Gilbert Laporte
Stefan Nickel
Francisco Saldanha da Gama *Editors*

Location Science

Second Edition

Springer

Location Science

Gilbert Laporte • Stefan Nickel •
Francisco Saldanha da Gama
Editors

Location Science

Second Edition

Springer

Editors
Gilbert Laporte
HEC Montréal
Montréal, QC, Canada

Stefan Nickel
Institute for Operations Research
Karlsruhe Institute of Technology
Karlsruhe, Germany

Francisco Saldanha da Gama
Faculty of Science
University of Lisbon
Lisbon, Portugal

ISBN 978-3-030-32176-5 ISBN 978-3-030-32177-2 (eBook)
https://doi.org/10.1007/978-3-030-32177-2

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

Preface

The first edition of the book *Location Science* was published as a response to a need expressed by the location science community to have a comprehensive book covering all main aspects in the field. The book was highly successful and its chapters have been downloaded more than 50,000 times. Given this success, we decided to publish the second edition.

The first edition contained 24 chapters and 644 pages. In this second edition, we decided to remove some chapters that were not central to the field of location science in order to add new topics. We also modified other chapters to better reflect the evolution of the field. All remaining chapters were updated. The result is a much improved book with 26 chapters and 788 pages.

We thank all the authors who accepted our challenge to be involved in the second edition of the book. Their commitment, dedication, and enthusiasm will certainly guarantee that this new edition will be as successful as the previous one.

Finally, we thank Mr Christian Rauscher and the Springer staff for their help and encouragement throughout this project.

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

Contents

Part II Advanced Concepts

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.

1.1 Introduction

Since the 1960s, 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
Canada Research Chain in Distribution Management, Interuniversity Research Centre on Enterprise Networks, Logistics, and Transportation (CIRRELT), HEC Montréal, Montréal, QC, Canada
e-mail: gilbert.laporte@cirrelt.ca

S. Nickel
Institute for Operations Research and Research Center for Information Technology (FZI), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
e-mail: stefan.nickel@kit.edu

F. Saldanha-da-Gama (✉)
Departamento de Estatística e Investigação Operacional, Centro de Matemática, Aplicações Fundamentais e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal
e-mail: fsgama@ciencias.ulisboa.pt

G. Laporte et al. (eds.), *Location Science*,
https://doi.org/10.1007/978-3-030-32177-2_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 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 in 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 and its extensions still attract the attention of the scientific community (see, for instance, Nam 2013, Görner and Kanzow 2016, Benko and Coroian 2018).

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

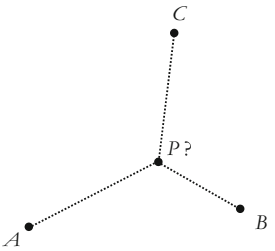


Fig. 1.1 Fermat's problem

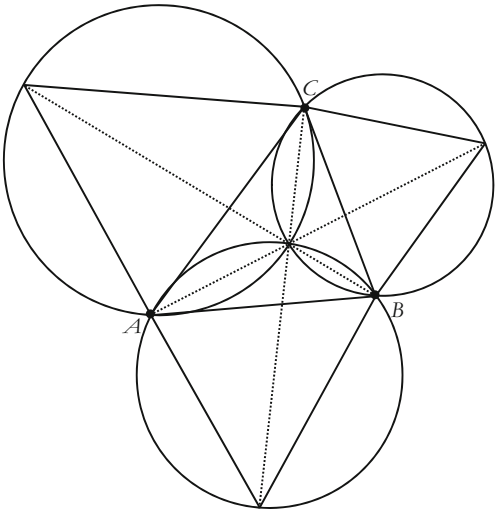
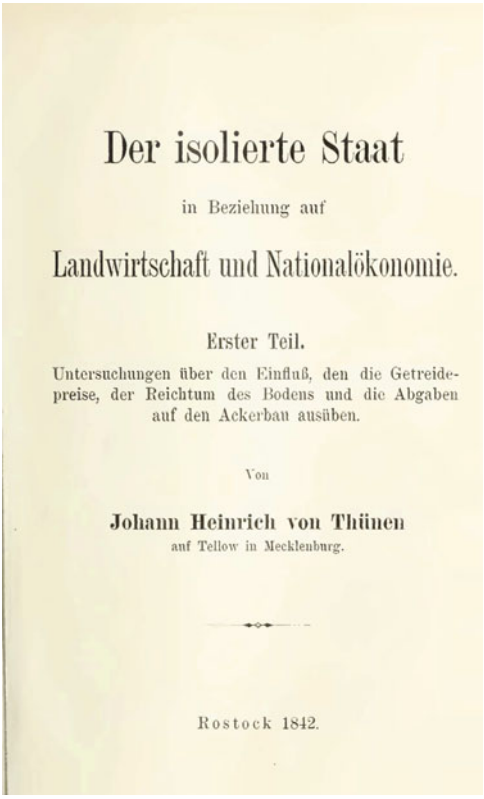


Fig. 1.2 Torricelli's geometric construction for the Fermat's problem

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 *Der 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 Thünen (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.



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

One should therefore expect that the different agriculture activities will evolve in concentric rings around the urban center (Fig. 1.4).

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 iron ore (collected at A) have to be combined with w_B tons of coal (collected at

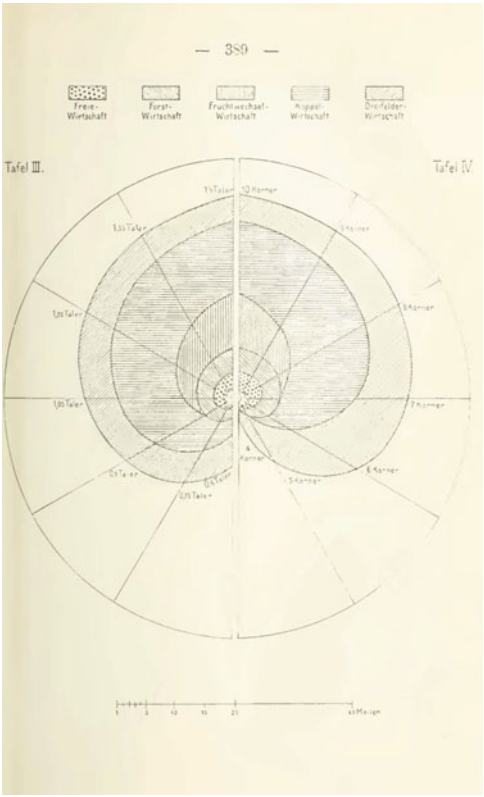


Fig. 1.4 Von Thünen's rings. From "Der Isolierte Staat" by Johann Heinrich von Thünen, Rostock 1842, page 389 (Source: University of Toronto—Robarts Library, https://archive.org/details/derisoliertestaa00thuoft)

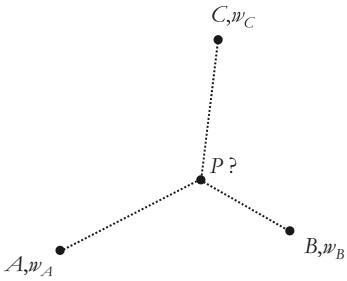


Fig. 1.5 Location problem proposed by Launhardt (1900) within an industrial context

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 $w_A d_A + w_B d_B + w_C d_C$.

This problem introduced by Launhardt is exactly what we now call the 3-node Weber problem. However, as pointed out by Pinto (1977), the problem was

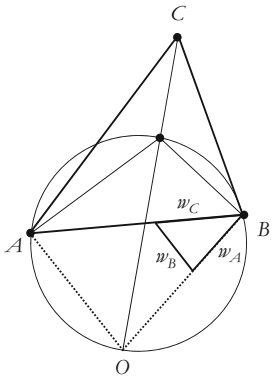


Fig. 1.6 Launhardt's geometric solution

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—the *weight triangle* as it is referred to by Weber (1909). 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.[2] 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 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).[3] The iterative procedure proposed in this work was revisited in depth more recently by Plastria (2011).

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

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

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 papers by Brimberg and Drezner 2013; Brimberg et al. 2014; Drezner et al. 2015a,b; Drezner and Salhi 2017).

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.

It is interesting to note that an important step toward the development of discrete facility location problems had been taken the previous year when Manne (1964) 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). This model would be revisited later by Balinski (1965) who introduced inequalities of the type $x_{ij} \leq x_{jj}$ ensuring that if a node i is allocated to a node j, then the latter corresponds to a facility and therefore it is assigned to itself. Such inequalities would be later considered by ReVelle and Swain (1970) when formulating the first MILP model for the discrete p-median problem. In the following year, 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 book by Eiselt and Marianov (2011) describes the works that can be considered to constitute the basis of Location Science.

Within a few decades, significant advances were 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). In this volume, we do not focus on computational geometry since there are excellent volumes covering the topic (e.g. Goodman et al. 2017)

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: p-center problems;
- Chapter 4: Fixed-charge facility location problems;
- Chapter 5: 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 6: Anti-covering problems.
- Chapter 7: Locating 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: Hierarchical facility location problems;
- Chapter 14: Competitive location models;
- Chapter 15: Location-routing and location-arc routing;
- Chapter 16: Location logistics in supply chain management;
- 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:

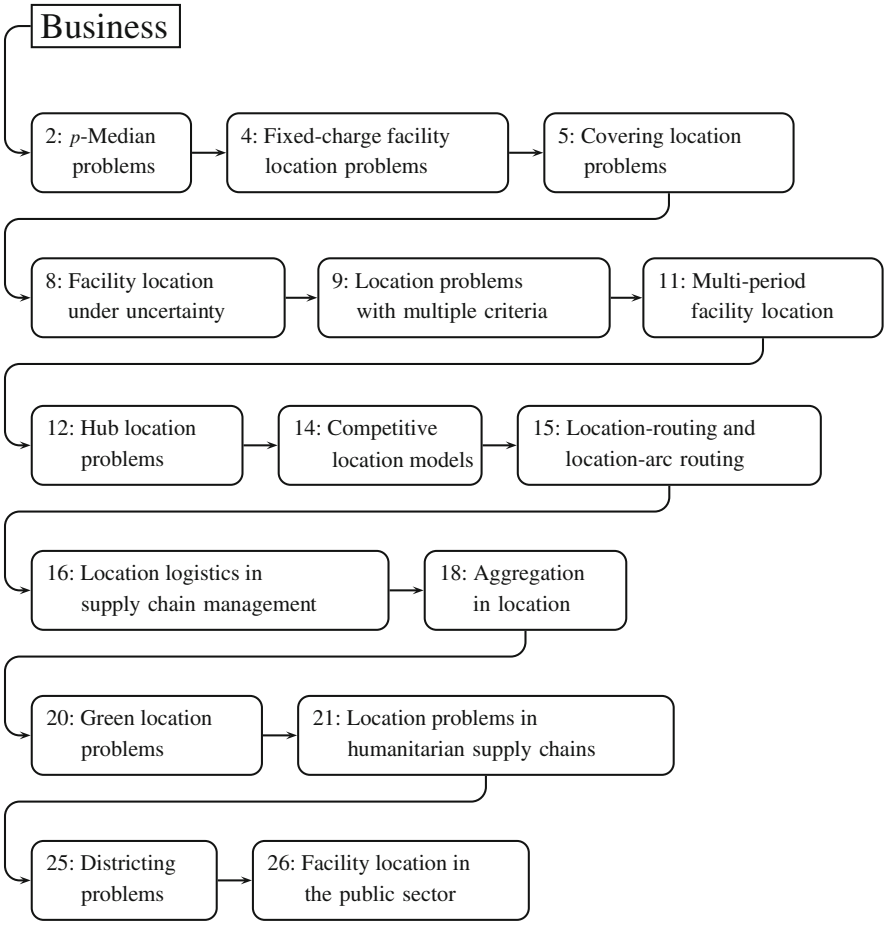
- Chapter 19: Location and geographic information systems;
- Chapter 20: Green location problems;
- Chapter 21: Location problems in humanitarian supply chains;
- Chapter 22: Location problems under disaster events;
- Chapter 23: Location problems in healthcare;
- Chapter 24: The design of rapid transit networks;
- Chapter 25: Districting problems;
- Chapter 26: Facility location in the public sector.

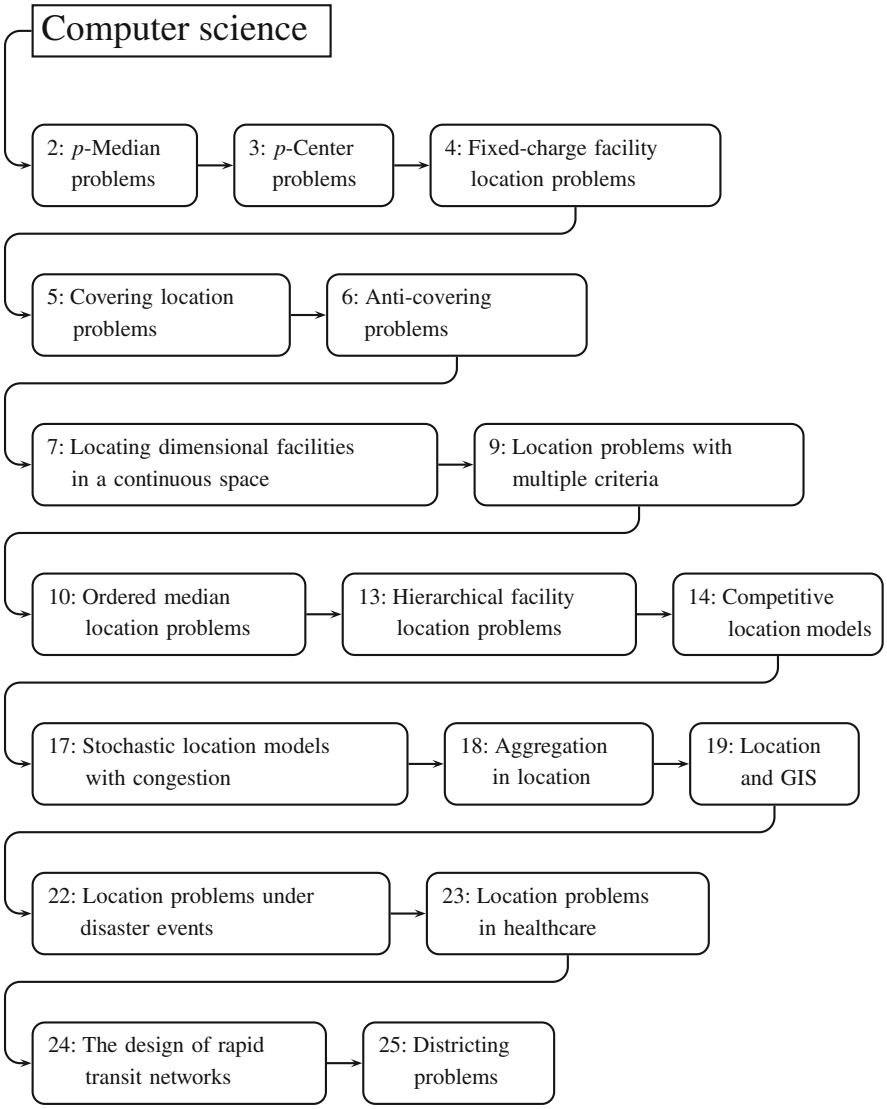
This second edition of *Location Science* should be viewed as a complement to the first edition. It covers topics that were not included in the first edition, such as hierarchical facility location, location problems capturing environmental concerns, location problems in humanitarian logistics, and location problems in the public sector. On the other hand, some topics that were sufficiently covered by the first edition are not part of the current volume. These include quadratic assignment problems and location problems in telecommunications. For such problems the reader should refer to the first edition of the book.

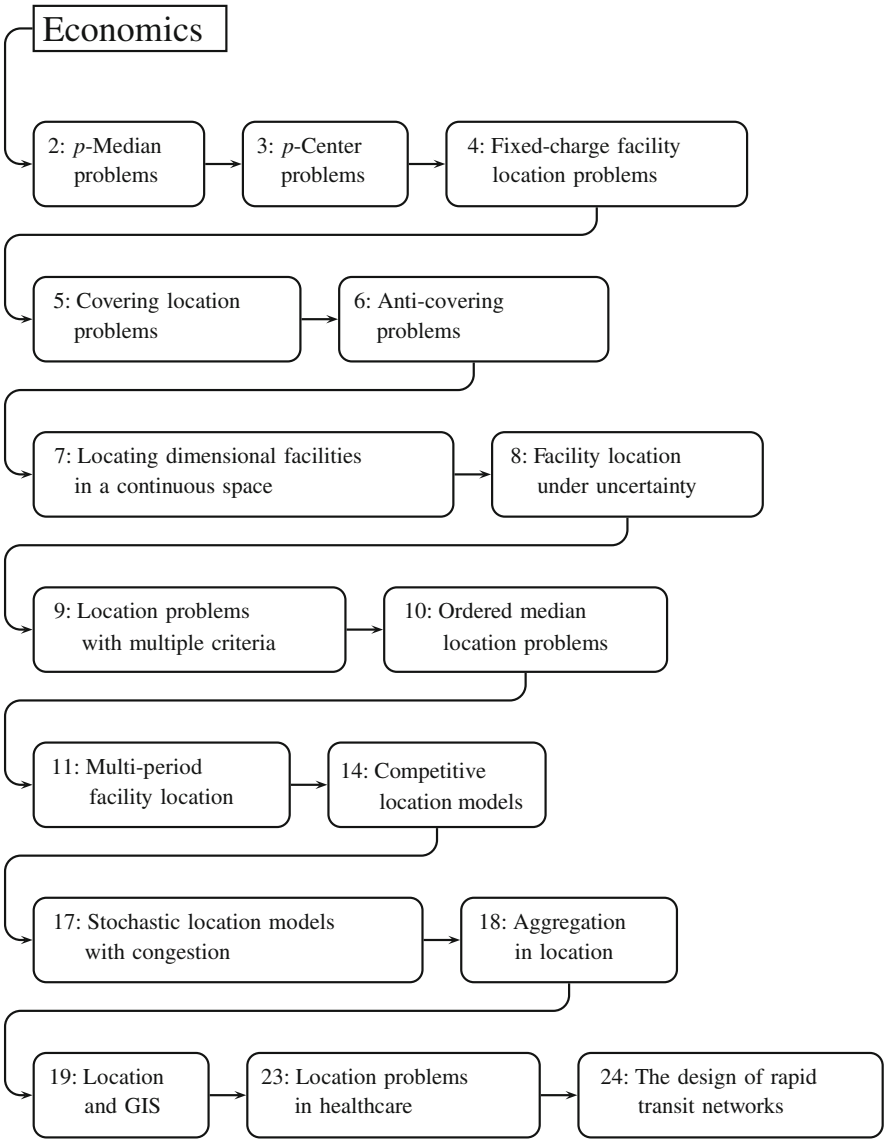
1.5 How to Use This Book

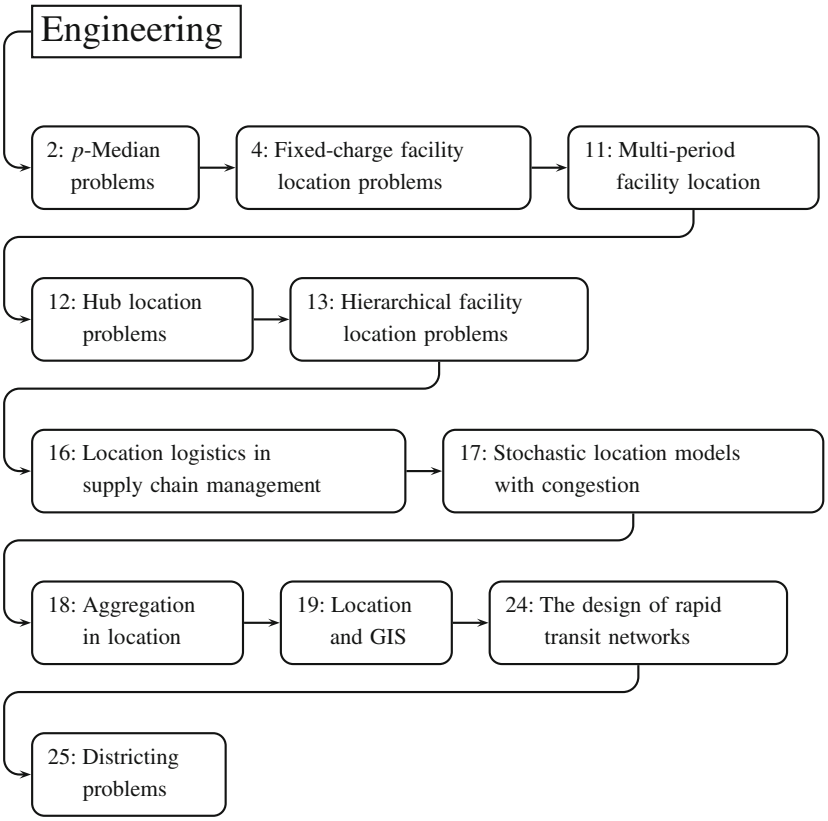
Problems, models, properties, and techniques from Location Science are taught to students enrolled in different programs. We have identified six types of post-graduate curricula having a strong location content: business, computer science, economics, engineering, geography and mathematics.

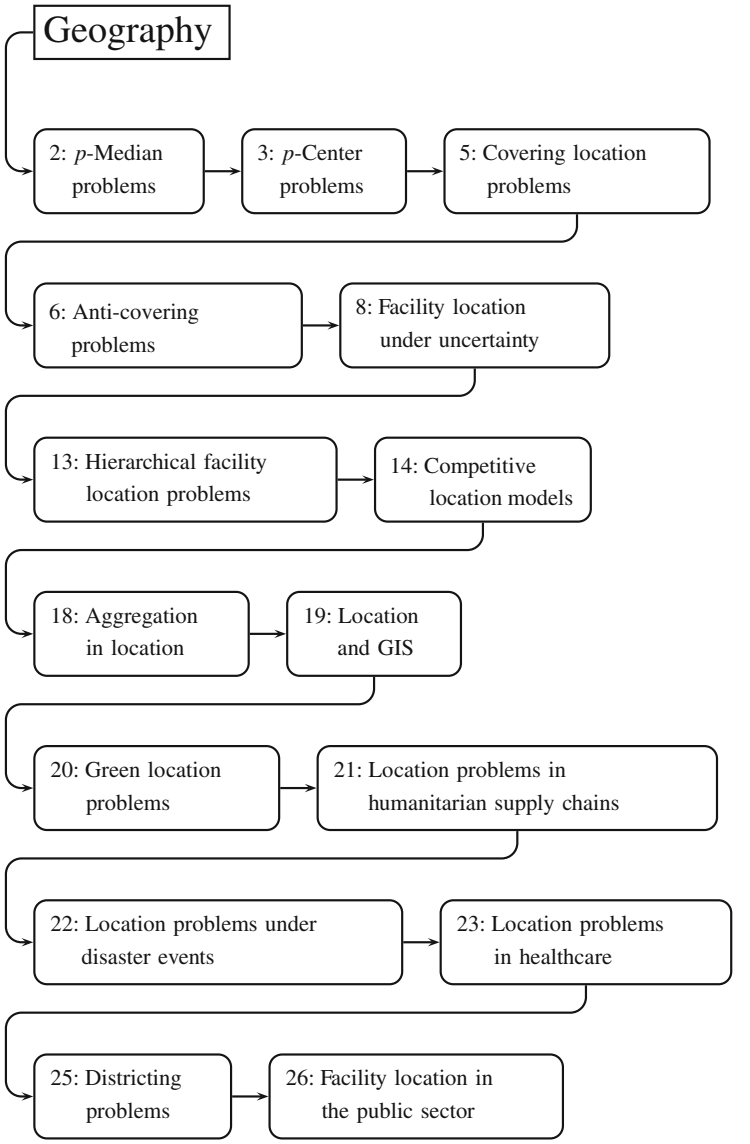
Depending on the audience, different contents emerge as the most appropriate. This book can be used with the purpose of organizing courses tuned for specialized targets by selecting specific combinations of chapters. Below, we offer some suggestions.

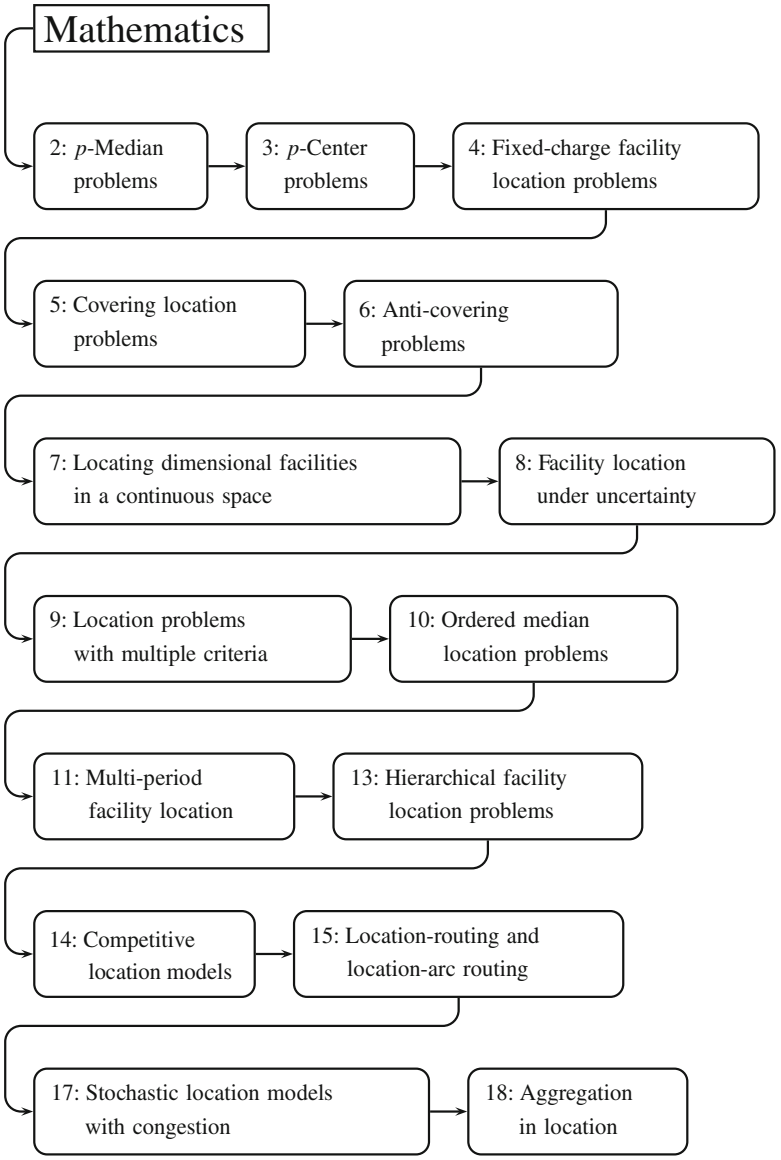

Business
2: p-Median problems
4: Fixed-charge facility location problems
5: Covering location problems
8: Facility location under uncertainty
9: Location problems with multiple criteria
11: Multi-period facility location
12: Hub location problems
14: Competitive location models
15: Location-routing and location-arc routing
16: Location logistics in supply chain management
18: Aggregation in location
20: Green location problems
21: Location problems in humanitarian supply chains
25: Districting problems
26: Facility location in the public sector


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

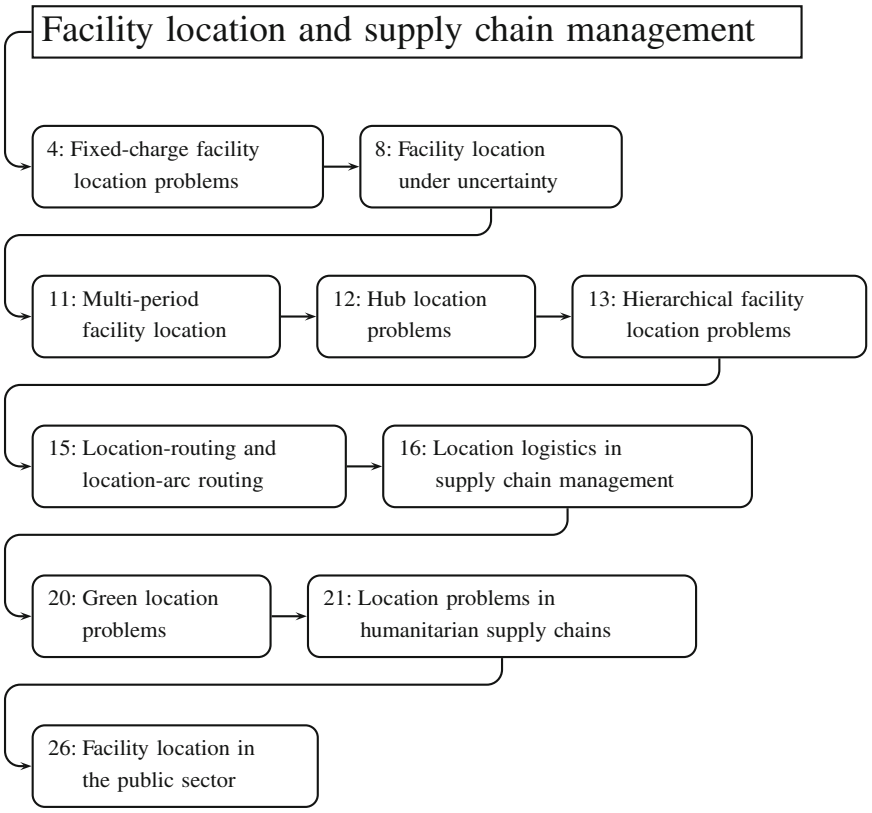

Economics
2: p-Median problems
3: p-Center problems
4: Fixed-charge facility location problems
5: Covering location problems
6: Anti-covering problems
7: Locating dimensional facilities in a continuous space
8: Facility location under uncertainty
9: Location problems with multiple criteria
10: Ordered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

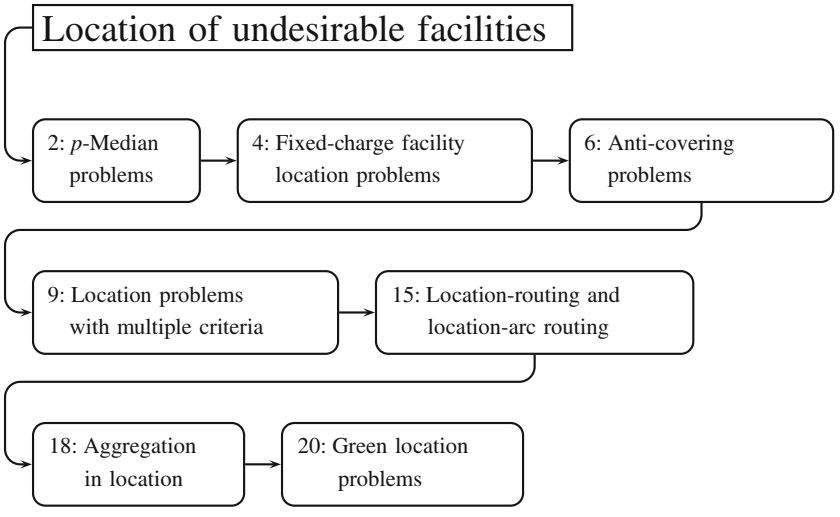

Engineering
2: p-Median problems
4: Fixed-charge facility location problems
11: Multi-period facility location
12: Hub location problems
13: Hierarchical facility location problems
16: Location logistics in supply chain management
17: Stochastic location models with congestion
18: Aggregation in location
19: Location and GIS
24: The design of rapid transit networks
25: Districting problems

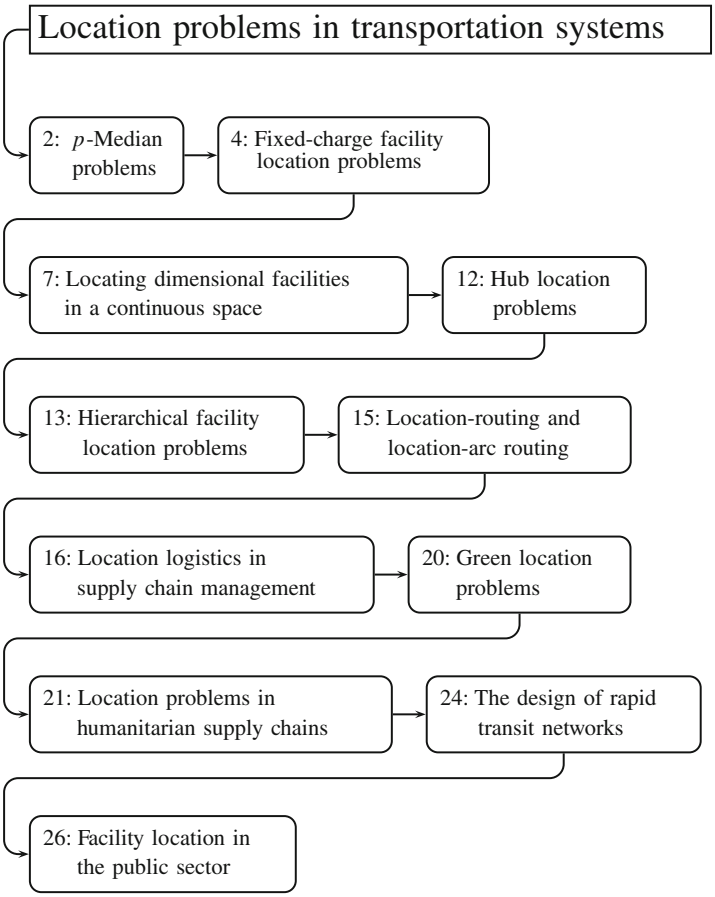

Geography
2: p-Median problems
3: p-Center problems
5: Covering location problems
6: Anti-covering problems
8: Facility location under uncertainty
13: Hierarchical facility location problems
14: Competitive location models
18: Aggregation in location
19: Location and GIS
20: Green location problems
21: Location problems in humanitarian supply chains
22: Location problems under disaster events
23: Location problems in healthcare
25: Districting problems
26: Facility location in the public sector

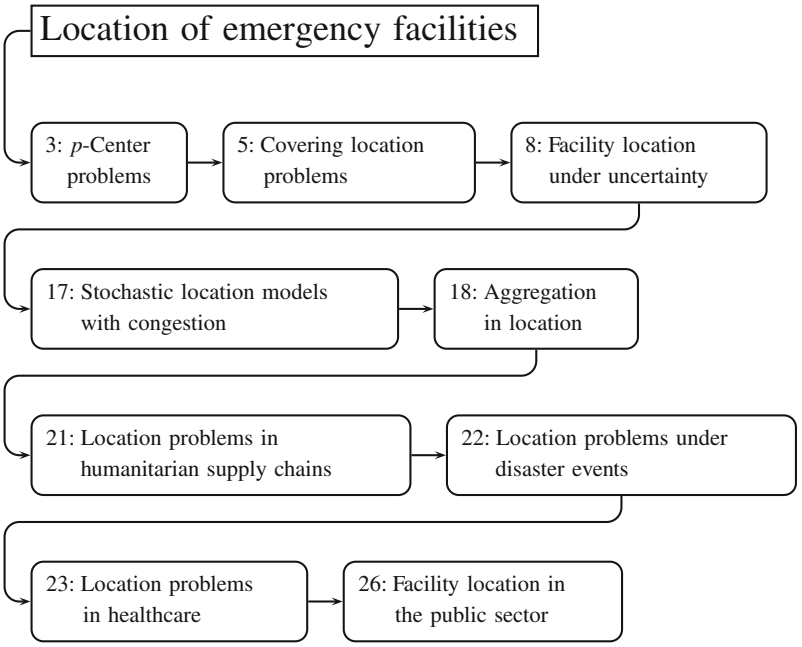


This book can also be used to build specialized courses in specific areas. Below, we provide examples in four areas: facility location and supply chain management, location of undesirable facilities, location of emergency facilities, and location in transportation systems.


Facility location and supply chain management
4: Fixed-charge facility location problems
8: Facility location under uncertainty
11: Multi-period facility location
12: Hub location problems
13: Hierarchical facility location problems
15: Location-routing and location-arc routing
16: Location logistics in supply chain management
20: Green location problems
21: Location problems in humanitarian supply chains
26: Facility location in the public sector


Location of undesirable facilities
2: p-Median problems
4: Fixed-charge facility location problems
6: Anti-covering problems
9: Location problems with multiple criteria
15: Location-routing and location-arc routing
18: Aggregation in location
20: Green location problems


Location problems in transportation systems
2: p-Median problems
4: Fixed-charge facility location problems
7: Locating dimensional facilities in a continuous space
12: Hub location problems
13: Hierarchical facility location problems
15: Location-routing and location-arc routing
16: Location logistics in supply chain management
20: Green location problems
21: Location problems in humanitarian supply chains
24: The design of rapid transit networks
26: Facility location in the public sector



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

Benko D, Coroian D (2018) A new angle on the Fermat-Torricelli point. Coll Math J 49:195–199

Block D, DuPuis EM (2001) Making the country work for the city. Am J Econ Soc 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
Drezner Z, Salhi S (2017) Incorporating neighborhood reduction for the solution of the planar p-median problem. Ann Oper Res 258:639–654
Drezner Z, Brimberg J, Mladenović N, Salhi S (2014) New heuristic algorithms for solving the planar p-median problem. Comput Oper Res 62:296–304
Drezner Z, Brimberg J, Mladenović N, Salhi S (2015) Solving the planar p-median problem by variable neighborhood and concentric searches. J Global Optim 63:501–514
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
Goodman JE, O'Rourke J, Tóth CD (eds) (2017) Handbook of discrete and computational geometry, 3rd edn. CRC Press, Boca Raton
Görner S, Kanzow C. (2016) On Newton's method for the fermat-weber location problem. J Optim Theory Appl 170:107–118
Hale TS, Moberg CR (2003) Location science research: a review. Ann Oper Res 123:21–35
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
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 (Bewley A, Trans., 1900). Lawrence Asylum Press, Madras
Lösch A (1944) The economics of location (Woglom WH, Trans., 1954). Yale University Press, New Haven
Manne AS (1964) Plant location under economies of scale-decentralization and computation. Manag. Sci 11:213–235
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. In: 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 (Wartenberg CM, Trans., 1966). Pergamon Press, Oxford
Weber A (1909) Theory of the location of industries (Friedrich CJ, Trans., 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
p-Median Problems

Alfredo Marín and Mercedes Pelegrín

Abstract One of the basic problems in the field of discrete location is the *p*-median problem. In this chapter we present and analyze several versions of the problem, but we can roughly define it as the choice of p facilities, among a set of n candidates, that minimize the cost of supplying a finite set of users. The p chosen facilities are usually called *medians*. Since the nature of the problem is combinatorial, integer programming is the common framework in which the problem is studied. Hence different formulations and their polyhedral properties constitute the kernel of this chapter. The study of the problem on a graph and heuristic procedures are treated in separate sections. Necessarily and unfortunately, we have to overlook many important references and results in the literature in the interest of legibility. Extensions of the problem, also of great interest, are covered in subsequent chapters and therefore are also ignored here. A companion problem of unquestionable importance, the Simple Plant Location Problem, is one of the main subjects of Chap. 4. Consequently, we have paid only little attention to it in our discussion.

2.1 Introduction

Discrete location problems consist of choosing a subset of locations, among a finite set of candidates, in which to establish facilities and then using these to satisfy the demand of a finite set of users. The choice of the locations must be made to minimize the sum of the fixed facility costs and of the cost of supplying the demand from the facilities.

Within this general framework, various problems can be identified as discrete location problems, most of which are studied and analyzed in this book. In this chapter we deal with a problem in the family of *median* problems. This term, in

A. Marín (✉) · M. Pelegrín
Universidad de Murcia, Facultad de Matemáticas, Campus de Espinardo, Murcia, Spain
e-mail: amarin@um.es; mariamercedes.pelegrin@um.es

G. Laporte et al. (eds.), *Location Science*,
https://doi.org/10.1007/978-3-030-32177-2_2

contrast with others like *center* and *equity*, refers to the definition of the cost to be minimized. When we speak about *median* (or *minisum*) problems we mean that the objective to be minimized depends in equal measure on the costs associated with each of the users.

