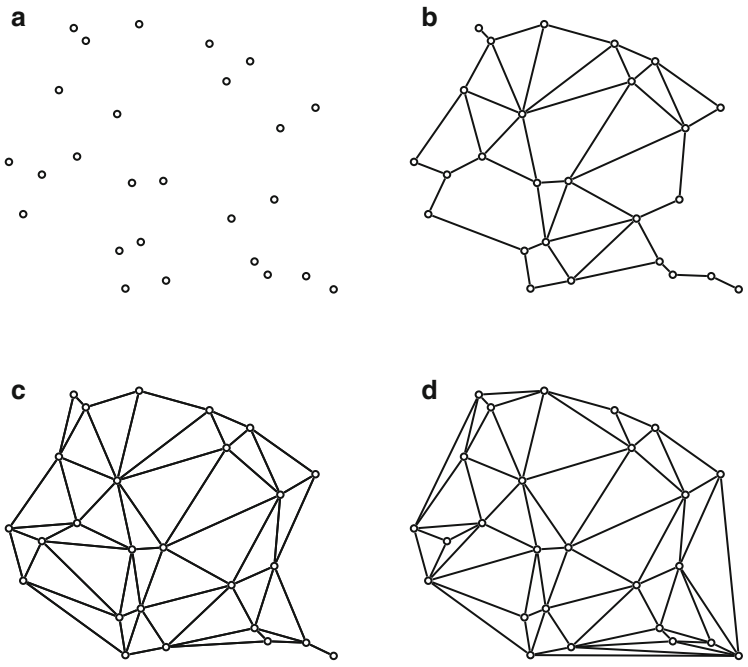


Fig. 23.2 Three different approximate contiguity graphs. (**a**) Point set of basic units. (**b**) Gabriel graph. (**c**) Voronoi-based graph. (**d**) Non-crossing edges graph

Finally, if the underlying road network is given, yet another possibility is to define two basic units as being adjacent, if the shortest path between the two does not contain another basic unit.

23.4.3.2 Geometric Measures

If no neighborhood information for basic units is given or can reasonably be derived, an alternative is to determine the overlap between the districts. For example, by computing the convex hull $ch(D_k)$ around each district D_k and defining a district to be contiguous if no basic unit of another district lies in its convex hull, i.e., $ch(D_k) \cap ch(D_l) = \emptyset, \forall\, l \neq k$ (Kalcsics et al. 2005; Jarrah and Bard 2012). One advantage of this approach is that convex districts usually prevent the crossing of routes of different districts, a characteristic that typically implies inefficient routes.

23.4.3.3 Mathematical Modelling

In districting models, contiguity is always treated as a hard constraint (except in Hanafi et al. 1999). One possibility to include it in a mathematical programming formulation is the following: Let $c_k \in J_c$ be the predetermined center of district k and $S \subseteq J \setminus \{N(c_k) \cup \{c_k\}\}$ be a subset of basic units that are not adjacent to c_k. If all elements of S are assigned to k, i.e., $S \subset D_k$, then at least one basic unit not in S that is adjacent to an element of S must also be assigned to k:

$$\sum_{j \in \bigcup_{i \in S} N(i) \setminus S} x_{kj} - \sum_{j \in S} x_{kj} \geq 1 - |S| \qquad \forall\, S \subseteq J \setminus \{N(c_k) \cup \{c_k\}\},$$

where x_{kj} is 1 if $j \in J$ is assigned to district k and 0 otherwise (Drexl and Haase 1999). The drawback of this formulation is, that it requires an exponential number of constraints (although it gives naturally rise to a cut generation approach, Ríos-Mercado and López-Pérez 2013). A second possibility that only needs a linear number of constraints is based on network flow constraints. Each basic unit has one unit of supply, and the district centers act as sinks. District k is contiguous iff there exists a flow from its basic units to c_k that only passes basic units in D_k:

$$\begin{aligned}
\sum_{i \in N(j)} f_{ji} - \sum_{i \in N(j)} f_{ij} &= x_{kj} && \forall\, j \in J \setminus \{c_k\} \\
\sum_{i \in N(j)} f_{ij} &\leq (n-2)\, x_{kj} && \forall\, j \in J \setminus \{c_k\} \\
\sum_{i \in N(c_k)} f_{i,c_k} &\leq n - 1,
\end{aligned}$$

where f_{ij} is the flow from basic unit i to j and $f_{c_k,j} = 0, \forall\, j \in N(c_k)$ (Shirabe 2009).

A simpler approach is to require that each district is a subtree of a shortest path tree $T(c_k)$ rooted at the district center c_k, where the edge lengths typically correspond to road distances or are all assumed to be 1. Then, for each basic unit j of district k, at least one of the adjacent basic units $i \in N(j)$ that immediately precedes j on some shortest path to the center c_k also has to be included in the district:

$$x_{kj} \leq \sum_{i \in S_j} x_{ki} \qquad \forall\, j \in J \setminus \{c_k\},$$

where $S_j = \{i \in N(j) \mid i$ immediately precedes j on some shortest path from j to $c_k\}$ (Zoltners and Sinha 1983; Mehrotra et al. 1998). Although this excludes some contiguous districts, these are unlikely to be compact, as they typically have large protrusions or indentations, or contain enclaves.

It is straight forward to extend all of the above constraints to the case where the choice of district centers is part of the optimization. For geometric contiguity measures obviously only informal mathematical formulations can be derived.

Remark 23.1 Only few authors try to derive approximate neighborhood graphs for point-like basic units. The majority simply does not consider contiguity at all and tries to obtain districts with little overlap through an appropriate compactness measure, see also Example 23.3.

23.4.4 Compactness

A district is said to be geographically compact if it is somewhat round-shaped and undistorted. The motivation for compact districts is almost identical to ensuring contiguity: to prevent gerrymandering or to reduce the day-to-day travel distances within the districts. Although being a very intuitive concept, a rigorous definition of compactness does not exist and, moreover, strongly depends on the geometric representation of basic units. In the context of political districting, typically measures based on the shape of districts are employed whereas in sales and distribution districting, distance-based measures are predominant. In the following, the most common ones for both approaches are presented.

23.4.4.1 Geometric Measures

If basic units are given as polygons, geometric approaches based on the area or perimeter of a district can be used to quantify compactness. Two common local measures are the Reock and Schwartzberg tests. The former calculates the ratio

of the district area to the area of the smallest enclosing circle, while the latter determines the ratio of the districts perimeter length to the circumference of a circle with equal area

$$cmp(D_k) = \frac{A(D_k)}{\pi r_{enc}^2} \quad \text{and} \quad cmp(D_k) = \frac{P(D_k)}{2\sqrt{\pi\, A(D_k)}},$$

where $A(\cdot)$ and $P(\cdot)$ denote the area and the length of the perimeter, respectively, of a district and r_{enc} the radius of the smallest enclosing circle (Young 1988). For the Reock (Schwartzberg) test, larger (smaller) ratios indicate greater compactness. Other measures relate the activity of a district with the total activity of all basic units within the smallest enclosing circle (Ricca and Simeone 2008) or determine the ratio of the squared diameter of a district and its area (Garfinkel and Nemhauser 1970). A common global measure for the compactness of a districting plan is based on the length of the boundary between districts, i.e., the total length of the perimeter of the districts in the interior (Bozkaya et al. 2003; Lei et al. 2012)

$$cmp(\mathscr{D}) = \sum_{k=1}^{p} P(D_k) - P(J).$$

Short inter-district boundaries typically result in compact districts. Numerous other measures have been discussed in the literature. Unfortunately, none of them is comprehensive; some fail to detect districts that are obviously noncompact, others assign a low rating to visibly compact districts (Niemi et al. 1990; Horn et al. 1993; Williams 1995).

To use geometric measures for basic units representing points or lines, one can try to give "shape" to the districts, for example through the smallest enclosing rectangle or circle, or through the convex hull. Instead of the convex hull, one can also use χ-shapes, which are polygons enclosing the point set that can provide a better fit to the points than the convex hull (Duckham et al. 2008). However, much more common are the following, distance-based measures:

23.4.4.2 Distance-Based Measures

Distance-based measures are used predominantly in applications where people have to travel within the districts, e.g., sales- or mailmen. This confers with the motivation of compact districts in these applications: to reduce the day-to-day travel times. Moreover, in these applications basic units typically represent points or lines, making geometric measures unapplicable in the first place. The most common group of local measures is based on the sum of distances between the center of a district

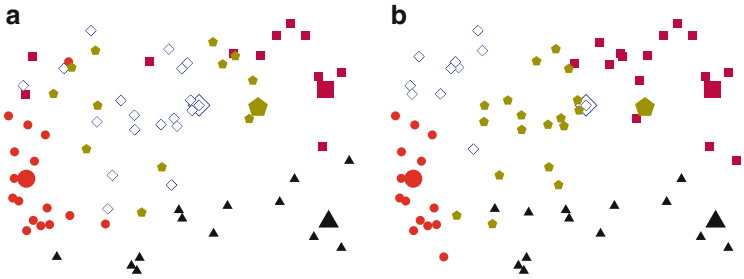


Fig. 23.3 Districting plans for two center-based compactness measures without contiguity. (**a**) Districts for $cmp_{ud}(\cdot)$. (**b**) Districts for $cmp_{wd^2}(\cdot)$

and its basic units. Variations exist in whether the distances are weighted with activity measures or not (w/u) and whether distances are squared or not (d^2/d)

$$cmp_{ud}(D_k) = \sum_{j \in D_k} d_{c_k,j} \qquad cmp_{ud^2}(D_k) = \sum_{j \in D_k} d^2_{c_k,j}$$

$$cmp_{wd}(D_k) = \sum_{j \in D_k} w_j \, d_{c_k,j} \qquad cmp_{wd^2}(D_k) = \sum_{j \in D_k} w_j \, d^2_{c_k,j}$$

(Bard and Jarrah 2009; Bergey et al. 2003; Hess and Samuels 1971; Zoltners and Sinha 2005). The second and forth measure are also known as the (weighted) moment of inertia (Hess et al. 1965). Although the four local compactness measures follow the same idea, the resulting districts may look considerably different as the following example shows.

Example 23.2 Consider a point set of $n = 75$ basic units that has to be partitioned into $p = 5$ districts, each having a predetermined center. The allowed relative deviation in terms of balance from the mean district size μ is 5 %, and contiguity is not explicitly imposed. Figure 23.3 shows the resulting districting plans that minimize the sum of the two center-based compactness measures $cmp_{ud}(\cdot)$ and $cmp_{wd^2}(\cdot)$ over all districts. The enlarged icons represent the district centers.

Having in mind that compactness acts as a proxy for travel times, the most natural measure is $cmp_{ud}(\cdot)$. However, we observe that there is a considerable overlap in the districts for this measure, especially between the districts represented by the diamond and pentagon shaped basic units. A much better visual separation is instead obtained for the weighted squared distance, $cmp_{wd^2}(\cdot)$, even if some district centers now lie outside their actual district (again, diamonds and pentagons). A large overlap between districts typically yields less efficient routes for sales persons. To underline this observation, we determine for each district the TSP tour through all basic units, including the center. The total lengths of the TSP tours for the two districting plans are: 92.78 and 73.56. The travel distances for the weighted squared distance are 20 % smaller than for $cmp_{ud}(\cdot)$. The results for $cmp_{wd}(\cdot)$ and $cmp_{ud^2}(\cdot)$ in terms of

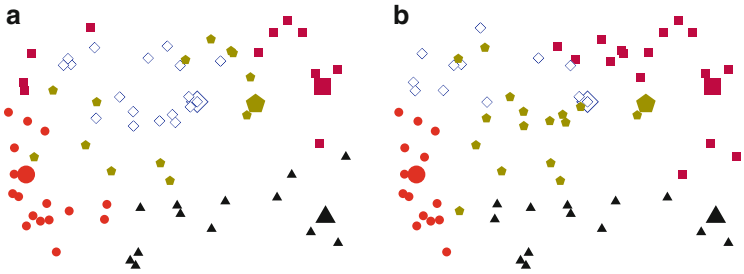


Fig. 23.4 Districting plans for two center-based compactness measures with contiguity. (**a**) Districts for $cmp_{ud}(\cdot)$. (**b**) Districts for $cmp_{wd^2}(\cdot)$

overlap and travel distances are between the other two measures, with the former being slightly better.

The situation is different if we try to enforce contiguity. Assume that an approximate neighborhood graph has been computed using the approach in Haugland et al. (2007). Using the contiguity constraints of Shirabe (2009), the resulting districting plans for $cmp_{ud}(\cdot)$ and $cmp_{wd^2}(\cdot)$ are shown in Fig. 23.4. The separation between the districts for $cmp_{ud}(\cdot)$ is clearer than before. However, even if the total length of the TSP tours reduces considerably (from 92.78 to 81.15), the districts consisting of the diamond, pentagon, and square shaped basic units are still distorted and will receive little approval from planners. (The square shaped district is connected since there exists an edge along the top of the point set.) For $cmp_{wd^2}(\cdot)$ the overlap is not much different from the previous plan, and the total travel distance even slightly decreased to 72.97. The main difference is that the centers are now all included in their districts, if only at the boundary.

This example illustrates the considerable differences between districting plans for different compactness measures and the influence of contiguity constraints. However, this is just a single example, and the observations cannot be generalized without further testing. Also, the length of a TSP tour is just an indicator for travel distances, as a sales person may not visit all customers on a single day.

The fact that squared distances produce compact but non-contiguous districts for fixed centers has been observed several times in the past (Hojati 1996; Schröder 2001). An important factor influencing the shape of districts is the spatial distribution of the district centers. If they are spread evenly, the differences between the measures in terms of district overlap will decrease, see Example 23.3. However, this uneven distribution is not unusual as sales force residences often concentrate in certain areas, e.g., larger cities, and sometimes even have the same address. Also the threshold for the allowed balance deviation has an impact on the compactness of solutions. The smaller the threshold value is, the larger the overlap between districts will get.

Instead of taking the sum, one could also take the maximum for each of the center-based measures (cf. Elizondo-Amaya et al. 2014; Ríos-Mercado and

Fernández 2009; Muyldermans et al. 2003). However, this leaves considerable freedom for assignments below the maximal distance and typically increases the overlap. A slightly different approach is based on the maximal pairwise distance and the weighted sum of pairwise distances

$$cmp_{mpw}(D_k) = \max_{i,j \in D_k, i \neq j} d_{ij} \qquad cmp_{spw}(D_k) = \sum_{i,j \in D_k, i \neq j} w_i \, w_j \, d_{ij}$$

(Ríos-Mercado and Salazar-Acosta 2011 and Blais et al. 2003, respectively).

In case of measures based on the sum (maximum) of distances, the global compactness of a districting plan is then usually also computed as the sum (maximum) over all districts. But sometimes also a sum-max combination is used or a convex combination of sum and max (Muyldermans et al. 2003; Silva de Assis et al. 2014; Butsch et al. 2014).

23.4.4.3 Mathematical Modelling

The majority of districting models has compactness as an objective function to be optimized. In addition, sometimes the maximal distance between a basic unit and its district center or between two basic unit of the same district is restricted (Benzarti et al. 2013). The appeal of distance-based measures is that they easily give rise to linear or, in case of pairwise distances, quadratic expressions. Therefore, these measures are sometimes also used for polygonal basic units, even if geometric measures could have been applied (Ríos-Mercado and Fernández 2009).

23.4.5 District Center

Strictly speaking, determining district centers is in most cases not an optimization criterion in itself. However, several measures for contiguity and compactness rely on district centers. Thus, if no centers are predefined for the districts, seeking district centers is part of the optimization process. Typically, a district center is the basic unit of the district that minimizes the respective compactness measure. But also the (weighted) center of gravity can be used to determine a district center. Note however that this center usually does not coincide with a basic unit, which is problematic if distance computations are based on road networks.

23.4.6 Other Criteria

There are a few other criteria for districting problems that are included from time to time in districting models. For example, for re-districting problems the changes in

allocation from the old to the new districting plan should be kept small (Silva de Assis et al. 2014). Especially in sales territory design, customers often have a preferred sales representative by whom they want to be serviced or vice-versa, i.e., customers have banned salesmen (cf. Ríos-Mercado and López-Pérez 2013). Another criterion concerns the number of districts. Typically, p is predetermined such that, for example, the expected workload in a district neither exceeds the working time restriction of a deliverer nor renders him underutilized. If however travel times within a district account for a large portion of the total working time, then it is not always possible to fix p a priori since travel times strongly depend on the shape of districts, i.e., their compactness. Therefore, sometimes p is a design criterion (cf. Muyldermans et al. 2003).

23.5 Solution Approaches

As with most optimization problems also for districting many different solution approaches have been proposed in the literature over the years. These approaches can roughly be divided in those that utilize a mathematical programming model and those that depend merely upon heuristics. Among the former, location-allocation and set partitioning methods have been discussed. The latter mainly focus on geometric algorithms, simple construction methods, and classical meta heuristics, like Tabu Search, GRASP, and Simulated Annealing. This section will present only a rough overview and description of the most common approaches. Detailed reviews can be found in Kalcsics et al. (2005) and Ricca et al. (2013).

23.5.1 Location-Allocation Methods

The first mathematical programming approach was proposed by Hess et al. (1965) for political districting. They had the idea to model the problem as a capacitated p-median facility location problem (see also Chap. 3). Basic units correspond to customers and their activity measure to their demand. The facilities to be located are the district centers, and the capacity of the facilities is chosen in such a way that the districts obtained by solving the problem are well balanced. Candidate locations for the facilities are all basic units. For an allowed relative deviation $\alpha > 0$ of the district size from the mean district size μ, the formulation of Hess et al. (1965) is

$$\text{minimize} \quad \sum_{i,j \in J} w_j \, d_{ij}^2 x_{ij} \tag{23.1}$$

$$\text{subject to} \quad \sum_{i \in J} x_{ij} = 1 \qquad \forall \, j \in J \tag{23.2}$$

$$\sum_{j \in J} w_j \, x_{ij} \;\geq\; (1-\alpha)\,\mu\, y_i \qquad \forall\, i \in J \tag{23.3}$$

$$\sum_{j \in J} w_j \, x_{ij} \;\leq\; (1+\alpha)\,\mu\, y_i \qquad \forall\, i \in J \tag{23.4}$$

$$\sum_{i \in J} y_i \;=\; p \tag{23.5}$$

$$y_i,\, x_{ij} \;\in\; \{0,1\} \qquad \forall\, i, j \in J, \tag{23.6}$$

where $x_{ij} = 1$ if basic unit j is assigned to basic unit i, 0 otherwise, and $y_i = 1$ if basic unit i is selected as district center, 0 otherwise. The objective function (23.1) maximizes the compactness of the districts using the center-based measure $cmp_{wd^2}(\cdot)$. Constraints (23.2), together with the integrality constraints on the x_{ij}-variables, model the unique and exclusive assignment criterion. Constraints (23.3) and (23.4) restrict the balance of the districts. Finally, Constraints (23.5) ensure that exactly p basic units are selected as district centers. As a result, all basic units allocated to the same basic unit i constitute a district with the basic unit as its center, i.e., there is a one-to-one correspondence between centers and districts. Note that the centers are just required to evaluate district compactness and have no meaning in itself.

Unfortunately, due to its NP-hardness, the practical use of this formulation is limited to instances with a few hundred basic units, which is rather small for typical sales districting problems. To this end, Hess et al. (1965) propose to use Cooper's location-allocation heuristic to solve the problem. In this heuristic, the simultaneous location and allocation decisions of the underlying facility location problem are decomposed into two independent phases, a location and an allocation phase, which are alternatingly performed until a satisfactory result is obtained. In the location phase, a set J_c of district centers is determined. A fairly simple and commonly used method is to solve in each district resulting from the last allocation phase a single facility location problem with the respective compactness measure as objective function (cf. Fleischmann and Paraschis 1988; George et al. 1997). To obtain an initial set of centers, one can determine new centers based on the solution of a Lagrangean subproblem (Hojati 1996). Alternatively, one can use any of the heuristics for the (uncapacitated) p-median problem or one of the heuristics mentioned below.

Once the centers have been fixed, the allocation phase determines a balanced assignment of basic units to district centers. This can be done by fixing $y_i = 1$ for all $i \in J_c$ in the above formulation. With present-day computers and MIP solvers, the resulting problem can be solved optimally even for large instances with 10,000 basic units or more within a short time. Even in the presence of contiguity constraints, several thousand basic units can be assigned in reasonable time (Ríos-Mercado and López-Pérez 2013). Alternatively, the allocation problem can be modeled as a minimum cost network flow problem allowing more flexibility for measuring and optimizing the balance and compactness of districts (George et al. 1997).

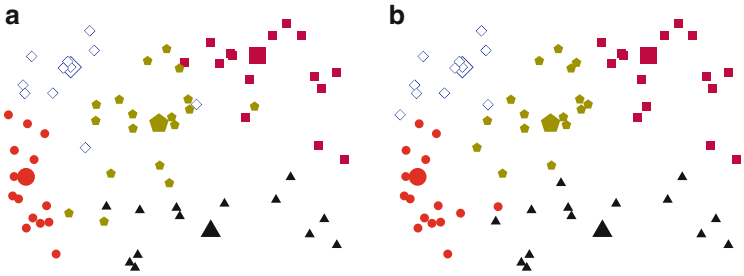


Fig. 23.5 Illustration of one iteration of the location-allocation procedure. (**a**) Location phase: new districts centers. (**b**) Allocation phase: new districts

Example 23.3 Consider again the example depicted in Fig. 23.3, but assume now that the district centers are flexible and the current ones are just a starting point. Based on the districting plan for the measure $cmp_{wd^2}(\cdot)$, the new centers that minimize $cmp_{wd^2}(\cdot)$ over each district are shown on the left-hand side in Fig. 23.5. The subsequent allocation phase yields the new districts shown on the right-hand side. The districts are visually much more compact and there is no overlap between the convex hulls of the districts.