The letter p in the term *p-median* refers to the number of locations to be chosen among the candidates, which is fixed beforehand. In other words, in the p-median problem a fixed number of p locations, usually called medians, must be chosen from the set of candidate facilities. Alternatively, it can be considered that p is the maximum number of locations that can be chosen. The cost to be minimized is calculated as the sum of the *allocation costs* of users to the medians. Let then $I = \{1, \ldots, m\}$ be the set of potential facilities and $J = \{1, \ldots, n\}$ the set of users to be supplied. The unit costs of supplying users from candidate facilities are arranged in a matrix $C = (c_{ij})$. We assume that *supplying costs* satisfy $c_{ij} \geq 0 \; \forall i \in I, j \in J$. The demand of a user $j \in J$ is denoted with $d_j > 0$; then, the allocation cost of j to a median $i \in I$ is given by $d_j c_{ij}$. In order to obtain the lowest overall cost, each user will be assigned to the median with minimum allocation cost.

Now we can formally define the p-median problem as follows. Suppose a matrix $C = (c_{ij})$ with non-negative entries, m rows denoted by $I = \{1, \ldots, m\}$ and called candidates facilities, and n columns denoted by $J = \{1, \ldots, n\}$ and called users. Given an n-dimensional vector (d_j) with positive entries and given $p \in \mathbb{Z}$, $1 \leq p \leq m - 1$, choose a subset $P \subseteq I$ of p rows of C in such a way that the *total cost* defined by $\sum_{j \in J} \min_{i \in P}\{d_j c_{ij}\}$ is minimized.

Figure 2.1 shows several examples of optimal solutions to p-median problems. Here $I = J$ is given by the same set of $n = 30$ points on the plane. Costs c_{ij} are given by the Euclidean distances between points and demands are assumed to be equal to one. In Fig. 2.1a we have taken $p = 2$ and drawn the best choice of 2 facilities (represented with squares) and the allocation of the 30 points to the corresponding closest facility. Different optimal solutions for $p = 3$, 4 and 5 are given also in Fig. 2.1b, c, and d, respectively.

Note that the kernel of the problem is the choice of the p facilities among the m candidates (a purely combinatorial subject, with $\binom{m}{p}$ possible solutions). Customers allocation to the facilities is trivially carried out by choosing, for each user $j \in J$, the facility in P with minimum allocation cost.

The p-median problem is strongly related with a problem that will be studied in Chap. 4, the Simple Plant Location Problem (SPLP)—also called Uncapacitated Facility Location Problem. In the SPLP, the number of facilities is not fixed *a priori*. Instead, a cost associated to each of the candidates is given, usually represented by $f_i \geq 0 \; \forall i \in I$. Then, given $C = (c_{ij})$ with non-negative entries, (d_j) with positive entries, and given the vector of non-negative costs $f = (f_i)$, SPLP aims to choose a subset $P \subseteq I$ of rows of C in such a way that $\sum_{i \in P} f_i + \sum_{j \in J} \min_{i \in P}\{d_j c_{ij}\}$ is minimized. SPLP is also a *minisum* problem, with a trade-off between costs associated to the facilities and allocation costs.

Despite its apparent simplicity, the p-median problem is NP-hard (Kariv and Hakimi 1979). Its origins can be traced back to Hakimi (1964, 1965), where the problem was defined on a graph, and ReVelle and Swain (1970), where an

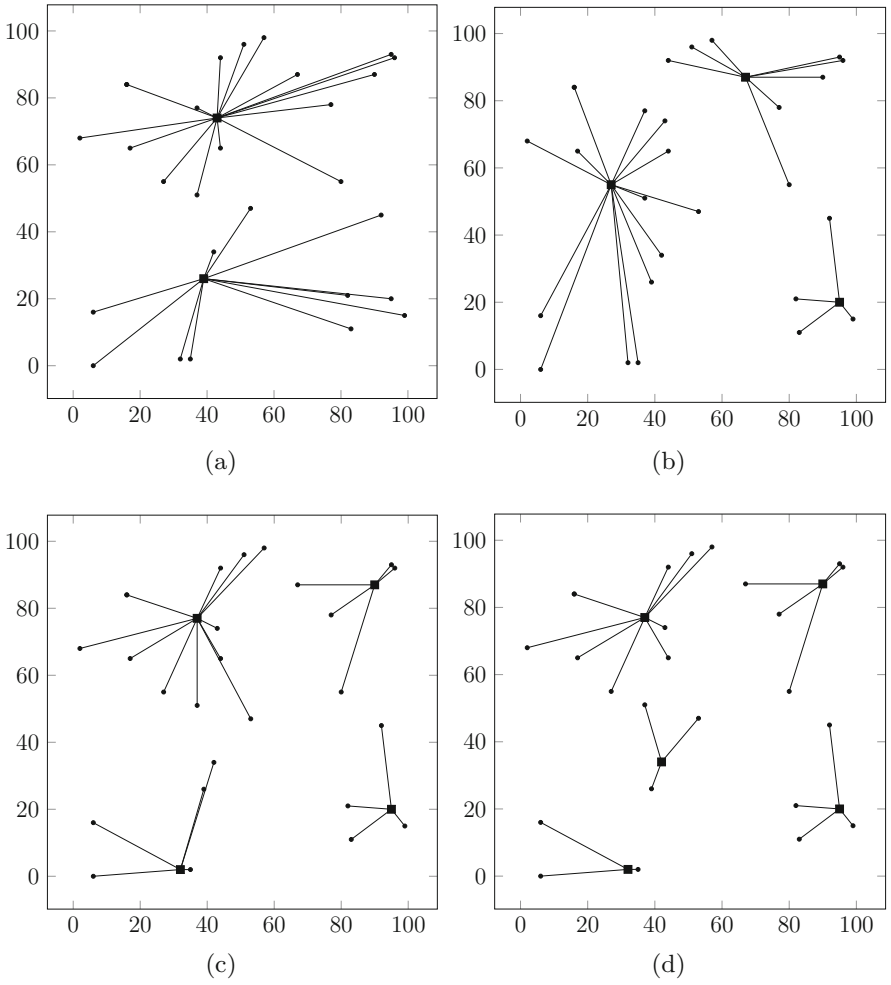


Fig. 2.1 Optimal solutions to the same instance of the p-median problem for different values of p. (**a**) $p = 2$. (**b**) $p = 3$. (**c**) $p = 4$. (**d**) $p = 5$

integer linear programming (ILP) formulation was proposed, inspired in Balinski (1965). Other related seminal papers are Hua et al. (1962), Kuehn and Hamburger (1963) and Manne (1964). Given its combinatorial nature, (mixed) integer linear programming (Nemhauser and Wolsey 1988; Wolsey 1998) has usually been the approach used to formulate and optimally solve the problem. The literature on the p-median problem is vast and it is not our aim to give an exhaustive list of papers. We focus our attention on recent results and suggest consulting Mirchandani (1990) and Reese (2006) as additional information sources.

We have organized the rest of the chapter as follows. In Sect. 2.2 several non-immediate applications, that show a wide range of possibilities of use, are presented.

In Sect. 2.3 the first integer linear programming formulations of the problem are introduced and analyzed. Section 2.4 deals with some of the most interesting available solution methods. Valid inequalities and facets for the polyhedra defined by the linear relaxations of different formulations are described in Sect. 2.5. We have included in a separate Sect. 2.6 the formulations and polyhedral results that arise when the p-median problem is solved on a (possibly non-complete) directed graph. Since solving large instances of the p-median problem is a difficult task, the literature on heuristic approaches is vast, and we try to give an idea of this vastness in Sect. 2.7, before closing the chapter with some final considerations.

2.2 Applications

In this section we present some applications of the p-median model taken from the literature. To emphasize its wide range of possibilities, we have selected applications outside the field of location of warehouses, plants, shelters or other kind of facilities, which is the natural interpretation of our problem.

Clustering was one of the first applications of the p-median problem. In the paper by Vinod (1969) it is said that *a large number of objects, persons, variables, symbols, etc. have been often to be grouped into a smaller number of mutually exclusive groups so that members within a group are similar to each other in some sense*. There is a limited number of groups, each of them having a distinguished member called centroid. The fitness of the partition depends on the average similarity of each object with the centroid of its group. The similarity between two pairs can be calculated from the input data and would correspond with costs ($d_j c_{ij}$) in our problem. The number of groups or clusters would be p and the centroids would be our medians.

Another application of the p-median problem, as presented in Vigneron et al. (2000), is the optimal placement of cache proxies in a computer network (see also Li et al. 1998). Nodes in a rooted tree network request a service that follows the path from the node to the root. When a proxy, located at a node of the tree, is found along this path, it satisfies the request. The location of p proxies in the nodes of the network in such a way that the sum of the distances from the nodes to the closest proxy in the corresponding path is minimized can be seen as a p-median problem. Vigneron et al. (2000) developed an algorithm to solve it on this special tree network topology.

We also include in this review of applications the so-called *Optimal Diversity Management Problem* (see Briant and Naddef 2004). Assume that a factory will manufacture a product that can, to some extent, be customized. For example, a car with t different improvements to be chosen or not by the users. The car becomes better and more expensive with each of these improvements, and then the users will not complain if they receive a car with more extras than required, at the same price. The factory cannot produce the 2^t different vehicles, so they decide to produce only p of the combinations and to deliver to each user the car with minimum cost among

those that include all the extras the user asked for. In this p-median problem, p is the number of different versions of the product that the factory can produce and $I = J$ is the set of all possible combinations of extras. Medians are the versions of the product that will be finally produced, and a combination of extras will be assigned to the median that will replace it when serving user requests. Replacing user request j by the version of the product i has a cost $d_j c_{ij}$.

A similar application is to determine p times for public vehicle departures on a temporal line, aiming at maximizing the total satisfaction of users. This served as the base for addressing the *Transit Network Timetabling and Scheduling Problem* in Mesa et al. (2014). In a public transit line, each vehicle performs a number of line runs or expeditions that have to be located in time. Users of the transit corridor have to be allocated to the line run that better fits their preferences, while fulfilling some capacity requirements. The formulation in Mesa et al. (2014) is a more complex version of the classical p-median that includes additional constraints.

Finally, in Goldengorin et al. (2012) (see also AlBdaiwi et al. 2011) the *cell formation* problem is established and studied as a p-median problem. A set of machines and their dissimilarities $d_j c_{ij}$ are given. It can be considered, for example, that when two machines process almost the same set of parts, there is a small dissimilarity between them (and can take part or the same cell). The problem is then to find p machines that are best representatives of p manufacturing cells, that is to say, the sum over the cells of the dissimilarities between these representatives and all other machines belonging to the same the cell has to be minimum. The problem can be considered as a special p-median problem on a graph, as defined in Sect. 2.6 below.

2.3 Integer Programming Formulations for the p-Median Problem

The *classical* ILP formulation for the p-median problem is

$$\text{(F1) minimize} \quad \sum_{i \in I} \sum_{j \in J} d_j c_{ij} x_{ij} \tag{2.1}$$

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

$$x_{ij} \leq y_i \quad \forall i \in I, j \in J \tag{2.3}$$

$$\sum_{i \in I} y_i = p \tag{2.4}$$

$$x_{ij} \in \{0, 1\} \quad \forall i \in I, j \in J \tag{2.5}$$

$$y_i \in \{0, 1\} \quad \forall i \in I. \tag{2.6}$$

Two sets of binary variables are used. On the one hand,

$$y_i = \begin{cases} 1 & \text{if candidate location } i \text{ is chosen as a median,} \\ 0 & \text{otherwise,} \end{cases} \quad \forall i \in I.$$

These variables are sometimes called *location variables*. Constraint (2.4) ensures that p candidate locations are chosen as facilities. Note that $y_i = 1$ when $i \in P$. On the other hand,

$$x_{ij} = \begin{cases} 1 & \text{if user } j \text{ is supplied from candidate facility } i, \\ 0 & \text{otherwise,} \end{cases} \quad \forall i \in I, j \in J.$$

The variables in this second set are sometimes called *allocation variables*. Constraints (2.2) guarantee that each user $j \in J$ is allocated to (supplied from) some candidate location $i \in I$. And constraints (2.3) prohibit allocations to candidate locations that were not chosen as facilities: when $y_i = 0$ (i.e., $i \notin P$), $x_{ij} = 0$ $\forall j \in J$, i.e., no user can be assigned to the location.

Allocation variables also serve to select the individual allocation costs that the solution entails and that are used to compute the total cost in linear combination (2.1).

Formulation (F1) contains $nm + m$ binary variables and $n + nm + 1$ linear constraints. A reduced formulation can be produced by replacing the set of nm constraints (2.3) by a set with only m constraints in the form

$$\sum_{j \in J} x_{ij} \leq n y_i \quad \forall i \in I. \tag{2.7}$$

Note that the effect of (2.7) when $y_i = 0$ is the same of (2.3), fixing to zero x_{ij} for all $j \in J$. In the case $y_i = 1$, the sum of n binary variables will be upperly bounded by n, thus producing no effect. We call (F2) formulation (F1) where constraints (2.3) have been replaced by (2.7).

Although formulation (F2) is more compact than formulation (F1), it has obvious disadvantages when a branch-and-bound procedure is used to solve the p-median problem, since summing up (2.3) for all $j \in J$, constraints (2.7) directly follow. This means that the polytope defined by the constraints of (F1) after relaxing the integrity of the variables, is included in the polytope analogously defined by the constraints of (F2). The consequence is that the lower bounds produced by (F1) will be better than those produced by (F2).

Several ways of reducing the size of (F1) without loss of quality in the formulation have been explored. First, it can be observed (see e.g. Church 2003) that a user will never be supplied from a facility if there are at least $m - p + 1$ candidates with strictly less associated supplying cost. To formalize this, we sort, for each user $j \in J$, the corresponding column in the cost matrix C to obtain $\hat{c}_{1j} \leq \hat{c}_{2j} \leq \cdots \leq \hat{c}_{mj}$. Then, some x-variables can be fixed to zero: $x_{ij} := 0$

$\forall i \in I : c_{ij} > \hat{c}_{m-p+1,j}$. Another possibility, see Church (2003), is to identify and match equivalent x-variables in the formulation. Consider two users $j_1 < j_2 \in J$, a candidate $i \in I$ and a scenario where $\Omega := \{\ell \in I : c_{\ell j_1} < c_{ij_1}\} = \{\ell \in I : c_{\ell j_2} < c_{ij_2}\}$. If, in an optimal solution, $x_{ij_1} = 1$, it follows that no candidate in Ω has been chosen as a facility, but i has been (since j_1 was assigned to i). Then, one facility to which j_2 is allocated with minimum cost is i as well. Consequently, $x_{ij_2} = 1$ is an optimal choice. On the other hand, $x_{ij_1} = 0$ means that either a candidate in Ω has been chosen as median or there are no medians in Ω and neither is i a median. In both cases, $x_{ij_2} = 0$. The conclusion is that x_{ij_1} and x_{ij_2} can be identified, and thus the size of the formulation can be reduced by replacing all x_{ij_2} with x_{ij_1}.

Following the same reasoning as in Cho et al. (1983a), we can handle formulation (F1) to rewrite constraints (2.2) and (2.3) in a different way. Note that, since (2.2) are equalities, the sums $\sum_{i\in I} x_{ij}\ \forall j \in J$ will be constant in any feasible solution to (F1). Hence using a large enough number, M, the alternative objective

$$\sum_{i\in I}\sum_{j\in J} d_j c_{ij} x_{ij} - \sum_{i\in I}\sum_{j\in J} M x_{ij} = \sum_{i\in I}\sum_{j\in J} \tilde{c}_{ij} x_{ij}$$

where $\tilde{c}_{ij} := d_j c_{ij} - M < 0\ \forall i \in I, j \in J$ can be utilized. The advantage of this function is that, since the coefficients are negative and we are minimizing, the x-variables will take value one in an optimal solution if they are not restricted by the constraints of the formulation. This means that constraints (2.2) can be relaxed to

$$\sum_{i\in I} x_{ij} \leq 1 \qquad \forall j \in J. \tag{2.8}$$

Consider now a different set of binary variables

$$y'_i = \begin{cases} 1 & \text{if candidate location } i \text{ is } \textbf{not} \text{ chosen as a facility,} \\ 0 & \text{otherwise,} \end{cases} \qquad \forall i \in I,$$

that is to say, $y'_i := 1 - y_i\ \forall i \in I$. Using this new set of variables, constraints (2.3) can be rewritten as

$$x_{ij} + y'_i \leq 1 \qquad \forall i \in I, j \in J. \tag{2.9}$$

Both sets of constraints, (2.8) and (2.9), are defined as sums of binary variables upperly bounded by 1. These *set packing* constraints can be analyzed, see Cánovas et al. (2000, 2002, 2003), Cho et al. (1983a,b), and Cornuéjols and Thizy (1982), to produce a tighter formulation, using the so-called *intersection* (or *conflict*) graph, where each node is associated with a variable, and nodes are neighbors if they share at least one constraint. Since this analysis is the same as that carried out for the SPLP, we refer the reader to Chap. 4 for a detailed analysis. The reformulation of

(F1) by means of (2.8) and (2.9) still contains constraint (2.4), which enables us to perform the polyhedral analysis of the formulation in a different way, see Sect. 2.5.

A different relaxation of (F1) can be carried out, that of the integrity of the x-variables. Constraints (2.5) can be replaced by

$$x_{ij} \geq 0 \quad \forall i \in I, j \in J. \tag{2.10}$$

To see this, observe that (2.2) and (2.10) imply $x_{ij} \in [0, 1]\ \forall i \in I, j \in J$. Now, consider a set $P \subseteq I$ of p facilities and the sets $A_j := \{i \in P : c_{ij} = \min_{\ell \in P} c_{\ell j}\}$. It is obvious that in any optimal solution where P is the set of chosen facilities, $\sum_{i \in A_j} x_{ij} = 1$ holds for all $j \in J$. Since all variables in the last sum have the same cost, an equivalent integer solution can be trivially obtained by fixing one of them to one and the rest to zero. After relaxing (2.5)–(2.10), the meaning of the x-variables can be re-established as $x_{ij} =$ *fraction of the demand of user* j *that is supplied by facility* i.

Consider now the version of the problem where $I = J$ and $c_{ii} = 0\ \forall i \in I$. This case has some special characteristics that allow to reformulate the problem. Whenever $y_i = 1$, the minimum possible allocation cost for point i will be 0, obtained by allocating i to itself. Then $y_i = 1 \Rightarrow x_{ii} = 1$. Since $y_i = 0 \Rightarrow x_{ii} = 0$, both variables can be identified, and y_i can be replaced by x_{ii} in the formulation. The resulting reduced formulation is given by

$$\text{(F3) minimize} \sum_{i \in I} \sum_{\substack{j \in I: \\ i \neq j}} d_j c_{ij} x_{ij}$$

$$\text{subject to} \qquad (2.2)$$

$$x_{ij} \leq x_{ii} \quad \forall i, j \in I : i \neq j \tag{2.11}$$

$$\sum_{i \in I} x_{ii} = p \tag{2.12}$$

$$x_{ij} \in \{0, 1\} \quad \forall i, j \in I : i \neq j \tag{2.13}$$

$$x_{ii} \in \{0, 1\} \quad \forall i \in I. \tag{2.14}$$

Again, constraints (2.13) can be relaxed to $x_{ij} \geq 0\ \forall i, j \in I: i \neq j$.

Under the given conditions, constraints (2.7) can be slightly improved. Note that the existence of p users that are going to be self-allocated guarantees that no more than $n - p + 1$ users will be allocated to the same facility. Hence the constant in the right hand side of (2.7) can be modified to yield the tighter constraints

$$\sum_{\substack{j \in J: \\ j \neq i}} x_{ij} \leq (n - p) x_{ii} \quad \forall i \in I. \tag{2.15}$$

In what follows we still assume $c_{ii} = 0\ \forall i \in I$ and $c_{ij} \geq 0\ \forall i, j \in I: i \neq j$ to produce a formulation based on a completely different set of variables. The ideas we are going to present come from Cornuéjols et al. (1980), where they were applied to the SPLP. A preprocessing of the data is required before proceeding. It is necessary, for each $j \in J$, to sort the entries of the j-th column of the cost matrix C, removing the multiplicities: $0 = \bar{c}_{1j} < \bar{c}_{2j} < \cdots < \bar{c}_{G_j j} = \max_{i \in I} c_{ij}$. Since we do not know *a priori* how many different supplying costs there are in column j of C, we use G_j to denote this number. A new set of binary variables, sometimes called *cumulative variables*, is defined as

$$z_{kj} = \begin{cases} 1 \text{ if the supplying cost of user } j \text{ is at least } \bar{c}_{kj} \\ \quad \text{(no matter which facility it is allocated to)}, \quad \forall j \in J, 2 \leq k \leq G_j. \\ 0 \text{ otherwise,} \end{cases}$$

Note that the variables z_{1j} have not been used, since by definition $z_{1j} = 1$ if the supplying cost of user j is at least $\bar{c}_{1j} = 0$, and this condition is always satisfied. Initially we will also use variables $y_i, \forall i \in I$, to keep track of the chosen facilities. Then consider a new formulation for the p-median problem given by

$$\text{(F4) minimize} \sum_{j \in J} \sum_{k=2}^{G_j} d_j (\bar{c}_{kj} - \bar{c}_{k-1,j}) z_{kj} \tag{2.16}$$

$$\text{subject to} \quad (2.4),\ (2.6)$$

$$z_{kj} + \sum_{\substack{i \in I: \\ c_{ij} < \bar{c}_{kj}}} y_i \geq 1 \qquad \forall j \in J, 2 \leq k \leq G_j \tag{2.17}$$

$$z_{kj} \in \{0, 1\} \qquad \forall j \in J, 2 \leq k \leq G_j. \tag{2.18}$$

In formulation (F4) we keep constraints (2.4) and (2.6) to account for the number of facilities. The difference between (F4) and the previously introduced formulations is that in (F4) there is no information in the variables about the allocation of users to facilities, but there is about the smallest allocation costs of the users when only chosen facilities are considered. Let us analyze constraints (2.17). The term $\sum_{i \in I:\ c_{ij} < \bar{c}_{kj}} y_i$ takes value zero only when no candidate with supplying cost less than $\bar{c}_{kj}$ (the k-th supplying cost for user j) has been selected as a facility. It is clear, then, that z_{kj}, as defined, must take value 1. Since the coefficients in the objective function (2.16) are strictly positive, in an optimal solution all variables will take value 0 unless the corresponding constraint (2.17) force them to take value 1. For this reason, z-variables can be relaxed to be positive continuous variables and constraints (2.18) can be simply removed.

For a given user $j \in J$, the sets of candidates inside a *radius* $\bar{c}_{kj}$, $B_k := \{i \in I : c_{ij} < \bar{c}_{kj}\}$, satisfy the strict inclusion relations $B_2 \subsetneq B_3 \subsetneq \cdots \subsetneq B_{G_j}$. This implies that, in any optimal solution, $z_{2j} \geq z_{3j} \geq \cdots \geq z_{G_j j}$, that is to say, the

appearance of vector $z_{\cdot j}$ will be $(1, \ldots, 1, 0, \ldots, 0)$. Assume the last 1 corresponds with variable z_{aj}. Then, in the objective function (2.16) the sum $\sum_{k=2}^{G_j} d_j(\bar{c}_{kj} - \bar{c}_{k-1,j})z_{kj}$ will be $\sum_{k=2}^{a} d_j(\bar{c}_{kj} - \bar{c}_{k-1,j})$. Taking into account that $\bar{c}_{1j} = 0$, the value of this telescopic sum will be $d_j\bar{c}_{aj}$, that is to say, the cost of allocating j to median a, as wished.

In Fig. 2.2 we see, using the same example as in Fig. 2.1d (where $I = J$, $d_j = 1$ $\forall j \in J$, and supplying costs are given by Euclidean distances between points), the effect of constraints (2.17) on user $j = 1$ assuming that the facilities of the optimal solution are given. Constraint (2.17), with $k = 2$, reads $z_{21} + y_1 \geq 1$. Since 1 is not a median, it follows that $z_{21} = 1$. Taking now $k = 3$, it reads $z_{31} + y_1 + y_2 \geq 1$, implying $z_{31} = 1$. Similarly, $z_{41} = z_{51} = 1$. Then, for $k = 6$, $z_{61} + y_1 + y_2 + y_3 + y_4 + y_5 \geq 1$ is satisfied since $y_5 = 1$. Due to the objective function, $z_{61} = z_{71} = \cdots = 0$, and that the cost of allocating point 1 to point 5 will be $(10.77-0)\cdot 1+(15.65-10.77)\cdot 1+(16.49-15.65)\cdot 1+(17.72-16.49)\cdot 1 = 16.49$, the distance between points 1 and 5.

A reduction in the size of (F4) can be made noting that constraints (2.17) when $k = 2$ read $z_{2j} + y_j \geq 1$ and these constraints are always satisfied as equalities by an optimal solution. Then y_i can be replaced with $1 - z_{2i}$ $\forall i \in I$.

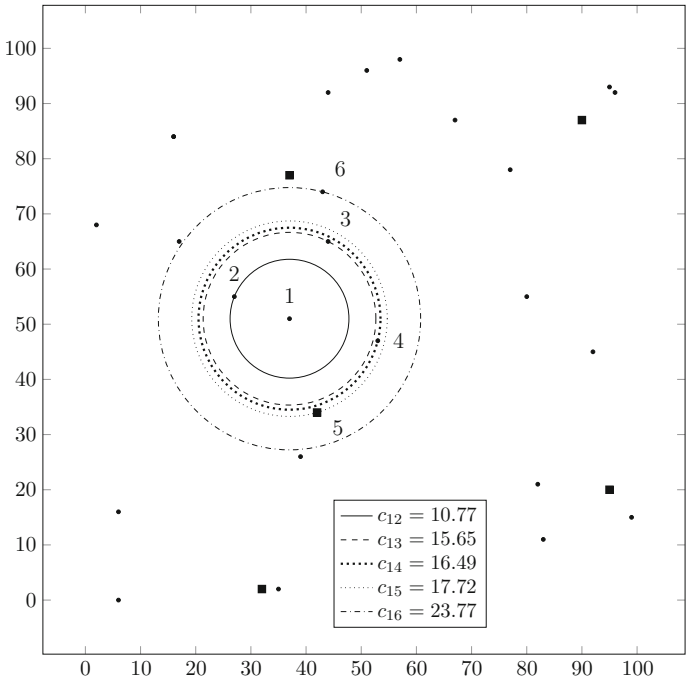


Fig. 2.2 Graphical representation of the role of the z-variables in formulation (F4) on the same example as in Fig. 2.1d

Regarding the size of (F4), observe that for each user $j \in J$, the number of z-variables in (F4) will be the number of different costs in the j-th column of C, minus one. Therefore, the total number of z-variables in the formulation will be in the set $\{0, \ldots, nm\}$. For each z-variable there is one constraint in family (2.17), thus the number of linear constraints will be in $\{1, \ldots, nm + 1\}$. In the worst case, when all costs in each column of C are distinct, the size of (F4) will be exactly the same as the size of (F1).

Although the size of (F4) can be smaller than the size of (F1), Cornuéjols et al. (1980) proved that both linear relaxations yield the same lower bound on the optimal value of the problem. There exist many works where formulations (F1)–(F3) have been used. However, references containing formulation (F4) are scarce, and almost limited to the study of the companion problem SPLP: Kolen (1983) used a version of formulation (F4) to solve the SPLP in polynomial time on a tree; Simão and Thizy (1989) studied the linear relaxation of a modification of (F4); (F4) for SPLP was also considered in Cornuéjols et al. (1990) and Kolen and Tamir (1990). Finally, Xu and Lowe (1993) compared the work of Simão and Thizy (1989) with a previous method in the literature to solve the SPLP.

2.4 Optimal Solution Procedures

Several exact algorithms for the p-median problem are available. We summarize some of them here, without intending to be exhaustive.

Galvão (1980) realized that solving the p-median problem within a branch-and-bound framework means solving many linear relaxations of subproblems of large size. He then devised a method to efficiently obtain good lower bounds instead of optimally solving the relaxed continuous subproblems. To this end, he considered formulation (F3), replaced the equality (2.2) by '$\geq$', relaxed constraints (2.13) and (2.14) and built the dual problem

$$
\begin{aligned}
\text{(F3D) maximize} \quad & p\sigma_{n+1} + \sum_{i \in I} \sigma_i \\
\text{subject to } & \sigma_i + \sigma_{n+1} - \sum_{\substack{j \in I: \\ j \neq i}} \pi_{ij} \leq 0 \quad \forall i \in I \\
& \sigma_j - \pi_{ij} \leq d_j c_{ij} \qquad \forall i, j \in I \\
& \pi_{ij} \leq 0 \qquad \forall i, j \in I : i \neq j \\
& \sigma_i \geq 0 \qquad \forall i \in I \\
& \sigma_{n+1} \leq 0.
\end{aligned}
$$

Table 2.1 A summary of the computational experience on exact solution methods up to date

Authors	Year	Computer	n	t (s)
Galvão	1980	Unknown	30	879
Church	2003	Sun Ultra Sparc 10	372	879
Avella et al.	2007	Compaq EVO W4000 PC Pentium IV 1.8 GHz, 1 GB RAM	5535	394
García et al.	2011	Intel CORE 2 CPU 6600 2.4 GHz, 3 GB RAM	85,900	66,000

The last three columns stand for the maximum size and time in seconds of the instances tested but do not necessarily correspond to one same instance

Noticing then that, in any optimal solution to (F3D),

$$\sigma_{n+1} \leq \min_{i \in I}\{-\sigma_i + \sum_{\substack{j \in I: \\ j \neq i}} \pi_{ij}\} \text{ and } \pi_{ij} = -\max\{0, \sigma_j - d_j c_{ij}\}\ \forall i, j \in I : i \neq j,$$

he designed a two-phase method to calculate good feasible solutions of (F3D) in an attempt to increase the objective value. In the first phase the value of σ_{n+1} was maximized and then the values of σ_i, $i \in I$, were maximized without modifying σ_{n+1}. Then he embedded this procedure, which produces good lower bounds in a short time, into the branch-and-bound algorithm and obtained good computational results. Table 2.1 gives an insight about the evolution of the sizes of the instances that could be solved with each exact method. Note that the best lower bound that can be produced with this approach is the one provided by the linear relaxation of (F3).

The use of formulations (F1) and (F3) with aggregated but weaker constraints (2.7) or (2.15), combined with the inclusion of (2.3) as valid inequalities, has served as an alternative strategy in several papers. As an example, in Church (2003) a subset of constraints (2.3), those corresponding to the candidates with minimum supplying cost with respect to each user, is initially incorporated in formulation (F3). The combination of this strategy and the matching of equivalent x-variables (see Sect. 2.3) also produced good computational results (see Table 2.1).

Beltrán et al. (2006) approached the p-median problem from a similar point of view. They initially considered formulation (F1) and the Lagrangian relaxation of constraints (2.2) and (2.4) by means of unrestricted multipliers v_j, $\forall j \in J$ and v_0, respectively. An overview on Lagrangian relaxation can be consulted in Guignard (2003). The advantage of relaxing equality constraints is that any optimal solution to a relaxed subproblem that also satisfies the relaxed constraints is an optimal solution of the primal problem. The disadvantage of relaxing all these constraints is that the optimal value of the dual problem is the same as the optimal value of the linear relaxation of the problem. The authors found first a good set of Lagrangian multipliers and used them as a starting point for a second problem relaxed in a Lagrangian fashion. In this case they added constraints $\sum_{i \in I} x_{ij} \leq 1$, $\forall j \in J$, to the relaxed subproblem, which becomes more difficult to solve but can yield better

lower bounds. The advantage of using the '$\leq$' version of the constraints is that all variables x_{ij} with non-negative coefficient $d_j c_{ij} + v_j$ in the relaxed subproblem can be fixed to zero. The subproblem is then easier to solve and can even be decomposed, since the non-removed variables could be grouped in subsets that do not relate each other. The final set of multipliers is then used as the starting point for a third and last relaxation obtained by adding one more constraint to the subproblem, namely $\sum_{i \in I} y_i \leq p$.

Avella et al. (2007) designed a branch-and-cut-and-price algorithm that was able to solve very large instances (see Table 2.1) of the p-median problem on a graph (see forthcoming formulation (F5)). Cuts were added based on new valid inequalities called $W - q$, *lifted odd hole* and *cycle inequalities*. Details of them are given in Sect. 2.6. Pricing was carried out by solving a master problem to optimality and using dual variables to price out the variables of the initial problem that were not considered in the master, adding new variables if necessary. The novelty of the approach was that constraints (2.20) were also relaxed and incorporated to the master problem when the corresponding column was. The authors also developed criteria to fix the values of some y-variables to zero when lower bounds calculated fixing y_i to one were greater than previously known upper bounds.

Finally, we summarize the solution method based on (F4) developed in García et al. (2011). Recall that, in (F4), given an optimal solution (y^*, z^*) and a fixed user $j \in J$, $z^*_{\cdot j}$ will have the shape $(1, \ldots, 1, 0, \ldots, 0)$. We have also a similar property of any optimal solution of the linear relaxation of (F4), $(\bar{y}, \bar{z})$: for all $j \in J$, $\bar{z}_{2j} \geq \bar{z}_{3j} \geq \ldots \geq \bar{z}_{G_j j}$. Therefore, if $\bar{z}_{aj} = 0$ for some a, then $\bar{z}_{kj} = 0$ for all $k > a$. Suppose we could know this optimal solution $(\bar{y}, \bar{z})$ beforehand. Since each z-variable only appears in one constraint, and the z-variables taking value zero have not been forced by the optimal values of the y-variables to take value 1, we could remove all variables and constraints associated with the null $\bar{z}$-values and the linear relaxation of this reduced formulation would provide us with the same optimal solution. Conversely, let us remove variables $z_{a+1,j}, \ldots, z_{G_j j}$, for a given $j \in J$, from the linear relaxation of (F4). If $\bar{z}_{aj} = 0$ in the optimal solution of the relaxed problem, this is done. Otherwise, if $\bar{z}_{aj} > 0$, it is possible that some of the removed variables had taken a positive value in the optimal solution. In this case, a has been wrongly selected and a larger value for it must be considered. The method proposed in García et al. (2011) then considered a first formulation with a very small set of z-variables and constraints, and added more variables and their corresponding constraints when needed. At every node of the branching tree, the final formulation of the predecessor node was used. The result was an exact branch-and-cut-and-price method that allowed the authors to solve the p-median problem with a drastically reduced formulation that required much fewer constraints and variables than formulations (F1)–(F4). This method performed extremely well on very large instances (see Table 2.1) with large values of p. Note that the larger the value of p, the smaller the allocation costs associated to the users and, consequently, the smaller the number of z-variables (and constraints) added to the initial reduced formulation.

2.5 Polyhedral Properties

In this section we present polyhedral properties of the formulations (F1) and (F3) or their modifications. It is worth mentioning that since the polyhedron of these p-median formulations is obtained from the polyhedron of the SPLP by adding only one constraint, all valid inequalities for the corresponding formulations of the SPLP are also valid for the p-median problem. Nevertheless, they do not usually define facets. In this section we focus on models that produce valid inequalities or facets for the p-median problem that are not necessarily valid for the SPLP. Basic knowledge on polyhedral theory is assumed in this section (we refer the interested reader to Nemhauser and Wolsey 1988)

A seminal paper in this field is de Farias (2001). The author considered a modified version of formulation (F1), with equalities (2.2) and (2.4) relaxed to inequalities of type '$\leq$'. He proved that the polyhedron so defined is fully dimensional, and found a family of facets by taking a subset J' of J with cardinality at least $p+1$ and disjoint nonempty subsets of I named I_j, $j \in J'$, with $\cup_{j\in J'} I_j \subsetneq I$. He showed that the constraints

$$\sum_{j\in J'}\sum_{i\in I_j} x_{ij} + \sum_{i\notin \bigcup_{\ell\in J'} I_\ell}\sum_{j\in J'} x_{ij} \leq p + (|J'| - p) \sum_{i\notin \bigcup_{\ell\in J'} I_\ell} y_i$$

are valid for the given formulation and define facets. We now present an example taken from de Farias (2001) with $n = 3$, $m = 4$, $p = 2$, $J' = J$, $I_j = \{j\}$, $j = 1, 2, 3$:

$$x_{11} + x_{22} + x_{33} + x_{41} + x_{42} + x_{43} \leq 2 + y_4.$$

Note that $y_4 = 0$ implies $x_{41} + x_{42} + x_{43} = 0$ and then $x_{11} + x_{22} + x_{33} \leq 2$ is valid since $p = 2$. On the other hand, in the case $y_4 = 1$, the inequality becomes trivial.

Consider now de Vries et al. (2003). Among different results on the polyhedral structure of the p-median problem, the authors generate a family of valid inequalities for (F3) of the form

$$\sum_{i\in R\cup S} x_{ii} - \frac{1}{r-p}\sum_{i\in R}\sum_{\substack{j\in R:\\ i\neq j}} x_{ij} - \frac{1}{r-p+1}\sum_{i\in S}\sum_{j\in R} x_{ij} \leq p - 1, \tag{2.19}$$

where R is a subset of $I = J$ of cardinality $r \geq p$, and S is a subset of $I \setminus R$. For example, take $m \geq 4$, $p = 2$, $R = \{1, 2, 3\}$ and $S = \{4\}$. The facet in family (2.19) would be

$$2x_{11} + 2x_{22} + 2x_{33} + 2x_{44}$$
$$\leq 2x_{12} + 2x_{21} + 2x_{13} + 2x_{31} + 2x_{23} + 2x_{32} + x_{41} + x_{42} + x_{43} + 2.$$

Observe that when all medians belong to the set {1, 2, 3}. To illustrate, assume that the two medians are 1 and 2. Then $2x_{11}+2x_{22}+2x_{33}+2x_{44} = 4$ and the inequality becomes $1 \leq x_{13} + x_{23}$, and it obviously holds. A second possibility is that the two medians are 1 and 4. Then it follows that $2 \leq 2x_{12} + 2x_{13} + x_{42} + x_{43}$. Since 2 and 3 must be supplied from 1 or 4, it also holds. Finally, if $x_{11} + x_{22} + x_{33} + x_{44} \leq 1$, the inequality becomes trivial. In de Vries et al. (2003) it is proven that inequalities (2.19) define facets when $r > p$, $S \neq \emptyset$ and $S \cup R \neq I$.

In Zhao and Posner (2011), a generalization of the family of facets (2.19) is developed. Here, a partition of I given by the sets $T_1,\ldots,T_r$, S and Q, with $r > p$ and $T_i \neq \emptyset, i = 1, \ldots, r$, $Q \neq \emptyset$, is required. Defining $T = \cup_{i=1}^{r} T_i$, $R \subseteq T \cup Q$ of cardinality r such that $|R \cap T_i| \leq 1, i = 1, \ldots, r$ and a bijection τ of R in the set $\{1, \ldots, r\}$, the new family of valid inequalities for (F3) is given by

$$\sum_{i \in T \cup S} x_{ii} - \frac{1}{r-p} \sum_{j \in R} \sum_{i \in T \setminus T_{\tau(j)}} x_{ij} - \frac{1}{r-p+1} \sum_{i \in S} \sum_{j \in R} x_{ij} \leq p - 1.$$

These inequalities define facets when $2 \leq p < r$ and $|Q| = 1$ or $|(T \cup S) \setminus R| \geq 1$. The authors also devised a heuristic procedure to separate these inequalities.

Also observe that Cánovas et al. (2007) introduce *dominance constraints* in the shape of $x_{ij_1} \leq x_{ij_2}$ that can be incorporated to formulation (F3). These inequalities can be used whenever $\{\ell \in I : c_{\ell j_2} < c_{ij_2}\} \subseteq \{\ell \in I : c_{\ell j_1} < c_{ij_1}\}$. We present additional polyhedral material after introducing a new version of the problem, in the next section.