In former times, when the exact solution of the allocation problem was unattainable for larger instances, the assignment problem was solved heuristically. Setting the tolerance α to zero and relaxing the integrality constraints on the assignment variables, i.e., $x_{ij} \in [0, 1]$, the resulting linear program is a classical transportation problem that can be solved efficiently using specialized network algorithms. However, solving the relaxed problem yields districts that are perfectly balanced but usually assign portions of basic units to more than one district, i.e., $\exists\, i,\ i' \in J_c$, $i \neq i'$, $j \in J$, such that $x_{ij},\ x_{i'j} > 0$. Such basic units are called splits. For an optimal basic feasible solution of the transportation problem, it is easy to prove that there are at most $p - 1$ splits (Hojati 1996). To restore the integrity of basic units, it is necessary to round for every split its fractional variables to one (one variable) or zero (the other variables). This yields disjoint districts but destroys their perfect balance. A simple split resolution rule is to assign a split to the district (center) that "owns" the largest share of the split (Hess and Samuels 1971). However, if there are just few basic units per district, this rule may produce very unbalanced districts. An optimal split allocation with a minimal maximal percentage deviation can be obtained in polynomial time by using tree partitioning methods; unfortunately, the problem of finding a split resolution with a minimal total deviation is NP-hard; see Schröder (2001) for details.

23.5.2 Set-Partitioning Models

As districting is essentially a partitioning problem, classical set partitioning approaches can be used to solve the problem. In a first step, balanced, contiguous, and compact candidate districts are generated in a heuristic fashion. In a second step, districts are selected from the set of candidates to optimize the overall balance of the district plan (Garfinkel and Nemhauser 1970; Mehrotra et al. 1998). Unfortunately, only small instances can be solved optimally with this approach. An advantage compared to location-allocation methods is however that almost any criterion can be applied on the generation of candidate districts.

23.5.3 Computational Geometry Methods

A very simple but efficient solution approach for basic units representing points is the successive dichotomies strategy (Kalcsics et al. 2005). The main idea is to recursively subdivide the problem geometrically using lines into smaller and smaller subproblems until an elementary level is reached, where the problem can be solved efficiently. Hence, the basic operation is to partition a subset J' of basic units into two subsets J'_l and J'_r by drawing a line within this set of points. Given a number of line directions, for each direction the position of the line is determined in such a way that the two resulting subproblems are best balanced. For every direction, the line is evaluated by a convex combination of its balance and its compactness (evaluated through the length of inter-district boundaries), and the best line is then used to divide the problem into two subproblems. This procedure is repeated until every subset corresponds to a single district. The strategy quickly determines a well-balanced districting plan with no overlap between districts. However, as it does not explicitly account for (road) distances, the resulting districts sometimes lack compactness. Moreover, it is difficult to include neighborhood information. Instead of using lines, other geometric concepts can be used. Alternatively, the process of subdividing a point set J' can be modeled and solved as a 2-facility location problem (Salazar-Aguilar et al. 2013a).

Example 23.4 Consider again the example in Fig. 23.3 and assume that the district centers are flexible. Figure 23.6 shows the districting plan obtained with the successive dichotomies algorithm using horizontal, vertical, and diagonal lines.

Another approach is based on weighted Voronoi diagrams on networks (for a definition of weighted Voronoi diagrams see Aurenhammar et al. 2013). Assume that the neighborhood graph G is given. For center-based measures the most compact solution is obtained by assigning each basic unit to the closest center. If the distances $\{d_{c_k,j} \mid c_k \in J_c\}$ are unique for each $j \in J$, then each district will also be connected. However, the resulting districts are often far from being balanced. To overcome this drawback, the idea is to modify the distances $d_{c_k,j}$

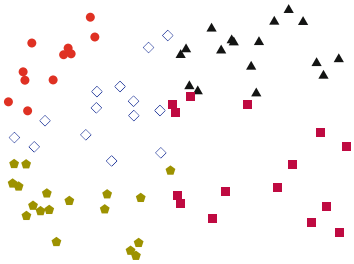


Fig. 23.6 Districting plan with the successive dichotomies algorithm

between basic units and centers in such a way that assignments to overly large districts are "penalized" and allocations to too small districts are "stipulated". There are basically two options to modify distances. The first adds a real-valued weight $r_k \in \mathbb{R}$ to each distance $d_{c_k,j}$ (Zoltners and Sinha 1983) and the second multiplies $d_{c_k,j}$ by a positive weight $r_k \in \mathbb{R}^+$ (Ricca et al. 2008). Hence, basic unit $j \in J$ is closer to center c_k than to center $c_l \in J_c$ if $d_{c_k,j} + w_k < d_{c_l,j} + w_l$ or $w_k\, d_{c_k,j} < w_l\, d_{c_l,j}$, respectively. Increasing (decreasing) the weight for a specific center c_k while keeping the other weights unchanged, will reduce (increase) the number of basic units assigned to c_k under the closest assignment rule and thus reduce (increase) the size of the district. To obtain balanced districts, the weights have to be updated iteratively until a satisfactory result is obtained. During the update, care has to be taken because some districts may turn out empty under additive weights or become disconnected for multiplicative weights if the weights are too uneven. For details on the update procedures see Zoltners and Sinha (1983) and Ricca et al. (2008). The partitions of the graph induced by these weights are the so-called additively and multiplicatively weighted Voronoi diagrams. Note that the approach using additive weights is in fact a Lagrangean relaxation where the balancing constraints have been relaxed.

Most districting problems are solved using discrete models. However, these problems (and a number of other logistics problems as well) can be converted into problems with continuous demand functions. Continuous demand approximations models are based on the spatial density and distribution of demand rather than on precise information on every demand point. Given continuous approximations, one can for example use Voronoi diagrams to compute or to smooth existing districts (Galvão et al. 2006), or determine perfectly balanced districts (Carlsson and Delage 2013).

23.5.4 Construction Methods

There exist several easy approaches for constructing a districting plan from scratch. One of the most popular ones is based on the multi-kernel growth methodology first introduced in Vickrey (1961). The general idea of this methodology is to select a certain number of basic units as "seed centers" and then assign to each seed

neighboring basic units in order of decreasing distance until the desired district size is reached. Variations exist with respect to the selection of seeds, whether districts grow simultaneously or sequentially around the seeds, and how to deal with enclaves of unassigned basic units which typically occur at the end of this greedy process (Bodin and Levy 1991; Williams 1995; Mehrotra et al. 1998; Bozkaya et al. 2003). The resulting districting plans are not always connected or balanced and typically serve as a starting point for a meta heuristic.

A different approach treats each basic unit initially as a single district and then merges iteratively pairs of districts until the prescribed number of districts is reached (Deckro 1977).

23.5.5 Meta Heuristics

There exists a wide range of meta heuristics that have been applied to districting problems: Simulated Annealing (D'Amico et al. 2002), Tabu Search (Ricca and Simeone 2008; Bozkaya et al. 2003), GRASP (Ríos-Mercado and Fernández 2009; Salazar-Aguilar et al. 2013b), and Genetic algorithms (Forman and Yue 2003; Bergey et al. 2003; Bação et al. 2005), just to name a few. A major advantage of these methods is their flexibility to include almost any practical criterion and measure for the design of districts.

23.6 Conclusions

Despite the large number of publications, it is striking that only few authors consider the districting problem independently from a practical background. Moreover, there is no consensus on which criteria are eligible and important and, on how to measure them appropriately. Thus, instead of devising yet another (variant of a) meta heuristic for a districting model with yet another measure for compactness or additional constraint, research should foremost concentrate on a common and generic framework for districting problems. And it should try to categorize the suitability of criteria and measures based on the availability of data, the geometric representation of the basic units, and the different types of applications.

Acknowledgements This work was partly supported by grant NI 521/6-1 of the German Research Foundation (DFG). This support is gratefully acknowledged.

References

Aurenhammar F, Klein R, Lee DT (2013) Voronoi diagrams and Delaunay triangulations. World Scientific, Singapore

Bação F, Lobo V, Painho M (2005) Applying genetic algorithms to zone design. Soft Comput 9:341–348

Baker J, Clayton E, Moore L (1989) Redesign of primary response areas for county ambulance services. Eur J Oper Res 41:23–32

Bard JF, Jarrah AI (2009) Large-scale constrained clustering for rationalizing pickup and delivery operations. Transp Res B Methodol 43:542–561

Benzarti E, Sahin E, Dallery Y (2013) Operations management applied to home care services: analysis of the districting problem. Decis Support Syst 55:587–598

Bergey P, Ragsdale C, Hoskote M (2003) A simulated annealing genetic algorithm for the electrical power districting problem. Ann Oper Res 121:33–55

Blais M, Lapierre S, Laporte G (2003) Solving a home-care districting problem in an urban setting. J Oper Res Soc 54:1141–1147

Bodin L, Levy L (1991) The arc partitioning problem. Eur J Oper Res 53:393–401

Bozkaya B, Erkut E, Laporte G (2003) A tabu search heuristic and adaptive memory procedure for political districting. Eur J Oper Res 144:12–26

Bozkaya B, Erkut E, Haight D, Laporte G (2011) Designing new electoral districts for the city of Edmonton. Interfaces 41:534–547

Butsch A, Kalcsics J, Laporte G (2014) Districting for arc routing. INFORMS J Comput 26:809–824

Carlsson J, Delage E (2013) Robust partitioning for stochastic multivehicle routing. Oper Res 61:727–744

D'Amico S, Wang S, Batta R, Rump C (2002) A simulated annealing approach to police district design. Comput Oper Res 29:667–684

Deckro R (1977) Multiple objective districting: a general heuristic approach using multiple criteria. Oper Res Q 28:953–961

Dixon RJ (1968) Democratic representation: reapportionment in law and politics. Oxford University Press, New York

Drexl A, Haase K (1999) Fast approximation methods for sales force deployment. Manag Sci 45:1307–1323

Duckham M, Kulik L, Worboys M, Galton A (2008) Efficient generation of simple polygons for characterizing the shape of a set of points in the plane. Pattern Recognit 41:3224–3236

Elizondo-Amaya M, Ríos-Mercado R, Díaz J (2014) A dual bounding scheme for a territory design problem. Comput Oper Res 44:193–205

Ferland J, Guénette G (1990) Decision support system for a school districting problem. Oper Res 38:15–21

Fleischmann B, Paraschis J (1988) Solving a large scale districting problem: a case report. Comput Oper Res 15:521–533

Forman S, Yue Y (2003) Congressional districting using a TSP-based genetic algorithm. In: Proceedings of the 2003 international conference on genetic and evolutionary computation: Part II, GECCO'03, pp 2072–2083

Galvão L, Novaes A, Souza de Cursi J, Souza J (2006) A multiplicatively-weighted Voronoi diagram approach to logistics districting. Comput Oper Res 33:93–114

Garfinkel R, Nemhauser G (1970) Optimal political districting by implicit enumeration techniques. Manag Sci 16:495–508

George J, Lamar B, Wallace C (1997) Political district determination using large-scale network optimization. Socio Econ Plan Sci 31:11–28

Glaze T, Weinberg C (1979) A sales territory alignment program and account planning system. In: Sales management: new developments from behavioral and decision model research. Marketing Science Institute, Cambridge, pp 325–343

Gross J, Yellen J (2003) Handbook of graph theory. CRC Press, Boca Raton
Hanafi S, Fréville A, Vaca P (1999) Municipal solid waste collection: an effective data structure for solving the sectorization problem with local search methods. INFOR 37:236–254
Handley L, Grofmann B (eds) (2008) Redistricting in comparative perspective. Oxford University Press, New York
Haugland D, Ho S, Laporte G (2007) Designing delivery districts for the vehicle routing problem with stochastic demands. Eur J Oper Res 180:997–1010
Hess S, Samuels S (1971) Experiences with a sales districting model: criteria and implementation. Manag Sci 18:41–54
Hess S, Weaver J, Siegfeldt H, Whelan J, Zitlau P (1965) Nonpartisan political redistricting by computer. Oper Res 13:998–1008
Hojati M (1996) Optimal political districting. Comput Oper Res 23:1147–1161
Horn D, Hampton C, Vandenberg A (1993) Practical application of district compactness. Polit Geogr 12:103–120
Howick R, Pidd M (1990) Sales force deployment models. Eur J Oper Res 48:295–310
Jarrah A, Bard J (2012) Large-scale pickup and delivery work area design. Comput Oper Res 39:3102–3118
Kalcsics J, Nickel S, Schröder M (2005) Towards a unified territorial design approach – applications, algorithms and GIS integration. TOP 13:1–74
Lei H, Laporte G, Guo B (2012) Districting for routing with stochastic customers. EURO J Transp Logist 1:67–85
Lewyn M (1993) How to limit gerrymandering. Florida Law Rev 45:403–486
Lin HY, Kao JJ (2008) Subregion districting analysis for municipal solid waste collection privatization. J Air Waste Manag Assoc 58:104–111
Lodish L (1975) Sales territory alignment to maximize profit. J Market Res 12:30–36
López-Pérez J, Ríos-Mercado R (2013) Embotelladoras ARCA uses operations research to improve territory design plans. Interfaces 43:209–220
Mehrotra A, Johnson E, Nemhauser G (1998) An optimization based heuristic for political districting. Manag Sci 44:1100–1114
Minciardi R, Puliafito PP, Zoppoli R (1981) A districting procedure for social organizations. Eur J Oper Res 8:47–57
Muyldermans L, Cattrysse D, Van Oudheusden D, Lotan T (2002) Districting for salt spreading operations. Eur J Oper Res 139:521–532
Muyldermans L, Cattrysse D, Van Oudheusden D (2003) District design for arc-routing applications. J Oper Res Soc 54:1209–1221
Nagel S (1965) Simplified bipartisan computer redistricting. Stanford Law Rev 17:863–899
Niemi R, Grofman B, Carlucci C, Hofeller T (1990) Measuring compactness and the role of a compactness standard in a test for partisan and racial gerrymandering. J Polit 52:1155–1181
Parker F (1990) Black votes count. The University of North Carolina Press, Chapel Hill
Puppe C, Tasnádi A (2008) A computational approach to unbiased districting. Math Comput Model 48:1455–1460
Ricca F, Simeone B (2008) Local search algorithms for political districting. Eur J Oper Res 189:1409–1426
Ricca F, Scozzari A, Simeone B (2008) Weighted Voronoi region algorithms for political districting. Math Comput Model 48:1468–1477
Ricca F, Scozzari A, Simeone B (2013) Political districting: from classical models to recent approaches. Ann Oper Res 204:271–299
Ríos-Mercado R, Fernández E (2009) A reactive GRASP for a commercial territory design problem with multiple balancing requirements. Comput Oper Res 36:755–776
Ríos-Mercado R, López-Pérez J (2013) Commercial territory design planning with realignment and disjoint assignment requirements. Omega 41:525–535
Ríos-Mercado R, Salazar-Acosta J (2011) A GRASP with strategic oscillation for a commercial territory design problem with a routing budget constraint. Advances in Soft Computing Lecture notes in computer science, 10th Mexican International Conference on Artificial Intelligence,

MICAI 2011, Puebla, Mexico, Proceedings, Part II, vol 7095, pp 307–318, Springer Berlin Heidelberg
Salazar-Aguilar M, González-Velarde J, Ríos-Mercado R (2013a) A divide-and-conquer approach for a commercial territory design problem. Computación y Sistemas 16:309–320
Salazar-Aguilar M, Ríos-Mercado R, González-Velarde J (2013b) GRASP strategies for a bi-objective commercial territory design problem. J Heuristics 19:179–200
Schoepfle O, Church R (1991) A new network representation of a "classic" school districting problem. Socio Econ Plan Sci 25:189–197
Schröder M (2001) Gebiete optimal aufteilen. Ph.D. thesis, Universität Karlsruhe. http://www.ubka.uni-karlsruhe.de/eva
Shirabe T (2009) Districting modeling with exact contiguity constraints. Environ Plan B Plan Des 36:1053–1066
Silva de Assis L, Morelato Franca P, Luiz Usberti F (2014) A redistricting problem applied to meter reading in power distribution networks. Comput Oper Res 41:65–75
Simchi-Levi D, Kaminsky P, Simchi-Levi E (2003) Designing & managing the supply chain: concepts, strategies & case studies, 2nd edn. McGraw-Hill/Irwin, New York
Skiera B, Albers S (1994) COSTA: Ein Entscheidungs-Unterstützungs-System zur deckungsbeitragsmaximalen Einteilung von Verkaufsgebieten. Z Betriebswirt 64:1261–1283
Tavares-Pereira F, Figueira J, Mousseau V, Roy B (2007) Multiple criteria districting problems: the public transportation network pricing system of the Paris region. Ann Oper Res 154:69–92
Vickrey W (1961) On the prevention of gerrymandering. Polit Sci Q 76:105–110
Webster G (2013) Reflections on current criteria to evaluate redistricting plans. Polit Geogr 32:3–14
Williams JC Jr (1995) Political redistricting: a review. Pap Reg Sci 74:13–40
Xu C, Yum TSP (2010) Patrol districting and routing with security level functions. In: Proceedings of the IEEE international conference on systems, man and cybernetics, pp 3555–3562
Young H (1988) Measuring the compactness of legislative districts. Legis Stud Q 13:105–115
Zhong H, Hall R, Dessouky M (2007) Territory planning and vehicle dispatching with driver learning. Transp Sci 41:74–89
Zhou G, Min H, Gen M (2002) The balanced allocation of customers to multiple distribution centers in a supply chain network: a genetic algorithm approach. Comput Ind Eng 43:251–261
Zoltners AA, Sinha P (1983) Sales territory alignment: a review and model. Manag Sci 29:1237–1256
Zoltners A, Sinha P (2005) Sales territory design: thirty years of modeling and implementation. Market Sci 24:313–331

Chapter 24
Location Problems Under Disaster Events

Maria Paola Scaparra and Richard L. Church

Abstract Facility systems may be vulnerable to a disaster, whether caused by intention, an accident, or by an act of nature. When disrupting events do occur, services may be degraded or even destroyed. This chapter addresses problems of disruption associated with facility based service systems. Three main questions often arise when dealing with a possible disaster: (1) how bad can it get? (2) is there a way in which we can protect our system from such an outcome? and (3) is there a way in which we can incorporate such issues in our future designs and plans? This chapter addresses each of these main questions with respect to several classic location problems. Specifically, it discusses recent location models under disaster events along three main streams of research: facility interdiction, facility protection, and resilient design.

Keywords Interdiction • Protection • Reliability

24.1 Introduction

Although Murphys law (if anything can go wrong, it will) does not always come true, it seems at least important to address what might go wrong when designing and operating infrastructures, such as service systems and supply chains. Whether intentional or accidental, disasters can render a system inoperable or inefficient for quite some time. For example, in 2011, flooding in Thailand was considered to be the worst in 50 years. This event disrupted supply chains around the world from computer storage disk manufacturing to cars. In that flood, a production facility for Honda was closed for more than 3 months, and a financial analyst estimated that floods would reduce profits at Toyota, Nissan, and Honda by more than a combined Y35bn (Soble 2011). Harm can also be intentional and simple. For example, a

M.P. Scaparra (✉)
Kent Business School, University of Kent, CT2 7ET Canterbury, UK
e-mail: m.p.scaparra@Kent.ac.uk

R.L. Church
Department of Geography, University of California, Santa Barbara, CA, USA
e-mail: rick.church@ucsb.edu

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5_24

10-day labor strike at the ports of Los Angeles and Long Beach had such an impact on some retailers in 2002 that it took 6 months before supply chains fully recovered (Reid and Gorman 2012). In response some retailers reduced their reliance on one port and one set of shipping routes, to where they now utilize multiple shipping routes and multiple ports to ensure that product flows will not be totally disrupted by one event. In Sacramento, CA, a fire started by an arsonist destroyed a railroad trestle in 2007. Trains that normally used this route had to detour more than 100 miles until the trestle was replaced (Peterson and Church 2008). In exerting a level of control and force, a drug cartel in 2013 bombed 18 electrical stations in Michoacan, one of the largest states in Mexico (Casey 2013). This event caused a blackout that affected more than half a million people for 15 h. As a final example of intentional disruption, snipers in April 2013 opened fire on a substation supplying power to Silicon Valley, California, and knocked out 17 giant transformers, nearly bringing the entire area to a complete blackout. U.S. Officials have stated that this was the most significant incident in domestic terrorism involving the grid that has ever occurred. In an unreported U.S. government analysis, researchers found that knocking nine key substations out of 55,000 substations on a scorching summer day could result in a coast-to-coast blackout (Smith 2014) and it is believe that protecting 100 key substations would be enough to mitigate such an attack. This gives credence to addressing the question of what is critical to protect. Overall, addressing such potential risks when designing and operating a system of facilities may lead to more resilient and efficient systems.

Facilities and associated transport networks are key elements in any production, supply, and service system. Traditional modeling approaches for facility location problems are based upon the assumption that systems will operate as designed. Virtually all modern textbooks on modeling production and supply systems ignore the problem of disruption when optimizing the location of a set of facilities. Church et al. (2004) demonstrated that a given deployment of facility resources, although optimal, could be significantly disrupted in service efficiency, while other close-to-optimal configurations were relatively resilient when subject to the same level of disruption. This work and the work of Snyder and Daskin (2005) were instrumental in establishing a need to handle facility reliability and vulnerability explicitly. Because of this there has been an increased interest in modeling the fragility of networks and facility systems over a wide range of possible events from natural disasters to intentional strikes.

Research in facility disruption is new and evolving. There are three major problems of interest. The first one is: how much impact can be expected? This problem involves the search for the most critical elements of a system, that is, those facilities which when removed from operation impact the system the most. The second important question is: how can such impacts be averted? One way of averting a crisis may be to fortify facilities against disaster. This may mean something simple like providing backup generators for power or providing enough security that it will ward off a would-be attacker. It could also mean moving the facility to a nearby site that is less vulnerable to something like flooding. The third main question is: how might facilities be configured so that the resulting system is both efficient in service

delivery and resilient when disrupted? This last question deals with the design of a new system, whereas the first two questions deal with an existing system. All of these are major issues and are addressed in this chapter.