2.6 p-Median Problem on a Graph and Additional Polyhedral Results

Many authors consider and analyze a particular case of the p-median problem defined on a directed graph (V, A). The set of nodes, V, represents users and also candidate locations for facilities. The set of arcs A, is used to express the possible allocations of users to facilities. Self-allocation is implicitly assumed or, in other words, a node is either chosen as a median or it must be allocated to another node. Note that this is equivalent to fixing some variables x_{ij} to zero in formulation (F3): $x_{ij} = 0$ if $(i, j) \notin A$. The same effect can be achieved by taking c_{ij} large enough in the objective function of (F3). Nevertheless, knowing beforehand that some variables have been removed from the formulation has some advantages that several authors have exploited. We explicitly state the following formulation of the p-median problem on a directed graph (V, A):

$$\text{(F5) minimize} \quad \sum_{(i,j) \in A} d_j c_{ij} x_{ij}$$

$$\text{subject to } x_{ii} + \sum_{\substack{j \in V: \\ (j,i) \in A}} x_{ji} = 1 \ \forall i \in V$$

$$x_{ij} \leq x_{ii} \qquad \forall (i,j) \in A \tag{2.20}$$

$$\sum_{i \in V} x_{ii} = p$$

$$x_{ij} \in \{0,1\} \qquad \forall (i,j) \in A \tag{2.21}$$

$$x_{ii} \in \{0,1\} \qquad \forall i \in V.$$

A different version of this formulation is considered by Avella and Sassano (2001) who do not make use of the x_{ii} variables. Instead, they pay attention to the fact that $n - p$ nodes must be allocated by means of an arc (i.e., they are not self-allocated) and then each feasible solution will correspond to a set of $n - p$ arcs in A. They then propose the following formulation:

$$\text{(F6) minimize} \quad \sum_{(i,j) \in A} d_j c_{ij} x_{ij}$$

$$\text{subject to} \quad (2.21)$$

$$x_{ij} + \sum_{\substack{\ell \in V: \\ (\ell,i) \in A}} x_{\ell i} \leq 1 \ \forall (i,j) \in A \tag{2.22}$$

$$\sum_{(i,j) \in A} x_{ij} = n - p. \tag{2.23}$$

Avella and Sassano (2001) consider the case where A is a complete digraph and develop two families of inequalities. The first family, the so-called $W - 2$ inequalities, only makes use of constraints (2.22). They can then be used for the SPLP. The shape of these constraints is

$$\sum_{(i,j) \in A \cap [((W \times W) \setminus H) \cup (\bar{W} \times U)]} x_{ij} \leq |W| - 2, \tag{2.24}$$

where $W \subseteq V$ and $3 \leq |W| \leq n - p + 1$, H is a subset of arcs of A in $W \times W$ such that $\forall w \in W$ there is exactly one arc in H with origin in w, and U is the set of nodes of W that are not destinations of any arc of H. Inequalities (2.24) are facets whenever $|U| \leq \max\{1, |W|-3\}$. We present here the example used in Avella and Sassano (2001) to illustrate this family. Consider the complete directed graph of eight nodes and the subgraph given in Fig. 2.3a. Here $|W| = 6$, $H = \{(1,3), (3,1), (2,4), (4,2), (5,3), (6,4)\}$ and $U = \{5,6\}$. It produces the inequality in the family (2.24) in the shape of

$$x_{12} + x_{14} + x_{15} + x_{16} + x_{21} + x_{23} + x_{25} + x_{26} + x_{32} + x_{34} + x_{35} + x_{36} + x_{41} + x_{43} +$$

$$x_{45} + x_{46} + x_{51} + x_{52} + x_{54} + x_{56} + x_{61} + x_{62} + x_{63} + x_{65} + x_{75} + x_{76} + x_{85} + x_{86} \leq 4.$$

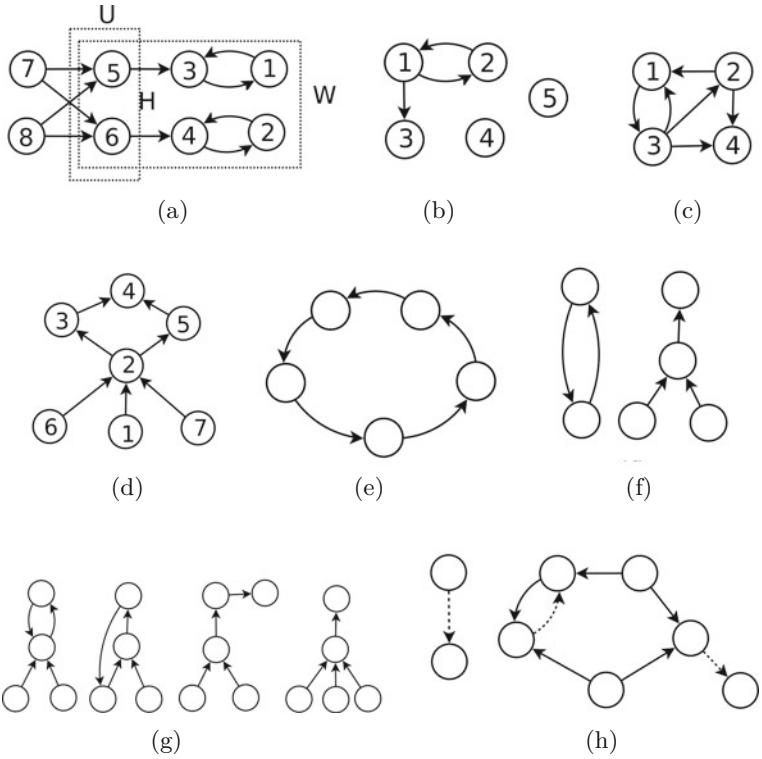


Fig. 2.3 Illustration of several inequalities and families of subgraphs. (**a**) $W-2$. (**b**) Cover. (**c**) $W-q$. (**d**) Odd hole inequalities. (**e**) $W-1$ and $|F|$ odd. (**f**) 2-cycle and Y-graph. (**g**) Graphs with Y-subgraphs. (**h**) Forbidden structure in Baïou and Barahona (2011)

Nodes 5 and 6 can be supplied or not from nodes not belonging to W. Take $x_{75} = x_{76} = 1$. Thus, the inequality becomes $x_{12}+x_{14}+x_{21}+x_{23}+x_{32}+x_{34}+x_{41}+x_{43} \leq 2$. Since no node in the set $\{1, 2, 3, 4\}$ can supply more than two other nodes in the set, it must be satisfied. Similar reasonings can be applied by taking other values of x_{75} and x_{76}.

The second family of inequalities in Avella and Sassano (2001), called *cover inequalities*, make use of constraint (2.23). They again consider A to be a complete digraph. Consider a set S of arcs and let $r(S)$ be the maximum number of arcs of S that can simultaneously take part in a solution for (F6). Let $F(S)$ be the collection of all subsets of A containing $r(S)$ arcs from S that form a solution for (F6). Choose at least one arc from each subset in $F(S)$ to create set $T(S)$. Then

$$\sum_{(i,j)\in S} x_{ij} - \sum_{(i,j)\in T(S)} x_{ij} \leq r(S) - 1$$

are valid inequalities for (F6). As an example, take the complete directed graph of five nodes, let S be the subset of arcs of Fig. 2.3b and $p = 2$. Then $r(S) = 2$ and $F(S) = \{\{(1,2),(1,3),(4,5)\},\{(1,2),(1,3),(5,4)\},\{(1,2),(1,3),(1,4)\},\{(1,2),(1,3),(1,5)\}\}$. Taking $T(S) = \{(4,5),(5,4),(1,4),(1,5)\}$, the inequality produced is $x_{12} + x_{21} + x_{13} \leq 1 + x_{45} + x_{54} + x_{14} + x_{15}$. In the case $x_{45} = x_{54} = x_{14} = x_{15} = 0$, all nodes other than 1 should be assigned to node 1, but in this case $p \neq 2$. Otherwise, the sum in the left hand side is bounded by 2, the value of $r(S)$.

Regarding (F5), valid inequalities and characterizations of the polyhedron in some particular cases have been obtained by several authors. We present the main results below.

In Avella et al. (2007), the so-called $W - q$ inequalities were derived. We show an example of such inequalities based on the graph of Fig. 2.3c. Let W be the set of nodes $\{1, 2, 3, 4\}$ and F the set of arcs $\{(2, 1), (3, 2), (1, 3), (2, 4), (3, 4)\}$. Note that arc $(3, 1)$ is not included in the set.

Consider the following valid inequalities:

$$\begin{array}{rr}
x_{21} & +x_{13} \leq 1, \\
\hline
x_{32} & +x_{21} \leq 1, \\
x_{32} & +x_{24} \leq 1, \\
\hline
x_{13} & +x_{32} \leq 1, \\
x_{13} & +x_{34} \leq 1, \\
\hline
x_{24} + x_{34} & \leq 1.
\end{array}$$

These valid inequalities are arranged in blocks and have been systematically built in the following way. Each block is devoted to one node $j \in W$. For each j, the sum of all variables corresponding to arcs of F that end in j is considered. Then, the sum is completed in several ways (one represented by each row of the block) by adding x_{jh} for all distinct h such that $(j, h) \in F$ (if any). In the example, this yields at most two inequalities for each block, since no more than two arcs of F leaves the same node. Note that this construction of the inequalities implies that every variable x_{ij} with $(i, j) \in F$ appears in two inequalities or in three when there are two arcs leaving node j. In order to complete those blocks that only have one inequality, we add a copy of $x_{24} + x_{34} \leq 1$ to the last block and $x_{21} \leq 1$ to the first one. Summing up the resulting set of eight inequalities, we obtain $3(x_{21} + x_{32} + x_{13} + x_{24} + x_{34}) \leq 8$. Dividing by 3 and rounding down the right-hand side, the following valid inequality in the family $W - 2$ is produced: $x_{21} + x_{32} + x_{13} + x_{24} + x_{34} \leq 2$. In the general case, consider a set of nodes $W \subseteq V$, an integer number $1 \leq q \leq |W| - 1$, and a set of arcs with both ends in W, $F \subset A \cap (W \times W)$, in such a way that no more than q arcs leave the same node. The valid inequality associated to W and F, in the family $W - q$, is then

$$\sum_{(i,j)\in F} x_{ij} \leq \left\lfloor \frac{q|W|}{q+1} \right\rfloor.$$

Avella et al. (2007) also studied odd-hole inequalities and lifted them. As an example, consider Fig. 2.3d. It is obvious that $x_{12} + x_{23} \leq 1$, $x_{23} + x_{34} \leq 1$, $x_{34} + x_{54} \leq 1$, $x_{54} + x_{25} \leq 1$ and $x_{25} + x_{12} \leq 1$. Summing up and rounding down, it follows that $x_{12} + x_{23} + x_{34} + x_{54} + x_{25} \leq 2$, named *odd-hole inequality* by the authors. Moreover, this kind of inequality can be lifted to $x_{12} + x_{23} + x_{34} + x_{54} + x_{25} + x_{62} + x_{72} \leq 2$ since arcs (1,2), (6,2) and (7,2) play the same role and only one of them can be taken in a feasible solution.

Baïou and Barahona (2008) consider the particular case of $W - q$ when $q = 1$ and $|F|$ is odd. This corresponds with oriented odd-cycles of k nodes C_k, like the one shown in Fig. 2.3e, that generate the inequalities

$$\sum_{(i,j)\in C_k} x_{ij} \leq \frac{k-1}{2}. \tag{2.25}$$

They prove that, when the graph does not contain either of the two subgraphs of Fig. 2.3f, the linear relaxation plus all the constraints in family (2.25) completely describe the polyhedron associated with formulation (F5). Graphs that do not contain these two structures are called *Y-free graphs*. They also describe a separation procedure for inequalities (2.25) through an auxiliary graph. Baïou and Barahona (2011) show that the family of graphs whose p-median polytope is integer (that is to say, the linear relaxation of formulation (F5) always produces an integer optimal solution) for all values of p are those containing none of any of the structures of Fig. 2.3g, nor any cycle of the type depicted in Fig. 2.3h. They also give additional polyhedral results in their recent paper, Baïou and Barahona (2016). Note that the structure of Fig. 2.3h is a cycle (continuous arcs) with an odd number of nodes with positive in-degree in the cycle; there are arcs (dotted) with origin in the nodes of in-degree two in the cycle and destination at nodes that either are not in the cycle or have out-degree other than two in the cycle; and there is an arc with its two nodes outside the cycle.

2.7 Heuristics

The literature on heuristics for p-median problems is vast. The account presented here does not pretend to be exhaustive and many interesting works on the topic may have been omitted. We invite the interested reader to consult other reviews for an overview of the problem from different perspectives. For instance, in Reese (2006) works are classified by solution method and are also listed by year; Mladenović et al. (2007) classify them into two classes, classical heuristics and metaheuristics, and describe the methods belonging to each group; Basu et al. (2015) focus on metaheuristics; finally, Irawan (2016) is devoted to aggregation methods, which reduce the number of demand points to obtain smaller problems.

2.7.1 Classical Heuristics

The first methodologies approaching p-median problems were heuristics. A simple one produces a feasible solution by starting from an empty set of medians and successively adding the candidate that yields the greatest decrease in the current solution value, until p candidates have been added to the set. This is known as the *greedy heuristic*. Even if Kuehn and Hamburger (1963) is usually cited as the earliest work on greedy heuristics for facility location, Cornuéjols et al. (1977) were the first to formally state the greedy heuristic for p-median problems. In the same vein, the *greedy drop* or simply *drop heuristic*, first devised by Feldman at al. (1966), starts with I as the initial set of medians and iteratively discards the candidate location whose closure produces the smallest increment of the objective function, until the initial set has been reduced to p candidates (see e.g. Whitaker 1981; Salhi and Atkinson 1995).

Other heuristics try to improve a given selection of p candidates. One of the oldest and most widely known of these heuristics allocates each user to the candidate in the initial selection with minimum supplying cost. By grouping users allocated to the same candidate, p neighborhoods are obtained. Then, a 1-median problem is solved for each neighborhood, yielding a new set of p (potentially) different medians. The process is iterated until the set of medians becomes steady. This heuristic is usually referred to as the *alternate heuristic*, and was first proposed by Maranzana (1964). Nevertheless, the idea was not new at the time and it is, in fact, a particular case of the k-means clustering, first conceived by Steinhaus (1957). Another heuristic of this type is the so-called *interchange heuristic* or *vertex substitution*, first proposed by Teitz and Bart (1968). The starting point is also a feasible set of p location candidates, and possible exchanges with the rest of the candidates are iteratively examined. A formal description of the interchange heuristic can be consulted in Whitaker (1983). The alternate and interchange heuristics have been compared empirically in several works. All of them conclude that the interchange heuristic finds better solutions but consumes more time (see e.g. Rushton and Kohler 1973; Rosing et al. 1979). This is probably why the alternate heuristic has received less attention and efforts have concentrated on improving the performance of the interchange heuristic. Countless attempts have been made in this direction, and here we mention some of them. Whitaker (1983) designed a variant of the interchange heuristic that uses a greedy initialization, called *fast interchange*; Densham and Rushton (1991) detailed specific speedup strategies and, later on, Densham and Rushton (1992) introduced GRIA (global regional interchange algorithm); Resende and Werneck (2003) presented an implementation of the fast interchange that performed especially well for large instances and reported speedups of up to three orders of magnitude over the original implementation of Whitaker. Finally, Lim and Ma (2013) introduced a parallel vertex substitution and reported speedups ranging from 10 to 57 times over the traditional algorithm.

2.7.2 Metaheuristics

The above-mentioned methods, together with dynamic programming, dual ascent and Lagrangean relaxation, can be considered as classical heuristics. These first heuristic approaches were followed by the development of metaheuristics in the 1990s. The list of works on metaheuristics for p-median problems is long. One can find well-known schemes, such as tabu search, variable neighborhood search, genetic algorithms, simulated annealing or neural networks, among others. As Mladenović et al. (2007) conclude in their review, empirical results show that metaheuristics represent an improvement in solution quality on large instances, where the performance of classical heuristics is poor. In the last decade the focus has been on solving larger and larger instances. Most effective algorithms usually combine features from different metaheuristics. In this section, we outline the most noteworthy attempts to produce scalable solution techniques. Table 2.2 summarizes some information on the accuracy and computational effort of these heuristics.

Resende and Werneck (2004) proposed a hybrid heuristic that has features of GRASP (greedy randomized adaptive search procedure), tabu search, scatter search and genetic algorithms. They empirically compared the procedure with six other methods and concluded that it was a valuable candidate for a general-purpose

Table 2.2 Summary of the available computational experience on metaheuristics

Authors	Year	Computer	n	t (s)	dev. (%)
Resende and Werneck	2004	SGI Challenge (196 MHz)	5934	8687	0.6
Hansen et al.	2009	Pentium 4 1800 MHz, 256 MB RAM	89,600	50,083	3.2
Avella et al.	2012	IntelCore 2Quad 2.6 GHz, 4 GB RAM, 64 bits	89,600	5779	54.7
Irawan and Salhi	2013	IntelCore i5-650 3.20 GHz, 4 GB RAM, 32 bits	89,600	4415	95.8
Irawan et al.	2014	IntelCore i5-6503.20 GHz, 4 GB RAM, 32 bits	89,600	3404	5.9
Salhi and Irawan	2015	IntelCore i5-650 3.20 GHz, 4.00 GB RAM, 32 bits	264,000	1,875,300	271.0
Janáček and Kvet	2016	IntelCore 2 Duo E6700 2.66 GHz, 3 GB RAM	3038	1102	9.7
Cebecauer and Buzna	2017	Brutus high-performance cluster of ETH Zurich	670,000	–[a]	4.0

The last three columns stand for the maximum size, time in seconds and deviations of the instances tested but do not necessarily correspond to one same instance. Deviations are calculated either with respect to the optimum or to the best objective known

[a]The authors set a time limit of several days and reported time efficiency with respect to the unaggregated problem

approach for the p-median problem. They used a varied testbed with instances of up to 5934 demand points and gave an account of the strengths and weaknesses of their approach, which they did not recommend for really large instances. Hansen et al. (2009) tackled the clustering problem as a large scale p-median model, using an approach based on the variable neighborhood search metaheuristic. They report better solutions in less time than with the state-of-the-art heuristics, even after upgrading these procedures with the same efficient strategies on instances of up to 89,600 nodes.

Avella et al. (2012) introduced a heuristic for large-scale instances that consists of three main components: subgradient column generation, a core heuristic, which computes an upper bound based on Lagrangean reduced costs, and an aggregation procedure that defines reduced size instances. They compared their approach with that of Resende and Werneck (2004) and Hansen et al. (2009) using the same testbed as these authors. They reported excellent results that have merited the recognition as state-of-the-art heuristic for years. Irawan and Salhi (2013) designed a hybrid heuristic for large-scale instances. The proposed approach was tested on the largest "BIRCH" instances of Hansen et al. (2009) (from 25,000 to 89,600 demand points). The authors claimed to have obtained better solutions than those of the algorithm by Avella et al. (2012), AV, and relatively similar to the ones of the algorithm by Hansen et al. (2009), HA. Nonetheless, improvement respect to AV in quality represents some decimals (in %) and they do not run AV nor HA, but take the times reported by Avella et al. (2012) and apply a transformation to estimate running times in their machine.

Irawan et al. (2014) presented a multiphase approach that incorporates aggregation, variable neighborhood search and an exact method. This heuristic proved to be faster than the one by Irawan and Salhi (2013) on the same testbed used in that previous work. This time, the algorithm is also compared with AV and HA, and times for these algorithms are again obtained by estimation. Regarding solution quality, the proposed heuristic compares with AV and HA in a similar way as that of Irawan and Salhi (2013). Salhi and Irawan (2015) introduced a data compression approach for very large facility location problems in the Euclidean space. They incorporated these techniques into two different methods for p-median problems, a multi-start and a reduced variable neighborhood search. After testing their approach, the authors concluded that it is very effective when applied to very large instances (up to 264,000 demand points in their experiments). Janáček and Kvet (2016) suggested an approximate approach based on the radius formulation (F4) and presented it as a compromise approach enabling a trade-off between accuracy and computational time. They compared their proposal with AV and the exact approach by García et al. (2011) on instances having up to 3038 demand points. Even though the results reported are not conclusive, their method seems to be a good candidate for some instances. Cebecauer and Buzna (2017) proposed the concept of adaptive aggregation that keeps the problem size in reasonable limits. They introduced a framework to approach facility location problems that iteratively adjusts the aggregation level during the solution process. They applied it to the p-median and compared its performance to the exact approach by García et al. (2011), obtaining promising results for benchmarks, which reach up to 670,000 demand points.

2.7.3 Approximation Heuristics

One of the drawbacks of many heuristics is that they do not provide any guarantee regarding the quality of the solution obtained. Since the p-median is a core problem in location, it is not surprising to find works that focus on guaranteeing good-quality approximations, even these days. One of the first works concerned with approximate solutions quality is Cornuéjols et al. (1977), who presented a worst-case analysis for relative errors of the Lagrangean relaxation, the greedy, the interchange and dynamic programming heuristics. Some of the heuristics mentioned above also provide a lower bound on the objective function, which gives an estimation of the quality of their solutions. When we have a precise assessment of the quality of the solution with respect to the optimum we speak about *approximation algorithms*. We define an α-approximation algorithm as a polynomial-time algorithm that computes a solution with cost at most α times that of an optimal solution. Most of the papers on approximation algorithms make some assumption regarding costs. When they are given by Euclidean distances, it is known that, for any $\epsilon > 0$, there exists a nearly linear-time $(1 + \epsilon)$-approximation algorithm, see Kolliopoulos and Rao (1999). When costs satisfy the triangle inequality, we speak about the *metric* p-median and the best current approximation factor is $2.675 + \epsilon$, obtained by Byrka et al. (2014). Moreover, Jain et al. (2002) proved that there is no α-approximation of the metric p-median with $\alpha < 1 + 2/e$, unless P = NP.

2.8 Conclusions

We have briefly presented different versions of the p-median problem, their formulations, solution methods, polyhedral properties and heuristic algorithms. We have focused on the basic models, without going into details about the properties of the Simple Plant Location Problem, a very similar problem that is well studied in Chap. 4. Neither have we paid attention to the many possible extensions of the problem, that make it more applicable and realistic, but which are covered in different chapters of this book (addition of a limit of capacity in the facilities, opening and closing facilities in different periods of time, stochastic demands, different objective functions and a long list of options). The p-median problem still receives considerable attention 50 years after its first appearance in the literature and is an exciting field of future research.

Acknowledgements This research has been partially supported by the research projects MTM2015-65915-R (MINECO, Spain), 19320/PI/14 (*Fundación Séneca*) and FUNDBBVA/016/005 (*Fundación BBVA*), and the PhD grant FPU15/05883 (MECD, Spain).

References

AlBdaiwi BF, Ghosh D, Goldengorin B (2011) Data aggregation for p-median problems. J Combin Optim 21:348–363

Avella P, Sassano A (2001) On the p-median polytope. Math Program A 89:395–411

Avella P, Sassano A, Vasilyev I (2007) Computational study of large scale p-median problems. Math Program A 109:89–114

Avella P, Boccia M, Salerno S, Vasilyev I (2012) An aggregation heuristic for large scale p-median problem. Comput Oper Res 39:1625–1632

Baïou M, Barahona F (2008) On the p-median polytope on Y-free graphs. Discrete Optim 5:205–219

Baïou M, Barahona F (2011) On the linear relaxation of the p-median problem. Discrete Optim 8:344–375

Baïou M, Barahona F (2016) On the p-median polytope and the directed odd cycles inequalities: triangle-free oriented graphs. Discrete Optim 22:206–224

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

Basu S, Sharma M, Ghosh PS (2015) Metaheuristic applications on discrete facility location problems: a survey. Opsearch 52:530–561

Beltrán C, Tadonki C, Vial JP (2006) Solving the p-median problem with a semi-Lagrangian relaxation. Comput Optim Appl 35:239–260

Briant O, Naddef, D (2004) The optimal diversity management problem. Oper Res 52:515–526

Byrka J, Pensyl T, Rybicki B, Srinivasan A, Trinh K (2014) An improved approximation for k-median, and positive correlation in budgeted optimization. In: Proceedings of 26th annual ACM-SIAM symposium on discrete algorithms, San Diego, pp 737–756

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 (2002) On the facets of the simple plant location problem. Discrete Appl Math 124:27–53

Cánovas L, Landete M, Marín A (2003) Facet obtaining procedures for set packing problems. SIAM J Discrete Math 16:127–155

Cánovas 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

Cebecauer M, Buzna L' (2017). A versatile adaptive aggregation framework for spatially large discrete location-allocation problems. Comput Ind Eng 111:364–380

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

Church RL (2003) COBRA: a new formulation of the p-median location problem. Ann Oper Res 122:103–120

Cornuéjols GP, Thizy JM (1982) Some facets of the simple plant location polytope. Math Program 22:50–74

Cornuéjols G, Fisher ML, Nemhauser GL (1977) Location of bank accounts to optimize float: an analytical study of exact and approximate algorithms. Manage Sci 22:789–810

Cornuéjols G, Nemhauser GL, Wolsey LA (1980) A canonical representation of simple plant location problems and its applications. SIAM J Algebra Discrete Methods 1:261–272

Cornuéjols G, Nemhauser GL, Wolsey LA (1990) The uncapacitated facility location problem. In: Mirchandani PB, Francis RL (eds) Discrete location theory. Wiley, New York, pp 119–171

de Farias IR (2001) A family of facets for the uncapacitated p-median polytope. Oper Res Lett 28:161–167

Densham PJ, Rushton G (1991) Designing and implementing strategies for solving large location-allocation problems with heuristic methods. Technical report 91-10, National Center for Geographic Information and Analysis, Buffalo

Densham PJ, Rushton G (1992) A more efficient heuristic for solving large p-median problems. Pap Reg Sci 71:307–329
de Vries S, Posner ME, Vohra RV (2003) Polyhedral properties of the K-median problem on a tree. Math Program 110:261–285
Feldman E, Lehrer FA, Ray TL (1966) Warehouse locations under continuous economies of scale. Manag Sci 12:670–684
Galvão RD (1980) A dual-bounded algorithm for the p-median problem. Oper Res 28:1112–1121
García S, Labbé M, Marín A (2011) Solving large p-median problems with a radius formulation. INFORMS J Comput 22:546–556
Goldengorin B, Krushinsky D, Slomp J (2012) Flexible PMP approach for large-size cell formation. Oper Res 60:1157–1166
Guignard M (2003) Lagrangean relaxation. Top 11:151–200
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 and some graph-related theoretic problems. Oper Res 12:462–475
Hansen P, Brimberg J, Urošević D, Mladenović N (2009) Solving large p- median clustering problems by primal-dual variable neighborhood search. Data Min Knowl Disc 19:351–375
Hua LK et al. (1962) Application of mathematical methods to wheat harvesting. Chin Math 2(7791):8
Irawan, CA (2016) Aggregation and non aggregation techniques for large facility location problems-a survey. Yugosl. J. Oper. Res. 25(3):313–341
Irawan CA, Salhi S (2013) Solving large p-median problems by a multistage hybrid approach using demand points aggregation and variable neighbourhood search. J Global Optim 63(3):537–554
Irawan CA, Salhi S, Scaparra MP (2014) An adaptive multiphase approach for large unconditional and conditional p-median problems. Eur J Oper Res 237(2):590–605
Jain K, Mahdian M, Saberi A (2002) A new greedy approach for facility location problems. In: Proceedings of 34th annual ACM symposium on theory of computing, New York, pp 731–740
Janáček J, Kvet M (2016) Sequential approximate approach to the p-median problem. Comput Ind Eng 94:83–92
Kariv O, Hakimi SL (1979) An algorithmic approach to network location problems. II: the p-medians. SIAM J Appl Math 37:539–560
Kolen A (1983) Solving covering problems and the uncapacitated plant location problem on trees. Eur J Oper Res 12:266–278
Kolen A, Tamir A (1990) Covering problems. In: Mirchandani PB, Francis RL (eds) Discrete location theory. Wiley, New York, pp 263–304
Kolliopoulos SG, Rao S (1999) A nearly linear-time approximation scheme for the Euclidean k-median problem. In: Nešetřil J (ed) Algorithms - ESA'99. Lecture notes in computer science, vol 1643. Springer, Berlin, pp 378–389
Kuehn AA, Hamburger MJ (1963) A heuristic program for locating warehouses. Manag Sci 11:643–666
Li, B, Deng X, Golin MJ, Sohraby K (1998) On the optimal placement of web proxies in the internet: the linear topology. In: High performance networking. Springer, Boston, pp 485–495
Lim GJ, Ma L (2013) GPU-based parallel vertex substitution algorithm for the p-median problem. Comput Ind Eng 64(1):381–388
Manne A (1964) Plant location under economics of scale – decentralization and computation. Manag Sci 11:213–235
Maranzana FE (1964) On the location of supply points to minimize transport costs. Oper Res Quart 15:261–270
Mesa JA, Ortega FA, Pozo MA (2014) Locating optimal timetables and vehicle schedules in a transit line. Ann Oper Res 222(1):439–455
Mirchandani P (1990) The p-median problem and generalizations. In: Mirchandani PB, Francis RL (eds) Discrete location theory. Wiley, New York, pp 55–117

Mladenović N, Brimberg J, Hansen P, Moreno-Pérez JA (2007) The p-median problem: a survey of metaheuristic approaches. Eur J Oper Res 179:927–939

Nemhauser GL, Wolsey LA (1988) Integer and combinatorial optimization. Wiley, New York

Reese J (2006) Solution methods for the p-median problem: an annotated bibliography. Networks 48:125–142

Resende MGC, Werneck RF (2003) On the implementation of a swap-based local search procedure for the p-median problem. In: Proceedings of 5th workshop on algorithm engineering and experiments, Philadelphia, pp 119–127

Resende MGC, Werneck RF (2004) A hybrid heuristic for the p-median problem. J Heuristics 10:59–88

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

Rosing KE, Hillsman EL, Rosing-Vogelaar H (1979) A note comparing optimal and heuristic solutions to the p-median problem. Geogr Anal 11:86–89

Rushton G, Kohler JA (1973) ALLOC: heuristic solutions to multifacility location problems on a graph. In: Rushton G, Goodchild M, Ostresh L (eds) Computer programs for location-allocation problems. University of Iowa, Department of Geography. Monograph, vol 6, pp 163–187

Salhi S, Atkinson RA (1995) Subdrop: a modified drop heuristic for location problems. Locat Sci 2:267–273

Salhi S, Irawan CA (2015) A quadtree-based allocation method for a class of large discrete Euclidean location problems. Comput Oper Res 55:23–35

Simão HP, Thizy JM (1989) A dual simplex algorithm for the canonical representation of the uncapacitated facility location problem. Oper Res Lett 8:279–286

Steinhaus H (1957) Sur la division des corps matériels en parties. Bull Acad Polon Sci 4(12):801–804

Teitz MB, Bart P (1968) Heuristic methods for estimating the generalized vertex median of a weighted graph. Oper Res 16:955–961

Vigneron A, Gao L, Golin MJ, Italiano GF, Li B (2000) An algorithm for finding a k-median in a directed tree. Inf Process Lett 74:81–88

Vinod HD (1969) Integer programing and the theory of grouping. J Am Stat Assoc 64:506–519

Whitaker RA (1981) A tight bound drop exchange algorithm for solving the p-median problem. Environ Plan A 12:669–680

Whitaker RA (1983) A fast algorithm for the greedy interchange of large-scale clustering and median location problems. INFOR Inf Syst Oper Res 21:95–108

Wolsey LA (1998) Integer programming. Wiley, New York

Xu N, Lowe TJ (1993) On the equivalence of dual method for two location problems. Transp Sci 27:194–199

Zhao W, Posner ME (2011) A large class of facets for the K-median polytope. Math Program 128:171–203

Chapter 3
p-Center Problems

Hatice Çalık, Martine Labbé, and Hande Yaman

Abstract A *p*-center is a minimax solution that consists of a set of *p* points minimizing the maximum distance between a demand point and a closest point belonging to that set. We present different variants of this 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.

3.1 Introduction

Minimizing the total or average distance that potential users have to travel to reach a facility may not be the right criterion when locating some types of facilities. Such measures tend to favor clients who are clustered in population centers to the detriment of clients who are spatially dispersed. Accessibility discrimination may have a negative impact on remote clients, for instance, 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. Çalık (✉)
Department of Computer Science, CODeS, KU Leuven, Gent, Belgium
e-mail: hatice.calik@kuleuven.be

M. Labbé
Department of Computer Science, Université Libre de Bruxelles, Brussels, Belgium

Inria Lille-Nord Europe, Lille, Villeneuve d'Ascq, France
e-mail: mlabbe@ulb.ac.be

H. Yaman
Faculty of Economics and Business, ORSTAT, KU Leuven, Leuven, Belgium

Department of Industrial Engineering, Bilkent University, Ankara, Turkey
e-mail: hande.yaman@kuleuven.be

G. Laporte et al. (eds.), *Location Science*,
https://doi.org/10.1007/978-3-030-32177-2_3

The 1-center location problem on a network consists of finding a vertex whose distance to all the other vertices is minimum. This problem 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 of this problem 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 so-called *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 case, 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. Player X chooses a point x in G; then player Y chooses a point y in G. As a result 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 Çalık (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. 3.2. Section 3.3 is devoted to the mixed integer linear programming models

and algorithms for solving p-center problems. Heuristics are discussed in Sect. 3.4 and some extensions of the p-center problem are considered in Sect. 3.5. Section 3.6 concludes the chapter.

3.2 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 the so-called *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 $O(n^2)$, where n denotes the number of nodes. 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 those authors 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. In fact, a solution can be found in the set of local centers, i.e., solutions to the problem where centers are restricted to be on edges. The objective function remains piecewise linear on each edge. Nevertheless, the slopes of the linear pieces depend on the vertex weights. A point x on an edge $\{v_k, v_m\}$ is 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 $w_i d(v_i, x) = w_i(d(v_i, v_k) + d(v_k, x)) = w_j d(x, v_j) = w_j(d(x, v_m) + d(v_m, v_j))$. Kariv and Hakimi (1979) showed that, on an edge, at most $3n - 2$ intersections points can determine a local minima. Their algorithm solves the weighted absolute center problem in $O(|E|nlogn)$ time.

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 an arbitrary v_i, the algorithm first determines the vertex v_j whose distance to v_i is maximum. Then it determines the node v_k whose distance to v_j is maximum. The path linking v_j and v_k is a longest one; its midpoint is the absolute center of the tree.

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 a general 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 $O(n^2logn)$ time when G is a tree.

Hochbaum and Shmoys (1985) developed a 2-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 to the minimum weight. In the unweighted case, this guarantee is 2. Recently, Garcia-Diaz et al. (2017) proposed a 3-approximation algorithm for the vertex p-center problem that performs better than the 2-approximation algorithms on benchmark instances.

3.3 Exact Methods

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

In the case of the vertex-restricted p-center problem on networks, 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 have 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 x_i than to any other point in 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 $J = V$.

Given the above observation, the remainder of this section is devoted to models and algorithms for solving the discrete p-center problem.

Several methods based on solving a finite number of instances of the set covering problem have been proposed. The set covering problem (see Chap. 5) is closely related to the p-center problem and can be stated as follows: Given a zero-one matrix $A = [a_{ji}]$ with some cost associated to each column, find a set of columns of minimum total cost covering all 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 is 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 procedure for the p-center problem 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 algorithm solves the set covering problem for the smaller value in R not yet considered 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. The authors also propose the reduction of the size of the set covering matrix by using standard matrix reductions and heuristic techniques. For the selection of the radius values to consider along the execution of the algorithm, they proposed using a bisection method and a binary search strategy instead of moving from one radius value to the next smaller one. Both methods perform the search by halving the search space at each iteration. The difference is that the search space of the bisection method is the real values between the smallest and largest radius values whereas it is the finite set of radius values for the binary search.

A mixed integer programming (MIP) formulation for the discrete p-center problem can be found in 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 located at $i \in I$ and 0 otherwise. The formulation can be stated as follows:

$$\text{Minimize} \quad z \tag{3.1}$$

$$\text{subject to} \quad \sum_{i \in I} d_{ji} x_{ij} \leq z \qquad \forall j \in J, \tag{3.2}$$

$$\sum_{i \in I} x_{ij} = 1 \qquad \forall j \in J, \tag{3.3}$$

$$x_{ij} \leq y_i \qquad \forall i \in I, j \in J, \tag{3.4}$$

$$\sum_{i \in I} y_i \leq p, \tag{3.5}$$

$$y_i \in \{0, 1\} \qquad \forall i \in I, \tag{3.6}$$

$$x_{ij} \in \{0, 1\} \qquad \forall i \in I, j \in J. \tag{3.7}$$

The objective function (3.1) together with (3.2) ensure that the objective value is greater than or equal to the maximum of the distances between demand points and the facilities they are assigned to. Constraints (3.3) establish the assignment of each demand point to exactly one facility. Constraints (3.4) avoid the assignment of demand points to locations with no facility. Constraint (3.5) restricts the number of facilities to p. Constraints (3.6) and (3.7) are the binary restrictions for the decision variables.

Daskin (2013) also presented a set covering based algorithm for the discrete p-center problem, in which the radius value 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. The idea behind this algorithm is similar to the bisection method of Garfinkel et al. (1977). The main difference is that the individual set covering problems of Garfinkel et al. (1977) consider the cardinality restrictions (3.5) as constraints, so, they become feasibility problems whereas these are tackled by the objective function in Daskin (2013) as aforementioned.

Ilhan and Pınar (2001) proposed a two-phase extension of the algorithm developed by Garfinkel et al. (1977). In the first phase, they solve the linear programming (LP) relaxation of the feasibility problem defined by (3.5), (3.6), and

$$\sum_{i \in I} a_{ji} y_i \geq 1, \quad \forall j \in J, \tag{3.8}$$

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 to consider using the lower bound obtained in the first phase. Finally, they 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:

$$\text{Minimize} \qquad r_1 + \sum_{k=2}^{K} (r_k - r_{k-1}) z^k \tag{3.9}$$

$$\text{subject to} \qquad (3.5) \text{ and } (3.6),$$

$$\sum_{i \in I} y_i \geq 1, \tag{3.10}$$

$$z^k + \sum_{i:d_{ji} < r_k} y_i \geq 1 \qquad \forall j \in J,\ k = 2, \ldots, K, \tag{3.11}$$

$$z^k \in \{0, 1\} \qquad k = 2, \ldots, K. \tag{3.12}$$

Constraint (3.10) eliminates the solutions with no open facility. Constraints (3.11) and the objective function (3.9) 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 two-phase idea of Ilhan and Pınar (2001) with the binary search strategy like Garfinkel et al. (1977) to select the radius values from the finite set, R, for solving the set covering problems at each iteration.

Calik and Tansel (2013) developed new IP formulations and a new exact algorithm for 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:

$$\text{Minimize} \qquad \sum_{k=1}^{K} r_k u_k \tag{3.13}$$

subject to (3.5) and (3.6),

$$\sum_{i:d_{ji} \leq r_k} y_i \geq u_k \qquad \forall j \in J,\ k = 1, \ldots, K, \tag{3.14}$$

$$\sum_{k=1}^{K} u_k = 1, \tag{3.15}$$

$$u_k \in \{0, 1\} \qquad k = 1, \ldots, K. \tag{3.16}$$

Constraint (3.15) 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 (3.13). Constraints (3.14) ensure that each customer is served within the selected radius by at least one facility. Constraints (3.16) 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 (3.14) are replaced with constraints (3.17) given below:

$$\sum_{i:d(i,j)\le r_k} y_i \ge \sum_{q=1}^{k} u_q, \quad \forall j \in J,\ k = 1, \ldots, K. \tag{3.17}$$

The semi relaxations of these formulations, in which the binary restriction of the y-variables is 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 where n ranges from 100 to 900 and p ranges from 5 to $(n/3)$. 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.

3.4 Heuristics

Mladenović et al. (2003) introduced the first metaheuristic 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 3038 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 metaheuristic 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 metaheuristic and integrated these bounds into 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 uncapacitated p-median instances of the OR-Library revealed that the utilization of these bounds decreases the solution times of Daskin's algorithm.