The main optimization models developed to answer these questions can be classified as follows:

I. *Interdiction models*. These models identify vulnerabilities of service/supply systems and quantify the impacts of potential losses of key components on a system ability to provide efficient service.
II. *Protection models*. These models optimize the allocation of protective resources among the facilities of already existent systems.
III. *Design models*. These models are used for planning new service and supply systems which are secure and resilient to disruptions.

In this chapter, we will provide a description of the seminal models in each class and outline how these models have then been further developed and extended to capture the additional complexities and interdependencies characterizing real service and supply systems. The description of the models is paralleled by a brief description of the solution methodologies which have been proposed for solving them.

The reminder of this chapter is organized as follows. Section 24.2 introduces the notation used throughout the chapter. Interdiction, protection and design models are described in Sects. 24.3, 24.4 and 24.5, respectively. In Sect. 24.6, we highlight future trends in modeling location problems under disaster events. Some conclusive remarks are finally provided in Sect. 24.7.

24.2 Notation

In the following description of location models under disruption, we assume that the reader is already familiar with the classic location problems introduced in the previous chapters (e.g., median, covering, fixed-charge and hub location problems). Here we briefly summarize the main notation used throughout the chapter.

Inputs

$I =$ Set of potential locations for the facilities, indexed by i
$J =$ Set of customers, indexed by j
$F =$ Set of facilities in an existing system
$d_j =$ Demand of customer j
$c_{ij} =$ Unitary cost for serving customer j from facility i
$N_j =$ Set of facilities covering customer j ($N_j \subseteq I$)
$p =$ Number of facilities to be located
$r =$ Number of facilities to be interdicted
$b =$ Number of facilities to be protected

Decision variables

$$y_i = \begin{cases} 1 & \text{if a facility is located at site } i \\ 0 & \text{otherwise} \end{cases}$$

$$s_i = \begin{cases} 1 & \text{if a facility located at } i \text{ is interdicted} \\ 0 & \text{otherwise} \end{cases}$$

$$z_i = \begin{cases} 1 & \text{if a facility located at } i \text{ is protected} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{ij} = \begin{cases} 1 & \text{if the demand of customer } j \text{ is supplied from facility } i \\ 0 & \text{otherwise} \end{cases}$$

$$u_j = \begin{cases} 1 & \text{if customer } j \text{ is covered before disruption} \\ 0 & \text{otherwise} \end{cases}$$

$$v_j = \begin{cases} 1 & \text{if customer } j \text{ is covered after disruption} \\ 0 & \text{otherwise} \end{cases}$$

24.3 Identifying Critical Facilities: Interdiction Models

Interdiction models date back a few decades and were originally designed to assess the impact of losing critical links in transportation networks for military applications (see, for example, Wollmer 1964 and Wood 1993). The first interdiction models within the facility location literature were introduced by Church et al. (2004) to identify the most critical facility assets in median and covering systems. The first problem, called the r-Interdiction Median Problem (r-IMP), can be seen as the antithesis of the p-median problem and aims at identifying the best set of r facilities to remove, among the existing ones, in order to maximize the overall demand-weighted cost for serving the customers from the remaining facilities (these are referred to as non-interdicted facilities). Similarly, the r-Interdiction Covering Problem (r-ICP) can be seen as the antithesis of the maximal covering problem and involves finding the subset of r facilities, which when removed, minimizes the total demand that can be covered within a specified distance or travel time. In essence, both models identify the subset of facilities whose loss has the greatest impact on service delivery, where the impact is measured either in terms of cost increase or in terms of lost coverage to mirror two different service protocols.

24.3.1 The r-Interdiction Median Problem

In addition to the notation introduced in Sect. 24.2, the mathematical formulation of r-IMP requires the definition of the set $T_{ij} = \{k \in F | d_{kj} > d_{ij}\}$ defined for each facility $i \in I$ and customer $j \in J$. T_{ij} represents the set of existing sites that are farther than i is from demand j. The r-IMP can be formulated in the following manner:

$$\text{maximize} \quad \sum_{i \in F} \sum_{j \in J} d_j c_{ij} x_{ij} \tag{24.1}$$

$$\text{subject to} \quad \sum_{i \in F} x_{ij} = 1 \quad \forall j \in J \tag{24.2}$$

$$\sum_{i \in F} s_i = r \tag{24.3}$$

$$\sum_{k \in T_{ij}} x_{kj} \leq s_i \quad \forall i \in F, j \in J \tag{24.4}$$

$$x_{ij} \in \{0, 1\} \quad \forall i \in F, j \in J \tag{24.5}$$

$$s_i \in \{0, 1\} \quad \forall i \in F. \tag{24.6}$$

The objective function (24.1) maximizes the demand-weighted total cost after the interdiction of r facilities. Constraints (24.2) ensure that each customer is assigned to a facility after interdiction. Constraints (24.3) stipulate that exactly R facilities are to be interdicted. Constraints (24.4) force each customer j to be assigned to its closest non-interdicted facility. Namely, this set of constraints prevents each customer j from being assigned to facilities which are further than facility i, unless facility i is interdicted. Finally, constraints (24.5) and (24.6) represent the binary restrictions on the assignment and interdiction variables, respectively. Note that the structure of the problem guarantees that there is always one optimal solution in which all the x_{ij} variables are binary, so that the integrality restrictions on these variables can be relaxed.

In the above model the parameter r, i.e., the number of facilities that are lost simultaneously in a particular event, is chosen as a metric of possible disruption. In other words, r is used to capture the possible extent of a disruptive event: small values are usually associated with low-impact but possibly frequent events, whereas larger values are associated with disruptions which may affect a large number of assets. Given the difficulty of estimating this parameter precisely, an analyst would normally solve each model over a range of facility losses, r, in order to capture the range of possible impacts to system operations. Using a loss parameter r makes sense in modeling worst case disruptive scenarios due to natural events; however, in a case of intentional disruption one may want to consider the fact that each facility may require different amounts of resources to be completely disabled. For this type

of case, one might want to cast disruption as a budget constrained process (see for example Losada et al. 2012b). However using an interdiction budget requires information that may be completely hidden from the system operator, including the costs of striking and the available budget itself. The use of cardinality constraints such as (24.3) can be seen as a surrogate to knowing exact budget values of the interdictor.

The r-IMP can be cast as an integer linear programming model which can be solved with general-purpose integer programming software. The above formulation of the r-IMP can be streamlined by consolidating redundant assignment variables under special proximity conditions. The resulting variable reduction of this consolidation mechanism, which was initially proposed by Church (2003) for the p-median problem, can be substantial. Scaparra and Church (2008a) report reductions of up to 80 % of the initial number of variables. The same authors also analyze and compare different formulations of the closest assignment constraints (24.4) to identify the most efficient formulation for the r-IMP. Although other approaches could be devised to solve the r-IMP, including decomposition methods or heuristics, solving the streamlined model by commercial software is usually quite effective, even for problem instances of significant size.

Clearly, the r-IMP makes some simplifying assumptions which may limit its practical applicability. For instance, it assumes that every strike or disruption is successful and always results in a complete impairment of the affected facility. In reality, the chances of losing a facility following a natural disaster or a man-made attack are based upon some probability. Church and Scaparra (2007a) introduced a probabilistic version of r-IMP where an attempted interdiction is successful only with a given probability. The same authors also show how to build a *reliability envelope* for identifying the range of possible impacts associated with losing one or more facilities. Losada et al. (2012b) further extended this probabilistic r-IMP by assuming that the probability of impairing a facility depends on the intensity of the disruption or on the amount of offensive resources used in the attack. In a further extension, Lei and Church (2011) address the issue of interdiction when not all demands are served by their closest facility after a disruption.

The r-IMP also assumes no restrictions on the facilities capacity, thus implying that after a disruption, the unaffected facilities have enough combined capacity to supply all the demand. This may not be a realistic assumption as most real supply systems usually operate with capacity limits. The capacitated version of the r-IMP can be found in Scaparra and Church (2012). Another interesting variation of the r-IMP which considers capacity restrictions is the partial interdiction problem introduced by Aksen et al. (2012). In this model, an interdicted facility may preserve part of its capacity; the capacity loss due to interdiction is commensurate to the intensity of the attack and the unmet demand after interdiction can be outsourced at some cost.

24.3.2 The r-Interdiction Covering Problem

The r-Interdiction Covering Problem (r-ICP) can be stated mathematically as follows:

$$\text{minimize} \quad \sum_{j \in J} d_j v_j \tag{24.7}$$

$$\text{subject to} \quad v_j \geq 1 - s_i \quad \forall j \in J, i \in N_j \cap F \tag{24.8}$$

$$\sum_{i \in F} s_i = r \tag{24.9}$$

$$v_j \in \{0, 1\} \quad \forall j \in J \tag{24.10}$$

$$s_i \in \{0, 1\} \quad \forall i \in F. \tag{24.11}$$

The objective function (24.7) minimizes the amount of customer demand which is covered after interdiction. Constraints (24.8) stipulate that a customer j must be covered unless all the facilities that currently cover it (i.e., the facilities in $N_j \cap F$) are interdicted. Constraints (24.9) force the number of facilities to be eliminated to equal r. The last two sets of constraints (24.10) and (24.11) are binary restrictions on the coverage and interdiction variables. Note that the binary integer restrictions are only needed for the s_i variables whereas the v_j variables automatically take on binary integer values in any optimal solution.

r-ICP instances of considerable size can generally be solved by commercial optimization packages without the need of resorting to more sophisticated approaches or heuristic techniques (Sevaux et al. 2015). Clearly, the same problem variations that have been considered for the r-IMP may be developed for the r-ICP so as to capture additional features such as probabilistic failures, capacity restrictions, and partial interdiction.

24.3.3 Other Interdiction Models

Although our focus so far has been on interdiction models for median and covering systems, an interdiction model counterpart can be devised for virtually every facility location problem proposed in the literature. As an example, Lei (2013) proposes the Hub Interdiction Median Problem which identifies the most critical hub facilities in hub–and–spoke systems.

24.4 Hardening Facilities: Protection Models

Interdiction models are a valuable tool for assessing facility criticality and worst-case scenario losses in case of disruption. However, it can be easily demonstrated that securing those facilities that are identified as the most critical in an optimal interdiction solution does not necessarily result in the most effective protection strategy (Church and Scaparra 2007b). Interdiction is a function of what is protected and this interdependency must be captured explicitly into a modeling framework to guarantee that limited protective resources are allocated in an optimal way. Most of the facility protection models existing in the literature incorporate an interdiction model as a tool for evaluating worst-case losses in response to protection plans. These models are expressed mathematically as bilevel optimization programs (Dempe 2002) which emulate a game played between a system *defender* (the leader) and a system *attacker* or *interdictor* (the follower). In this bilevel structure, the upper level problem involves decisions on which facilities to harden, whereas the lower level problem identifies which unprotected facilities to attack to inflict maximum damage.

In the following, we show how the model presented for the r-IMP in the previous section can be embedded within a protection model to optimize security investments in median systems (Scaparra and Church 2008a).

24.4.1 The r-Interdiction Median Problem with Fortification

The bilevel formulation of the r-IMP with Fortification (r-IMPF) is as follows.

$$\text{minimize} \quad H(z) \tag{24.12}$$

$$\text{subject to} \quad \sum_{i \in F} z_i = b \tag{24.13}$$

$$z_i \in \{0,1\} \quad \forall i \in F, \tag{24.14}$$

where

$$H(z) = \max \sum_{i \in F} \sum_{j \in J} d_j c_{ij} x_{ij} \tag{24.15}$$

$$\text{s.t.} \quad s_i \leq 1 - z_i \tag{24.16}$$

$$(24.2) - (24.6).$$

The leader objective (24.12) is to minimize the highest possible level of demand-weighted service cost, H, following the disruption on r facilities by allocating b protective resources (24.13). The worst-case cost H is computed in the follower

problem, which is simply the r-IMP problem defined in Sect. 24.3 with the additional constraints (24.16). These constraints, which link the upper level protection variables and the lower level interdiction variables, prevent the interdiction of any protected facility.

It is important to note that in the above model protection resources can be cast with a budget constraint and facility varying protection costs (Aksen et al. 2010). It is also possible to add the costs of protection as a an additional term in the objective, where the costs of protection and costs of worst case operation are simultaneously minimized. In either case (as formulated or as an added objective term), one would generally want to solve a series of such problems in order to determine tradeoff curves of system impacts versus protection resources. The above form can be used to identify both supported and unsupported non-dominated solutions whereas the latter will be effective in solving for only supported non-dominated solution. In any case, one would want to understand exactly what protection provides in terms of reducing impacts of interdiction as compared to the added costs of protection.

Bilevel programs are generally very difficult to solve (Moore and Bard 1990), especially when integer variables appear in both levels and when the upper level variables parametrize the feasible region of the lower level problem, as it is the case in r-IMPF. Common approaches to solve bilevel integer programs include reformulation into single level problems and decomposition methods. Examples of casting r-IMPF as a single level problem can be found in Church and Scaparra (2007b) and Scaparra and Church (2008b). However, these single level models require a complete enumeration of all the possible ways of interdicting r out of the $|F|$ existing facilities and therefore become quickly intractable as the value of the parameters $|F|$ and r increases. Scaparra and Church (2008a) propose an implicit enumeration algorithm to solve the bilevel r-IMPF. The approach is based upon the observation that an optimal protection plan must include at least one of the critical facilities identified by solving a simple r-IMP. The recursive use of this property allows a significant reduction of the number of protection strategies that must be evaluated in an enumeration scheme. To date, this algorithm remains one of the most effective methods for solving this type of protection/interdiciton models and has been successfully applied to problems in different settings as well (e.g., the network protection models in Cappanera and Scaparra 2011).

Since its appearance, the r-IMPF has spurred a significant amount of research and several different variants to the original problem have been proposed in the literature. As an example, Liberatore et al. (2010) introduced a stochastic version of r-IMPF where the number of possible losses r is uncertain, to reflect the fact that the extent of a disruption is usually not known with certainty. In a follow up paper, Liberatore and Scaparra (2011) compared the model proposed for the above stochastic problem with two regret-based models to identify robust protection strategies in uncertain environments.

Aksen et al. (2010) proposed a budget-constrained version of the r-IMPF with flexible capacity expansion. In particular, they replaced the cardinality constraint

(24.13) with a budget constraint and assume that the facilities have different protection costs and flexible capacity (i.e., the capacity can be expanded to accommodate the demand of customers previously assigned to interdicted facilities).

Another interesting variation of the r-IMPF is the problem investigated by Liberatore et al. (2012), which optimizes protection plans in the face of large area disruptions. The problem includes capacitated facilities, partial interdiction (interdiction reduces the amount of demand that can be served by a facility) and correlated disruptions (when a facility is hit, nearby facilities are affected as well). The problem was formulated as a tri-level program, and solved by dualization integrated in the implicit enumeration algorithm devised by Scaparra and Church (2008a) for the r-IMPF.

All the problems cited so far are static which means that they do not consider the effect of disruptions over time. In reality, disrupted facilities may have different recovery times and the duration over which system operations are degraded should be considered when modeling worst-case disruption scenarios. To redress this shortcoming, Losada et al. (2012a) proposed a different protection model for median systems where protection does not necessarily prevent facility failure altogether, but speeds up recovery time following a potential disruption. The resulting model also incorporates the possibility of multiple disruptions over time and is solved using three different decomposition approaches.

An underlying assumption of the r-IMPF and all its variations is that protection is always successful and, therefore, protected facilities are never interdicted in a worst-case scenario. Bricha and Nourelfath (2013) relaxed this assumption and proposed a model where a protected facility is immune to disruption only with a given probability. The initial model was then extended to consider protection against concerted attacks by multiple interdictors.

Whereas most of the focus has been on protection models for median systems, Zhu et al. (2013) proposed a game theoretical model to identify optimal defense strategies for an uncapacitated fixed-charge location model. In this model, the defender has several investment strategies (or levels of investment) available and aims at minimizing the expected damage to the systems along with the protection expenditure. Similarly, the interdictor can choose different attack levels on each facility and aims at maximizing a utility function, which combines damage and attack expenditures.

24.5 Planning Robust Systems: Design Models

Hardening existing facilities can be an effective way of mitigating the impact of facility failures. An alternative approach is to incorporate the risks of potential failures in the initial design of a system by identifying location strategies which are both cost-efficient and robust to external disruptions. Several studies have demonstrated that significant improvements in reliability can often be obtained without significant increases in operating costs (Snyder and Daskin 2005).

Location models for planning reliable systems can be broadly grouped into two main categories which reflect different risk attitudes of the decision maker: risk-averse and risk-neutral.

24.5.1 Planning for a Risk-Averse Designer

The models in this category identify location strategies for coping with the worst case in terms of facility loss or disruption. They therefore capture the perspective of a risk-averse decision maker and are suitable for hedging against deliberate disruptions and strategic risks. These models typically embed an interdiction model in a multi-level structure where the upper-level model identifies the optimal location of the facilities, whereas the lower-level model endogenously generates worse-case scenario losses.

We illustrate how such location-interdiction models can be formulated by presenting the Maximal Covering Location-Interdiction Problem (MCLIP). The idea is to couple the classical Maximal Covering Location problem with the r-ICP presented in Sect. 24.3 to identify the location of p facilities which maximizes a weighted combination of (1) the initial coverage and (2) the minimum coverage level following the loss of the most critical r facilities (O'Hanley and Church 2011).

The MCLIP model can be formulated as follows:

$$\text{maximize} \quad \alpha \sum_{j \in J} d_j u_j + (1-\alpha) H(y) \tag{24.17}$$

$$\text{subject to} \quad \sum_{i \in I} y_i = p \tag{24.18}$$

$$\sum_{i \in N_j} y_i \geq u_j \quad \forall j \in J \tag{24.19}$$

$$y_i \in \{0,1\} \quad \forall i \in I \tag{24.20}$$

$$u_j \in \{0,1\} \quad \forall j \in J, \tag{24.21}$$

where

$$H(y) = \min \sum_{j \in J} d_j v_j \tag{24.22}$$

$$\text{subject to} \quad \sum_{i \in I} s_i = r \tag{24.23}$$

$$v_j \geq y_i - s_i \quad \forall j \in J, i \in N_j \tag{24.24}$$

$$s_i \in \{0,1\} \quad \forall i \in I \tag{24.25}$$

$$v_j \in \{0,1\} \quad \forall j \in J. \tag{24.26}$$

The upper level objective (24.17) is to maximize the weighted sum of covered demand before and after interdiction by locating p facilities (24.18). The demand covered before interdiction is determined by constraints (24.19), whereas the worst-case demand-weighted coverage after interdiction, $H(y)$, is computed in the lower level problem (24.22) – (24.26). This is a simple modification of the r-ICP problem (24.7) – (24.11), where constraints (24.8) are replaced by (24.24). These constraints state that customer j must be covered after disruption ($v_j = 1$) unless all the open facilities covering customer j are interdicted.

Bilevel location-interdiction problems such as the MCLIP are even more difficult to solve than the protection-interdiction problems discussed in Sect. 24.4 and some efficient approaches devised for protection models, such as the implicit enumeration algorithm for r-IMPF, are not applicable to them. In O'Hanley and Church (2011), the MCLIP is solved by a decomposition method using *supervalid inequalities*.

Another example of location/interdiction models can be found in Parvaresh et al. (2012) for p-hub median problems. In this case, the bilevel model is solved heuristically via simulated annealing and tabu search.

Note that design and protection decisions may be coupled within the same modeling framework. Examples of risk-averse design models including the option of hardening some of the facilities to be located can be found in Aksen et al. (2011), Aksen and Aras (2012) and Shishebori and Jabalameli (2013).

24.5.2 Planning for a Risk-Neutral Designer

In this class of models, facilities are assumed to fail at random and the objectives typically deal with expected costs or performances.

Although the first paper to consider unreliable facilities which fail with a given probability appeared more than a couple of decades ago (Drezner 1987), a renewed interest in this type of problems has only emerged more recently with the reliability problems investigated by Snyder and Daskin (2005): the Reliability p-Median Problem (RPMP) and the Reliability Fixed-Charge Location Problem (RFLP). Both problems aim at locating a set of facilities so as to minimize the costs incurred by the system when all the facilities are operational and the expected transportation costs after facilities failures.

In the RPMP model, each open facility may fail with the same fixed probability π, failures are independent and several facilities can fail simultaneously. If customer j is not served by any facility, either because all open facilities fail or because it is too costly to receive service by the closest operational facility, the system incurs a lost-sale cost per unit of demand. To model this situation, the set I of potential locations for the facilities is augmented with a dummy emergency facility. Let m be the cardinality of the augmented set $|I|$ and the index of the emergency facility. The emergency facility m never fails and has unitary service cost c_{mj} to customer j. As facility m is forced to open, $p + 1$ facilities must be located instead of p as in standard p-median problems.

To formulate RPMP, the following assignment variables are defined:

$$x_{ijl} = \begin{cases} 1 & \text{if customer } j \text{ is assigned to facility } i \text{ at level } l \\ 0 & \text{otherwise} \end{cases}$$

The idea behind the RPMP formulation is that each customer is assigned to facilities depending upon their operational status. Accordingly, several assignment levels can be associated with each customer. Level-0 assignments are those made to primary facilities that serve the customers under normal circumstances. Level-l assignments ($l > 0$) are those made to alternative facilities that can serve a customer if the l closer facilities have failed.