Other than metaheuristics, 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 set of facilities to close are obtained from the embedded sub-graphs of the original graph. Through an analysis of the properties of these embedded sub-graphs, initial lower and upper bounds were obtained, two polynomial time algorithms were proposed and procedures to verify the optimality of the solutions were developed. The algorithms provided optimal solutions for several special cases. 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 also the absolute p-center 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 benchmark 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.

Irawan et al. (2016) propose two metaheuristics for the vertex p-center problem and they adapt them for the conditional variant (see below).

3.5 Variants

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

3.5.1 The Capacitated p-Center Problem

The first variant concerns problems with capacitated facilities. There are few studies on this variant. Bar-Ilan et al. (1993) introduced a 10-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.

Özsoy and Pınar (2006) proposed an exact algorithm to solve the capacitated p-center problem. The idea is to see if all nodes can be assigned within a given distance and update lower and upper bounds on the optimal 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 Özsoy and Pınar (2006) to obtain bounds on the optimal radius, Albareda-Sambola et al. (2010) proposed a second subproblem that maximizes 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 better. 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.

Quevedo-Orozco and Ríos-Mercado (2015) proposed an iterated greedy local search with variable neighborhood descent and reported improvement over the algorithm of Scapparra et al. (2004).

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 Özsoy and Pınar (2006).

3.5.2 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 $O(log\ n)$ 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.

3.5.3 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.

3.5.4 The Fault Tolerant p-Center Problem

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), Chechik and Peleg (2015), Fernandes et al. (2018) study approximation algorithms and Chen and Chen (2013) presents two optimal algorithms. Espejo et al. (2015) focus on a variant where they minimize the maximum distance from a customer to each second closest facility. They propose several formulations, a preprocessing algorithm, and valid inequalities.

3.5.5 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. Martínez-Merino et al. (2017) introduced the probabilistic p-center problem where they considered the K largest assignment distances. They provided several formulations and a variable neighborhood search heuristic.

3.6 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, Çalık (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 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 multiple allocation and locating multiple facilities with identical capacities at a node. Hence investigating the capacitated p-center problem on tree networks with non-identical capacities, at most one center at each node and/or single allocation might be a worthwhile undertaking.

Finally, developing different exact approaches and metaheuristic algorithms for the variants of the fault tolerant p-center problem and the p-center problem with uncertain parameters might also 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

Çalık 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. Theor Comput Sci 566:12–25
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
Espejo I, Marín A, Rodríguez-Chía AM (2015) Capacitated p-center problem with failure foresight. Eur J Oper Res 247:229–244
Fernandes CG, de Paula SP, Pedrosa LL (2018) Improved approximation algorithms for capacitated fault-tolerant k-center Algorithmica 80:1041–1072
Garcia-Diaz J, Sanchez-Hernandez J, Menchaca-Mendez R, Menchaca-Mendez R (2017) When a worse approximation factor gives better performance: a 3-approximation algorithm for the vertex k-center problem. J Heuristics 23:349–366
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. Bilkent University, Department of Industrial Engineering, Technical Report. http://www.ie.bilkent.edu.tr/~mustafap/pubs
Irawan CA, Salhi S, Drezner Z (2016) Hybrid meta-heuristics with VNS and exact methods: application to large unconditional and conditional vertex p-centre problems. J Heuristics 22:507–537
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. Inf 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
Martínez-Merino LI, Albareda-Sambola, M, Rodríguez-Chía AM (2017) The probabilistic p-center problem: planning service for potential customers. Eur J Oper Res 262:509–520
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 Inf 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

Özsoy, FA, Pınar, MÇ (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

Quevedo-Orozco DR, Ríos-Mercado RZ (2015) Improving the quality of heuristic solutions for the capacitated vertex p-center problem through iterated greedy local search with variable neighborhood descent. Comput Oper Res 62:133–144

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 by 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 4
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 in 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.

4.1 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 these costs.

E. Fernández (✉)
Department of Statistics and Operations Research, University of Cádiz, Cádiz, Spain
e-mail: elena.fernandez@uca.es

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*,
https://doi.org/10.1007/978-3-030-32177-2_4

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. 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 an 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. 4.2 we introduce our notation and we provide an overview of the problems we study. Section 4.2 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. 4.3. 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. 4.4 and 4.5. 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. 4.4.1 studies some properties derived from linear programming duality, whereas Sect. 4.4.2 presents a formulation for the UFLP based on its supermodular property and relates it with the so-called radius based formulations. Finally, Sect. 4.5 gives some polyhedral results related to the UFLP. The chapter closes in Sect. 4.6 with some comments.

4.2 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:

$$\text{minimize } z = \sum_{i \in I} f_i y_i + \sum_{i \in I} \sum_{j \in J} c_{ij} x_{ij} \tag{4.1}$$

$$\text{subject to} \sum_{i \in I} x_{ij} = 1 \qquad j \in J \tag{4.2}$$

$$\sum_{j \in J} d_j x_{ij} \leq q_i y_i \qquad i \in I \tag{4.3}$$

$$y_i \in \{0, 1\} \qquad i \in I \tag{4.4}$$

$$x_{ij} \in \{0, 1\} \qquad i \in I,\ j \in J. \tag{4.5}$$

Constraints (4.2) guarantee that each customer is served from one facility, while constraints (4.3) 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 (4.4) and (4.5) define the domains of the decision variables. In the above formulation inequalities (4.3) can be substituted by the two sets:

$$\sum_{j \in J} d_j x_{ij} \leq q_i \qquad i \in I \tag{4.6}$$

$$x_{ij} \leq y_i \qquad i \in I, j \in J. \tag{4.7}$$

Now the set of knapsack constraints (4.6) enforce that facility capacities are not violated, whereas inequalities (4.7) relate the two sets of decision variables. While constraints (4.3) are equivalent to (4.6) and (4.7) when the binary condition of the y variables (4.4) is enforced, the compact set of constraints (4.3) dominates (4.6) and (4.7) when the integrality of the location variables is relaxed to $0 \leq y_i \leq 1$, $i \in I$.

Formulation (4.1)–(4.5) 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 (4.1)–(4.5) define 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

$$0 \leq x_{ij} \leq 1, \qquad i \in I, j \in J. \tag{4.8}$$

With the above definition of the allocation decision variables, constraints (4.2) 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 (4.1)–(4.5) we "only" have to change the domain of the allocation variables x. It then follows that that (4.1)–(4.4) together with (4.8) 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:

$$TP(S) \quad \text{minimize } z = \sum_{i \in S} \sum_{j \in J} (c_{ij}/d_j) s_{ij} \tag{4.9}$$

$$\text{subject to} \sum_{i \in S} s_{ij} \geq d_j \qquad j \in J \tag{4.10}$$

$$\sum_{j \in J} s_{ij} \leq q_i \qquad i \in S \tag{4.11}$$

$$s_{ij} \geq 0 \qquad i \in S,\ j \in J. \tag{4.12}$$

In formulation (4.9)–(4.12) 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

$$GAP(S) \quad \text{minimize } z = \sum_{i \in S} \sum_{j \in J} c_{ij} x_{ij} \tag{4.13}$$

$$\text{subject to} \sum_{i \in S} x_{ij} = 1 \qquad j \in J \tag{4.14}$$

$$\sum_{j \in J} d_j x_{ij} \leq q_i \qquad i \in S \tag{4.15}$$

$$x_{ij} \in \{0, 1\} \qquad i \in S,\ j \in J. \tag{4.16}$$

So far we have presented FLPs as minimization problems in which both types of decisions incur costs. Nevertheless, 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. 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), b_{ij} is typically 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

$$\text{maximize } z = -\sum_{i \in I} f_i y_i + \sum_{i \in I} \sum_{j \in J} b_{ij} x_{ij}. \tag{4.17}$$

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 (4.17) can be rewritten as

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

Hence objective (4.17) reduces to

$$\sum_{j\in J} d_j s_j - \min\left[\sum_{i\in I} f_i y_i + \sum_{i\in I}\sum_{j\in J} c'_{ij}x_{ij}\right]. \tag{4.18}$$

Since the first term in (4.18) is a constant, the maximization FLP is equivalent to a minimization FLP.

4.2.1 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

$$SPSFLP \quad \text{minimize} \quad \sum_{i\in I}\sum_{k\in K_i} p_{ki}z_{ki} \tag{4.19}$$

$$\text{subject to} \quad \sum_{i\in I}\sum_{k\in K_i} a_{ijk}z_{ki} = 1 \qquad j \in J \tag{4.20}$$

$$\sum_{k\in K_i} z_{ki} = y_i \qquad i \in I \tag{4.21}$$

$$y_i \in \{0,1\} \qquad i \in I \tag{4.22}$$

$$z_{ki} \in \{0,1\} \qquad i \in I,\ k \in K_i. \tag{4.23}$$

Constraints (4.20) ensure that each customer is assigned to exactly one facility. Constraints (4.21) 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 (4.20), constraints (4.21) can be written as $\leq$ inequalities and will still be satisfied as equalities. Constraints (4.22) and (4.23) 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 simply 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 (4.21), 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.

4.3 Solution Algorithms for Fixed-Charge Facility Location

In this section we overview the available algorithms 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) where 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; 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 by Wentges (1996) and Van Roy (1986), whereas branch-and-price has been applied by Díaz and Fernández (2002) and Klose and Görtz (2007). Some more recent works are Barahona and Chudak (2005), Sankaran (2007), Sharma and Berry (2007), Ghiani et al. (2012), and Zhen et al. (2012). In the paper of An et al. (2017) the authors give an alternative formulation for the FLP based on multi-commodity flows whose integrality gap is constant, i.e, its linear relaxation approximates the optimum value within a constant. 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), and 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).

4.3.1 Lagrangean Relaxation

We next present a Lagrangean relaxation of model SFLP in which the assignment constraints (4.2) 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

$$L_{SFLP}(\pi) = \quad \text{minimize} \sum_{i\in I}\left(f_i y_i + \sum_{j\in J} c_{ij}x_{ij}\right) + \sum_{j\in J} u_j\left(1 - \sum_{i\in I} x_{ij}\right) \tag{4.24}$$

$$\text{subject to} \sum_{j\in J} d_j x_{ij} \le q_i y_i \qquad i \in I \tag{4.25}$$

$$x_{ij} \in \{0,1\} \qquad i \in I,\ j \in J \tag{4.26}$$

$$y_i \in \{0,1\} \qquad i \in I. \tag{4.27}$$

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

$$KP(i): \quad \text{maximize} \quad \sum_{j\in J}\left(c_{ij} - \pi_j\right)x_{ij} \tag{4.28}$$

$$\text{subject to} \quad \sum_{j\in J} d_j x_{ij} \le q_i \tag{4.29}$$

$$x_{ij} \in \{0,1\} \quad j \in J. \tag{4.30}$$

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 4.1 *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 (4.20) in $SPSFLP$ in a Lagrangean fashion:

$$L_{SPSFLP}(\pi) = \text{minimize} \quad \sum_{i\in I}\sum_{k\in K} p_{ki} z_{ki} + \sum_{j\in J} \pi_j \left(1 - \sum_{i\in I}\sum_{k\in K_i} a_{ijk} z_{ki}\right) \tag{4.31}$$

$$\text{subject to} \quad \sum_{k\in K_i} z_{ki} \le y_i \qquad i \in I \tag{4.32}$$

$$z_{ki} \ge 0 \qquad i \in I,\ k \in K_i \tag{4.33}$$

$$y_i \in \{0, 1\} \qquad i \in I. \tag{4.34}$$

The objective function (4.31) 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 $p_{k(i)i} - \sum_{j\in T_{k(i)}} \pi_j \ge 0$ then $y_i = 0, z_{ki} = 0, k \in K_i$.

Note that for each feasible solution $(\hat{z}, \hat{y})$ to (4.32)–(4.34), 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 (4.25). 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. $\sum_{j\in J} d_j \hat{x}_{ij} \le q_i \hat{y}_i$.

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

$$\begin{aligned} \sum_{i \in I} \pi_i + \text{minimize} \quad & \sum_{j \in J} \left(f_i y_i + \sum_{j \in J} c_{ij} x_{ij} \right) - \sum_{j \in J} \sum_{i \in I} \pi_j x_{ij} && (4.35) \\ \text{subject to} \quad & \sum_{j \in J} d_j x_{ij} \leq q_i y_i && i \in I \\ & x_{ij} \in \{0, 1\} && i \in I, j \in J \\ & y_i \in \{0, 1\} && i \in I, \end{aligned}$$

which is indeed $L_{SFLP}(\pi)$. □

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

Proposition 4.1 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).

4.3.2 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 dual variables associated with constraints (4.21) and (4.20), respectively. Then in order to know whether there exists a z variable of the overall formulation which, 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 $\sum_{j \in J} d_j a_{ijk} \leq q_i$. Therefore, 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

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

4.4 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 (4.3) 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. Hence, by replacing constraints (4.3) by (4.7) we obtain the following valid formulation for the UFLP:

$$UFLP \quad \text{minimize} \quad \sum_{i \in I} f_i y_i + \sum_{i \in I} \sum_{j \in J} c_{ij} x_{ij} \tag{4.36}$$

$$\text{subject to} \quad \sum_{i \in I} x_{ij} = 1 \qquad j \in J \tag{4.37}$$

$$x_{ij} \leq y_i \qquad i \in I, j \in J \tag{4.38}$$

$$0 \leq x_{ij} \qquad i \in I, \ j \in J \tag{4.39}$$

$$y_i \in \{0, 1\} \qquad i \in I. \tag{4.40}$$

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, and to Posta (2014) or Fischetti et al. (2017) for insight on the difficulty of the benchmark instances in the UFL library *UflLib,* some of which remain unsolved.

4.4.1 Bounds for UFLP Derived from LP Duality

Consider the LP relaxation of UFLP expressed in standard form, for which constraints (4.38) 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 (4.37), (4.38) and

the upper bound constraints, respectively. Then, the dual of the LP relaxation of UFLP is

$$DUFLP \quad \text{maximize} \quad \sum_{j\in J} u_j - \sum_{i\in I} t_i \tag{4.41}$$

$$\text{subject to} \quad \sum_{j\in J} w_{ij} - t_i \le f_i \qquad i \in I \tag{4.42}$$

$$u_j - w_{ij} \le c_{ij} \qquad i \in I,\ j \in J \tag{4.43}$$

$$w_{ij} \ge 0 \qquad i \in I,\ j \in J \tag{4.44}$$

$$t_i \ge 0 \qquad i \in I. \tag{4.45}$$

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 $u_j \ge \min_{i\in I} c_{ij}$ 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 $\sum_{j\in J} \left(u_j - c_{ij}\right)^+ - f_i \le 0$ 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 (4.42) are

$$(f_i - \sum_{j\in J} \left(u_j - c_{ij}\right)^+) y_i = 0 \qquad i \in I. \tag{4.46}$$

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 $y_i > 0$,

(4.46) holds provided that $\sum_{j\in J}\left(u_j - c_{ij}\right)^+ = f_i$. For the integer UFLP the complementarity slackness conditions (4.46) give the guidelines for primal-dual heuristics. Two alternative strategies may be applied: (i) 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 (ii) 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

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

$$\sum_{j\in J} \min_{i'\in S} c_{i'j} - \sum_{i\in I}\left(\sum_{j\in J}\left(\min_{i'\in S} c_{i'j} - c_{ij}\right)^+ - f_i\right)^+ =$$

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

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 $\sum_{j\in J}\left(\min_{i'\in S} c_{i'j} - c_{ij}\right)^+ - f_i \le 0$ 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 $u_j \ge \min_{i\in I(u)} c_{ij}$. 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).

4.4.2 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., Sect. III.3.1 in Nemhauser and Wolsey 1988) and introduce some additional notation.

Definition 4.1 Let N be a finite set, and Z a real-valued function on the subsets of N. The function Z is *supermodular* if $Z(S) + Z(T) \leq Z(S \cup T) + Z(S \cap T)$, $\forall S, T \subseteq N$.

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 4.1 *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 4.2 *$c(.)$ is supermodular and non-increasing.*

Proof We will use the characterization of supermodular functions of Lemma 4.1(b). 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 $c(S \cup \{i\}) - c(S) = \sum_{j \in J} \left[\min_{i' \in S \cup \{i\}} c_{i'j} - \min_{i' \in S} c_{i'j} \right] \leq 0$. □

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 4.1 can be rewritten as

$$c(T) \geq c(S) + \sum_{i \in T \setminus S} [c(S \cup \{i\}) - c(S)] = c(S) + \sum_{i \in T \setminus S} [c(S \cup \{i\}) - c(S)], \forall S, T \subset I. \tag{4.47}$$

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 4.1 *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

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

Since c is non-increasing supermodular, by (4.47), we also have

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

Thus, for all $S \subseteq I$

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

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 4.1, the UFLP can be stated as the following MIP (see Nemhauser and Wolsey 1981):

$$\text{minimize} \quad \eta \tag{4.48}$$

$$\text{subject to} \quad \eta \geq \sum_{i \in I} f_i y_i + c(S) + \sum_{e \notin S} \rho_i(S) y_i \qquad \forall S \subseteq I^* \tag{4.49}$$

$$\eta \geq 0 \tag{4.50}$$

$$y_i \in \{0, 1\} \qquad i \in I, \tag{4.51}$$

where $I^* = I \cup \{i^*\}$ and i^* is a fictitious facility such that (i) $c_{i^*k} > \max_{i \in I} c_{ij}$, for all $j \in J$; and (ii) $\sum_{j \in J} c_{i^*j} > \max_{i \in I}(f_i + \sum_{j \in J} c_{ij})$. 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 substituting objective (4.48) and constraints (4.49) by (4.52) and (4.53), respectively, where

$$\text{minimize} \quad \sum_{i \in I} f_i y_i + \sum_{j \in J} \eta^j, \tag{4.52}$$

$$\text{and} \quad \eta^j \geq \min_{i \in S} c_{ij} + \sum_{i \notin S} \left[\min_{i' \in S \cup \{i\}} c_{i'j} - \min_{i' \in S} c_{i'j} \right] y_i, \quad \forall S \subseteq I^*, j \in J. \tag{4.53}$$

The following observation indicates that only a polynomial number of constraints (4.53) is required to obtain a valid formulation for the UFLP.

Remark 4.1 For $S \subset I$ and $j \in J$ given, the right-hand side of their associated constraint (4.53) 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 (4.53) 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_{rj} the r-th index according to that ordering. That is, $c_{i_1 j} \leq c_{i_2 j} \leq \cdots \leq c_{i_m j} \leq c_{i_{m+1} j}$, where $c_{i_{m+1} j} = c_{i^* j}$ is the allocation cost of customer j to the fictitious facility i^*. For simplicity, when the index j is clear from the context we just write i_r to denote the r-th ordered element.

Theorem 4.2 *The UFLP can be formulated as*

$$(SUFLP) \quad v_S = \text{minimize} \quad \sum_{i \in I} f_i y_i + \sum_{j \in J} \eta^j \tag{4.54}$$

$$\text{subject to} \quad \eta^j \geq c_{i_r j} + \sum_{i \in I} (c_{ij} - c_{i_r j})^- y_i \quad r = 1, \ldots, m+1, \ j \in J \tag{4.55}$$

$$\eta^j \geq 0 \qquad j \in J \tag{4.56}$$

$$y_i \in \{0, 1\} \qquad i \in I. \tag{4.57}$$

The proof which is based on Remark 4.1 is left to the reader. Formulation (4.54)–(4.57) 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 constraints (4.55) are nothing but Benders cuts. Thus formulation (4.54)–(4.57) admits an alternative interpretation in terms of a Benders type reformulation. The interested reader is addressed to Magnanti and Wong (1990) for an extensive description of the application of Benders reformulations to the UFLP.

Modern implementations of Benders decomposition, in which Benders cuts are embedded within branch-and-cut enumeration methods, have been recently developed for the UFLP and some extensions. In particular, Fischetti et al. (2016) deals with the UFLP and its extension to separable quadratic allocation costs. The reformulation is particularly successful for large scale instances of the classical UFLP with linear costs, since the huge number of allocation variables is replaced with a linear number of continuous variables that model the customer allocation cost directly. Fischetti et al. (2017) have addressed the multiple allocation capacitated case, both for the classical objective with linear costs (MFLP) and when the objective includes convex but non-separable quadratic terms.

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 r-th 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

$$(RUFLP) \quad v_R = \text{minimize} \quad \sum_{i \in I} f_i y_i + \sum_{j \in J}(c_{i_1 j} + \sum_{r=2}^{m}(c_{i_r j} - c_{i_{r-1} j})z_{rj}) \tag{4.58}$$

$$\text{subject to} \quad z_{rj} + \sum_{\substack{i \in I \\ c_{ij} < c_{i_r j}}} y_i \geq 1 \qquad r = 1, \ldots, m+1, \ j \in J \tag{4.59}$$

$$z_{rj} \in \{0, 1\} \qquad j \in J, r = 1, \ldots, m+1 \tag{4.60}$$

$$y_i \in \{0, 1\} \qquad i \in I. \tag{4.61}$$

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 (4.55) guarantee that (z, y) satisfies constraints (4.59) 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 $r^* = \arg\min\{c_{i_r j} : y_{i_r} = 1\}$.

4.5 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. 4.2 facility location problems can also be modeled as maximization problems in which the expression of the objective function is (4.17). 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

$$(KUFLP) \quad \text{maximize } z = \sum_{i\in I} f_i \bar{y}_i + \sum_{i\in I}\sum_{j\in J} p_{ij}x_{ij} \tag{4.62}$$

$$\text{subject to} \sum_{i\in I} x_{ij} \leq 1 \qquad j \in J \tag{4.63}$$

$$x_{ij} + \bar{y}_i \leq 1 \qquad i \in I,\ j \in J \tag{4.64}$$

$$x_{ij} \in \{0,1\} \qquad i \in I,\ j \in J \tag{4.65}$$

$$y_i \in \{0,1\} \qquad i \in I. \tag{4.66}$$

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 $m^* \leq m$ and $n^* \leq n$, 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), 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 coefficient 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.

4.5.1 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 4.3 (Cornuéjols et al. 1977) *The fractional feasible solution* $(x, \bar{y})$ *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 prove that condition (a) of Theorem 4.3 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.

4.5.2 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}

$$\sum_{i \in I^C} b_{ij} x_{ij} + \sum_{i \in I^C} \bar{y}_i \leq 2k - \lceil k/t \rceil, \tag{4.67}$$

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 (4.67) is a facet.

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

Theorem 4.4 (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 $m \times n$-adjacency matrix.

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

Theorem 4.5 (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,*

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

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 4.6 (Cornuéjols and Thizy 1982) *The inequality*

$$x_{21} + x_{32} + x_{13} + \sum_{i=1}^{3} x_{ii} + \sum_{i=1}^{3} \bar{y}_i \leq 4$$

is facet-defining for P^{33}.

Theorem 4.7 (Cornuéjols and Thizy 1982) *The inequality*

$$x_{15} + x_{13} + x_{41} + \sum_{i=1}^{5} x_{ii} + \sum_{i=1}^{4} x_{(i+1)i} + \sum_{i=1}^{5} \bar{y}_i \leq 7$$

is facet-defining for P^{55}.

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

Theorem 4.8 (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,

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

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

Theorem 4.9 (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 4.10 (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,

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

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

Theorem 4.11 (Cánovas et al. 2002) *Let B be the cyclic* $(2k+1, 2)$ *matrix,* $k \ge 1$, *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,

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

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.

Recently an exponentially large family of valid inequalities called *homogeneous inequalities* has been introduced in Galli (2018). Homogeneous inequities generalize the valid inequalities in Theorems 4.4–4.8, namely, those whose coefficients are binary. Necessary and sufficient conditions for homogeneous inequalities to be facet-defining for P^{mn} are given in the mentioned paper.

4.5.3 Lifting Procedures

The procedures that transform a valid inequality (facet) of a polyhedron $P^{m^*n^*}$ into a valid inequality (facet) of a higher dimensional polyhedron P^{mn}, $m \ge m^*$ and $n \ge n^*$, are called lifting procedures. Such results invite the study of smaller polyhedra. The following result indicates how to lift all the facets in the previous section. Apart from the results in this section, other lifting procedures for general set packing models can be found in Cánovas et al. (2003).

Theorem 4.12 (Cho et al. 1983b) *Let*

$$\sum_{i\in P}\sum_{j\in D} \pi_{ij}x_{ij} + \sum_{i\in P} \mu_i\bar{y}_i \le \pi_0 \tag{4.68}$$

be a facet-defining inequality of $P^{m^*n^*}$. *Then, (4.68) 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 4.13 (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. $\pi_{ij} = \mu_i = q - 1, \quad i \in P \cup S, j \in D_i$,
II. $\pi_{ij} = \mu_i = q - 2, \quad i \in T, j \in D_i$.

Theorem 4.14 (Galli 2018) *Let*

$$\sum_{i \in I} \sum_{j \in J} \pi_{ij} x_{ij} + \sum_{i \in P} \mu_i \bar{y}_i \leq \pi_0$$

be a valid inequality of P^{mn}. *Let* P *be an arbitrary subset of* I *and* $\pi_j^+(P) = \max_{i \in P}\{\pi_{ij}\}$. *The augmented inequality*

$$\sum_{i \in I} \sum_{j \in J} \pi_{ij} x_{ij} + \sum_{j \in J} \pi_j^+(P) x_{(m+1)j} + \sum_{i \in I} \mu_i \bar{y}_i + (\sum_{i \in I} \mu_i) y_{m+1} \leq \pi_0$$

is valid of $P^{(m+1)n}$.

4.6 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; Gao 2012; Albareda-Sambola et al. 2013, 2017), 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), just to mention a few possibilities. Some of these extensions are addressed in other chapters of this book.

A wider view of FLPs is provided from the perspective of Multilevel Facility Location (MFL), which defines a large class of problems that is receiving increasing attention and generalizes FLPs. In MFL the set of potential facilities is partitioned in several levels and the goal is to determine the facilities to open at each level, and the assignment of customers to possible multiple sequences of open facilities, so as to optimize a given objective function. The interested reader is referred to Contreras et al. (2018) for a comprehensive overview on MFL, and to Contreras et al. (2017, 2019) for recent approaches to solving some problems of this class.

Acknowledgements This work was partly supported by the Spanish Ministry of *Economía y Competitividad* through grants MTM2015- 68097-P, MTM2015-63779-R 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, 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 36:597–611

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

Albareda-Sambola M, Landete M, Monge JF, Sainz-Pardo JL (2017) Introducing capacities in the location of unreliable facilities. Eur J Oper Res 259:175–188

An HC, Singh M, Svensson O (2017) LP-based algorithms for capacitated facility location. SIAM J Comput 46(1):272–306

Babayev (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 Discrete Math 23:665–679

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

Balcik B, Beamon M (2008) Facility location in humanitarian relief. Int J Logist Res Appl Lead 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, white plains, New York, IBM Data Processing Division

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

Barceló J, Casanovas J (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 Spektr 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 Discrete 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. Discrete Appl Math 124:27–53

Cánovas L, Landete M, Marín A (2003) Facet obtaining procedures for set packing problems. SIAM J Discrete Math 16:127–155

Chen X, Chen Z, Zang W (2012) Total dual integrality in some facility location problems. SIAM J Discrete 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(3):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 Discrete Math 1:163–177
Cornuéjols GP, Nemhauser GL, Wolsey LA (1990) The uncapacitated facility location problem. In: Mirchandani PB, Francis RL (ed) Discrete location theory. Wiley, 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. Naval Res Logist Quart 16:331–343
Delmaire H, Díaz JA, Fernández E, Ortega M (1999a) Reactive GRASP and tabu search based heuristics for the single source capacitated plant location problem. Inf Syst Oper Res 37:194–225
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
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 (ed) (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 Trans 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(3):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
Fischetti M, Ljubic I, Sinnl M (2016) Benders decomposition without separability: a computational study for capacitated facility location problems. Eur J Oper Res 253(3):557–569
Fischetti M, Ljubic I, Sinnl M (2017) Redesigning benders decomposition for large scale facility location. Manag Sci 63:2146–2162
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
Galli L, Letchford AN, Miller SJ (2018) New valid inequalities and facets for the simple plant location problem. Eur J Oper Res 269(3):824–833
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 Trans 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 Discrete Math 1:247–271
Hamacher HW, Labbé M, Nickel S, Sonneborn T (2004) Adapting polyhedral properties from facility to hub location problems. Discrete 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 aceleration 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 Trans 39:41–55
Klincewicz JG, Luss H (1986) A Lagrangean relaxation heuristic for the capacitated facility location with sigle-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
Khumawala BM (1972) An efficient branch and bound algorithm for the warehouse location problem. Manag Sci 18:718–731
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 (ed) Discrete location theory. Wiley, 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 Discrete Math 11:279–301
Nemhauser GL, Wolsey LA (1988) Integer and combinatorial optimization. Wiley-Interscience series in discrete mathematics and 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
Ortiz-Astorquiza C, Contreras I, Laporte G (2017) Formulations and approximation algorithms for multilevel uncapacitated facility location. INFORMS J Comput 29:767–779
Ortiz-Astorquiza C, Contreras I, Laporte G (2018) Multi-level facility location problems. Eur J Oper Res 267:791–805
Ortiz-Astorquiza C, Contreras I, Laporte G (2019) An exact algorithm for multi-level uncapacitated facility location. Transp Sci 53:917–1212
Owen SH, Daskin MS (1998) Strategic facility location: a review. Eur J Oper Res 111(3):423–447
Pirkul H (1987) Efficient algorithms for the capacitated concentrator location problems. Comput Oper Res 14:197–208

Posta M, Ferland JA, Michelon P (2014) An exact cooperative method for the uncapacitated facility location problem. Math Program Comput 6:199–231

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(1):36–43

Spielberg K (1969a) Plant location wih 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

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 Meth 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 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 she is located less than a certain distance from the nearest 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 most relevant problems in this field: the Set Covering Problem, the Maximal Covering Location Problem, and related problems, In addition, it is introduced a general model that has as particular cases the main covering location models. The most important 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.

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 she is located less than a certain distance from the nearest 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

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*,
https://doi.org/10.1007/978-3-030-32177-2_5

name of *covering location problems* and, when the previous condition holds, it is said that the customer is covered. In this chapter, we will refer to them simply as covering problems.

The first mentions to covering problems in the literature can be found in two papers. One is 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. The other is Hakimi (1965), where it must be decided the minimum number of police patrols required to protect a highway network. 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 exist 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 minimizing the number of facilities to be located. A quick analysis of a solution to the set covering problem will usually show that it is possible to cover an important percentage of the demand with just a few facilities, and that full coverage can be achieved only by locating a large number of them. 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 number of 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 widespread 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), Steiner matrices (Feo and Resende 1989), and humanitarian logistics (Li et al. 2018).

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, Chung (1986) examined 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 algorithms (exact and heuristic) for the set covering problem. Plastria (2002) provides an exhaustive review of continuous covering models, and it is a perfect complement to this chapter. More recently, Berman et al. (2010) considered some of the latest trends by reviewing gradual coverage, cooperative coverage, and variable radius coverage models, and Snyder (2011) who reviewed the seminal covering models plus some extensions. The most recent

general survey is that of Farahani et al. (2012), which contains an exhaustive list of models and reviews more than 150 papers on covering problems in the area of facility location. Murray (2016) is a more recent survey that focuses on maximal covering problems. More a tutorial than a survey, Daskin (2013) constitutes an excellent introduction to the basic properties of covering models.

At this point, it must be said that there exist many different models involving covering, and that the goal of this chapter is not to review 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. The rest of this chapter is organized as follows: the main models are presented 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 is widely used for 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 as well as several other basic location problems that 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 the field of 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 through different procedures depending on the model under consideration, as we will see later.

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 costs of the centers to be opened in $i \in I$ are the same (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 constant $e_i \in$

$\mathbb{Z}^+$. Particularly, if $e_i = 1$, then y_i is a binary variable that takes value one if and only if a facility is located at site i.

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 $K = \{1, \ldots, h\}$ that can be different depending on the particular model that is considered. Associated to variables w_{jk} are fixed costs $g_{jk} \in \mathbb{R}$. 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) contains two terms. The first sum is the total fixed cost of opening y_i facilities at site $i \in I$. The second sum is 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) counts the number of open facilities that cover 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 that are covering j beyond the minimum number b_j. Due to the condition that we imposed on the g-values, the w-variables that take 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. A cost g_{j1} will be paid if demand point j is covered by at least $b_j + 1$ centers; an 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 the y-variables and impose that at most e_i centers can be installed at site i. Constraints (5.5) state that the variables w are binary.

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 with the number of facilities that over-cover customers. By giving particular values to the constants in (COV), different existing 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 less than 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 to minimize the number of open centers. Therefore, the variables w_{jk} can be removed from the model by replacing the equalities in (5.3) with 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. Therefore, $e_i = 1$ $\forall i \in I$. Given the importance of this model, its classical formulation is explicitly provided:

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

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 extensively studied. The interested reader can consult the survey by Balinski (1965). Other interesting seminal papers are those of 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 often happens 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 that 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 1.

Redundant Covering Location Problem: The Redundant Covering Location Problem (RCLP) was studied by 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 that 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 that cover j), only variables w_{j1} would be necessary if equalities (5.3) were replaced with 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 sufficiently large cost.

Hierarchical Covering Location Problem (HCLP): An objective function that allows the simultaneous minimization of the number of opened facilities and the maximization of the number of previously existing facilities that are kept (within the minimum total number of facilities) was introduced by Plane and Hendrick (1977) to study the location of fire stations. The coefficients a_{ij} are equal to one if and only if focal point i can be served by a pumper company at location j within 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 was 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 the number of facilities to be located is now 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, proposed in Church and ReVelle (1974), is to combine mandatory covering of some demand points (assume that 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 handled 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 more 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 indicate 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) proposed a different criterion for coverage which can also be viewed as a particular case of (COV). More 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) described in 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, $\sum_{k=1}^{\ell} g_{jk} = -t_j(1-q^{\ell})$, which is the correct contribution of j to the objective function when j is covered by ℓ facilities and $w_{jk} \leq w_{j,k+1}\ \forall k$. Note that this last inequality is satisfied implicitly because $q^k \geq q^{k+1}$ means that coefficients $\{g_{jk}\}_k$ are sorted in non-decreasing 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., the servers are independent or they have the same failure probabilities) have been relaxed, for example, by 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 the fraction of the day that the service is needed at 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 obtain 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 the 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 obtain it as a particular case of (COV): for each demand point j, a vector $D_j = (D_{1j}, \ldots, D_{G_j j})$ 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 obtain the right optimal value. This formulation has been 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. 4 (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	δ[c]	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
[c]$\delta := \min\{b_j \in \mathbb{Z} \,/\, \left(\frac{F_j}{b_j}\right)^{b_j} \leq 1-\alpha\}$

5.3 Theoretical Results

The Set Covering Problem is NP-hard (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, polyhedral analysis, and relation to other problems.

When solving the 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 $\{i \in I : a_{ij_2} = 1\} \subseteq \{i \in I : a_{ij_1} = 1\}$, that is, if all the facilities covering demand point j_2 can also cover j_1. Similarly, a facility i_1 that covers a set of demand points that can be all covered by a cheaper facility i_2 will never be used: if $f_{i_1} \geq f_{i_2}$ and $\{j \in J : a_{i_1 j} = 1\} \subseteq \{j \in J : a_{i_2 j} = 1\}$, 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 of cost (with those columns with equal cost sorted in decreasing order of 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 iteratively 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. More specifically, 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 by 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) perform a similar study for the case of facets with coefficients in $\{0, 1, 2, 3\}$ and right-hand side equal to 3. The polyhedral structure of a problem that includes simultaneously covering, partitioning and packing constraints is studied in Kuo and Leung (2016). Finally, a different type of result can be found in Aguilera et al. (2017), where vertex adjacencies for the polyhedron of the SCP are described, including a sufficient condition for adjacency.

The connection of the SCP to other classical problems has also been studied. 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 the 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.

Fewer theoretical results can be found for the Maximal Covering Location Problem, which is known to be NP-hard (Megiddo et al. 1983). The 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

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

with d_{ij} the distances from the original problem and s 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) developed a branch-and-bound method that exploits the structure of the SCP formulation and solutions. Later, Etcheberry (1977) used a branch-and-bound strategy where the branching is done on constraints and not on variables. The lower bounds of the tree are calculated by 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), who proposed a branch-and-bound method whose main elements are a dual ascent procedure and subgradient optimization. This algorithm was latter 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) proposed 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) developed 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) developed a procedure applied to a Lagrangian dual problem at each node that combines subgradient optimization with primal and dual heuristics that tighten the upper and lower bounds. These strengthened bounds allow to fix some variables. In general, Lagrangian methods are the most extended and effective algorithms in the literature for these problems. More recently, Avella et al. (2009) proposed a cutting plane algorithm where the separation algorithm is solved exactly on a subproblem defined by a subset of the original constraints and variables of the set covering problem formulation. In Haddadi (2017) Benders decomposition is used to solve to optimality set covering problems that "almost" satisfy a certain consecutive ones property.

On the contrary, not many exact algorithms have been developed for the Maximal Covering Location Problem. Downs and Camm (1996) obtained a primal solution by using the greedy heuristic of Church and ReVelle (1974). They used 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 through subgradient optimization. If an integer solution is not obtained, branch-and-bound is applied.

5.5 Approximate Algorithms

As it happens with any hard optimization problem, heuristic algorithms are more frequently used than exact methods. 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 by 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 applied to the 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 implemented 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 was provided. Later, Daskin (1983) described a heuristic for the Maximum Expected Covering Location Problem that 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, the heuristic determines whether there exists an interval in which the current best solution is improved. By iterating this procedure, the interval [0,1] is partitioned, and a heuristic solution is given for each of the resulting subintervals. In MCLP, Galvão and ReVelle (1996) developed 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 the 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 the SCP and the MCLP.

5.6 Lagrangian Relaxation

Among the many different methods that have been developed for covering models, here we highlight Lagrangian Relaxation (LR) for several reasons. First, LR can be used as a heuristic method but additionally can also provide good lower bounds, which can be embedded within a branch-and-bound framework to yield 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. Finally, LR usually produces very good results within a reasonable 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

$$\begin{array}{lll} \text{(COVLR}(v)\text{)} & \text{Minimize} & \sum_{i\in I} \left(f_i + \sum_{j\in J} v_j a_{ij} \right) y_i + \sum_{j\in J}\sum_{k\in K} \left(g_{jk} - v_j \right) w_{jk} \\ & \text{subject to} & \text{(5.2), (5.4), and (5.5),} \end{array}$$

and then adding constant $-\sum_{j\in J} v_j b_j$, we will obtain 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

$$\begin{array}{lll} \text{(COVLRy}(v)\text{)} & \text{Minimize} & \sum_{i\in I} \left(f_i + \sum_{j\in J} v_j a_{ij} \right) y_i \\ & \text{subject to} & \text{(5.2), (5.4),} \end{array}$$

and

$$\begin{array}{lll} \text{(COVLRw}(v)\text{)} & \text{Minimize} & \sum_{j\in J}\sum_{k\in K} \left(g_{jk} - v_j \right) w_{jk} \\ & \text{subject to} & \text{(5.5).} \end{array}$$

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

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

If, as in most of the models that we considered, the g_{jk}-values are sorted in non-decreasing 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 be:

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

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 non-decreasing order:

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

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

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

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

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

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 that 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

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

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 non-decreasing order, the optimal solution to $(\text{COV})_j$ 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,

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

5.7 Continuous Covering Location Problems

When speaking about continuous covering, the set of candidates where facilities can be located is not discrete but a full continuous space. Because of the nature of these problems, most of them are in the plane or, if height or 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 of finding the smallest circle in the plane that contains all the points of a given set that 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 not be 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 that 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 exists a similar property when the facilities can be located on any part of a network (Church and Meadows 1979). Church (1984) shows an analogous result 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

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 they have received. 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 have introduced. As far as we know, this is the first attempt to develop such an unified approach for the study of set covering problems.

Set covering problems having received so much attention, it seems that the number of theoretical results is relatively small. These results reduce basically to some preprocessing rules and to the study of some facets. 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 The research of the authors has been partially supported by the research project 19320/PI/14 (*Fundación Séneca, Región de Murcia, Spain*). Alfredo Marín has also been supported by the research projects MTM2015-65915-R (MINECO, Spain) and "Cost-sensitive classification. A mathematical optimization approach" (*Fundación BBVA*).

References

Aguilera NE, Katz RD, Tolomei PB (2017) Vertex adjacencies in the set covering problem. Discrete Appl Math 218:40–56

Al-Sultan KS, Hussain MF, Nizami JS (1996) A genetic algorithm for the set covering problem. J Oper Res Soc 47(5):702–709

Avella P, Boccia M, Vasilyev I (2009) Computational experience with general cutting planes for the set covering problem. Oper Res Lett 37(1):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 subradient-based branch-and-bound procedure for set covering. Oper Res 44(6):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(4):710–760

Balinski ML (1965) Integer programming: methods, uses, computations. Manag Sci 12(3):253–313

Batta R, Mannur NR (1990) Covering-location models for emergency situations that require multiple response units. Manag Sci 36(1):16–23

Batta R, Dolan JM, Krishnamurthy NN (1989) The Maximal expected covering location problem: revisited. Transp Sci 23(4):277–287

Beasley JE (1987) An algorithm for the set covering problem. Eur J Oper Res 31(1):85–93
Beasley JE (1990) A Lagrangian heuristic for set-covering problems. Nav Res Log 37(1):151–164
Beasley JE, Chu PC (1996) A genetic algorithm for the set covering problem. Eur J Oper Res 94(2):392–404
Beasley JE, Jørnsten K (1992) Enhancing an algorithm for set covering problems. Eur J Oper Res 58(2):293–300
Bell T, Church RL (1985) Location-allocation theory in archaeological settlement pattern research: some preliminary applications. World Archaeol 16(3):354–371
Bellmore M, Ratliff HD (1971) Set covering and involutory bases. Manag Sci 18(3):194–206
Berge C (1957) Two theorems in graph theory. Proc Natl Acad Sci USA 43(9):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 Algebra Discr 7(3):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(5):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(2):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(5):591–599
Chung C (1986) Recent applications of the maximal covering location planning (M.C.L.P.) model. J Oper Res Soc 37(8):735–746
Church RL (1984) The planar maximal covering location problem. J Reg. Sci 24(2):185–201
Church RL, Meadows ME (1979) Location modeling utilizing maximum service distance criteria. Geogr Anal 11(4):358–373
Church RL, ReVelle C (1974) The maximal covering location problem. Pap Reg Sci Assoc 32(1):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(4):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(2):105–112
Chvátal V (1979) A greedy heuristic for the set-covering problem. Math Oper Res 4(3):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 Algebra Discrete Methods 1(3):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(2):116–126
Current JR, Storbeck JE (1988) Capacitated covering problems. Environ Plann B 15(2):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, propeties and heuristic solution. Transp Sci 17(1):48–70
Daskin MS (2013) Covering problems. In: Networks and discrete location. Models, algorithms and applications, 2nd edn. Wiley, New York, pp 124–192

Daskin MS, Stern EH (1981) A hierarchical objective set covering model for emergency medical service vehicle deployment. Transp Sci 15(2):137–152
Daskin MS, Haghani AE, Khanal M, Malandraki C (1989) Aggregation effects in maximum covering problems. Ann Oper Res 18(1):115–140
Downs BT, Camm JD (1996) An exact algorithm for the maximal covering problem. Nav Res Log 43(3):435–461
Drezner Z (1981) On a modified one-center problem. Manag Sci 27(7):848–851
Drezner Z, Wesolowsky GO, Drezner T (2004) The gradual covering problem. Nav Res Log 51(6):841–855
Dwyer FP, Evans JR (1981) A branch and bound algorithm for the list selection problem in direct mail advertising. Manag Sci 27(6):658–667
Eaton DJ, Sánchez HML, Lantigua RR, Morgan J (1986) Determining ambulance deployment in Santo Domingo, Dominican Republic. J Oper Res Soc 37(2):113–126
Elzinga D, Hearn D (1972) Geometric solutions for some minimax location problems. Transp Sci 6(4):379–394
Etcheberry J (1977) The set-covering problem: a new implicit enumeration algorithm. Oper Res 25(5):760–772
Farahani RZ, Asgari N, Heidari N, Hosseininia M, Goh M (2012) Covering problems in facility location: a review. Comput Ind Eng 62(1):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(6):674–688
Galvão RD, ReVelle C (1996) A Lagrangean heuristic for the maximal covering location problem. Eur J Oper Res 88(1):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(2):377–389
Galvão RD, Chiyoshia FY, Morabito R (2005) Towards unified formulations and extensions of two classical probabilistic location models. Comput Oper Res 32(1):15–33
García S, Labbé M, Marín A (2011) Solving large p-median problems with a radius formulation. INFORMS J Comput 23(4):546–556
Garey MR, Johnson DS (1979) Computers and intractibility: a guide to the theory of NP-completeness. WH Freeman, New York
Guignard M (2003) Lagrangean relaxation. TOP 11(2):151–200
Haddadi S (2017) Benders decomposition for set covering problems almost satisfying the consecutive ones property. J Comb Optim 33(1):60–80
Hakimi SL (1965) Optimum distribution of switching centers in a communication network and some related graph theoretic problems. Oper Res 13(3):462–475
Hogan K, ReVelle C (1986) Concepts and applications of backup coverage. Manag Sci 32(11):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(7):1129–1140
Klastorin TD (1979) On the maximal covering location problem and the generalized asssignment problem. Manag Sci 25(1):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(1):36–81
Kuo YH, Leung JMY (2016) On the mixed set covering, packing and partitioning polytope. Discrete Optim 22:162–282
Lemke CE, Salkin HM, Spielberg K (1971) Set covering by single branch enumeration with linear programming subproblems. Oper Res 19(4):998–1022

Li XP, Ramshani M, Huang Y (2018) Cooperative maximal covering models for humanitarian relief chain management. Comput Ind Eng 119:301–308
Lorena LAN, Lopes FB (1994) A surrogate heuristic for set covering problems. Eur J Oper Res 79(1):138–150
Mannino C, Sassano A (1995) Solving hard set covering problems. Oper Res Lett 18(1):1–5
Megiddo N, Zemel E, Hakimi SL (1983) The maximum coverage location problem. SIAM J Algebra Discrete Methods 4(2):253–261
Murray AT (2016) Maximal coverage location problem: impacts, significance, and evolution. Int Reg Sci Rev 39(1):5–27
Neebe AW (1988) A procedure for locating emergency-service facilities for all possible response distances. J Oper Res Soc 39(8):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(4):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(1):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(3):192–200
Roth R (1969) Computer solutions to minimum-cover problems. Oper Res 17(3):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(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(2):93–110
Storbeck JE, Vohra RV (1988) A simple trade-off model for maximal and multiple coverage. Geogr Anal 20(3):220–230
Toregas C, ReVelle C (1972) Optimal location under time or distance constraints. Pap Reg Sci Assoc 28(1):133–144
Toregas C, ReVelle C (1973) Binary logic solutions to a class of location problem. Geogr Anal 5(2):145–175
Toregas C, Swain A, ReVelle C, Bergman L (1971) The location of emergency service facilities. Oper Res 19(6):1363–1373
Vasko FJ, Wolf FE, Stott KL (1989) A set covering approach to metallurgical grade assignment. Eur J Oper Res 38(1):27–34

Part II
Advanced Concepts

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.

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
Instituto de Matemáticas de la Universidad de Sevilla, Universidad de Sevilla, Sevilla, Spain
e-mail: ecarrizosa@us.es

B. G.-Tóth (✉)
Department of Computational Optimization, Institute of Informatics, University of Szeged, Szeged, Hungary
e-mail: boglarka@inf.szte.hu

G. Laporte et al. (eds.), *Location Science*,
https://doi.org/10.1007/978-3-030-32177-2_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 and Barkhi (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 hazadarous 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. For covering models for which interactions are not measured via planar distances, but network distances instead (typically shortest-path distances) the works (Berman et al. 1996; Berman and Huang 2008; Berman and Wang 2011; Colebrook and Sicilia 2013) give a good overview.

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. (2000, 2002, 2008), 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. (2003, 2010), Karasakal and Karasakal (2004), Brimberg et al. (2015). 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.

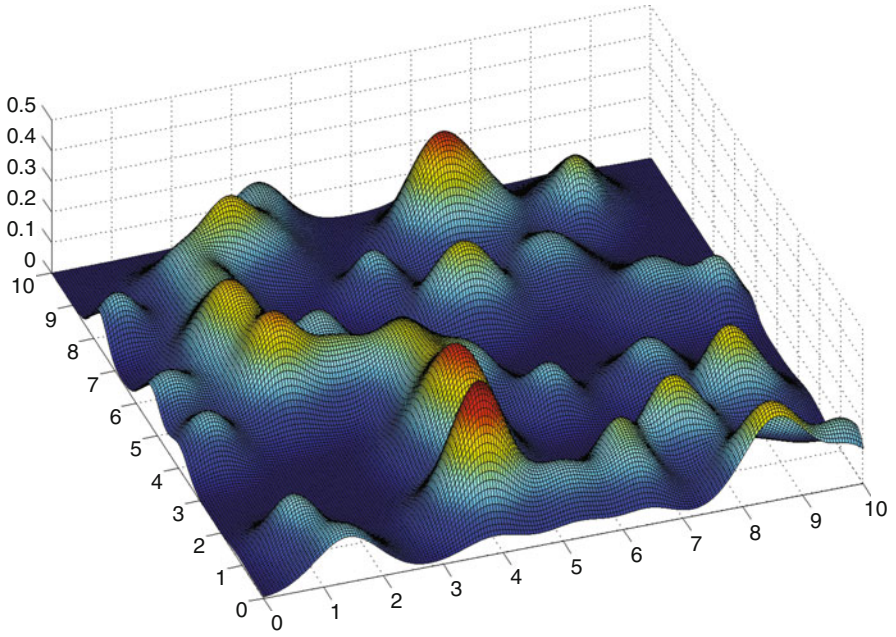


Fig. 6.1 Pdf of a mixture of 50 bivariate gaussians

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),

$$\begin{aligned} C^F(\mathbf{x}) &\leq \delta \text{ ifandonlyif} \\ c_{ij}^F(x_i,x_j) &\leq \delta \ \ \forall i,\ j,\ i\neq j, \text{ ifandonlyif} \\ \|x_i-x_j\| &\geq (\varphi_{ij}^F)^{-1}(\delta) \ \ \forall i,\ j,\ i\neq j. \end{aligned}$$

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; Saboonchi et al. 2014; Sayyady and Fathi 2016), that has been combined with the p-center, p-median and Max-Sum diversity objectives into a bi-objective problem in Tutunchi and Fathi (2019), Sayyady et al. (2015), Colmenar et al. (2018), respectively. 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. In Drezner et al. (2019) the same problem on the plane was solved by a Voronoi based heuristic.

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 probability density functions (pdf), 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.

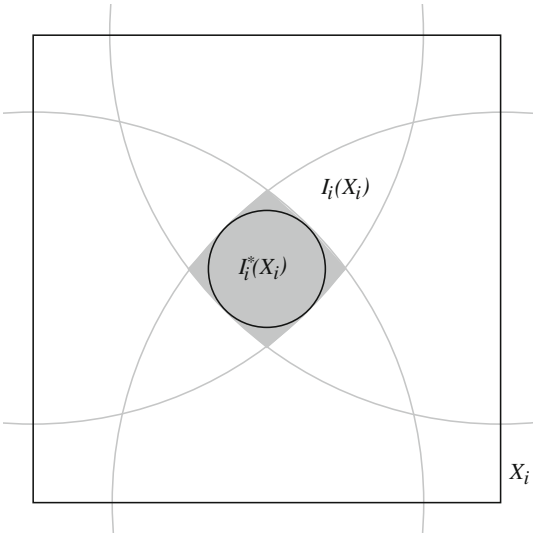


Fig. 6.2 Intersection of covered areas from $\text{ext}(X_i)$ giving the region which is covered by all points in the box. The integral is computed over the inscribed circle of this region, $I_i^*(X_i)$

This leads to

$$\min_{\mathbf{x}\in\mathbf{X}} C(\mathbf{x}) \geq \int_{\bigcup_{i=1}^{p} I_i(X_i)} d\mu(a) \geq \sum_{i=1}^{p} \int_{I_i^*(X_i)} d\mu(a) - \sum_{\substack{i,j=1 \\ i<j}}^{p} \int_{I_i^*(X_i)\bigcap I_j^*(X_j)} d\mu(a).$$

In what follows, the so obtained lower 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 outlined 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

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	12,282.9
	$1.5R$	3754.7	18,167.5
	$2R$	3675.1	>8 h
100	R	20,026.1	>8 h
	$1.5R$	>8 h	>8 h
	$2R$	>8 h	>8 h

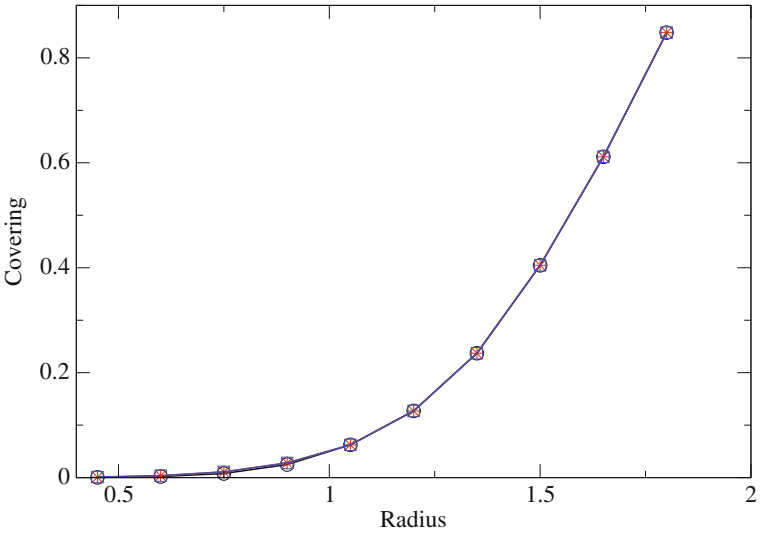


Fig. 6.3 Pareto frontier of the problem of maximizing the radius and minimizing covering

performed. When at least two out of the three instances could not reach the desired accuracy in 8 h, the message "$>$8 h" 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 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 small 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.

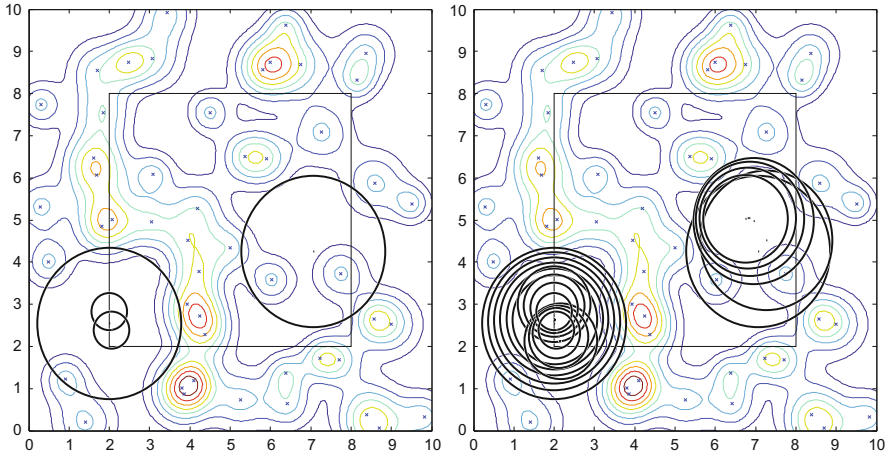


Fig. 6.4 Optimal covering for extreme radii (left) and all radii (right)

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. To mention a few extensions (Wei and Murray 2014, 2017), include spatial uncertainty minimization (Niblett and Church 2015), introduce the disruptive anti-covering location problem.

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 $c^F_{jk}(x_j, x_k)$ 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), the Hungarian National Research, Development and Innovation Office—NKFIH (OTKA grant PD115554), 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. Discret 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 Glob Optim 23:139–154
Blanquero R, Carrizosa E (2008) Continuous location problems and big triangle small triangle: constructing better bounds. J Glob 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
Brimberg J, Juel H, Körner MC, Shöbel A (2015) On models for continuous facility location with partial coverage. J Oper Res Soc 66:33–43
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- Oper Res 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 32:101–118
Colebrook M, Sicilia J (2013) Hazardous facility location models on networks. In: Batta R, Kwon C (eds) Handbook of OR/MS models in Hazardous materials transportation. Springer, New York, pp 155–186
Colmenar JM, Martí R, Duarte A (2018) Heuristics for the bi-objective diversity problem. Expert Sys Appl 108:193–205
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 Logist 51:841–855
Drezner Z, Kalczynski P, Salhi S (2019) The planar multiple obnoxious facilities location problem: a Voronoi based heuristic. Omega 87:105–116. https://doi.org/10.1016/j.omega.2018.08.013
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. Inf Syst Oper Res 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, 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 Meth Oper Res 74:281–310
Mladenović N, Plastria F, Urošević (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
Niblett MR, Church RL (2015) The disruptive anti-covering location problem. Eur J Oper Res 247:764–773
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, 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. Discret 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
Saboonchi B, Hansen P, Perron S (2014) MaxMinMin p-dispersion problem: a variable neighborhood search approach. Comput Oper Res 52:251–259
Sayyady F, Fathi Y (2016) An integer programming approach for solving the p-dispersion problem. Eur J Oper Res 253:216–225
Sayyady F, Tutunchi GK, Fathi Y (2015) p-Median and p-dispersion problems: a bi-criteria analysis. Comput Oper Res 61:46–55
Schilling VJ DA, 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
Tutunchi GK, Fathi Y (2019) Effective methods for solving the Bi-criteria p-Center and p-Dispersion problem. Comput Oper Res 101:43–54
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, Berlin
Wei, R, Murray, AT (2014) A multi-objective evolutionary algorithm for facility dispersion under conditions of spatial uncertainty. J Oper Res Soc 65:1133–1142
Wei, R, Murray, AT (2017). Spatial uncertainty challenges in location modeling with dispersion requirements. In: Thill JC (ed) Spatial analysis and location modeling in Urban and regional systems. Springer, Berlin, pp 283–300
Yao J, Murray AT (2014) Serving regional demand in facility location. Pap Reg Sci 93:643–662

Chapter 7
Locating Dimensional Facilities in a Continuous Space

Anita Schöbel

Abstract Many applications in data analysis such as regression, projective clustering, or support vector machines can be modeled as location problems in which the facilities to be located are not represented by points but as dimensional structures. Examples for one-dimensional facilities are straight lines, line segments, or circles while boxes, strips, or balls are two-dimensional facilities. In this chapter we discuss the location of lines and circles in the plane, the location of hyperplanes and hyperspheres in higher dimensional spaces and the location of some other dimensional facilities. We formulate the resulting location problems and point out applications in statistics, operations research and data analysis. We identify 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 years.

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 data 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 is also given in Blanquero et al. (2009).

Within the last 10 years, the topic has received new focus in the field of data science leading to new results and approaches. In this chapter, we review the new findings and present a unified understanding of the subject which is now possible

A. Schöbel (✉)
Technical University Kaiserslautern and Fraunhofer ITWM, Kaiserslautern, Germany
e-mail: schoebel@mathematik.uni-kl.de

G. Laporte et al. (eds.), *Location Science*,
https://doi.org/10.1007/978-3-030-32177-2_7

since the field has become more mature. We hence not only present a list of problems 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, computational geometry, and data science. 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 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

In classical facility location one looks for a point-shaped new facility. In our case we look for a dimensional facility X such as a line, a hyperplane, a circle or a square. The location of a dimensional facility is a natural generalization of locating a point. As in classical location problems we have given

- a finite set $V = \{v_1, \ldots, v_n\} \subseteq \mathbb{R}^D$ of *data points* (also called *existing facilities*) 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 $\mathbb{R}^D$. The distance measure is used for determining the *residuals*, i.e., the distances from the data points to the new facility X. Finally, we need
- a *globalizing function* $g : \mathbb{R}^n \to \mathbb{R}$ combining the weighted residuals to one global number.

We look for a new facility X which minimizes the globalizing function g of the weighted distances to the data 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 globalizing functions g are the sum, i.e., $g_1(y_1, \ldots, y_n) = \sum_{j=1}^{n} y_j$ or the maximum $g_{\max}(y_1, \ldots, y_n) = \max_{j=1,\ldots,n} y_j$. The resulting problems are called *minsum* (or *median*) location problem and *minmax* (or *cen-*

ter) location problem, respectively. Also, other globalizing functions such as the centdian, or more general, ordered median 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 (data) point v. As common in the literature the distance between a point v and a set X is determined by projecting the point v on the set X and then taking the distance from v to the projected point, i.e.,

$$d(X, v) = \min_{x \in X} d(x, v). \tag{7.2}$$

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 which have mostly been studied in the literature. The most common distances in location theory are *norm distances*. A norm distance is derived from a norm, 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, *gauge distances* which are 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. In particular in statistics, the *vertical distance* is used which is neither a norm nor a gauge. We will see that it gives nevertheless 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 constitutes the most common definition of residuals in regression.

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 data 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 data points. In the right part a circle minimizing the sum of Euclidean distances to the data points is depicted. The lengths of the thin lines in both examples correspond to the residuals, i.e., to the distances from the data points to the line (or to the circle, respectively). Note that the distance between $v \in X$ and X is zero—this happens twice in the right part of the figure where the minsum circle passes through two of the data 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

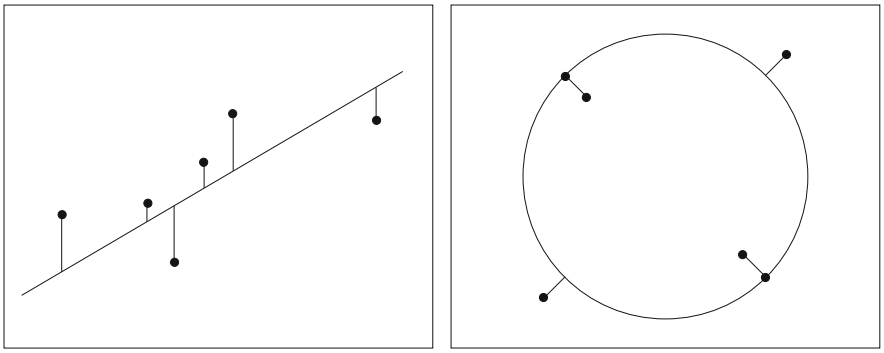


Fig. 7.1 Two illustrations for locating dimensional facilities, both with six demand points. Left: A line minimizing the maximum vertical distance. Right: A circle minimizing the sum of Euclidean distances

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 data 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 algorithmically useful.

7.3 Locating Lines and Hyperplanes

Given a set of data points $V \subseteq \mathbb{R}^D$ the hyperplane location problem is to find a hyperplane H minimizing the distances to the data points in V. In this section we consider such hyperplane location problems for different types of distances and different globalizing functions.

Note that line location deals with finding a line in $\mathbb{R}^2$ minimizing the distances to a set of two-dimensional data 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 in different fields: Operations research, computational geometry, and in statistics and data science. 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; Baldomero-Naranjo et al. 2018). If the set of red and blue data points are not linearly separable, one may look for a hyperplane which minimizes the maximum distance to the data points on the wrong side. This problem can be solved as a restricted hyperplane location problem (see Carrizosa and Plastria 2008; Plastria and Carrizosa 2012).

In *statistics*, classical linear regression asks for a hyperplane which minimizes the residuals of a set of data points, usually the sum of squared vertical distances between the data points and the hyperplane. Orthogonal regression (also called total least squares, see Golub and van Loan 1980) calls for a hyperplane minimizing the sum of squared *Euclidean* distances as residuals.

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 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, or the least quantiles of squares, introduced in Bertsimas and Shioda (2007). 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 define the residuals, i.e., which is used to measure the distance from the data points to the hyperplane. The vector $\lambda \in \mathbb{R}^n$ specifies the ordered median function g_λ used for modeling the respective estimator. The meaning of the λ notation is extensively discussed in Nickel and Puerto (2005) or in

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)
Least r-quantile of squares	$d = d_{ver}^2$	$\lambda = (\underbrace{0, \ldots, 0}_{r-1}, 1, \underbrace{0, \ldots, 0}_{n-1})$

Chap. 10 of this book. More applications to classification and regression are pointed out in Bertsimas and Shioda (2007), Blanco et al. (2018).

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 $d(H_{a,b}, v) = \min\{d(x, v) : a^t x + b = 0, x \in \mathbb{R}^D\}$. 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 and the third case comprise the case of a hyperplane which is vertical itself. Its distance to a point v is infinity unless the hyperplane passes through v. If not all data 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. We remark that the vertical distance is the

most commonly used measure for determining the size of the residuals in regression theory and in statistics.

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 presented 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.,*

$$\gamma^\circ(v) = \sup\{v^t x : \gamma(x) \leq 1, x \in \mathbb{R}^D\}.$$

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

Transforming a point to a hyperplane: $T_H(v^1, \ldots, v^D) := H_{v^1, \ldots, v^{D-1}, 1, v^D}$

Transforming a hyperplane to a point: $T_P(H_{a^1, \ldots, a^{D-1}, 1, b}) := (a^1, \ldots, a^{D-1}, b)$

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 the following result.

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 $(a^1, \ldots, a^{D-1}, b)$.

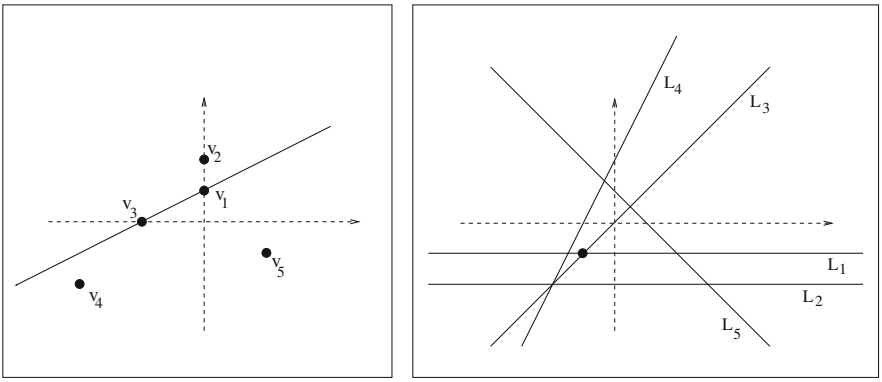


Fig. 7.2 Left: Five data points and a line in primal space. Right: The same situation in dual space corresponds to five lines and one point

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 data points (depicted in the left part of the figure), namely $v_1 = (0, \frac{1}{2})$, $v_2 = (0, 1)$, $v_3 = (-1, 0)$, $v_4 = (-2, -1)$ and $v_5 = (1, -\frac{1}{2})$. In the dual interpretation the data 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 data 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 $H_{-1,1,-1} = \{(x^1, x^2) : x^2 = x^1 + 1, x \in \mathbb{R}^D\}$ which passes through the three data 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* in which we use the sum of all residuals as globalizing function. It is defined as follows: Given a set of data 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 (or median hyperplane) w.r.t. the distance d*. Let us assume throughout this section that there are $n > D$ affinely independent data 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 the following index sets

$$\begin{aligned} J_{a,b}^{>} &:= \{j \in \{1, \ldots, n\} : v_j^t a + b > 0\} \\ J_{a,b}^{<} &:= \{j \in \{1, \ldots, n\} : v_j^t a + b < 0\} \\ J_{a,b}^{=} &:= \{j \in \{1, \ldots, n\} : v_j^t a + b = 0\}. \end{aligned} \tag{7.5}$$

We furthermore set

$$W_{a,b}^{>} := \sum_{j \in J_{a,b}^{>}} w_j, \quad W_{a,b}^{=} := \sum_{j \in J_{a,b}^{=}} w_j, \quad W_{a,b}^{<} := \sum_{j \in J_{a,b}^{<}} w_j$$

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 the following property which says that every minsum hyperplane splits the data points into two sets of almost equal weights.

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*

$$W^{>}_{a,b} \leq \frac{W}{2} \text{ and } W^{<}_{a,b} \leq \frac{W}{2} \tag{7.6}$$

Note that the halving property (7.6) is equivalent to

$$W^{>}_{a,b} \leq W^{<}_{a,b} + W^{=}_{a,b} \text{ and } W^{<}_{a,b} \leq W^{>}_{a,b} + W^{=}_{a,b}. \tag{7.7}$$

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*: There exists an optimal minsum hyperplane which passes through at least D of the data points and these points are affinely independent. Since D affinely independent points uniquely determine a hyperplane, the set of all $\binom{n}{D}$ such hyperplanes contains at least one optimal hyperplane and hence is a finite dominating set.

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 data points.*

Proof (Sketch of Proof) We can rewrite the objective function $f_1(H_{a,b})$ to

$$f_1(H_{a,b}) = \sum_{j \in J^{>}_{a,b}} w_j(v_j^t a + b) + \sum_{j : \in J^{<}_{a,b}} w_j(-v_j^t a - b) \tag{7.8}$$

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 $J^{\geq}, J^{\leq} \subseteq \{1, \ldots, n\}$ specifying which data points should be below (or on) and above (or on) the hyperplane:

$$R(J^{\geq}, J^{\leq}) := \Big\{ (a^1, \ldots, a^{D-1}, b) : v_j^t a + b \geq 0 \text{ for all } j \in J^{\geq} \\ v_j^t a + b \leq 0 \text{ for all } j \in J^{\leq} \Big\}.$$

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 $R(J^{\geq}, J^{\leq})$, 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 Distance

We now turn our attention to the location of hyperplanes with respect to a norm $\|\cdot\|$, i.e., the residuals are given as $d(v, H) = \min\{\|v - x\| : x \in H\}$. We can use Lemma 7.2 for computing the residuals and obtain the following objective function

$$f_1(H_{a,b}) = \sum_{j=1}^{n} w_j \frac{|v^t a + b|}{\|a\|^\circ} \tag{7.9}$$

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 distance and $H_{a,b}$ be a minsum hyperplane w.r.t. the distance d. Then*

$$W^+_{a,b} \leq \frac{W}{2} \text{ and } W^-_{a,b} \leq \frac{W}{2}$$

Also the incidence property of Theorem 7.2 still holds.

Theorem 7.4 (FDS for Minsum Hyperplanes) *(Schöbel 1999a; Martini and Schöbel 1998, 1999) Let d be a norm distance derived from norm $\|\cdot\|$ and let $n \geq D$. Then there exists a minsum hyperplane w.r.t. the distance d that passes through D affinely independent data points. If and only if $\|\cdot\|$ is a smooth norm, we have that* all *minsum hyperplanes pass through D affinely independent data points.*

Proof (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 $R(J^{\leq}, J^{\geq})$ in dual space. On any such region we obtain that the objective function (7.9) can be rewritten as

$$\begin{aligned} f_1(H_{a,b}) &= \sum_{j \in J^{>}_{a,b}} w_j \frac{v_j^t a + b}{\|a\|^\circ} + \sum_{j : \in J^{<}_{a,b}} w_j \frac{-v_j^t a - b}{\|a\|^\circ} \\ &= \frac{1}{\|a\|^\circ} \left(\sum_{j \in J^{>}_{a,b}} w_j (v_j^t a + b) + \sum_{j : \in J^{<}_{a,b}} w_j (-v_j^t a - b) \right), \end{aligned}$$

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 $R(J^{\leq}, J^{\geq})$, i.e., again at a hyperplane passing through D affinely independent data 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 Gauge Distance

In general, the results of Theorems 7.4 and 7.3 do not hold for gauges. 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 (7.7)) 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 distance and $H(a, b)$ be a minsum hyperplane w.r.t. the distance d. Then we have*

$$\sum_{j \in H^{<}_{a,b}} \frac{w_j}{\gamma^\circ(a)} \leq \sum_{j \in H^{>}_{a,b} \cup H^{=}_{a,b}} \frac{w_j}{\gamma^\circ(a)}$$

$$\sum_{j \in H^{>}_{a,b}} \frac{w_j}{\gamma^\circ(-a)} \leq \sum_{j \in H^{<}_{a,b} \cup H^{=}_{a,b}} \frac{w_j}{\gamma^\circ(-a)}.$$

For gauge distances it does also not hold that there always exists an optimal minsum hyperplane passing through D of the data 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 a gauge distance and let $n \geq D$. Then there exists a minsum hyperplane w.r.t. the distance d that passes through $D-1$ affinely independent data 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 use the maximum of the residuals as globalizing function. That is, we look for a hyperplane $H_{a,b}$ which minimizes

$$f_{\max}(H_{a,b}) = \max_{j=1,\dots,n} w_j d(H_{a,b}, v_j).$$

A hyperplane H minimizing $f_{\max}(H)$ is called *minmax hyperplane (or center hyperplane)* w.r.t. the distance 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 do not distinguish between vertical, norm- and gauge distances here.

Minmax point 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, relations to transversal theory may be exploited. We hence start with a link to computational geometry.

7.3.4.1 Relation to Transversal Theory

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 $\mathscr{M} = \{M_j(r), j = 1, \dots, n\}$ 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 data points, we can hence equivalently look for the smallest possible $r \geq 0$ such that a hyperplane transversal for the sets $M_j(r)$, $j = 1, \dots, n$ exists. As an example, in Fig. 7.3 we search a line minimizing the maximum rectangular distance to the five given data 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 $f_{\max}(l) \leq r$.

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. the*

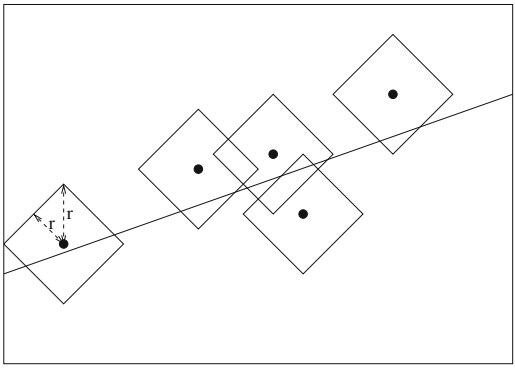


Fig. 7.3 A line transversal l to the five sets (each of them with radius r) exists, hence the objective function value of this line satisfies $f_{\max}(l) \leq r$

distance d that is at the same (maximum) distance from $D+1$ affinely independent data 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 data points.*

Proof (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, this result also holds for gauges. This was shown for $D = 2$ in Schöbel (1999a) and for arbitrary finite dimensions D 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 is 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 data 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 data points in V, i.e., there does not exist any other pair of parallel hyperplanes enclosing all data points and passing through the same data 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 data points. More effort is necessary to determine the hyperplanes being at maximum distance from $D+1$ of the data points for the minmax case. For $D = 2$ and norm distances these are parallel to one edge of the convex hull of the data points (Schöbel 1999a).

7.3.5.2 Linear Programming for Hyperplane Location with Vertical and Block Norm Distance

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

$$\text{minimize} \quad \sum_{j=1}^{n} w_j d_j \tag{7.10}$$

$$\text{subject to} \quad d_j \geq v_j^T a + b \text{ for } j = 1, \ldots, n \tag{7.11}$$

$$d_j \geq -v_j^T a - b \text{ for } j = 1, \ldots, n \tag{7.12}$$

$$d_j \geq 0 \text{ for } j = 1, \ldots, n \tag{7.13}$$

$$a^D = 1 \tag{7.14}$$

$$b, a^i \in \mathbb{R} \text{ for } i = 1, \ldots, D-1. \tag{7.15}$$

For the minmax problem, the objective (7.10) has to be replaced by the minmax 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 globalizing 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 data 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 (in linear time) 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).

The problem of locating a hyperplane with respect to a block norm distance can also be formulated as one large integer linear program (instead of the mentioned G linear programs) as done in Blanco et al. (2018).

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 to be described here could 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 data 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

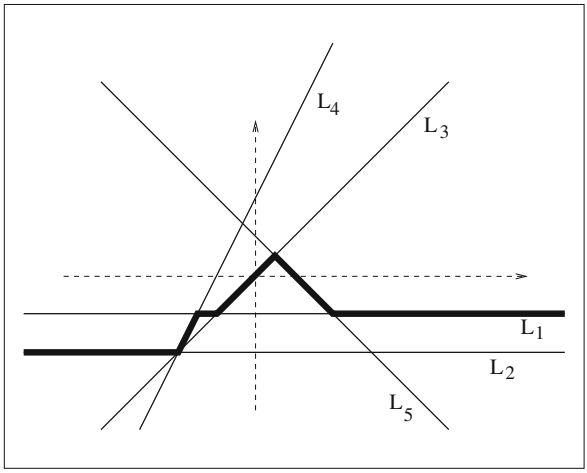


Fig. 7.4 The median trajectory for the example of Fig. 7.2

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 the uniform-gap on a circle problem (Yamamoto et al. 1988). That is, the order of $h(n)$ is at least $O(n \log n)$. 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 globalizing 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 a norm distance and let $n \geq 2$. Then there exists a solution l^* to the ordered line location problem w.r.t. the distance d that satisfies at least one of the following conditions:*

- *l^* passes through two of the data points.*
- *l^* passes through one of the data points and is at same weighted distance from two of the data points.*
- *l^* is at the same weighted distance from three of the data points.*
- *There exist two pairs of data 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 data points of each of the two pairs.

Proof (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 norm distances: Again, we look at the regions in dual space in which the order of the distances from the line to the data 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 $R(J^{\geq}, J^{\leq})$ of the proof of Theorem 7.4, we obtain quasiconcavity on the resulting (smaller) cells. This yields that the data 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 $K = |\{j = 1,\ldots,n : \lambda_j \neq 0\}|$.

Recently, a formulation with second-order cone constraints for ordered hyperplane location problems with arbitrary norm distances has been developed in Blanco et al. (2018). In the same paper, the authors also propose a formulation as mixed-integer linear program for the special case of ordered median hyperplane location problems with block norm distances.

A special case concerns the k-centrum line location problem, in which the sum of distances from the line to the k most distant data 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 data 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 data 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 data points, one may also consider an obnoxious problem in which the new facility should be as far away from the data points as possible. A rather general approach for obnoxious line location is presented in Lozano et al. (2015) 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 data 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 of the Euclidean anti-median line. The case of locating an obnoxious plane (i.e., finding the widest empty slab through a set of data 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 data point is served by its closest line. We may minimize the sum of distances

$$f_1(H_1, \ldots, H_p) = \sum_{j=1}^{n} w_j \min_{q=1,\ldots,p} d(H_q, v_j) \tag{7.16}$$

or the maximum distance

$$f_{\max}(H_1, \ldots, H_p) = \max_{j=1,\ldots,n} w_j \min_{q=1,\ldots,p} d(H_q, v_j) \tag{7.17}$$

from the data points to their closest hyperplanes, or we may use any other globalizing function. 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 hyperplanes 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 data 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 data points V. Let d be the vertical distance or a norm distance.*

- *If $n \geq D$ then there exists an optimal solution to the p-minsum-hyperplane location problem in which each hyperplane passes through D data 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 $D+1$ data points.*

Hence, enumeration approaches based on such an FDS are possible, however, the number of candidates to be enumerated is of order $O(n^D)$.

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 data 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.

Another possible approach is to use blockwise coordinate descent similar to the idea of Cooper's algorithm (Cooper 1964) and proceed iteratively: Start with a random set of p hyperplanes, determine the sets V_q for all $q = 1, \ldots, p$, re-optimize within these sets and repeat. The procedure converges to a local optimum. For a more detailed analysis of the convergence properties we refer to Jäger and Schöbel (2018).

Finally, the problem of finding p lines in the plane is studied as classification problem 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 data point v_j is assigned. Applying their basic formulation to the linear program (7.10)–(7.15) 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 as globalizing function in its objective.

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). Results of this type have been generalized to block norms and 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 data 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 data points. Plastria and Carrizosa (2012) consider the more general case of requiring a specified subset of data points below or above the hyperplane with applications in support vector machines.