The RPMP model is as follows.

$$\text{minimize} \quad \sum_{j \in J} d_j \sum_{l=0}^{p} \left[\sum_{i \in I \setminus m} c_{ij} \pi^l (1-\pi) x_{ijl} + c_{mj} \pi^l x_{mjl} \right] \tag{24.27}$$

$$\text{subject to} \quad \sum_{i \in I} x_{ijl} + \sum_{t=0}^{l-1} x_{mjt} = 1 \quad \forall j \in J, l = 0, \ldots, p \tag{24.28}$$

$$\sum_{l=0}^{p} x_{ijl} \leq 1 \quad \forall i \in I, j \in J \tag{24.29}$$

$$x_{ijl} \leq y_i \quad \forall i \in I, j \in J, r = 0, \ldots, p \tag{24.30}$$

$$\sum_{i \in I} y_i = p + 1 \tag{24.31}$$

$$y_m = 1 \tag{24.32}$$

$$y_i \in \{0, 1\} \quad \forall i \in I \tag{24.33}$$

$$x_{ijl} \in \{0, 1\} \quad \forall i \in I, j \in J, l = 0, \ldots, p. \tag{24.34}$$

The objective function (24.27) minimizes the demand-weighted expected transportation and lost-sales costs. These are computed as a function of the assignment variables by taking into account that each customer j is served by its level-l facility i if the l closer facilities have failed, which occurs with probability π^l, and facility i has not failed, which occurs with probability $1 - \pi$ for each $i \in I \setminus m$ and with probability 1 if $i = m$. Constraints (24.28) state that each customer j must be assigned to some facility at each level l, unless j has been assigned to the emergency facility at level $t < l$. Constraints (24.29) prevent the assignment of a customer to a given facility at more than one level. Constraints (24.30) prohibit the assignment to facilities which are not open, whereas constraint (24.31) state that exactly p facilities must be opened in addition to the emergency facility, which is forced to be open by constraint (24.32). Constraints (24.33) and (24.34) are standard integrality constraints (note that the integrality constraints on the assignment variables x_{ijl} can be relaxed).

The original RPMP model presented in Snyder and Daskin (2005) is slightly more general than model (24.27) – (24.34) in two aspects: (1) some of the facilities may be considered completely reliable and (2) the objective is to minimize the weighted sum of normal costs and expected failure costs. The authors show that by varying the weights of the resulting bi-objective model, one can generate a trade-off curve for identifying good compromise solutions. This type of analysis demonstrates that large reductions in failure costs can often be attained with only minor increases in operation costs.

The Reliability Fixed-Charge Location Problem, which we do not report for the sake of brevity, can be formulated in a similar way to RPMP. Both problems can be tackled by Lagrangian relaxation (Snyder and Daskin 2005). Efficient metaheuristic approaches have also been devised for RPMP by Alcaraz et al. (2012), which report very good results for large scale instances.

One of the major limitations of this structure for reliability models is that it relies on the assumption that all facilities fail with the same probability. Without this assumption, calculating expected transportation costs becomes significantly more complicated due to the need of expressing probability products using high-degree polynomials. Site-dependent probabilities were considered for the first time by Berman et al. (2007) but the resulting model is highly non-linear and is only solved heuristically. Several attempts at modelling heterogeneous facility failure probabilities using a linear mixed integer program have appeared in recent years (see for example Cui et al. 2010 and Lei and Tong 2013). Particularly noteworthy is the *probability chains* linearization technique proposed by O'Hanley et al. (2013) for solving the RPMP with site-dependent probabilities. The technique, which is general and can be extended to other model classes as well, is based on the idea of using a specialized network flow structure for evaluating compound probability terms. Empirical experiments indicate that this technique is quite effective in solving reliability models of significant size.

Other important issues in modeling location problems with unreliable facilities are correlation and informational uncertainty. Correlation concerns the extent to which the failure of one facility affects the operational status of other facilities. In many real situations neighboring facilities may be exposed to similar hazards and, therefore, fail simultaneously. Examples of models with correlated disruptions can be found in Li and Ouyang (2010) and Berman et al. (2013). Informational uncertainty relates to the information available to customers about the operational state of the facilities. It is clear that optimal location patterns and optimal service costs may differ if customers do not have prior information about the state of the facilities and must travel to different facilities before they can receive service. The role of information in reliable facility design is analyzed in Berman et al. (2009) and Berman et al. (2013).

Finally, as for the bilevel design models discussed in the previous section, location and hardening decisions can be combined into a probabilistic design model for identifying reliable and cost-efficient configurations of hardened and unhardened facilities (see, for example, Lim et al. 2010 and Li et al. 2013).

24.5.2.1 Scenario-Indexed Models

When the uncertainty associated with disruptions can be captured by a finite set of scenarios, we can resort to scenario-indexed models. Within the context discussed in this chapter, such models are an alternative way for writing two-stage stochastic mixed integer programs. The non-anticipative first-stage decisions concern the location of the facilities and are made in the presence of uncertainty about the realization of future disruption scenarios. The second-stage (recourse) decisions, which are conditional to the first-stage decisions, involve the assignment of customers to facilities in response to specific disruption scenarios.

Below we show a scenario-indexed model for the p-median problem, where the objective is to minimize the expected service cost over all failure scenarios. Let Ω be the set of disruption scenarios such that $a_{i\omega} = 1$ if facility i fails in scenario ω. The probability that scenario ω occurs is denoted by π_ω. The assignment decision variables are defined for each scenario as follows:

$$x_{ij\omega} = \begin{cases} 1 & \text{if customer } j \text{ is assigned to facility } i \text{ in scenario } \omega \\ 0 & \text{otherwise} \end{cases}$$

The scenario-indexed model is then:

$$\text{minimize} \quad \sum_{\omega \in \Omega} \pi_\omega \sum_{i \in I} \sum_{j \in J} d_j c_{ij} x_{ij\omega} \tag{24.35}$$

$$\text{subject to} \quad \sum_{j \in J} x_{ij\omega} \leq (1 - a_{i\omega}) y_i \quad \forall i \in I, \omega \in \Omega \tag{24.36}$$

$$\sum_{i \in I} x_{ij\omega} = 1 \quad \forall j \in J, \omega \in \Omega \tag{24.37}$$

$$\sum_{i \in I} y_i = P \tag{24.38}$$

$$y_i \in \{0, 1\} \quad \forall i \in I \tag{24.39}$$

$$x_{ij\omega} \in \{0, 1\} \quad \forall i \in I, j \in J, \omega \in \Omega \tag{24.40}$$

The objective function (24.35) minimizes the demand-weighted expected cost across all scenarios. Constraints (24.36) prevent the assignment of customer j to facility i in scenario ω if either i is not open or if it is open but not available in scenario ω. Constraints (24.37) guarantee that each customer is assigned to some facility in every scenario. The remaining constraints are standard cardinality and integrality constraints.

The expected performance criterion used in problem (24.35)–(24.40) yields solutions that may perform poorly in certain scenarios. Solutions which are effective no matter what scenario is realized can be obtained by incorporating robustness measures into the model. An example is the β-robustness measure introduced by

Snyder and Daskin (2006). Let z^*_ω be the optimal cost for scenario ω. By adding the following constraint

$$\sum_{i \in I} \sum_{j \in J} d_j c_{ij} x_{ij\omega} \leq (1 + \beta) z^*_\omega \quad \forall \omega \in \Omega, \tag{24.41}$$

it is possible to generate least-cost solutions whose relative regret in each scenario is no more than β, for a given $\beta \geq 0$.

The β-robustness measure has been used in Peng et al. (2011) to design reliable multi-echelon supply chain networks. Other risk measures to generate robust solutions in scenario planning models include the α-reliable minimax regret (Daskin et al. 1997) and the α-reliable mean-excess regret (Chen et al. 2006). In α-reliable minimax models, the maximum regret is computed only over a subset of scenarios, called the *reliability set*, whose total probability is at least α. The α-reliable mean-excess regret, which is closely related to the conditional value-at-risk (CVaR) objective of portfolio optimization (Rockafellar and Uryasev 2000), further extends the α-reliable concept by ensuring that solutions perform reasonably well even in the scenarios which are not included in the reliability set. Typically, the objective function of these models minimizes a weighted sum of the maximum regret over the reliability set and the conditional expectation of the regret over the scenarios excluded from the reliability set. Although these measures have not been explicitly used in facility location problems with disruptions, their application is quite straightforward and certainly deserves future investigation.

When uncertainty can be captured by a finite set of scenarios and a scenario-indexed model can be considered, it is easy to modify the model in a way that the models discussed in Sect. 24.5.2 cannot. As an example, capacity restrictions can be easily modeled by replacing constraints (24.36) with

$$\sum_{j \in J} d_j x_{ij\omega} \leq (1 - a_{i\omega}) q_i y_i \quad \forall i \in I, \omega \in \Omega, \tag{24.42}$$

where q_i is the capacity of facility i.

Partial disruptions can also be captured by simply redefining $a_{i\omega}$ as the proportion of facility i capacity which is lost in scenario ω to model the case where disruptions only reduce the capacity but do not completely disable a facility.

One major drawback of scenario-indexed models is that they can become very large if there are many scenarios (consider for example all the possible ways in which subsets of facilities can fail). To obviate this difficulty, the scenario space can be approximated using sampling techniques such as Sample Average Approximation (Kleywegt et al. 2002). Another alternative is to construct the scenario set empirically by using historical data or expert judgement. As an example, Rawls and Turnquist (2010) use a scenario planning approach to optimize facility locations and emergency resource stockings in the face of natural disasters. In their case study, the scenarios of concern are constructed by using historical records from a sample of 15 hurricanes.

24.6 Future Trends

The research to date on facility location problems with disruption, although groundbreaking, is still evolving. The impetus for such work has come from disasters such as 9/11, the Fukushima nuclear power plant destruction in Japan, and the more recent power disruption in Michoacan, Mexico. As such problems are often represented as a two person game (defender-attacker) or a three person game (defender-attacker-defender), they can be quite mathematically complex and difficult to solve. Because of this, work is needed to expand the range of problem sizes that can be addressed by such model structures.

The work discussed here is based upon the simplest of service systems involving the p-median and maximal covering problems. Although these problems and extensions can be used in many system designs, lifeline systems such as electrical generation and transmission, water supply and distribution, and communication networks of switches and lines, all present a level of complexity that has yet to be addressed in an efficient and comprehensive way. Systems are interconnected in many ways. A failure (or an attack) of one system component may lead to the failure of another. Such cascading failures have been documented in electrical and communication systems. In addition, the failure of an electrical system component may render a portion of a communication system inoperable. Connections between such systems have still to be adequately modeled as well. In addition, most models capturing disruption ignore the temporal component. Few have addressed the possible duration of a disrupting event as well as how best to cope with it and restore the initial operational level. This too, is an area where more research is needed.

Facilities are but one component in a production and distribution system. Recent flooding in Thailand demonstrated that inventories for key parts, like those for computer disk drives, could be disrupted to the extent that the retail price for storage drives almost doubled for a short period of time. Fully addressing such vulnerabilities requires the modeling of facility production and inventory levels simultaneously.

There are two principal ways in which resilient design has been approached: scenario based with robust optimization, and bilevel optimization. Work is needed to test the efficacy of each approach. For example, can a small number of scenarios be used to adequately define and couch possible outcomes as compared to the use of a bilevel optimization problem involving a defender-attacker? In addition, can simulation models be used in an efficient manner to identify system vulnerabilities? Further, it is important to develop better models to estimate risk.

Finally, the models developed to date to handle interdiction, fortification and reliable design are far more complex than their base-level counterparts, adding a level of computational difficulty that is a new research area. But, one must ask the question: can simpler models be developed which adequately address such uncertainties?