7.3.7.4 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 data 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 hyper-

sphere S which minimizes the distances to the data points in V. The most common hypersphere is the surface of the Euclidean unit ball (i.e., a classical circle in two dimensions), but the problem is also interesting for more general hyperspheres derived from unit balls of other norms. In this section we consider such hypersphere location problems for different types of norms and different globalizing functions.

Note that circle location deals with finding a circle in $\mathbb{R}^2$ minimizing the distances from its circumference to a set of data 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 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), 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 archaeology 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 archaeology,
- 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 *hypersphere of 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 the distance from v to its closest point on S, i.e.,

$$d(S, v) = \min_{y \in S} d(y, v)$$

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 $n = 3$ data points in the plane ($D = 2$) is a surprisingly interesting problem. Within a wider context it has been studied in Alonso et al. (2012a,b). Here, the circumcircle of a set of three data points is investigated (which is the optimal minmax or minsum circle for the three data 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, i.e., we use the sum of all residuals between the data points and the hypersphere as globalizing function. Given a distance d derived from norm $\|\cdot\|$, the goal hence is to find a hypersphere $S = S_{x,r}$ of norm $\|\cdot\|$ which minimizes

$$f_1(S_{x,r}) = \sum_{j=1}^{n} w_j d(S_{x,r}, v_j) = \sum_{j=1}^{n} w_j |d(x, v_j) - r|. \tag{7.18}$$

For the Euclidean case in the plane, (7.18) reduces to the location of a circle in the plane. It 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 of some 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) *Consider the hyperplane location problem* (7.18) *with respect to a norm. Then the following hold.*

- *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. Similar to the index sets (7.5) used for hyperplane location, we define index sets to distinguish data points outside, on, and inside the hypersphere

$$J_{x,r}^{>} := \{j \in \{1, \ldots, n\} : d(x, v_j) > r\}$$
$$J_{x,r}^{<} := \{j \in \{1, \ldots, n\} : d(x, v_j) < r\}$$
$$J_{x,r}^{=} := \{j \in \{1, \ldots, n\} : d(x, v_j) = r\}$$

and let

$$W^{>}_{x,r} := \sum_{j \in J^{>}_{x,r}} w_j, \quad W^{=}_{x,r} := \sum_{j \in J^{=}_{x,r}} w_j, \quad W^{<}_{x,r} := \sum_{j \in J^{<}_{x,r}} w_j.$$

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. a norm distance. Then*

$$W^{>}_{x,r} \leq \frac{W}{2} \text{ and } W^{<}_{x,r} \leq \frac{W}{2} \tag{7.19}$$

Proof (Sketch of Proof) If we increase the radius from r to $r + \epsilon$ the distance to data points with indices in $J^{>}_{x,r}$ decreases by ϵ, and the distance to data points with indices in $J^{<}_{x,r}$ 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 from hyperplane location problems to hypersphere location problems, this is unfortunately not true for the determination of a finite dominating set. This can already bee seen in the Euclidean case for $D = 2$, i.e., for locating a circle in the plane: Here, the generalization of Theorem 7.4 would be that there always exists an optimal Euclidean circle passing through three of the data points. However, this turns 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 of the data points. The only incidence property that can be shown for arbitrary norms is the following.

Lemma 7.7 *Let d be a norm distance. Then there exists a minsum hypersphere w.r.t. the distance d which passes through at least one point $v \in V$.*

Proof (Sketch of Proof) Let $S_{x,r}$ be a hypersphere. Fix its center point x and assume without loss of generality that the data 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 (7.18) 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 data points in at least two data 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 Lemma 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 data points of V.*

The result can be 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 data points cannot be a local minimum.

An algorithmic consequence of the Theorem 7.11 is that there exists an optimal circle with center point x being on a bisector of two of the data 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 data points 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).

7.4.3.2 Location of Minsum Circles and Hyperspheres with Block Norm Distance

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 data points $v \in V$ and add the bisectors for all pairs of data points in V. The intersection points of these lines form a finite dominating set which can be tested within $O(n^3)$ time, see Körner (2011), Brimberg et al. (2011a).

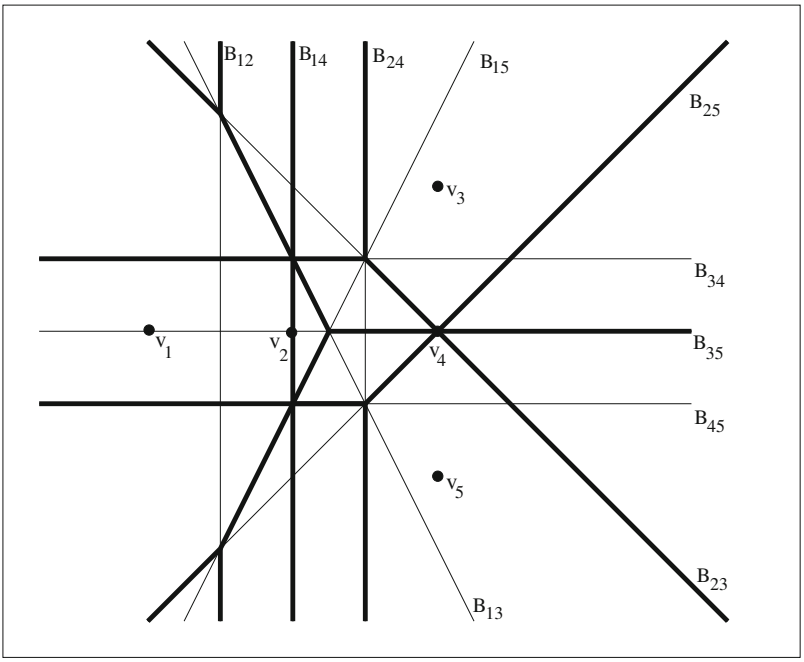


Fig. 7.5 The Euclidean bisectors for five data points. The notation B_{ij} indicates that the corresponding line is the bisector for data points v_i and v_j. The parts of the bisectors which may contain a center point of a minsum circle are marked in bold

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 norm distances 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.

$$\text{minimize} \quad \sum_{j=1}^{n} w_j \left(z_j^+ + z_j^-\right)$$

$$\text{subject to} \quad \sum_{g=1}^{G} \alpha_{g,j} = r + z_j^+ - z_j^- \text{ for } j = 1, \ldots, n$$

$$\sum_{g=1}^{G} \alpha_{g,j} e_g = x - v_j \text{ for } j = 1, \ldots, n$$

$$z_j^+, z_j^- \geq 0 \text{ for } j = 1, \ldots, n$$
$$\alpha_{g,j} \geq 0 \text{ for } g = 1, \ldots, G, j = 1, \ldots, n$$
$$r \geq 0$$
$$x \in \mathbb{R}^D.$$

7.4.4 The Minmax Hypersphere Location Problem

We now turn our attention to the location of a minmax hypersphere using the maximum of the residuals as globalizing function. That is, we look for a hypersphere which minimizes the maximum weighted distance to the set V of data points. Given a norm distance d, the goal hence is to find a hypersphere $S = S_{x,r}$ which minimizes

$$f_{\max}(S_{x,r}) = max_{j=1}^{n} w_j d(S_{x,r}, v_j) = \sum_{j=1}^{n} w_j |d(x, v_j) - r|. \tag{7.20}$$

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 data points it may be interpreted as finding an annulus of minimal width covering all data 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

$$\max_{j=1,\ldots,n} d(x, v_j) - \min_{j=1,\ldots,n} d(x, v_j),$$

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. 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 data 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 data points has two data points on its inner circumference and two data 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 data 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 data 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 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 data 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) Let d be the rectangular distance. Then there exists a minmax circle whose center point is a center point of a smallest enclosing square of the data points.*

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.

More 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 data 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 data points of V and passing through the same data 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 residuals as globalizing function*, 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 data 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), Nievergelt (2004).

7.4.5.2 Locating Euclidean Concentric Circles

In a recent paper (Drezner and Brimberg 2014), introduce the following 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 data 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 globalizing function, as well as for using minsum and minmax. 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 data points separately

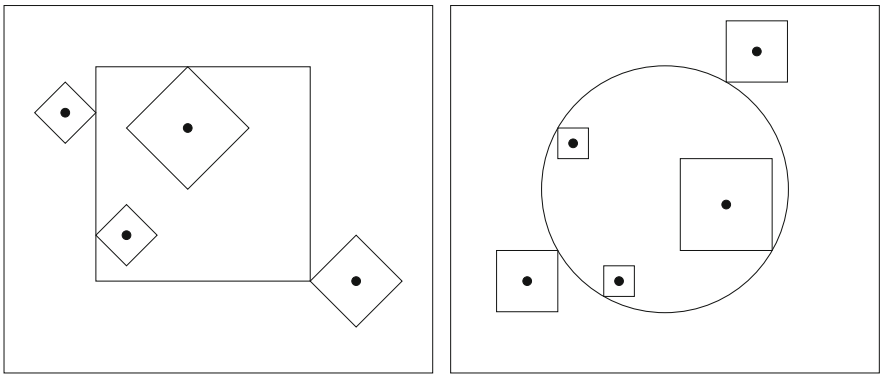


Fig. 7.6 Locating a circle of norm k_1 with respect to another norm k_2. Left: The unit circle of the maximum norm is to be located, distances are measured w.r.t. the rectangular norm. Right: The Euclidean circle is to be located, distances are measured w.r.t. the maximum norm

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 Locating a Hypersphere of One Norm Measuring Distances with Respect to Another Norm

In two dimensions, the circle location problem 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 data points in a set V are minimized, where the residuals are measured with respect to the same norm $\|\cdot\|$. 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 data points V in the plane with positive weights $w_j > 0$, the goal of *generalized hypersphere location* is to locate and scale a hypersphere of norm k_1 such that the sum of weighted distances to the data points is minimized, where the 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 data 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 in $D = 2$ dimensions 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 data points. However, if both norms k_1 and k_2 are block norms, a finite dominating set can 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 data 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.

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.

7.5.2 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 data points $V = \{v_1, \ldots, v_n\}$, and that partitions the data 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 data points. 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). 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).

7.5.3 Two-Dimensional Facilities

Covering problems are the most common problems in which the location of full-dimensional facilities is considered. There exist many papers about covering data points by a circle (i.e., locating one point x such that all data 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 globalizing function is minimized, has been considered in the literature. If there exists a location for X such that all data 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 data 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 globalizing function has been studied in Savas et al. (2002), Kelachankuttu et al. (2007) and Sarkar et al. (2007). In these papers, 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. (2015a) both for the minsum and the minmax globalizing 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 data 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 data points that are within an acceptable travel distance r from x. Thus, costs will be associated only to those data 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. (2015a) where the behavior of the optimal solution with respect to the threshold distance r is studied.

Line location with the partial coverage globalizing function is equivalent to locating a strip of given width and has recently been considered in Brimberg et al. (2015b).

7.5.4 General Approaches for Locating Dimensional Facilities

Blanquero et al. (2009) and Mallozzi et al. (2019) both deal with the location of a variety of dimensional facilities such as segments, arcs of circumferences, arbitrary convex and non-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/or an angle of rotation. The objective they follow is very general, including most globalizing functions used in location theory.

Blanquero et al. (2009) also allow to model obnoxious or semi-obnoxious location problems as follows: The set of data points 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 a data 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.

Mallozzi et al. (2019) deal with the location of p dimensional facilities of very general shapes and the allocation of them to some given demand. Instead of a distance measure, utility functions are used. The resulting location-allocation problem is discretized and tools from optimal mass transport are used for its solution.

7.6 Conclusions

For the location of dimensional facilities we can draw the following conclusions.

- The location of a zero-dimensional facility (i.e., a point) and of a full-dimensional facility of convex shape with respect to a norm is a convex problem.
- 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 globalizing 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, hyperplane and hypersphere location problems with ordered median globalizing function are 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.

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.

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

Enumeration may be enhanced by the halving property which can be used to directly discard candidates of an FDS. 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), Big-Cube-Small-Cube (Schöbel and Scholz 2010) or Big-Arc-Small-Arc (Drezner et al 2018) 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 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 globalizing 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 Assoc Comput Mach 51:605–635

Alonso J, Martini H, Spirova M (2012a) Minimal enclosing discs, circumcircles, and circumcenters in normed planes (part i). Comp Geom-Theor Appl 45:258–274

Alonso J, Martini H, Spirova M (2012b) Minimal enclosing discs, circumcircles, and circumcenters in normed planes (part ii). Comp Geom-Theor Appl 45:350–369

Baldomero-Naranjo M, Marténez-Merino LI, Rodríguez-Chía AM (2018) Exact and heuristic approaches for support vector machine with l1 ramp loss. EWGLA XXIV

Bennet K, Mangasarian O (1992) Robust linear programming discrimination of two linearly inseparable sets. Optim Methods Softw 1:23–34

Bertsimas D, Shioda R (2007) Classification and regression via integer optimization. Oper Res 55:252–271
Blanco V, Puerto J, Salmerón, R (2018) A general framework for locating hyperplanes to fitting set of points. Comput Oper Res 95:172–193
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 Logist 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. Discret 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 (2015a) On models for continuous facility location with partial coverage. J Oper Res Soc 66(1):33–43
Brimberg J, Schieweck R, Schöbel A (2015b) Locating a median line with partial coverage distance. J Glob Optim 62(2):371–389
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. ACM, New York, pp 300–309
Cheng SW (1996) Widest empty L-shaped corridor. Inf 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
Cooper L (1964) Heuristic methods for location-allocation problems. SIAM Rev 6:37–53
Crawford J (1983) A non-iterative method for fitting circular arcs to measured points. Nucl Instrum Methods Phys Res 211:223–225
Das G, Mukhopadhyay D, Nandy S (2009) Improved algorithm for the widest empty 1-corner corridor. Inf 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. Discret 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. Inf 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, 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. In: Drezner Z, Hamacher H (eds) Facility location: applications and theory. Springer, Berlin, chap 1, 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
Drezner T, Drezner Z, Schöbel A (2018) The Weber obnoxious facility location model: a Big Arc Small Arc approach. Comput Oper Res 98:240–250
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. Discret 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 Inc., 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. Discret Comput Geom 20:389–402
Gluchshenko O (2008) Annulus and center location problems. PhD 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. Jahrbuch der Deutsch Math Verein 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
Jäger S, Schöbel A (2018) The blockwise coordinate descent method for integer programs. Math Meth Oper Res 2019:1–25. Preprint Institut für Numerische und Angewandte Mathematik, Georg-August-Universität Göttingen: 2018-15
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 Trans 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. Lect Notes Comput Sci 447:276–286
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 Glob Optim 51:115–132
Körner MC, Martini H, Schöbel A (2012) Minisum hyperspheres in normed spaces. Discret Appl Math 16:2221–2233
Krempasky T (2012) Locating median lines and hyperplanes with a restriction on the slope. PhD Thesis, Universität Göttingen, Göttingen
Le V, Lee D (1991) Out-of-roundness problem revisited. IEEE Trans Pattern Anal Mach Intell 13:217–223
Lee D, Ching Y (1985) The power of geometric duality revisited. Inf 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 Algorithm 9:1–17
Lozano AJ, Mesa J, Plastria F (2015) Location of weighted anti-ordered median straight lines with euclidean distances. Discrete Appl Math. 182:122–133 https://doi.org/10.1016/j.dam.2013.04.016
Mallozzi L, Puerto J, Rodríguez-Madrena M (2019) On location-allocation problems for dimensional facilities. J Optim Theory Appl 182(2):730–767
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. Discret 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 Assoc Comput Mach 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 Algebra Discr 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
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, 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
Robins G, Shute C (1987) The Rhind mathematical papyrus. An ancient Egyptian text. British Museum, London
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, Schöbel A (2012) Properties and algorithms for line location with extensions. In: Proceedings of the 28th European Workshop on computational Geometry, 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. No. 25 in applied optimization series. Kluwer, Dordrecht
Schöbel A (1999b) Solving restricted line location problems via a dual interpretation. Discret Appl Math 93:109–125
Schöbel A (2003) Anchored hyperplane location problems. Discret 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, June 2–8
Swanson K, Lee DT, Wu V (1995) An optimal algorithm for roundness determination on convex polygons. Comp 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, Piscataway, pp 690–700
Wesolowsky G (1972) Rectangular distance location under the minimax optimality criterion. Transport Sci 6:103–113
Wesolowsky G (1975) Location of the median line for weighted points. Environ Plann 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. Inf Process Lett 18:123–128

Chapter 8
Facility Location Under Uncertainty

Isabel Correia and Francisco Saldanha-da-Gama

Abstract This chapter covers some of the existing knowledge on facility location under uncertainty. The goal is to provide the reader with essential tools for modeling and tackling problems in the area. To a large extent, the focus is put on discrete facility location problems. Several issues related with uncertainty are discussed. A distinction is made between problems in the areas of robust optimization, stochastic programming and chance-constrained programming. The presentation is complemented with several other aspects of relevance such as multi-stage stochastic programming models, scenario generation, and solution techniques. Several well-known facility location problems are used throughout the chapter for illustrative purposes.

8.1 Introduction

Many facility location problems involve strategic decisions that must hold for a considerable amount of time, during which uncontrolled changes may occur in the conditions underlying the problem. For example, 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 and thus make

I. Correia (✉)
Departamento de Matemática e Centro de Matemática e Aplicações, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal
e-mail: isc@fct.unl.pt

F. Saldanha-da-Gama
Departamento de Estatística e Investigação Operacional, Centro de Matemática, Aplicações Fundamentais e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal
e-mail: fsgama@ciencias.ulisboa.pt

G. Laporte et al. (eds.), *Location Science*,
https://doi.org/10.1007/978-3-030-32177-2_8

decisions that can somehow anticipate it. This can be accomplished by embedding uncertainty in mathematical models developed for supporting decision making processes.

The review papers by Louveaux (1993) and Snyder (2006) show that much work has been done within the context of facility location under uncertainty. The different sources of uncertainty we may face in these problems have led to the development of different research branches. One of them consists of so-called problems with congestion. In this case, the customers' requests for service have a probabilistic behavior. If a facility is busy when a new request arrives then we say that "congestion" occurs. This is the topic covered by Chap. 17. Another important research direction regards unexpected disruptions in the network structures, e.g., in the facilities or in the transportation channels. This is a topic addressed in detail in Chap. 22. In the current chapter, we focus on a third perspective: we consider the aspects emerging from uncertainty associated with the parameters of a facility location problem such as the demand levels or transportation costs. We show how uncertainty can be embedded in optimization models aiming at supporting a decision making process. For illustrative purposes, we work with several well-known problems. We focus on a discrete setting, i.e., we assume that there is a finite set of candidate locations for the facilities. This is motivated by the practical relevance that this setting has gained overt time, which stems from 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 the basic concepts of robust and stochastic optimization. Important references in these fields include Birge and Louveaux (2011) and Shapiro et al. (2009) for stochastic programming; 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 an overview of the contents presented.

8.2 Uncertainty Issues

Basic information underlying a facility location problem often includes demand levels, travel time, 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 and thus they are jointly represented by a random vector. 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 robustness measures, as 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, a probability can often be associated with each scenario. Depending on the problem, we may have a finite or an infinite number of scenarios. As will be discussed later, this impacts the models and techniques that can be used.

One important feature that influences the optimization model to be considered for a specific problem regards the attitude of the decision maker towards risk. Two attitudes are usually 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 a decision that 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 a 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 decisions that must be implemented before uncertainty is revealed—also called the here-and-now decisions; 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 implementation before uncertainty is revealed. Regarding the allocation or distribution decisions, it will depend on the specific problem being studied whether they are *ex ante* or *ex post* decisions. In the following sections we refer to 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. In this case, two classical objectives are often considered: minmax cost and minmax regret.

For illustrative purposes, we consider the well-known 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 problem can be formulated mathematically 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 objective 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 specifies all 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 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 this case, 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 a very relevant aspect: when the optimal objective function differs significantly across the different scenarios, the relative regret is a more appropriate robustness measure (see also Kouvelis and Yu 1997). In this case, (8.13) should be replaced with

$$\frac{\sum_{i \in J} \sum_{j \in J} d_{j\omega} a_{ij\omega} x_{ij\omega} - v^*_\omega}{v^*_\omega} \leq v, \quad \omega \in \Omega. \tag{8.14}$$

Serra and Marianov (1998) developed a heuristic for this problem.

A different problem is studied 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 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

[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).

in model (8.1)–(8.5). Furthermore, the location variables y_i are no longer necessary, since the variables x_{ii} $(i \in J)$ can play their role.

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 such 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 both cases considering the uncapacitated facility location problem (UFLP), whose well-known mathematical 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 unit cost for supplying the demand of customer $j \in J$ from facility $i \in I$ and d_j is 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 $\mathscr{U}_j^B = [\overline{d}_j - \epsilon\Delta_j, \overline{d}_j + \epsilon\Delta_j]$, $0 \leq \epsilon \leq 1$. 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; Δ_j is a scaling factor.

A particular case of box uncertainty arises when $\Delta_j = \overline{d}_j$ $(j \in J)$, which leads to the intervals $\mathscr{U}_j^B = [\overline{d}_j(1-\epsilon), \overline{d}_j(1+\epsilon)]$, $j \in J$. Denote $\mathscr{U}^B = \mathscr{U}_1^B \times \cdots \times \mathscr{U}_{|J|}^B$ and d the vector of demands, $d = (d_1, \ldots, d_{|J|})'$. We can write

$$\mathscr{U}^B = \{d \in \mathbb{R} \mid -1 \leq \frac{d_j - \overline{d}_j}{\epsilon \overline{d}_j} \leq 1,\ \forall j \in J\},$$

i.e., the multi-dimensional unit box is given by the absolute normalized deviations (Baron et al. 2011). We can now formulate the so-called robust counterpart of model

(8.15)–(8.19). To do so, we start by considering an auxiliary variable v, which allows us to rewrite the objective function of the problem as

$$\text{Minimize} \quad v. \tag{8.20}$$

The following constraint must now be included in the model:

$$\sum_{i \in I} f_i y_i + \sum_{i \in I} \sum_{j \in J} c_{ij} d_j x_{ij} \leq v. \tag{8.21}$$

By considering an augmented constraint for (8.21), namely

$$\sum_{i \in I} f_i y_i + \max_{d \in \mathscr{U}^B} \left\{ \sum_{i \in I} \sum_{j \in J} c_{ij} d_j x_{ij} \right\} \leq v, \tag{8.22}$$

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 their worst values simultaneously. This is often not realistic.

Nikoofal and Sadjadi (2010) avoid the too conservative solutions often arising from considering box uncertainty by imposing a maximum total scaled variation for the uncertain parameters. The authors consider 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 constraint

$$\sum_{i,j \in J,\, i<j} (a_{ij} - \underline{a}_{ij}) / (\overline{a}_{ij} - \underline{a}_{ij}) \leq L,$$

where L denotes a maximum level imposed for the total scaled variation. This type of constraint avoids the situation in which all or several parameters take their extreme values simultaneously.

Another alternative for overcoming the above-mentioned drawback when using box uncertainty is to consider ellipsoidal sets. Baron et al. (2011) apply this idea to a facility location problem with a time-varying (uncertain) demand. The location of the facilities and their operating capacity are *ex ante* decisions and should hold for the entire planning horizon, during which the demands must be satisfied. The goal is

to maximize the overall profit. We illustrate the process using the UFLP. Ellipsoidal uncertainty can be embedded in a model by defining the following uncertainty set

$$\mathscr{U}^E = \{d \in \mathbb{R}^{|J|} \mid \sum_{j \in J} \left[\frac{d_j - \overline{d}_j}{\epsilon \overline{d}_j} \right]^2 \leq L^2\} = \left\{ d \in \mathbb{R}^{|J|} \mid (d - \overline{d})^T \Lambda^{-1} (d - \overline{d}) \leq L^2 \right\},$$

with d being the demand vector already presented, L being a parameter and $\Lambda_{|J| \times |J|}$ being a diagonal matrix whose generic entry is $\sigma_j = \epsilon \overline{d}_j$. Since Λ is a positive definite matrix, the set $\mathscr{U}^E$ defines an ellipsoid. As pointed out by Baron et al. (2011), the set induced by $L = 1$ is the largest ellipsoid contained in $\mathscr{U}^B$ while the set induced by $L = \sqrt{|J|}$ is the smallest ellipsoid containing $\mathscr{U}^B$.

Under ellipsoidal uncertainty the augmented constraint for (8.21) is similar to (8.22) but replacing $\mathscr{U}^B$ with $\mathscr{U}^E$. Denote $V_j = \sum_{i \in I} c_{ij} x_{ij}$ and $V = (V_1, \ldots, V_{|J|})'$. The augmented constraint can be written as $\sum_{i \in I} f_i y_i + \max_{d \in \mathscr{U}^E} V'd \leq v$.

The problem that consists of finding a value $d \in \mathscr{U}^E$ maximizing $V'd$ can be easily solved by standard optimization techniques. The optimal solution is $V'\overline{d} + L\sqrt{V'\Lambda V}$. This leads to the following robust counterpart of (8.21):

$$\sum_{i \in I} f_i y_i + \sum_{j \in J} \overline{d}_j V_j + L \sqrt{\sum_{j \in J} \sigma_j^2 V_j^2} \leq v, \tag{8.24}$$

The non-linearity in the above expression is typically handled by introducing a new variable, $W = \sqrt{\sum_{j \in J} \sigma_j^2 V_j^2}$, which allows casting the problem as a conic programming problem (see Baron et al. (2011) and the references therein for further details).

In all problems discussed above, no probabilities were associated with the scenarios. However, in some situations, a probability π_ω can be associated to scenario $\omega \in \Omega$. A well-known robustness measure in this case, is the expected cost, which is equivalent to the expected regret (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 (whose cardinality is endogenously determined) during a planning horizon previously defined. 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 than or equal to a parameter 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:

$$\text{Minimize} \quad \sum_{\omega\in\Omega}\sum_{i\in J}\sum_{j\in J}\pi_\omega d_{j\omega}a_{ij\omega}x_{ij\omega} \tag{8.25}$$

$$\text{subject to} \quad (8.8)\text{–}(8.12)$$

$$\sum_{i\in J}\sum_{j\in J} d_{j\omega}a_{ij\omega}x_{ij\omega} \le (1+\alpha)v^*_\omega, \quad \omega\in\Omega. \tag{8.26}$$

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 Lagrangean relaxation based procedures 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) who 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 equal to 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). This is a value

indicating how much the decision maker would be willing to pay to obtain perfect information. For an expected cost minimization problem, 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 can refer to Kouvelis and Yu (1997) for further details.

8.4 Stochastic Facility Location Problems

A facility location problem under uncertainty can often be cast within a stochastic programming modeling framework if we know the joint probability distribution of the underlying random vector. 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 stochastic 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 at which the 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:

$$\text{Minimize} \quad \sum_{i\in I} f_i y_i + \mathrm{Q}(y) \tag{8.27}$$

$$\text{subject to} \quad \sum_{i\in I} y_i \geq 1 \tag{8.28}$$

$$y_i \in \{0,1\}, \quad i \in I, \tag{8.29}$$

with $\mathrm{Q}(y) = \mathbb{E}_{\Xi}\left[Q(y,\xi)\right]$, and $Q(y,\xi)$ denoting the optimal value of the following problem:

$$\text{Minimize} \quad \sum_{i\in I}\sum_{j\in J} c_{ij} d_j x_{ij} \tag{8.30}$$

$$\text{subject to} \quad \sum_{i\in I} x_{ij} = 1, \quad j \in J \tag{8.31}$$

$$x_{ij} \leq y_i, \quad i \in I,\ j \in J \tag{8.32}$$

$$x_{ij} \geq 0, \quad i \in I,\ j \in J. \tag{8.33}$$

Model (8.30)–(8.33) 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 according to the different observations of the random vector. For this reason, they are referred to as recourse decisions. Regarding the variables y_i associated with the location of the facilities they correspond to *ex ante* (first-stage) decisions and hence 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 toward risk. Later in this section, we discuss another possible attitude and the corresponding consequences from a modeling point of view. Finally, due to the presence of Constraint (8.28) 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 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 since we can consider scenario-indexed parameters and variables. Denote by $c_{ij\omega}$ the unit 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:

$$\text{Minimize} \quad \sum_{i \in I} f_i y_i + \sum_{\omega \in \Omega} \pi_\omega \left(\sum_{i \in I} \sum_{j \in J} c_{ij\omega} d_{j\omega} x_{ij\omega} \right) \tag{8.34}$$

$$\text{subject to} \quad (8.28), (8.29)$$

$$\sum_{i \in I} x_{ij\omega} = 1, \quad j \in J, \ \omega \in \Omega \tag{8.35}$$

$$x_{ij\omega} \leq y_i, \quad i \in I, \ j \in J, \ \omega \in \Omega \tag{8.36}$$

$$x_{ij\omega} \geq 0, \quad i \in I, \ j \in J, \ \omega \in \Omega. \tag{8.37}$$

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, facilities are assumed to be uncapacitated. When this is not the case, 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

[2] A decision should depend only on the information available at the time it is made (see Rockafellar and Wets 1991).

if we replace (8.32) with

$$\sum_{j \in J} d_j x_{ij} \leq q_i y_i, \quad i \in I. \tag{8.38}$$

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 of dealing with this issue is to penalize the non-satisfied demand, which makes sense from a practical point of view. In fact, such a penalty may correspond, for example, to a lost opportunity cost or to outsourcing. Denote by ψ_j the demand of customer $j \in J$ which is not supplied from the open facilities and denote by μ_j the corresponding unit penalty cost. Note that ψ_j is also a random variable since it depends on the occurring realization of the random vector Ξ. We can still consider the first stage problem (8.27)–(8.29). However, the second stage problem must be rewritten as follows:

$$\text{Minimize} \quad \sum_{i \in I} \sum_{j \in J} c_{ij} d_j x_{ij} + \sum_{j \in J} \mu_j \psi_j \tag{8.39}$$

$$\text{subject to} \quad (8.33), (8.38)$$

$$\sum_{i \in I} x_{ij} + \frac{\psi_j}{d_j} = 1, \quad j \in J \tag{8.40}$$

$$\psi_j \geq 0, \quad j \in J. \tag{8.41}$$

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, capacities must be set in advance before uncertainty is disclosed—they correspond to *ex ante* decisions. A unit cost g_i is assumed for the capacity to be installed at location $i \in I$. Additionally, 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 unit 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 introduced, representing the capacity to be installed at location $i \in I$. With the inclusion of revenues, it is no longer necessary to consider constraint (8.28). Furthermore, it may not be rewarding to satisfy all the demand; the trade-off between revenues and costs will determine 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):

$$\text{Minimize} \quad \sum_{i \in I} f_i y_i + \sum_{i \in I} g_i z_i + \mathrm{Q}(y, z) \tag{8.42}$$

$$\text{subject to} \quad (8.29)$$

$$z_i \geq 0, \quad i \in I, \tag{8.43}$$

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:

$$\text{Minimize} \quad \sum_{i \in I} \sum_{j \in J} \left(c_{ij} - r_j\right) d_j x_{ij} \tag{8.44}$$

$$\text{subject to} \quad \sum_{i \in I} x_{ij} \leq 1, \quad j \in J \tag{8.45}$$

$$(8.32), (8.33)$$

$$\sum_{j \in J} d_j x_{ij} \leq z_i, \quad i \in I. \tag{8.46}$$

Considering the above problem, Louveaux and Peeters (1992) assume that stochasticity is captured by a finite number of scenarios and propose a dual-based procedure for tackling the extensive form of the deterministic equivalent.

A different type of model emerges when the distribution decisions (represented by x-variables) become first-stage decisions. In this case, penalties are paid in the second stage for surplus or shortage inventory. In addition to the notation already presented, we denote by ϕ_j the inventory surplus at customer $j \in J$ and by λ_j the corresponding unit cost. Assuming deterministic distribution costs (they are now associated with an *ex ante* decision), we can formulate the stochastic facility location problem as follows:

$$\text{Minimize} \quad \sum_{i \in I} f_i y_i + \sum_{i \in I} \sum_{j \in J} c_{ij} x_{ij} + \mathrm{Q}(x) \tag{8.47}$$

$$\text{subject to} \quad (8.29), (8.32), (8.33),$$

with $\mathrm{Q}(x) = \mathbb{E}_{\Xi}[Q(x, \xi)]$, and $Q(x, \xi)$ denoting the optimal value of the following problem:

$$\text{Minimize} \quad \sum_{j \in J} \lambda_j \phi_j + \sum_{j \in J} \mu_j \psi_j \tag{8.48}$$

$$\text{subject to} \quad \psi_j - \phi_j = d_j \left(1 - \sum_{i \in I} x_{ij}\right), \qquad j \in J \tag{8.49}$$

$$\psi_j, \phi_j \geq 0, \qquad j \in J. \tag{8.50}$$

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 and Holmberg and Tuy 1999).

So far in this section, we have assumed that the allocation and distribution decisions are made simultaneously (the latter determining the former), either after or before uncertainty is disclosed. Nevertheless, in some problems these decisions are made separately. Let us assume that the allocation of the customers to the facilities is a here-and-now decision but the exact 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 shipping quantities dependent on the observed values of the stochastic parameters. The same type of situation occurs in service-providing companies that need to segment the customers *a priori* by allocating each customer to a server or facility. In this case, we need to explicitly consider allocation decision variables. In particular, we use the binary variable w_{ij} equal to 1 if and only if customer $j \in J$ is allocated to facility $i \in I$. The single-allocation version of the problem was introduced by Laporte et al. (1994), who proposed the following optimization model:

$$\text{Minimize} \quad \sum_{i \in I} f_i y_i + \sum_{i \in I} \sum_{j \in J} b_{ij} w_{ij} + \mathrm{Q}(w) \tag{8.51}$$

$$\text{subject to} \quad w_{ij} \leq y_i, \quad i \in I,\ j \in J \tag{8.52}$$

$$\sum_{i \in I} w_{ij} \leq 1, \quad j \in J \tag{8.53}$$

$$y_i, w_{ij} \in \{0, 1\}, \quad i \in I,\ j \in J, \tag{8.54}$$

with $\mathrm{Q}(w) = \mathbb{E}_{\Xi}[Q(w, \xi)]$, and $Q(w, \xi)$ denoting the optimal value of the following problem:

$$\text{Minimize} \quad \sum_{i \in I} \sum_{j \in J} \left(c_{ij} - r_j\right) d_j x_{ij} \tag{8.55}$$

$$\text{subject to} \quad x_{ij} \leq w_{ij}, \quad i \in I,\ j \in J \tag{8.56}$$

$$\sum_{j \in J} d_j x_{ij} \leq q_i, \quad i \in I \tag{8.57}$$

$$x_{ij} \geq 0, \quad i \in I,\ j \in J. \tag{8.58}$$

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. Moreover, a service level of 100% is not imposed—a customer may not be served by the system (constraints (8.53)). Laporte et al. (1994) consider a finite set of scenarios for capturing the stochasticity and solved 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 next, is important to show that deriving a compact model for the deterministic equivalent problem is not always straightforward (or even possible) as it could seem at a first glance when considering the contents presented so far in this section.

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 q_i of customers who 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. However, depending on how uncertainty is revealed, 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 is paid. 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 ℓ_i of customers allocated to it to justify its establishment. The problem can be conceptually formulated as follows:

$$\text{Minimize} \quad \sum_{i \in I} f_i y_i + \mathbb{E}_{\Xi}\left[\text{Service cost} + \text{Outsourcing cost}\right] \tag{8.59}$$

$$\text{subject to} \quad \sum_{i \in I} x_{ij} = 1, \quad j \in J \tag{8.60}$$

$$x_{ij} \leq y_i, \quad i \in I,\ j \in J \tag{8.61}$$

$$\ell_i y_i \leq \sum_{j \in J} x_{ij}, \quad i \in I \tag{8.62}$$

$$y_i, x_{ij} \in \{0, 1\}, \quad i \in I,\ j \in J. \tag{8.63}$$

Denote by ξ_j the demand of customer $j \in J$, which is assumed to be a random variable following 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 who request the service (refereed 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 since 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), investigate two possible outsourcing actions. We focus on 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 customer 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 the facility is s (i.e., $\eta_i = s$). We have $\mathbb{P}_i(s) = (1/s) \times \min\{q_i, s\}$.

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:

$$\begin{aligned}\mathbb{E}_\xi(\text{service cost}) &= \sum_{i\in I}\sum_{s=0}^{z_i} \mathbb{P}_x(\eta_i = s) \times \mathbb{E}(\text{Service cost}|\eta_i = s) \\ &= \sum_{i\in I}\sum_{s=0}^{z_i} \left[\mathbb{P}_x(\eta_i = s) \sum_{j\in J} \mathbb{P}(\xi_j = 1|\eta_i = s)\mathbb{P}_i(s)c_{ij}x_{ij} \right], \end{aligned} \tag{8.64}$$

$$\begin{aligned}\mathbb{E}_\xi(\text{Outsourcing cost}) &= \sum_{i\in I} \mathbb{P}_x(\eta_i = s) \times \mathbb{E}_\xi(\text{outsourcing cost}|\eta_i = s) \\ &= \sum_{i\in I} g_i \left(\sum_{s=q_i+1}^{z_i} \mathbb{P}_x(\eta_i = s)(s - q_i) \right). \end{aligned} \tag{8.65}$$

A close look at the above expressions reveals that even for tiny instances of the problem they are not tractable. In fact, the number of scenarios is huge even for a small number of customers because a scenario is defined not only by the set of customers requesting the service but also by the order the requests arrive. Nevertheless, for the homogeneous case, i.e., $p_j = p$, $j \in J$, it is possible to go farther and derive a compact formulation for the 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 $\mathbb{P}(\xi_j = 1|\eta_i = s) \times \mathbb{P}_i(s) = \min\{q_i, s\}/t$, which does not depend on x. Accordingly, the expected service cost (8.64) 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 discretizing the 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 if a facility is located at $i \in I$ and t customers in total are allocated to it ($t = \ell_i, \ldots, |J|$) and 0 otherwise. Also define x_{ij}^t as a binary variable equal to 1 if and only if customer $j \in J$ is allocated to facility $i \in I$ which has t customers allocated to it ($t = \ell_i, \ldots, |J|$). Using the new variables, we can formulate a deterministic equivalent problem:

$$\text{Minimize} \quad \sum_{i \in I} \sum_{t=\ell_i}^{|J|} y_i^t g_i \left[\sum_{s=q_i+1}^{t} \zeta_{tps}(s - q_i) \right] + \sum_{i \in I} \sum_{j \in J} \left(c_{ij} \sum_{t=\ell_i}^{|J|} x_{ij}^t \left[\sum_{s=0}^{t} \zeta_{tps} \frac{\min\{q_i, s\}}{t} \right] \right) \tag{8.66}$$

$$\text{subject to} \quad \sum_{i \in I} \sum_{t=\ell_i}^{|J|} x_{ij}^t = 1, \quad j \in J \tag{8.67}$$

$$\sum_{j \in J} x_{ij}^t = t y_i^t, \quad i \in I \tag{8.68}$$

$$\sum_{t=\ell_i}^{|J|} y_i^t \leq 1, \quad i \in I \tag{8.69}$$

$$y_i^t \in \{0, 1\}, \quad i \in I, \ t = \ell_i, \ldots, |J| \tag{8.70}$$

$$x_{ij}^t \in \{0, 1\}, \quad i \in I, \ j \in J, \ t = \ell_i, \ldots, |J|. \tag{8.71}$$

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 this model. The authors also explore the advantages of the homogeneous case for the alternative outsourcing action they consider. This work would be later extended in two different ways. Bieniek (2015) showed that tractable expressions can be obtained for the recourse functions when other probability distributions are considered (not necessarily discrete) as long as the assumption of homogeneity among customers is kept. Albareda-Sambola et al. (2017) proposed a heuristic algorithm for tackling the general problem (heterogeneous demand probabilities). The procedure consists of two phases. First, a GRASP algorithm is used for building two pools of solutions—one based upon quality and another upon diversity. Second, a path relinking procedure is devised for connecting solutions from both pools hoping that better feasible solutions can be found during the process.

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 objective in which the recourse function is expanded to account for variance. Taking, as an example, model (8.27)–(8.33) this consists of defining

$$\mathrm{Q}(y) = \mathbb{E}_{\Xi}\left[Q(y,\xi)\right] - \lambda \mathrm{Var}_{\Xi}\left[Q(y,\xi)\right]. \tag{8.72}$$

Such a modeling framework in facility location is far from new (see Jucker and Carlson 1976). Nevertheless, this type of model has a clear disadvantage: it often results in a non-linear large-scale mixed-integer model. Different possibilities for overcoming this difficulty are discussed by Louveaux (1993).

Stochastic programming approaches for discrete facility location problems have attracted much attention in the recent years. Some papers not mentioned so far include those by Ravi and Sinha (2004), Lin (2009), Wang et al. (2011), Kiya and Davoudpour (2012), and Álvarez-Miranda et al. (2015).

Hybridizing between stochastic programming with robust optimization has been also considered in the context of facility location. Alumur et al. (2012) explored this possibility by using a robustness measure embedded 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, it is 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). A minmax regret formulation is then proposed for the overall problem.

Another work combining the flavor of two-stage stochastic programming with robust optimization is due to Marques and Dias (2018) who study a multi-period facility location problem. Uncertainty is associated with fixed and assignment costs as well as to the customers that exist in each period. The authors seek the minimization of the total expected cost but impose a constraint on the maximum regret allowed in each scenario.

In the context of logistics systems 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).

One work worth pointing out is that of Hinojosa et al. (2014) who studied a stochastic facility location problem with location decisions made at an operational level, i.e., location decisions are *ex post* decisions. The multi-product problem considered in that 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 which 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. The single-commodity version of this problem would be investigated by Fernández et al. (2019) from the perspective of a risk-averse decision maker. In particular, the conditional value at risk is to be minimized.

As in the preceding section, when using a stochastic programming model, it is important to evaluate its relevance compared to a more simplified deterministic one. Although no robust measure exists for asserting such relevance, two measures are often used to provide 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 of the single-scenario problem for every scenario). In many cases this is cumbersome, namely when the number of scenarios is large or even infinite. The VSS emerges as an alternative and can be obtained in two steps: (1) the expected value problem is solved. This is the deterministic problem obtained when the random variables are replaced by their expectation; (2) the stochastic problem is considered and the difference between its optimal value and the value of the solution obtained in (1) 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 ensuring a certain service level, i.e., ensuring that with some pre-specified probability, the overall demand does not exceed the capacity of the operating facilities.

In order to exemplify this paradigm, 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 a service level constraint as follows:

$$\text{Minimize} \quad \sum_{i \in I} f_i y_i + \sum_{i \in I} \sum_{j \in J} c_{ij} x_{ij} \tag{8.73}$$

$$\text{subject to} \quad (8.16)\text{–}(8.18)$$

$$\mathbb{P}\left[\sum_{j \in J} d_j x_{ij} \leq q_i y_i \right] \geq \alpha_i, \quad i \in I \tag{8.74}$$

$$x_{ij} \in \{0, 1\}, \quad i \in I, \ j \in J. \tag{8.75}$$

For every $i \in I$, the corresponding chance constraint sets $q_i y_i$ equal to the α_i-quantile of the distribution of the demand assigned to facility i. In other words, the constraint stipulates that the probability of observing a demand assigned to the facility not exceeding the capacity of the facility is at least α_i. Typically, high values are assumed for α_i (e.g., 0.90 or 0.95).

One desirable feature of such a model is the possibility of finding a deterministic equivalent formulation, i.e., replacing the probabilistic constraints by deterministic (equivalent) ones. Unfortunately, this is not always straightforward. One successful example for the problem we are considering is due to Lin (2009). The author assumes independent demands following a Poisson or a Gaussian distribution. For illustrative purposes, we detail 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, (8.74) becomes equivalent to

$$\sum_{\ell=0}^{q_i y_i} e^{-\mu_i} \frac{\mu_i^\ell}{\ell!} \geq \alpha_i, \quad i \in I \tag{8.76}$$

which, in turn, has a deterministic equivalent of the form

$$\sum_{j \in J} \lambda_j x_{ij} \leq v_i y_i, \quad i \in I. \tag{8.77}$$

In this model, $v_i = \mathbb{E}[\Upsilon]$, where Υ is a random variable following a Poisson distribution with an expectation equal to the largest value ensuring that $\mathbb{P}(\Upsilon \leq q_i) \geq \alpha_i$. As detailed by Lin (2009), the value v_i can easily be 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 (8.74) with (8.77) the resulting problem becomes a single-source capacitated facility location problem which can

be tackled by any appropriate method (see Chap. 4). Lin (2009) also explore 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, which is used as part of a heuristic.

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

$$\sum_{i \in I_j} y_i \geq 1, \quad j \in J, \tag{8.78}$$

where I_j denotes the set of locations covering customer $j \in J$. The probabilistic version of these constraints is the following:

$$\mathbb{P}\left[\text{At least one location is available for serving customer } j\right] \geq \alpha, \quad j \in J. \tag{8.79}$$

These constraints have as a deterministic equivalent,

$$\sum_{i \in I_j} y_i \geq \beta, \tag{8.80}$$

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 $\pi^{\sum_{i \in I_j} y_i}$. Therefore, the probability that at least one location among those covering customer $j \in J$ can serve it immediately is given by $1 - \pi^{\sum_{i \in I_j} y_i}$ which, together with (8.79) leads to the deterministic equivalent just presented.

For other applications of facility location problems with chance constraints we refer the reader to Kınay et al. (2018, 2019) as well as to the references therein.

8.6 Challenges and Further Readings

Despite all the work we can find focusing 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 readings.

8.6.1 Multi-Stage Stochastic Programming Models

In most of the stochastic facility location problems discussed above, a single moment in time for uncertainty to be disclosed was assumed. In many situations, this is not the case. Instead, we may observe uncertainty being progressively revealed in a succession of points in time. When this is the case, 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 that of Hernández et al. (2012), who consider a multi-period problem with stochastic demands. The problem consists of (1) determining the locations and dimensions of a preset number of new jails in Chile; (2) deciding 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 that combines 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. This work would be complemented by Escudero et al. (2018), who developed two matheuristics for the problem. One is based upon cluster Lagrangean decomposition (Escudero et al. 2016) whereas the other is based upon a so-called sequential partial linear relaxation which is a scheme that optimizes a decreasing stage-based relaxation of the integrality constraints of the variables for obtaining tighter lower bounds to the original problem.

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 in the context of facility location include those by Ahmed and Garcia (2004), Aghezzaf (2005), Correia et al. (2018), and Marques and Dias (2018).

[3]The downside risk is a measure of how much the return on investment is below a target initially imposed.

8.6.2 Algorithms

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 and integer programming, such as heuristics and relaxation-based procedures.

Regarding robust facility location problems, the minmax structure often considered makes them harder to solve than the corresponding minisum deterministic problems. The reader can refer to Snyder (2006) for a deeper discussion of this issue. That paper presents a sketch of the procedure typically followed for tackling minmax regret problems. Although some general procedures have been proposed for such problems (e.g., Mausser and Laguna 1998, for minmax regret linear problems with interval uncertainty) in most cases, tailored procedures, exact or approximate, must be developed to efficiently tackle the problems. Analytic results and polynomial time algorithms have also been proposed but only for problems with an 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 a compact model can be derived for the extensive form of the deterministic equivalent, realistic instances often induce a large-scale mixed-integer linear programming problem not manageable by a general solver. In this case, specific algorithms, 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. Alonso-Ayuso et al. (2003) introduce the so-called branch-and-fix coordination scheme for tackling a problem in the context of logistics systems. The proposed technique 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 procedure for multi-stage stochastic mixed-integer linear programming problems was introduced by Escudero et al. (2009, 2010). In those papers, the branch-and-fix coordination scheme proposed by Alonso-Ayuso et al. (2003) was extended to solve multi-stage problems with integer variables. As mentioned above, Hernández et al. (2012) embed such approach within a heuristic procedure.

When exact algorithms 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 scheme. Sample average approximation (SAA) was introduced by Kleywegt et al. (2001) and it is one such example which has become quite popular. Applications of this procedure to stochastic facility location were proposed by Kiya and Davoudpour (2012), Romauch and Hartl (2005) and Santoso et al. (2005). Sampling schemes 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 a research direction worth exploring.

Armas et al. (2017) apply a so-called simheuristic to the stochastic UFLP. Uncertainty is assumed for the transportation costs. The algorithm integrates simulation and a metaheuristic. In particular, the authors integrate an iterative local search with Monte Carlo simulation (MCS). This type of procedure may be quite promising for tackling more complex stochastic facility location problems.

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 algorithms developed by Carrøe and Schultz (1999) and Escudero et al. (2012). To the best of our knowledge, the first type of algorithm 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. In some cases, experts may be inquired in terms of plausible scenarios as well as their occurrence probabilities. This may call for the use of subjective probabilities by means of eliciting probability distributions (O'Hagan 1998; Casement and Kahle 2017; Oakley 2017).

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 scheme 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 consider 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 a 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, which 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 without 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 has been 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

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. This keeps being 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

Albareda-Sambola M, Fernández E, Saldanha-da-Gama F (2017) Heuristic solutions to the facility location problem with general Bernoulli demands. INFORMS J Compu 29:737–753

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-Meth 46:529–543

Álvarez-Miranda E, Fernández E, Ljubić I (2015) The recoverable robust facility location problem. Transp Res B-Meth 79:93–120

Armas J, Juan AA, Marquès JM, Pedroso JP (2017) Solving the deterministic and stochastic uncapacitated facility location problem: from a heuristic to a simheuristic. J Oper Res Soc 68:1161–1176

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

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

Bieniek M (2015) A note on the facility location problem with stochastic demands. Omega 55:53–60

Birge JR, Louveaux FV (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 Meth Oper Res 58:331–349

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

Casement CJ, Kahle DJ (2017) Graphical prior elicitation in univariate models. Commun Stat-Simul C 47:2906–2924

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

Correia I, Nickel S, Saldanha-da-Gama F. (2018) A stochastic multi-period capacitated multiple allocation hub location problem: formulation and inequalities. Omega 74:122–134

Current JR, 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) Lagrangean decomposition for large-scale two-stage stochastic mixed 0−1 problems. Top 20:347–374

Escudero LF, Garín MA, Unzueta A (2016) Cluster Lagrangean decomposition in multistage stochastic optimization. Comput Oper Res 67:48–62

Escudero LF, Garín MA, Pizarro C, Unzueta A (2018) On efficient matheuristic algorithms for multi-period stochastic facility location-assignment problems. Comput Optim Appl 70:865–888

Fernández E, Hinojosa Y, Puerto J, Saldanha-da-Gama F (2019) New algorithmic framework for conditional value at risk: application to stochastic fixed-charge transportation. Eur J Oper Res 277:215–226

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. Transport 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 22:1123–1147

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. Manage Sci 47:295–307

Jucker JV, Carlson C (1976) The simple plant-location problem under uncertainty. Oper Res 24:1045–1055

Kınay OB, Kara BY, Saldanha-da-Gama F, Correia I (2018) Modeling the shelter site location problem using chance constraints: a case study for Istanbul. Eur J Oper Res 270:132–145

Kınay OB, Kara BY, Saldanha-da-Gama F (2019) On multi-criteria chance-constrained capacitated single-source discrete facility location problems. Omega 83:107–122

Kiya F, Davoudpour H (2012) Stochastic programming approach to re-designing a warehouse network under uncertainty. Transport 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, 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. Transport 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
Marques MC, Dias JM (2018) Dynamic location problem under uncertainty with a regret-based measure of robustness. Int Trans Oper Res 25:1361–1381
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 Tech 50:391–397
Oakley J (2017) SHELF: Tools to Support the Sheffield Elicitation Framework (R package). https://github.com/OakleyJ/SHELF
O'Hagan A (1998) Eliciting expert beliefs in substantial practical applications. J Roy Stat Soc D-Sta, 47:21–35
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 (2004) Hedging uncertainty: approximation algorithms for stochastic optimization problems. Lect Notes Compu Sci 3064:101–115
ReVelle CS, Hogan K (1989) The maximum availability location problem. Transport 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. Lect Notes Compu Sci 3777:180–189
Rosenhead J, Elton M, Gupta SK (1972) Robustness and optimality as criteria for strategic decisions. Oper Res Quart 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. Loc Sci 6:383–394

Serra D, Ratick S, ReVelle CS (1996) The maximum capture problem with uncertainty. Environ Plann 23:49–59

Shapiro A, Dentcheva D, Ruszczyński A (2009) Lectures on stochastic programming: modeling and theory. MPS-SIAM Series on Optimization, 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. Transport Sci 17:168–180

Chapter 9
Location Problems with Multiple Criteria

S. Nickel, J. Puerto, and A. 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.

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, Seville, 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*,
https://doi.org/10.1007/978-3-030-32177-2_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 no longer have the canonical order of $\mathbb{R}$. 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 $\{x : f(x) \leq \rho\}$ 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 $q \in \mathscr{Q}$, 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 (Refer to Fig. 9.1) Let us consider the points $a_1 = (0, 0)$, $a_2 = (8, 3)$, $a_3 = (-3, 5)$ and the functions $f^1(x) = \|x - a_1\|_1$, $f^2(x) = \|x - a_2\|_\infty$, $f^3(x) =$

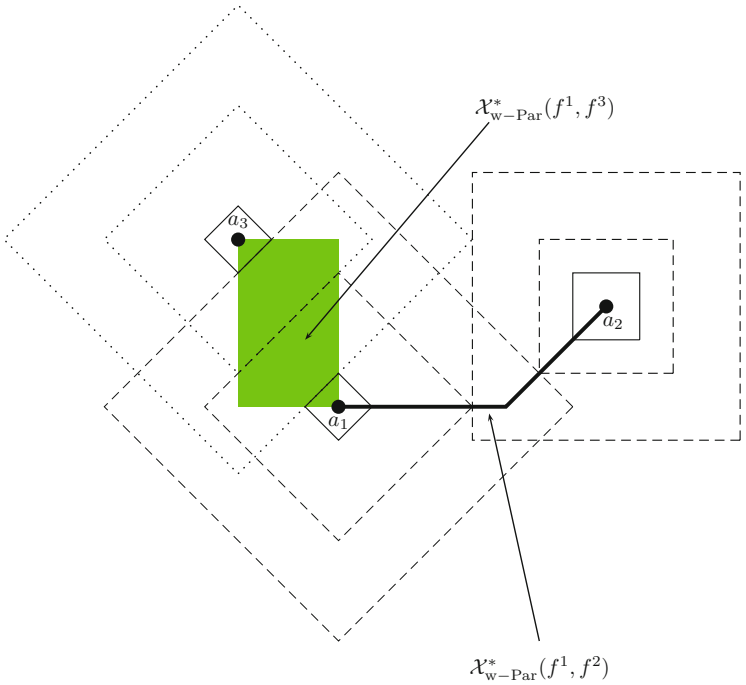


Fig. 9.1 Illustration of Remark 9.1

$\|x - a_3\|_1$. 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 filled 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 3-criteria solution sets already characterized. In order to do that, let

$$C_\infty(\mathbb{R}^+_0, \mathbb{R}^2) := \left\{\varphi \,|\, \varphi : \mathbb{R}^+_0 \to \mathbb{R}^2, \varphi \text{ continuous}, \lim_{t\to\infty} \|\varphi(t)\|_2 = \infty\right\},$$

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.

For a set $S \subseteq \mathbb{R}^2$ we define the enclosure of S by

$$\text{encl}\,(S) := \Big\{x \in \mathbb{R}^2 \,:\, \exists\, \varepsilon > 0 \text{ with } B(x, \varepsilon) \cap S = \emptyset\,,\ \exists\, t_\varphi \in [0, \infty) \text{ with } \varphi(t_\varphi) \in S \text{ for all } \varphi \in C_\infty(\mathbb{R}^+_0, \mathbb{R}^2) \text{ with } \varphi(0) = x\Big\},$$

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

$$\mathscr{X}^{\text{gen}}_{\text{w-Par}}\left(f^1, f^2, f^3\right) := \bigcup_{p=1}^{2} \bigcup_{q=p+1}^{3} \mathscr{X}^*_{\text{w-Par}}\left(f^p, f^q\right).$$

We use "gen" since this set will generate the set $\mathscr{X}^*_{\text{w-Par}}\left(f^1, f^2, f^3\right)$. The next theorem provides useful geometric information to build $\mathscr{X}^*_{\text{w-Par}}\left(f^1, f^2, f^3\right)$. Its proof can be found in Rodríguez-Chía and Puerto (2002).

Theorem 9.2

$$\mathscr{X}^*_{\text{w-Par}}(f^1, f^2, f^3) = \text{encl}\left(\mathscr{X}^{\text{gen}}_{\text{w-Par}}\left(f^1, f^2, f^3\right)\right) \cup \mathscr{X}^{\text{gen}}_{\text{w-Par}}\left(f^1, f^2, f^3\right).$$

Remark 9.2 It is worth noting that the region $encl\left(\mathscr{X}^{\text{gen}}_{\text{w-Par}}\left(f^1, f^2, f^3\right)\right)$ 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.

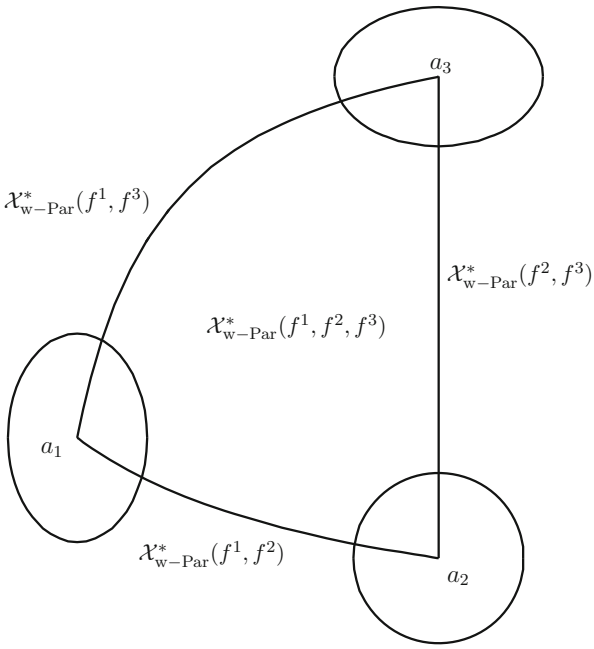


Fig. 9.2 Illustration of Theorem 9.2

Now we are in the right position to show the main result about the geometrical structure of $\mathscr{X}^*_{\text{w-Par}}(f^1, \ldots, f^Q)$.

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:

(i) 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.

(ii) 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.

(iii) 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 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 filled region, which in this case is the convex hull of a_1, a_2 and a_3.

Example 9.4 (Refer to 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

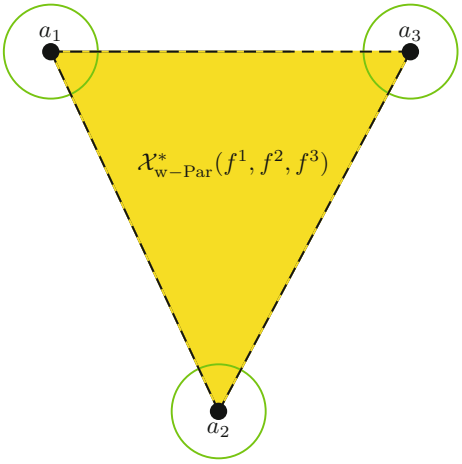


Fig. 9.3 Illustration of Remark 9.3.i

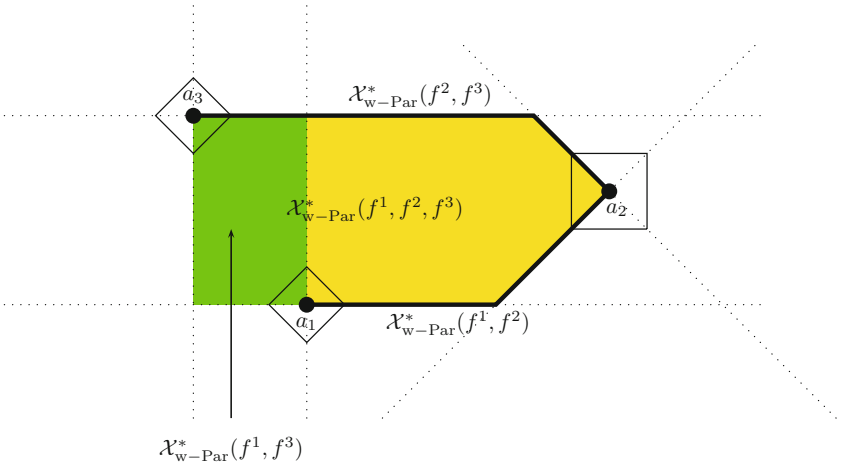


Fig. 9.4 Illustration of Remark 9.3.ii

and a_2, $\mathcal{X}^*_{\text{w-Par}}(f^2, f^3)$ is the thick path joining a_2 and a_3, and $\mathcal{X}^*_{\text{w-Par}}(f^1, f^3)$ is the filled rectangle with a_1 and a_3 as opposite extreme points. Therefore, by Theorem 9.2, $\mathcal{X}^*_{\text{w-Par}}(f^1, f^2, f^3)$ is the filled region surrounded by the union of the three previous sets. Note that this region is the union of two full dimensional elementary convex sets.

Example 9.5 (Refer to 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, $\mathcal{X}^*_{\text{w-Par}}(f^1, f^2) = R_1$, $\mathcal{X}^*_{\text{w-Par}}(f^1, f^3) = R_2 \cup R_4$, $\mathcal{X}^*_{\text{w-Par}}(f^2, f^3) =$

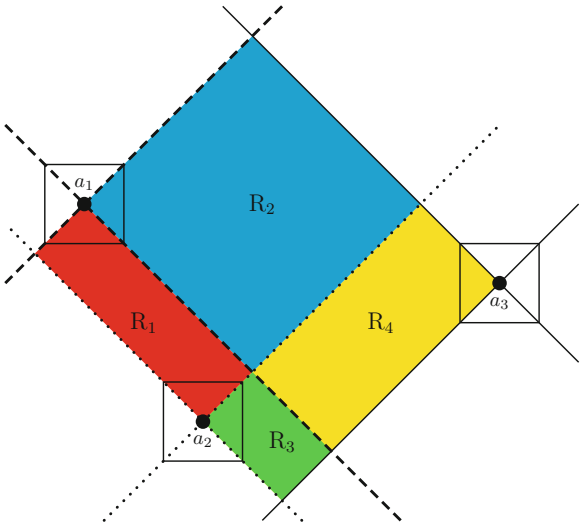


Fig. 9.5 Illustration of Remark 9.3.iii

$R_3 \cup R_4$. By Theorem 9.2, $\mathscr{X}^*_{\text{w-Par}}(f^1, f^2, f^3) = R_1 \cup R_2 \cup R_3 \cup R_4$. Note that in this example $\mathscr{X}^*_{\text{w-Par}}(f^1, f^2, f^3) = \mathscr{X}^*_{\text{w-Par}}(f^1, f^2) \cup \mathscr{X}^*_{\text{w-Par}}(f^1, f^3) \cup \mathscr{X}^*_{\text{w-Par}}(f^2, f^3)$.

9.2.1 *Polyhedral Planar Minisum Location Problems*

Consider a set of demand points $A := \{a_1, \ldots, a_M\} \subseteq \mathbb{R}^2$. For $i \in \mathscr{M} := \{1, 2, \ldots, M\}$, let $B_i \subset \mathbb{R}^2$ 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) may 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.

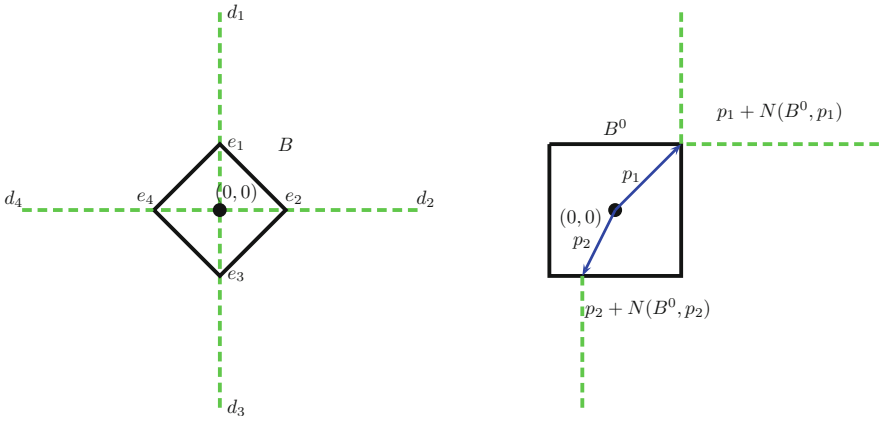


Fig. 9.6 Illustration of the unit ball for the ℓ_1-norm, its dual ball and two normal cones of this dual ball

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 $\mathrm{C}_\pi = \bigcap_{i \in \mathscr{M}} (a_i + N(B_i^o, p_i))$. Adopting the definition introduced by Durier and Michelot (1985), a nonempty convex set C is called an elementary convex set if there exists a family π such that $\mathrm{C}_\pi = \mathrm{C}$. 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 2-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 $\mathscr{IP}$. 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.

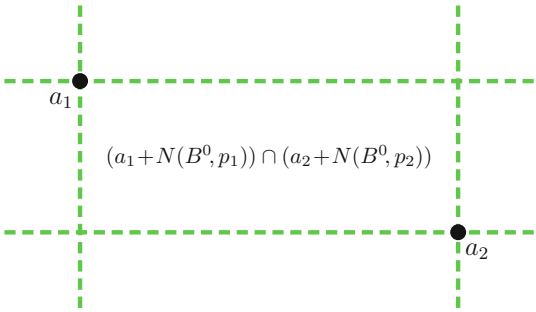


Fig. 9.7 Illustration of an elementary convex set for the ℓ_1-norm

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 $\mathbb{R}^2$,

$$\min_{x\in^2}\left(f^1(x):=\sum_{i=1}^{M}w_i^1\gamma_i(x-a_i),\ f^2(x):=\sum_{i=1}^{M}w_i^2\gamma_i(x-a_i)\right),$$

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 C $\in\mathscr{C}$ with $x\in$ 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. Three cases may occur:

Case 1: $x\in$ 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 $y \in \mathscr{X}^*_{\mathrm{Par}}\ \forall\, y \in \mathrm{C}$, hence $\mathrm{C} \subseteq \mathscr{X}^*_{\mathrm{Par}}$.

Case 2: $x \in \overline{ab} := \mathrm{conv}(\{a, b\}) \subset \mathrm{bd}(\mathrm{C})$ and $a, b \in \mathrm{Ext}(\mathrm{C})$. We can choose $y \in \mathrm{int}(\mathrm{C})$ and two cases can occur:

Case 2.1: $y \in \mathscr{X}^*_{\mathrm{Par}}$. Hence we can continue as in Case 1.
Case 2.2: $y \notin \mathscr{X}^*_{\mathrm{Par}}$. Therefore using the linearity we first obtain

$$\bigcap_{q=1}^{2} L_{\leq}(f^q, f^q(z)) \neq \bigcap_{q=1}^{2} L_{=}(f^q, f^q(z)) \quad \forall\, z \in \mathrm{int}(\mathrm{C}).$$

Second, since $x \in \mathscr{X}^*_{\mathrm{Par}}$, 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 $\mathrm{C} \not\subseteq \mathscr{X}^*_{\mathrm{Par}}$ and $\overline{ab} \subseteq \mathscr{X}^*_{\mathrm{Par}}$.

Case 3: $x \in \mathrm{Ext}(\mathrm{C})$. We can choose $y \in \mathrm{int}(\mathrm{C})$ and two cases can occur

Case 3.1: If $y \in \mathscr{X}^*_{\mathrm{Par}}$, we can continue as in Case 1.
Case 3.2: If $y \notin \mathscr{X}^*_{\mathrm{Par}}$, we choose $z_1, z_2 \in \mathrm{Ext}(\mathrm{C})$ such that $\overline{xz_1}, \overline{xz_2}$ are faces of C,

- If z_1 or z_2 are in $\mathscr{X}^*_{\mathrm{Par}}$, we can continue as in Case 2.
- If z_1 and z_2 are not in $\mathscr{X}^*_{\mathrm{Par}}$, then using the linearity in the same way as before we obtain that $(\mathrm{C} \setminus \{x\}) \cap \mathscr{X}^*_{\mathrm{Par}} = \emptyset$.

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 $\mathscr{X}^*_{\mathrm{Par}}$ 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 $\mathscr{X}^*_{\mathrm{Par}}$, 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 $\mathscr{X}^*_{\mathrm{Par}}$, 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 $\mathscr{X}^*_{2,1} := \arg\min_{x \in \mathscr{X}^*(f^2)} f^1(x)$.

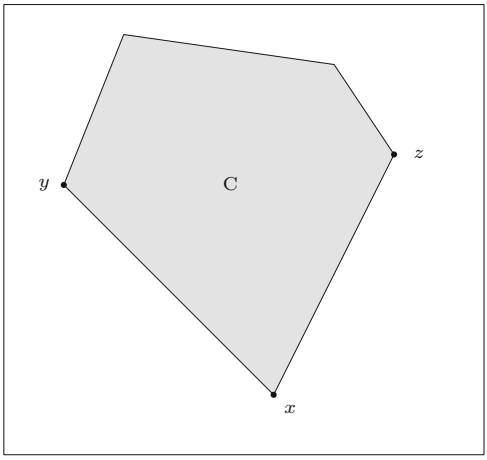


Fig. 9.8 Illustration to $y, x, z \in Ext(\mathrm{C})$ in counterclockwise order

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 $x \in \mathscr{X}^*_{\mathrm{Par}}$:

(S1) $\mathrm{C} \subseteq \mathscr{X}^*_{\mathrm{Par}}$, i.e., C is contained in the chain.
(S2) $\overline{xy}$ and $\overline{xz}$ are candidates for $\mathscr{X}^*_{\mathrm{Par}}$ and $\mathrm{int}(C) \not\subset \mathscr{X}^*_{\mathrm{Par}}$.
(S3) $\overline{xy}$ is candidate for $\mathscr{X}^*_{\mathrm{Par}}$ and $\overline{xz}$ is not contained in $\mathscr{X}^*_{\mathrm{Par}}$.
(S4) $\overline{xz}$ is candidate for $\mathscr{X}^*_{\mathrm{Par}}$ and $\overline{xy}$ is not contained in $\mathscr{X}^*_{\mathrm{Par}}$.
(S5) Neither $\overline{xy}$ nor $\overline{xz}$ are contained in $\mathscr{X}^*_{\mathrm{Par}}$.

We denote by $sit(\mathrm{C}, x)$ the situations (S1, S2, S3, 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* $x \in \mathscr{X}^*_{\mathrm{Par}}$ *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 $\mathscr{X}^*_{\mathrm{Par}}\left(f^1, f^2\right)$.

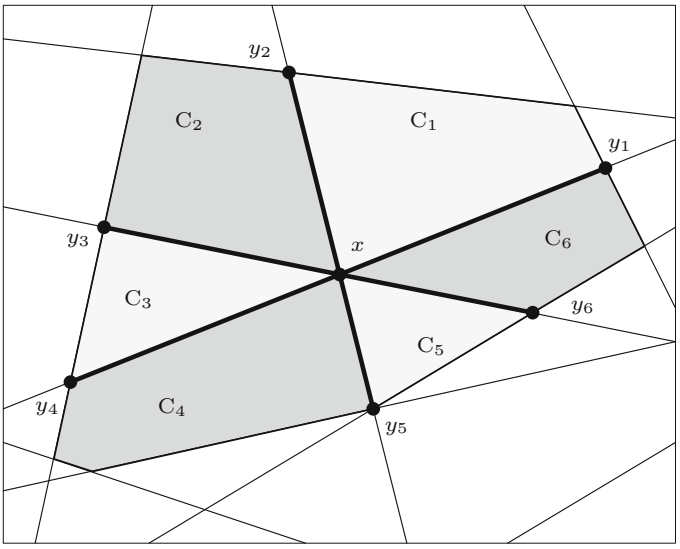


Fig. 9.9 Illustration to Lemma 9.1 with $P_x = 6$

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 $x \in \mathscr{X}^*_{1,2} \cap \mathscr{IP}$.

Step 4. Scan the list of cells adjacent to x until we get situation $S1$ for a cell C or two consecutive cells, C, $\overline{\text{C}}$, in situations C$\in \{S2, S3\}$ and $\overline{\text{C}} \in \{S2, S4\}$, respectively.

Step 5. If situation S1 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 $y \in \mathscr{IP} \cap C$ 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: $\mathscr{X}^*_{\text{Par}}\left(f^1, f^2\right)$. □

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

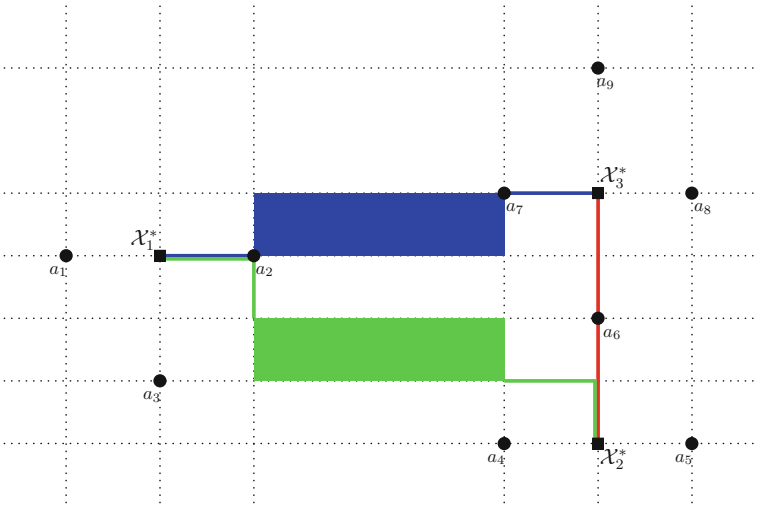


Fig. 9.10 Illustration to Algorithm 9.1

is $O(M^3 G_{\max}^2 \log(G_{\max}))$, since we have to consider at most $O(M^2 G_{max}^2)$ cells and the determination of $sit(.,.)$ can be done in $O(M \log(G_{max}))$ time. Hence, the overall complexity is $O(M^3 G_{\max}^2 \log(G_{\max}))$. Notice that the polynomial complexity of this algorithm allows an efficient computation of the solution set.

Example 9.6 Consider a problem with 9 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 $\{(-3, 0), (3, 0), (0, -4), (11, -6), (17, -6), (14, -2), (11, 2), (17, 2), (14, 6)\}$. Consider three median objective functions f^q, $q = 1, 2, 3$, namely those induced by the weights-vectors $w^1 = (2, 2, 1, 0, 0, 0, 0, 0, 0)$, $w^2 = (0, 0, 0, 2, 2, 1, 0, 0, 0)$ 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 $\mathcal{X}_1^* = \{(0, 0)\}$, $\mathcal{X}_2^* = \{(14, -6)\}$ and $\mathcal{X}_3^* = \{(14, 2)\}$, 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

$$\mathcal{X}_{\text{Par}}^* \left(f^1, f^3 \right) = \overline{(0, 0)\,(3, 0)} \cup \text{conv}(\{(3, 0)\,, (3, 2)\,, (11, 2)\,, (11, 0)\}) \cup \overline{(11, 2)\,(14, 2)},$$

$$\mathcal{X}_{\text{Par}}^* \left(f^2, f^3 \right) = \overline{(14, 2)\,(14, -6)},$$

$$\mathcal{X}^*_{\text{Par}}\left(f^1, f^2\right) = \overline{(0,0)\,(3,0)} \cup \overline{(3,0)\,(3,-2)} \cup \text{conv}(\{(3,-2)\,,(3,-4)\,,(11,-4)\,,(11,-2)\}) \cup \overline{(11,-4)\,(14,-4)} \cup \overline{(14,-4)\,(14,-6)}.$$

9.2.1.2 Three-Criteria Case

In this section we consider the 3-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 1-criterion solutions by

$$\mathcal{X}^{\text{gen}}_{\text{Par}}\left(f^1, f^2, f^3\right) := \bigcup_{q=1}^{3} \mathcal{X}^*(f^q) \cup \bigcup_{q=1}^{2} \bigcup_{p=q+1}^{3} \mathcal{X}^*_{\text{Par}}\left(f^p, f^q\right).$$

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.

Lemma 9.2 *The following inclusion of sets holds:*

$$cl\left(encl\left(\mathcal{X}^{gen}_{Par}\left(f^1, f^2, f^3\right)\right)\right) \subseteq \mathcal{X}^*_{s\text{-}Par}\left(f^1, f^2, f^3\right).$$

The interested reader is referred to Nickel et al. (2005b) for a detailed proof of this result.

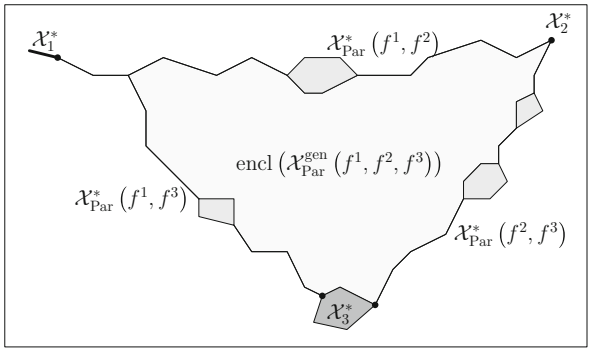


Fig. 9.11 The enclosure of $\mathcal{X}^{\text{gen}}_{\text{Par}}\left(f^1, f^2, f^3\right)$

Remark 9.4 Since $\mathscr{X}^{*}_{\text{Par}}\left(f^i, f^j\right) = \mathscr{X}^{*}_{\text{w}-\text{Par}}\left(f^i, f^j\right)$ for any $i, j \in \{1, 2, 3\}$, we have that:

$$\text{encl}\left(\mathscr{X}^{\text{gen}}_{\text{Par}}\left(f^1, f^2, f^3\right)\right) = \text{encl}\left(\mathscr{X}^{\text{gen}}_{\text{w}-\text{Par}}\left(f^1, f^2, f^3\right)\right).$$

Finally we obtain the following theorem which provides a subset as well as a superset of $\mathscr{X}^{*}_{\text{Par}}\left(f^1, f^2, f^3\right)$.

Theorem 9.5 *The following inclusions of sets hold:*

$$\begin{aligned}\text{encl}\left(\mathscr{X}^{\text{gen}}_{\text{Par}}\left(f^1, f^2, f^3\right)\right) &\subseteq \mathscr{X}^{*}_{\text{Par}}\left(f^1, f^2, f^3\right)\\ &\subseteq \mathscr{X}^{\text{gen}}_{\text{Par}}\left(f^1, f^2, f^3\right) \cup \text{encl}\left(\mathscr{X}^{\text{gen}}_{\text{Par}}\left(f^1, f^2, f^3\right)\right)\\ &= \mathscr{X}^{*}_{\text{w}-\text{Par}}\left(f^1, f^2, f^3\right).\end{aligned}$$

Proof Using Lemma 9.2 and Theorem 9.2 we have the following chain of inclusions that proves the thesis of the theorem.

$$\begin{aligned}\text{encl}\left(\mathscr{X}^{\text{gen}}_{\text{Par}}\left(f^1, f^2, f^3\right)\right) &\subseteq \mathscr{X}^{*}_{\text{s}-\text{Par}}\left(f^1, f^2, f^3\right)\\ &\subseteq \mathscr{X}^{*}_{\text{Par}}\left(f^1, f^2, f^3\right) \subseteq \mathscr{X}^{*}_{\text{w}-\text{Par}}\left(f^1, f^2, f^3\right)\\ &\subseteq \mathscr{X}^{\text{gen}}_{\text{Par}}\left(f^1, f^2, f^3\right) \cup \text{encl}\left(\mathscr{X}^{\text{gen}}_{\text{Par}}\left(f^1, f^2, f^3\right)\right).\end{aligned}$$

□

Now it remains to consider the Pareto-optimality of the set $\mathscr{X}^{\text{gen}}_{\text{Par}}\left(f^1, f^2, f^3\right)$ 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., $y \notin \text{col}_3(C)$ and the inclusion $\subseteq$ is proved.