24.7 Conclusions

This chapter has reviewed the research that has evolved over the last decade concerning facility disruption. Disruptions can be thought as arising out of intention (e.g., terrorism), by accident, or by a natural disaster. It has covered three main areas of related research: models of facility interdiction, combined models of facility interdiction and protection, and models of resilient design. These models are designed to address the three basic questions that concern systems planners and operators when facing reality: (1) how much can a service system be degraded in its efficiency when disrupted; (2) how might resources be allocated to protect against such possible events; (3) how might a new system be designed so that it is naturally resilient? Although past work has been based principally on the application of such models using hypothetical data, they have demonstrated that small changes in levels of protection can be effective at improving a system's ability to cope with a disaster. Further, it has been shown that equal if not better facility deployment results when taking into account possible levels of disruption (whether intentional or natural). Ignoring disaster may come at a cost that is too high when compared to addressing such possibilities in operation (interdiction/fortification) and design. In fact, the value in modeling for disruption is that one can capture levels of impact and determine whether to ignore them or make system adjustments. This area of research is still evolving and future work is needed in applying such concepts to a wide range of lifeline systems, including power generation and distribution, food production and distribution, and water supply systems.

References

Aksen D, Aras N (2012) A bilevel fixed charge location model for facilities under imminent attack. Comput Oper Res 39:1364–1381

Aksen D, Piyade N, Aras N (2010) The budget constrained r-interdiction median problem with capacity expansion. Cent Eur J Oper Res 18:269–291

Aksen D, Aras N, Piyade N (2011) A bilevel p-median model for the planning and protection of critical facilities. J Heuristics 19:373–398

Aksen D, Şengül Akca S, Aras N (2012) A bilevel partial interdiction problem with capacitated facilities and demand outsourcing. Comput Oper Res 41:346–358

Alcaraz J, Landete M, Monge JF (2012) Design and analysis of hybrid metaheuristics for the reliability p-median problem. Eur J Oper Res 222:54–64

Berman O, Krass D, Menezes MBC (2007) Facility reliability issues in network p-median problems: strategic centralization and co-location effects. Oper Res 55:332–350

Berman O, Krass D, Menezes MBC (2009) Locating facilities in the presence of disruptions and incomplete information. Decis Sci 40:845–868

Berman O, Krass D, Menezes MBC (2013) Location and reliability problems on a line: impact of objectives and correlated failures on optimal location patterns. Omega 41:766–779

Bricha N, Nourelfath M (2013) Critical supply network protection against intentional attacks: a game-theoretical model. Reliab Eng Syst Safe 119:1–10

Cappanera P, Scaparra MP (2011) Optimal allocation of protective resources in shortest-path networks. Transp Sci 45:64–80

Casey N (2013) Mexican cartel retaliates against civilians. Wall Street J 260(43):1–5
Chen G, Daskin MS, Shen Z-M, Uryasev S (2006) The α-reliable mean-excess regret model for stochastic facility location modeling. Nav Res Log 53:617–626
Church RL (2003) COBRA: a new formulation of the classic p-median location problem. Ann Oper Res 122:103–120
Church RL, Scaparra MP (2007a) Analysis of facility systems reliability when subject to attack or a natural disaster. In: Murray AT, Grubesic TH (eds) Critical infrastructure. Springer, Berlin/Heidelberg, pp 221–241
Church RL, Scaparra MP (2007b) Protecting critical assets: the r-interdiction median problem with fortification. Geogr Anal 39:129–146
Church RL, Scaparra MP, Middleton RS (2004) Identifying critical infrastructure: the median and covering facility interdiction problems. Ann Assoc Am Geogr 94:491–502
Cui T, Ouyang Y, Shen Z-M (2010) Reliable facility location design under the risk of disruptions. Oper Res 58:998–1011
Daskin MS, Hesse SM, Revelle CS (1997) α-Reliable p-minimax regret: a new model for strategic facility location modeling. Location Sci 5:227–246
Dempe S (2002) Foundations of bilevel programming. Kluwer Academic Publishers, Dordrecht
Drezner Z (1987) Heuristic solution methods for two location problems with unreliable facilities. J Oper Res Soc 38:509–514
Kleywegt AJ, Shapiro A, Homem-de-Mello T (2002) The sample average approximation method for stochastic discrete optimization. SIAM J Optim 12:479–502
Lei TL (2013) Identifying critical facilities in hub-and-spoke networks: a hub interdiction median problem. Geogr Anal 45:105–122
Lei TL, Church RL (2011) Constructs for multilevel closest assignment in location modeling. Int Reg Sci Rev 34:339–367
Lei TL, Tong D (2013) Hedging against service disruptions: an expected median location problem with site-dependent failure probabilities. J Geogr Syst 15:491–512
Li X, Ouyang Y (2010) A continuum approximation approach to reliable facility location design under correlated probabilistic disruptions. Transp Res B Methodol 44:535–548
Li Q, Zeng B, Savachkin A (2013) Reliable facility location design under disruptions. Comput Oper Res 40:901–909
Liberatore F, Scaparra MP (2011) Optimizing protection strategies for supply chains: comparing classic decision-making criteria in an uncertain environment. Ann Assoc Am Geogr 101:1241–1258
Liberatore F, Scaparra MP, Daskin MS (2010) Analysis of facility protection strategies against an uncertain number of attacks: the stochastic R-interdiction median problem with fortification. Comput Oper Res 38:357–366
Liberatore F, Scaparra MP, Daskin MS (2012) Hedging against disruptions with ripple effects in location analysis. Omega 40:21–30
Lim M, Daskin MS, Bassamboo A, Chopra S (2010) A facility reliability problem: formulation, properties, and algorithm. Nav Res Log 57:58–70
Losada C, Scaparra MP, O'Hanley JR (2012a) Optimizing system resilience: a facility protection model with recovery time. Eur J Oper Res 217:519–530
Losada C, Scaparra MP, Church RL, Daskin MS (2012b) The stochastic interdiction median problem with disruption intensity levels. Ann Oper Res 201:345–365
Moore J, Bard J (1990) The mixed integer linear bilevel programming problem. Oper Res 38:911–921
O'Hanley JR, Church RL (2011) Designing robust coverage networks to hedge against worst-case facility losses. Eur J Oper Res 209:23–36
O'Hanley JR, Scaparra MP, Garcia S (2013) Probability chains: a general linearization technique for modeling reliability in facility location and related problems. Eur J Oper Res 230:63–75
Parvaresh F, Husseini SMM, Golpayegany SAH, Karimi B (2012) Hub network design problem in the presence of disruptions. J Intell Manuf 25:755–774

Peng P, Snyder LV, Lim A, Liu Z (2011) Reliable logistics networks design with facility disruptions. Transp Res B Methodol 45:1190–1211
Peterson SK, Church RL (2008) A framework for modeling rail transport vulnerability. Growth Change 39:617–641
Rawls CG, Turnquist MA (2010) Pre-positioning of emergency supplies for disaster response. Transp Res B Methodol 44:521–534
Reid T, Gorman S (2012) Los Angeles port strike triggers fears, lobbying by businesses. Reuters (on-line edition), 2 December 2012
Rockafellar RT, Uryasev S (2000) Optimization of conditional value-at-risk. J Risk 2:21–41
Scaparra MP, Church RL (2008a) A bilevel mixed-integer program for critical infrastructure protection planning. Comput Oper Res 35:1905–1923
Scaparra MP, Church RL (2008b) An exact solution approach for the interdiction median problem with fortification. Eur J Oper Res 189:76–92
Scaparra MP, Church RL (2012) Protecting supply systems to mitigate potential disaster: a model to fortify capacitated facilities. Int Reg Sci Rev 35:188–210
Sevaux M, Sörensen K, Martí R (2015) Metaheuristics: a comprehensive guide to the design and implementation of effective optimisation strategies. Springer/New York, LLC
Shishebori D, Jabalameli MS (2013) A new integrated mathematical model for optimizing facility location and network design policies with facility disruptions. Life Sci J 10:1896–1906
Smith R (2014) Nation's power grid vulnerable to sabotage. Wall Street J 263:1–6
Snyder LV, Daskin MS (2005) Reliability models for facility location: the expected failure cost case. Transp Sci 39:400–416
Snyder LV, Daskin MS (2006) Stochastic p-robust location problems. IIE Trans 38:971–985
Soble J (2011) Honda suffers as Thai floods shut plant. Financial Times, October 21, 2011
Wollmer R (1964) Removing arcs from a network. Oper Res 12:934–940
Wood RK (1993) Deterministic network interdiction. Math Comput Model 17:1–18
Zhu Y, Zheng Z, Zhang X, Cai K (2013) The r-interdiction median problem with probabilistic protection and its solution algorithm. Comput Oper Res 40:451–462

About the Editors

Gilbert Laporte obtained his Ph.D. in Operations Research from the London School of Economics in 1975. He is professor of Operations Research at HEC Montréal, and Canada Research Chair in Distribution Management. He has been Editor of Transportation Science, Computers & Operations Research and INFOR. He has authored or coauthored 15 books, as well as more than 450 scientific articles in combinatorial optimization, mostly in the areas of vehicle routing, location and timetabling. He has received many scientific awards including the Pergamon Prize (United Kingdom) in 1987, the 1994 Merit Award of the Canadian Operational Research Society, and the CORS Practice Prize on three occasions. He has been a member of the Royal Society of Canada since 1998 and Fellow of INFORMS since 2005. In 2009, he received the Robert M. Herman Lifetime Achievement Award in Transportation Science from the Transportation Science and Logistics Society of INFORMS. In 2014, he obtained the Lifetime Achievement in Location Analysis Award from the Section on Location Analysis of INFORMS.

G. Laporte et al. (eds.), *Location Science*, DOI 10.1007/978-3-319-13111-5

Stefan Nickel obtained his Ph.D. in Mathematics at the Technical University of Kaiserslautern, Germany in 1995. He is a full professor and one of the directors of the Institute for Operations Research at the Karlsruhe Institute of Technology (KIT). Stefan Nickel is also member of the scientific advisory board, as well as the management board, of the Fraunhofer Institute for Applied Mathematics (ITWM) in Kaiserslautern, Germany. In 2011 he became one of the directors of the Karlsruhe Service Research Institute (KSRI) and of the Research Center for Information Technology (FZI) in Karlsruhe. He has authored or coauthored four books as well as around 100 scientific articles, mainly in the areas of location, supply chain management, health care and logistics. In addition, he had numerous research contracts with well-known companies such as BASF, Lufthansa, Miele and SAP. Stefan Nickel has been the Editor-in-Chief of Computers & Operations Research since October 2006 and is a member of the editorial board of Health Care Management Science. He has coordinated the Health Care working group within the German OR society (GOR) and was the president of the GOR from 2013 to 2014. Moreover, he has been coordinator of the EURO working group on locational analysis.

Francisco Saldanha da Gama is Professor of Operations Research at the Department of Statistics and Operations Research at the Faculty of Science of the University of Lisbon, where he received his Ph.D. in 2002. He has extensively published in scientific international journals, mostly in the areas of location theory, supply chain management, logistics and combinatorial optimization. Together with Stefan Nickel, he has been awarded the EURO prize for the best EJOR review paper (2012) and the Elsevier prize (2012) for the EJOR top cited article in the years 2007 to 2011, both for the paper entitled "Facility Location and Supply Chain Management—A Review". He is member of various international scientific organizations such as the EURO Working Group on Location Analysis of which he is one of the past coordinators. He is currently Area Editor of Computers & Operations Research. His research interests include stochastic mixed integer optimization, location theory and project scheduling.