In order to prove $\supseteq$, 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 obtain the following algorithm for the 3-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: $\mathscr{X}_{\text{Par}}^{*}\left(f^1, f^2, f^3\right)$. □

Example 9.7 (Refer to Example 9.6) In Fig. 9.12, 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 $\mathscr{X}_{\text{w-Par}}^{*}\left(f^1, f^2\right)$

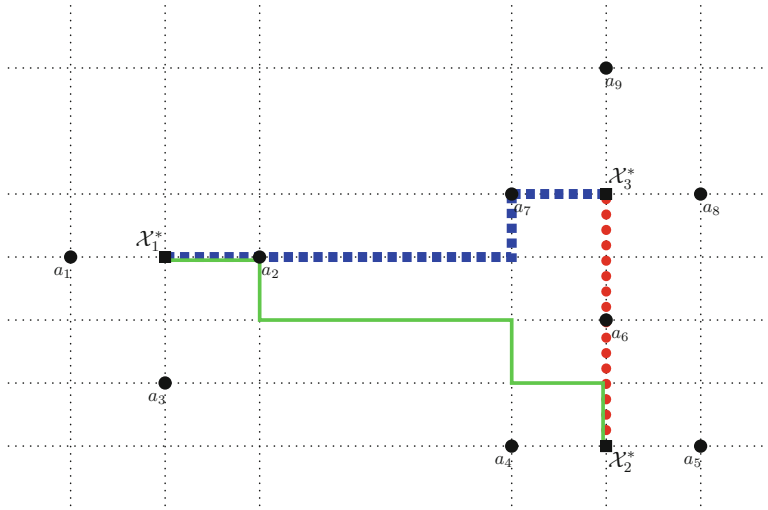


Fig. 9.12 Illustration of $\mathscr{X}_{\text{Par}}^{\text{gen}}\left(f^1, f^2, f^3\right)$ and $\mathscr{X}_{\text{Par}}^{*}\left(f^1, f^2, f^3\right)$ for the problem introduced in Example 9.6

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 no 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. $\bigcup_{\substack{p,q,r\in\mathcal{Q}\\ p<q<r}} \text{cl}\left(\text{encl}\left(\mathscr{X}_{\text{Par}}^{\text{gen}}\left(f^p, f^q, f^r\right)\right)\right) \subseteq \mathscr{X}_{\text{Par}}^{*}\left(f^1, \ldots, f^Q\right).$

II. $\mathscr{X}_{\text{Par}}^{*}\left(f^1, \ldots, f^Q\right) \subseteq \bigcup_{\substack{p,q,r\in\mathcal{Q}\\ p<q<r}} \mathscr{X}_{\text{Par}}^{\text{gen}}\left(f^p, f^q, f^r\right) \cup \bigcup_{\substack{p,q,r\in\mathcal{Q}\\ p<q<r}} \text{encl}\left(\mathscr{X}_{\text{Par}}^{\text{gen}}\left(f^p, f^q, f^r\right)\right) = \mathscr{X}_{\text{w-Par}}^{*}\left(f^1, \ldots, f^Q\right).$

Proof

(1) Let $x \in \bigcup_{\substack{p,q,r\in\mathscr{Q}\\ p<q<r}} \mathrm{cl}\left(\mathrm{encl}\left(\mathscr{X}_{\mathrm{Par}}^{\mathrm{gen}}\left(f^p, f^q, f^r\right)\right)\right)$. This is equivalent to

$$x \in \mathrm{cl}\left(\mathrm{encl}\left(\mathscr{X}_{\mathrm{Par}}^{\mathrm{gen}}\left(f^p, f^q, f^r\right)\right)\right) \quad \text{for some} \quad p, q, r \in \mathscr{Q}, p < q < r.$$

Then, by Lemma 9.2, $x \in \mathscr{X}^*_{\mathrm{s-Par}}\left(f^p, f^q, f^r\right)$ for some $p, q, r \in \mathscr{Q}$, $p < q < r$. Applying characterization (9.4), this is equivalent to $L_{\leq}(f^p, f^p(x)) \cap L_{\leq}(f^q, f^q(x)) \cap L_{\leq}(f^r, f^r(x)) = \{x\}$ for some $p, q, r \in \mathscr{Q}$, $p < q < r$ and since $x \in L_{\leq}(f^q, f^q(x))$ for all $q \in \mathscr{Q}$ it follows that $\bigcap_{q=1}^{Q} L_{\leq}(f^q, f^q(x)) = \{x\}$. Finally, again by (9.4), $x \in \mathscr{X}^*_{\mathrm{s-Par}}\left(f^1, \ldots, f^Q\right)$, which implies that $x \in \mathscr{X}^*_{\mathrm{Par}}\left(f^1, \ldots, f^Q\right)$.

(2) 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 $\bigcap_{q=1}^{Q} L_<(f^q, f^q(x)) = \emptyset$. By Helly's theorem, there exists $p, q, r \in \mathscr{Q}$, $p < q < r$, such that, $L_<(f^p, f^p(x)) \cap L_<(f^q, f^q(x)) \cap L_<(f^r, f^r(x)) = \emptyset$. By characterization (9.2), this is equivalent to $x \in \mathscr{X}^*_{\mathrm{w-Par}}\left(f^p, f^q, f^r\right)$ for some $p, q, r \in \mathscr{Q}$, $p < q < r$ and, by Theorem 3.2 in Rodríguez-Chía and Puerto (2002), this implies that $x \in \mathscr{X}^{\mathrm{gen}}_{\mathrm{Par}}\left(f^p, f^q, f^r\right) \cup \mathrm{encl}\left(\mathscr{X}^{\mathrm{gen}}_{\mathrm{Par}}\left(f^p, f^q, f^r\right)\right)$ for some $p, q, r \in \mathscr{Q}$, $p < q < r$. Finally, this can be equivalently written as

$$x \in \bigcup_{\substack{p,q,r\in\mathscr{Q}\\ p<q<r}} \mathscr{X}^{\mathrm{gen}}_{\mathrm{Par}}\left(f^p, f^q, f^r\right) \cup \bigcup_{\substack{p,q,r\in\mathscr{Q}\\ p<q<r}} \mathrm{encl}\left(\mathscr{X}^{\mathrm{gen}}_{\mathrm{Par}}\left(f^p, f^q, f^r\right)\right).$$

□

In the Q-criteria case the crucial region is now given by the cells $C \in \mathscr{C}$ with

$$C \subseteq \bigcup_{\substack{p,q,r\in\mathscr{Q}\\ p<q<r}} \mathscr{X}^{\mathrm{gen}}_{\mathrm{Par}}\left(f^p, f^q, f^r\right) \setminus \bigcup_{\substack{p,q,r\in\mathscr{Q}\\ p<q<r}} \mathrm{encl}\left(\mathscr{X}^{\mathrm{gen}}_{\mathrm{Par}}\left(f^p, f^q, f^r\right)\right)$$

$$= \bigcup_{\substack{p,q\in\mathscr{Q}\\ p<q}} \mathscr{X}^*_{\mathrm{w-Par}}\left(f^p, f^q\right) \setminus \bigcup_{\substack{p,q,r\in\mathscr{Q}\\ p<q<r}} \mathrm{encl}\left(\mathscr{X}^{\mathrm{gen}}_{\mathrm{Par}}\left(f^p, f^q, f^r\right)\right).$$

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 $\mathrm{int}(C)$. Finally we obtain the following theorem, which can be proven using the same reasoning as in the 3-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 by $\mathscr{C}$, 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. For every p,q and r with $p<q<r$ set
$\mathscr{X}^{\text{gen}}_{\text{Par}}\left(f^p,f^q,f^r\right):=\mathscr{X}^*_{\text{w-Par}}\left(f^p,f^q\right)\cup\mathscr{X}^*_{\text{w-Par}}\left(f^p,f^r\right)\cup\mathscr{X}^*_{\text{w-Par}}\left(f^q,f^r\right)$, 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 $O(Q^3)$ 3-criteria problems which dominates all other elementary operations of the algorithm. Each one of them has the same complexity as the 2-criteria problem. Thus, the overall complexity is $O(M^3G^2_{\max}Q^3(\log G_{\max})+M^2G^2_{\max}Q\log Q)=O(M^3G^2_{\max}Q^3(\log G_{\max})$.

We would like to conclude this section pointing that the multi-facility versions of the problems analyzed in this section have been scarcely studied in the literature, although an exception is the paper by Nickel (1997).

9.2.2 *Other References in Continuous Multicriteria Location Problems*

Along this section we have presented a complete description of the set of weak Pareto, Pareto, and strict Pareto locations for a general planar Q-criteria location problem based on the characterization of three criteria problems. The geometrical description and the characterizations of these sets allow the reader to get a general

idea of the multicriteria continuous location problem. In addition, one can also find more references and an overview on other location problems in the survey by Nickel et al. (2005a). Finally, Farahani et al. (2010) provides a review on results and developments in multicriteria location problems in three categories including bi-objective, multi-objective and multi-attribute problems and their solution methods.

In the following we list some interesting recent references in this field: The planar single-facility multiobjective location problem is also studied using the maximum norm in Alzorba et al. (2015) and using ℓ_1-norm in Alzorba et al. (2017). A scalarization proximal point method for solving a very general unconstrained multiobjective problem where the functions are locally Lipschitz and quasiconvex is studied in Apolinário et al. (2016), this methodology is applied to location problems. In Elleuch and Frikha (2018), a facility location decision which involves both qualitative and quantitative criteria is considered, the authors combined two methods, preference-ranking organisation method for enrichment evaluation (PROMETHEE) and a linear programming model, using the stretching and shrinking graphs method. Bhattacharya (2018) proposes a new mathematical model for locating k-obnoxious facilities that was solved by a nonlinear programming iterative algorithm.

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

$$f(x) := \begin{pmatrix} f^1(x) \\ \vdots \\ f^Q(x) \end{pmatrix} := \begin{pmatrix} \sum_{v_i \in V} w_i^1 \, (d(x, v_i) + d(v_i, x)) \\ \vdots \\ \sum_{v_i \in V} w_i^Q \, (d(x, v_i) + d(v_i, x)) \end{pmatrix}$$

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 $\mathcal{X}_{par}^*(S) := \{x \in S : f(x) \in W_{par}\}$. If $S = P(G)$ we simply write $\mathcal{X}_{par}^*$. A point $x \in \mathcal{X}_{par}^*(S)$ is called a Pareto location with respect to S, and the elements of $\mathcal{X}_{par}^*(V)$ are called Pareto nodes or Pareto vertices.

Computing $\mathcal{X}_{par}^*(V)$ can simply be done by pairwise comparison of the nodes. For $\mathcal{X}_{par}^*$ 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 $\mathcal{X}_{par}^*(V) = \mathcal{X}_{par}^*$. 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.

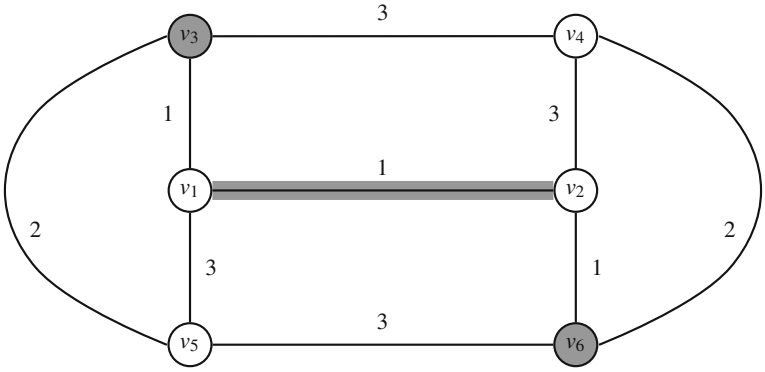


Fig. 9.13 Network of Example 9.8

Example 9.8 (See Fig. 9.13) 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

$$D = \begin{pmatrix} 0 & 1 & 1 & 4 & 3 & 2 \\ 1 & 0 & 2 & 3 & 4 & 1 \\ 1 & 2 & 0 & 3 & 2 & 3 \\ 4 & 3 & 3 & 0 & 5 & 2 \\ 3 & 4 & 2 & 5 & 0 & 3 \\ 2 & 1 & 3 & 2 & 3 & 0 \end{pmatrix}.$$

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 $\{v_6, v_2\},\ \{v_6, v_5\},\ \{v_6, v_4\}$

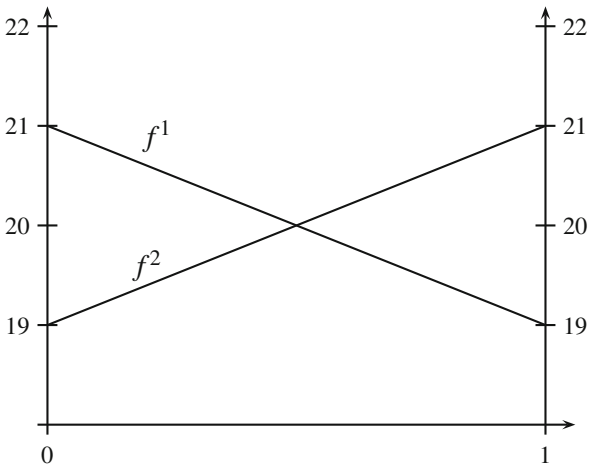


Fig. 9.14 Objective functions on the edge $\{v_1, v_2\}$ in Example 9.8

- 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 $\{v_1, v_2\}$

$$\Rightarrow \mathscr{X}^*_{par} = \{v_3\} \cup \{v_6\} \cup (\{v_1, v_2\},\ (0, 1)) .$$

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

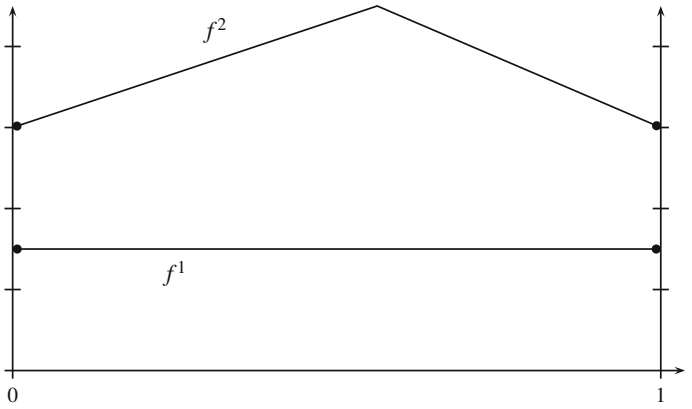


Fig. 9.15 Concavity on an edge with one objective function constant

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 imply that any location on the edge is 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 an interval with a small or null length) and all points are dominated by the left endnode. 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

$$t^1 := \max\{t \in [0, 1] : f^1(v_i) = f^1\big(\big(\{v_i, v_j\}, t\big)\big)\}$$

and

$$t^2 := \min\{t \in [0, 1] : f^2(v_j) = f^2\big(\big(\{v_i, v_j\}, t\big)\big)\}.$$

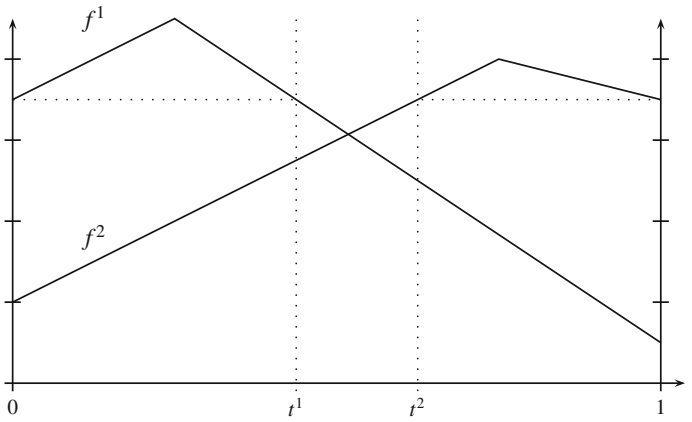


Fig. 9.16 Derivation of t^1 and t^2

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 $\mathscr{X}^*_{par}(e)$ for the bi-criteria median problem on a network)

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 $\mathscr{X}^*_{par}(e) := \{v_i\} \cup \{v_j\} \cup \left(\{v_i, v_j\},\ (t^1, t^2)\right)$
ELSE $\mathscr{X}^*_{par}(e) := \{v_i\} \cup \{v_j\}$
Step 7. STOP.

Output: $\mathscr{X}^*_{par}(e)$

To analyze the complexity of this algorithm, we need the following definition: A point $x = \left(\{v_i, v_j\}, t\right)$, $t \in [0, 1]$ on one edge $e = \{v_i, v_j\}$ is called a bottleneck

point for f^q if there exists a vertex v_k with $w_k^q > 0$, 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 $O(|E||V| \log |V|)$.

Example 9.9 Consider the following network (Fig. 9.17):
with distance matrix

$$D = \begin{pmatrix} 0 & 1 & 2 & 2 \\ 1 & 0 & 2 & 1 \\ 2 & 2 & 0 & 1 \\ 2 & 1 & 1 & 0 \end{pmatrix}.$$

We first compute

	v_1	v_2	v_3	v_4
f^1	10	7	8	6
f^2	7	8	9	9

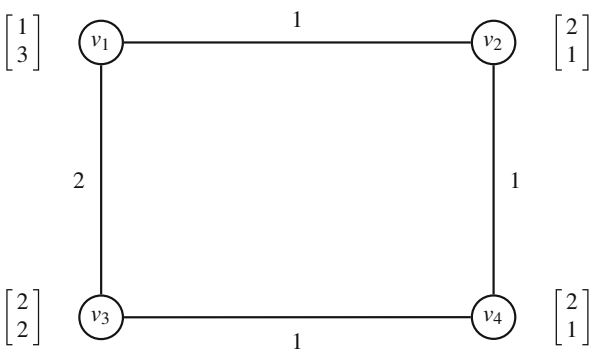


Fig. 9.17 Network of Example 9.9

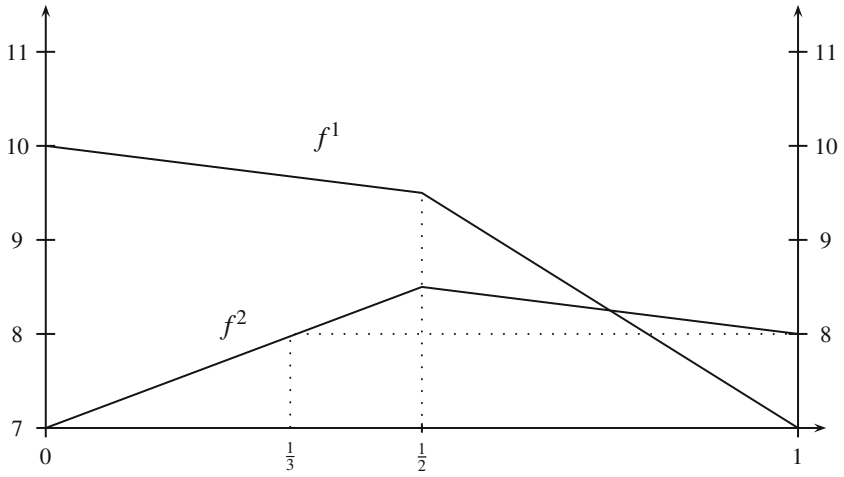


Fig. 9.18 Computing $\mathscr{X}^*_{par}(\{v_1, v_2\})$

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 Fig. 9.18.

$$t^1 = \max\left\{ t \in [0,1] :\ f^1(v_1) = f^1\left(\{v_1, v_2\}, t\right) \right\} = 0$$

$$t^2 = \min\left\{ t \in [0,1] :\ f^2(v_2) = f^2\left(\{v_1, v_2\}, t\right) \right\} = \frac{1}{3}$$

$$(\text{in } [0, \frac{1}{2}], \quad f^2(x) \equiv 7 + 3t,\ 7 + 3t = 8 \Leftrightarrow t = \frac{1}{3})$$

$$\mathscr{X}^*_{par}(e) = \{v_1\} \cup \{v_2\} \cup \left(\{v_1, v_2\}, \left(0, \frac{1}{3}\right) \right)$$

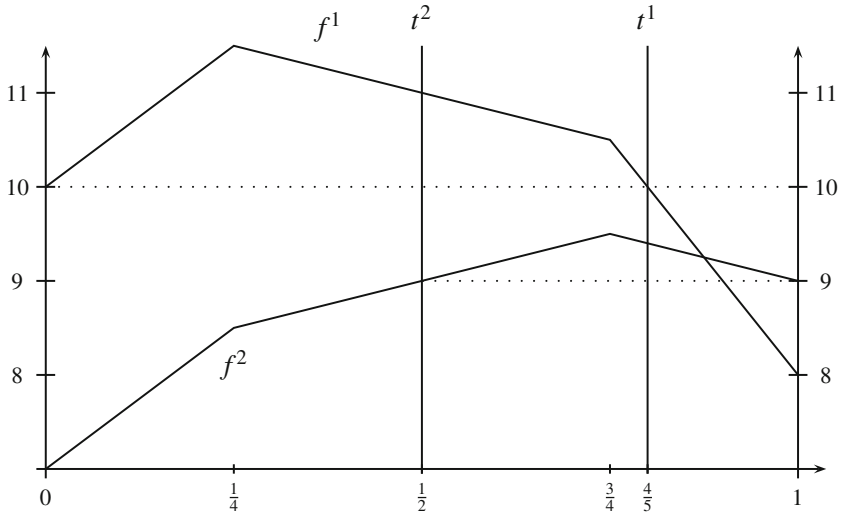


Fig. 9.19 Computing $\mathscr{X}^*_{par}(\{v_1, v_3\})$

- $e = \{v_2, v_4\}$. $f^1(v_2) = 7 > f^1(v_4) = 6$ 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 $v_3 \Rightarrow \mathscr{X}^*_{par}(e) = \{v_4\}$.
- $e = \{v_1, v_3\}$. (Fig. 9.19) $B_{13} = \left\{ \left(\underbrace{\{v_1, v_3\}}_{b^1_{13}},\ \frac{1}{4} \right), \left(\underbrace{\{v_1, v_3\}}_{b^2_{13}},\ \frac{3}{4} \right) \right\}$

$$f\left(b^1_{13}\right) = \begin{pmatrix} 11.5 \\ 8.5 \end{pmatrix}, \quad f\left(b^2_{13}\right) = \begin{pmatrix} 10.5 \\ 9.5 \end{pmatrix}$$

$$t_1 = \frac{4}{5}, \quad t_2 = \frac{1}{2}$$

$$\mathscr{X}^*_{par}(e) = \{v_1\} \cup \{v_3\}$$

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

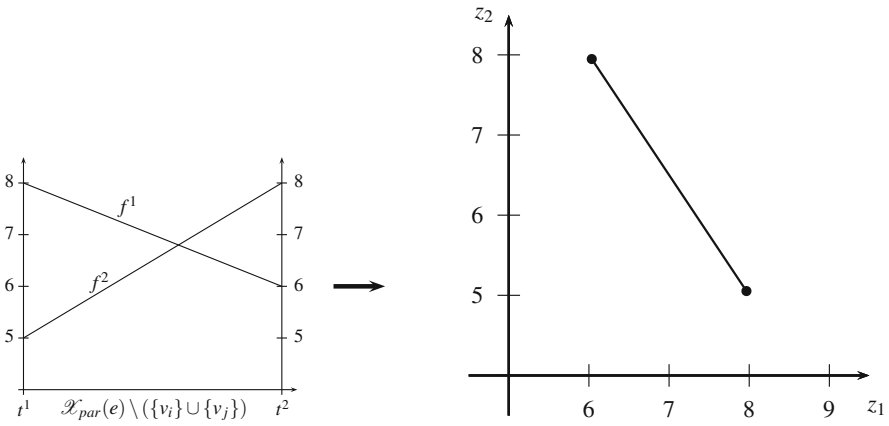


Fig. 9.20 Mapping $\mathscr{X}^*_{par}(e)$ to the objective space

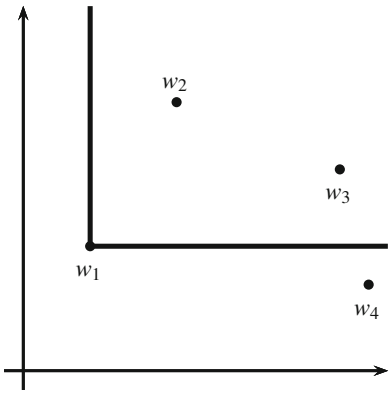


Fig. 9.21 w_1 is dominating w_2 and w_3

the edge. Again we would like to skip the technical details and proofs and refer the reader to Hamacher et al. (1999).

In the objective space, a point w dominates all other points in $w + \mathbb{R}^2_+ \setminus \{0\} := \left\{w + y : y \in \mathbb{R}^2_+ \setminus \{0\}\right\}$ (see Fig. 9.21).

In order to obtain $\mathscr{X}^*_{par}$ we draw 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

$$\bigcup \left\{(x, y) \in P : y \leq y' \text{ for all } (x, y') \in P\right\}.$$

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 $(z^1_{\max},\ z^2_l)$.

Step 4. Compute the lower envelope L of $\mathrm{IM}(f)$, 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 $\mathscr{X}^*_{par} := f^{-1}(L)$.

Output: $\mathscr{X}^*_{par}$

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 $\mathrm{IM}(f)$ such that we can get $\mathscr{X}^*_{par}$ from 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 (Fig. 9.23).

Example 9.10 (Example 9.9 Cont.) We first draw $\mathrm{IM}(f)$ and add the horizontal line segments. Finally, we get $\mathscr{X}^*_{par} = P\left(\{v_2, v_4\}\right) \cup \left(\{v_1, v_2\},\ \left[0, \frac{1}{3}\right)\right)$.

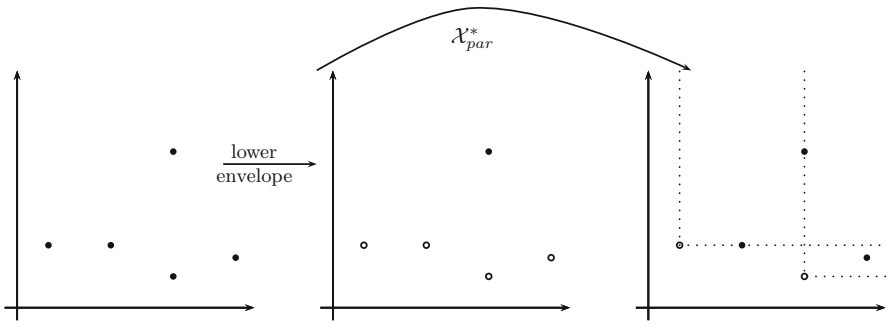


Fig. 9.22 Using the lower envelope to delete dominated solutions

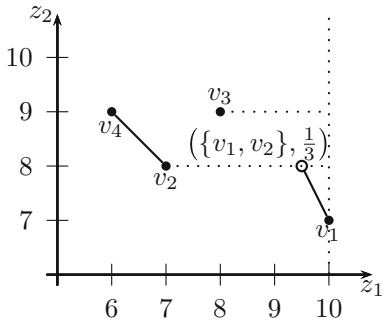


Fig. 9.23 Computing $\mathscr{X}^*_{par}$ for Example 9.9

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 $\mathscr{Q}_1$ and $\mathscr{Q}_2$ be a *partition* of $\mathscr{Q}$, such that $f^q(v_i) \geq f^q(v_j)$ for all $q \in \mathscr{Q}_1$ and $f^q(v_i) < f^q(v_j)$ for all $q \in \mathscr{Q}_2$. Of course, $\mathscr{Q}_1 \neq \emptyset$, $\mathscr{Q}_1 \cap \mathscr{Q}_2 = \emptyset$ and $\mathscr{Q}_1 \cup \mathscr{Q}_2 = \mathscr{Q}$. Also in case 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 $\big(\{v_j, v_l\},\ [t_r,\ t_{r+1}]\big)$. Take now a closed linear subedge from another edge $e_k = \{v_k, v_m\}$, then we get $\big(\{v_k, v_m\},\ [s_p,\ s_{p+1}]\big) \subseteq \mathscr{X}^*_{par}(e_k)$. This leads to

$$f^q\left(\left(\{v_k, v_m\},\ s\right)\right) = b_p^q + m_p^q s \quad \text{for all} \quad q \in \mathscr{Q},\ s \in [s_p, s_{p+1}],$$

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 $\left(\{v_k, v_m\},\ s\right)$, $s \in [s_p,\ s_{p+1}]$

$$\Leftrightarrow \quad b_p^q + m_p^q s \le b_r^q + m_r^q t \quad \text{for all} \quad q \in \mathscr{Q},$$

where at least one inequality is strict. Now we define the polyhedron

$$\mathscr{F} := \left\{(s,t) :\ m_r^q t - m_p^q s \ge b_p^q - b_r^q,\ \forall q \in \mathscr{Q}\right\} \cap \big([s_p,\ s_{p+1}] \times [t_r,\ t_{r+1}]\big).$$

We have two cases: If $\mathscr{F} = \emptyset$, 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, $\mathscr{F} \neq \emptyset$ is taken as a feasible solution of the two 2-variable linear programs

$$\mathrm{LB} = \min\{t :\ (s,t) \in \mathscr{F}\}, \quad \mathrm{UB} = \max\{t :\ (s,t) \in \mathscr{F}\}.$$

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 $q \in \mathscr{Q}$, then there is no dominance. Otherwise $\mathscr{X}^*_{par}(e_l) := \mathscr{X}^*_{par}(e_l) \setminus \big(\{v_j, v_l\},\ [\mathrm{LB},\ \mathrm{UB}]\big)$. 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 $\mathscr{X}^*_{par}(e)$ for all $e \in E$ and set $\mathscr{X}^*_{par} := \bigcup_{e \in E} \mathscr{X}^*_{par}(e)$

Step 2. Compare all v_i and all edges, where all f^q, $q \in \mathscr{Q}$ are constant

Step 3. For all Pareto linear subedges do a pairwise comparison as described above and reduce $\mathscr{X}^*_{par}$ accordingly.

Output: $\mathscr{X}^*_{par}$

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 1971a). 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:*

$$\{a, b\} \subseteq \mathcal{X}^*_{par} \text{ if and only if } L[a, b] \subseteq \mathcal{X}^*_{par} \ .$$

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 $w^q(V') := \sum_{v_i \in V'} w_i^q, \forall q \in \mathcal{Q}$.

Using Theorem 9.9 together with two lemmata from Goldman (1971b) and the above definition of $W(V)$ we can prove the following result which paves the way for solving Q-criteria median problems on a tree.

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 1971a). 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 9.11 Consider the tree depicted in Fig. 9.24. We solve the following instance of a 3-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 $\frac{1}{2}W = \begin{pmatrix} 25 \\ 31 \\ 30 \end{pmatrix}$.

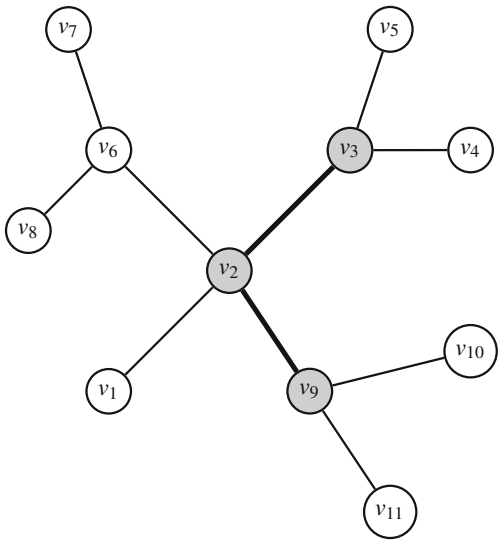


Fig. 9.24 Tree of Example 9.11. The bold edges and nodes indicate the set of Pareto locations

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"*.

- 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 survey by 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 et al. (2005b), Nickel and Puerto (2005). In Colebrook and Sicilia (2007b) the location of undesirable facilities on multicriteria location problems on networks is looked into 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 Kalcsics et al. (2015) found polynomial cases of multi-facility multicriteria location problems on networks. In Kalcsics et al. (2014), the authors discuss the multicriteria p-facility median location problem on networks with positive and negative weights; providing an efficient algorithm to solve the bicriteria 2-facility problem and a polynomial algorithm for the general problem when the number of facilities and criteria is fixed.

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. The methodology developed was extended to the capacitated version by Arora and Arora (2010).

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 such as 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 non-dominated solutions with non-integer values and even non-supported non-dominated 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 to compute those functions when the dimension, n, is fixed. 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 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:

$$\lambda^q c_B^q B^{-1}\overline{N} - \lambda^q c_{\overline{N}}^q = 0, \quad q \in \mathscr{Q}.$$

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

$$\text{Minimize} \sum_{i=1}^{M} \sum_{j=1}^{N} c_{ij}(\lambda_B^Q) x_{ij} \tag{9.16}$$

$$\text{subject to} \sum_{j=1}^{N} x_{ij} = 1, \quad i \in I, \tag{9.17}$$

$$x_{ij} \leq y_j, \quad i \in I,\ j \in J, \tag{9.18}$$

$$\sum_{j=1}^{N} y_j = p, \tag{9.19}$$

$$x_{ij} \in \{0, 1\},\ y_j \in \{0, 1\}, \quad i \in I,\ j \in J. \tag{9.20}$$

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.

9.4.4 Other References in Discrete Location Problems

In the previous two subsections, the entire set of Pareto locations is characterized using rational generating functions of integer points in polytopes and supported Pareto-optimal solutions are identified by solving binary linear problems. These approaches provide the reader with a general idea of the tools needed to characterize the set of Pareto optimal solutions in discrete location problems. Some additional references can be found in Nickel et al. (2005a). Also Farahani et al. (2010) reviews results and developments in multicriteria location problems.

In the following we list some interesting recent references in this field: Özpeynirci (2017) introduces new properties that restrict the possible locations

of the non-dominated points necessary for computing the nadir points and applied this methodology to multiobjective integer location problems. Pecci et al. (2017) study the multiobjective co-design problem of optimal valve placement and operation in water distribution networks. The resulting optimization problem is a multiobjective mixed integer nonlinear optimization problem. The multi-objective competitive location problem with distance-based attractiveness for two facilities is introduced in Wang et al. (2018). The multiobjective version of the obnoxious p-median problem was studied in Colmenar et al. (2018). That paper obtains high-quality approximations to the efficient front of the bi-objective case using a Multi-Objective Memetic Algorithm. Karatas and Yakici (2018) presents a novel methodology for solving multi-objective facility location problems with the focus on public emergency service stations, considering the p-median problem, the maximal coverage location problem and the p-center 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 MTM2016-74983-C2-01/02-R (Ministry of Science, Innovation and Universities\FEDER, Spain).

References

Alzorba S, Günther C, Popovici N (2015) A special class of extended multicriteria location problems. Optimization 64(5):1305–1320

Alzorba S, Günther C, Popovici N, Tammer C (2017) A new algorithm for solving planar multiobjective location problems involving the Manhattan norm. Eur J Oper Res 258(1):35–46

Apolinário HCF, Papa Quiroz EA, Oliveira PR (2016) A scalarization proximal point method for quasiconvex multiobjective minimization. J Global Optim 64(1):79–96

Arora S, Arora SR (2010) Multiobjective capacitated plant location problem. Int J Oper Res 7(4):487–505

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 Amer Math Soc 16:957–979

Bhattacharya U (2018) A mathematical model for locating k-obnoxious facilities on a plane. Int J Oper Res 31(3):384–402

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 21(4):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
Colmenar J, Martí R, Duarte A (2018) Multi-objective memetic optimization for the bi-objective obnoxious p-median problem. Knowl-Based Syst 144:88–101
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. In: 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 Spectr 22:425–460
Ehrgott M, Gandibleux X (2002) Multiple criteria optimization. In: State of the art annotated bibliographic surveys. Kluwer, Boston
Elleuch MA, Frikha A (2018) Combining the promethee method and mathematical programming for multi-objective facility location problem. Int J Multicrit Decis Mak 7(3/4):195–216
Farahani RZ, Steadieseifi M, Asgari N (2010) Multiple criteria facility location problems: a survey. Appl Math Model 34(7):1689–1709
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 (1971a) Optimal center location in simple networks. Transport Sci 5:212–221
Goldman AJ (1971b) Optimal center location in simple networks. Transport 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. Inform Process Lett 33:169–174
Kalcsics J, Nickel S, Pozo MA, Puerto J, Rodríguez-Chía AM (2014) The multicriteria p-facility median location problem on networks. Eur J Oper Res 235(3):484–493
Kalcsics J, Nickel S, Puerto J, Rodríguez-Chía AM (2015) Several 2-facility location problems on networks with equity objectives. Networks 65(1):1–9
Karatas M, Yakici E (2018) An iterative solution approach to a multi-objective facility location problem. Appl Soft Comput 62:272–287
Nickel S (1995) Discretization of planar location problems. Fachbereich mathematik, PhD Dissertation, 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
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 Optimiz Theory App 126:657–683
Özpeynirci O (2017) On nadir points of multiobjective integer programming problems. J Global Optim 69(3):699–712
Pecci F, Abraham E, Stoianov I (2017) Scalable Pareto set generation for multiobjective co-design problems in water distribution networks: a continuous relaxation approach. Struct Multidiscip Optim 55(3):857–869
Puerto J, Fernández F (1999) Multi-criteria minisum facility location problems. 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
Wang SC, Lin CC, Chen TC, Hsiao H (2018) Multi-objective competitive location problem with distance-based attractiveness for two facilities. Comput Electr Eng 71:37–250
Warburton A (1983) Quasiconcave vector maximization: connectedness of the sets of pareto-optimal and weak pareto-optimal alternatives. J Optimiz Theory App 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 and 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 solution 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 problem 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.

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, this modeling tool allows to distinguish the roles played by the different parties

J. Puerto
IMUS, Universidad de Sevilla, Seville, Spain
e-mail: puerto@us.es

A. M. Rodríguez-Chía (✉)
Dpto. Estadística e Investigación Operativa, Universidad de Cadiz, Cadiz, Spain
e-mail: antonio.rodriguezchia@uca.es

G. Laporte et al. (eds.), *Location Science*,
https://doi.org/10.1007/978-3-030-32177-2_10

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 situations where the driving force in a distribution problem can fall in any of 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 us, 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 to the demand point in the s-th position a smaller weight than the one assigned to demand point in the t-th position. (Note that these weights do not penalize site j but instead they compensate site j' because these weights 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 its 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), Blanco et al. (2013, 2014), Espejo et al. (2009), Nickel et al. (2005), Drezner (2007), Drezner and Nickel (2009a,b) and Rodríguez-Chía et al. (2010), 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; Labbé et al. 2017; Martínez-Merino et al. 2017; Deleplanque et al. 2018; Marín et al. 2009, 2010; Puerto et al. 2011, 2014, 2013; Redondo et al. 2016; Turner et al. 2015) 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; Rozanov and Tamir 2018; Turner and Hamacher 2011). The interested reader is also referred to Chap. 7 in this book and Blanco et al. (2018) for some applications to the location of extensive facilities.

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 one to formulate and solve location problems in different solution spaces (continuous, networks and discrete) 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 solution spaces: 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 one 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 $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. 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 the introduction, 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* $f_\lambda(x), f_\mu(x) \in \mathcal{OMf}(n)$.

(1) $f_\lambda(x)$ *is a continuous function.*
(2) $f_\lambda(x)$ *is a symmetric function , i.e., for any* $x \in \mathbb{R}^n$ $f_\lambda(x) = f_\lambda(sort_n(x))$.
(3) $f_\lambda(x)$ *is a convex function iff* $\lambda_1 \leq \ldots \leq \lambda_n$.
(4) If c_1 *and* c_2 *are constants, then the function* $c_1 f_\lambda(x) + c_2 f_\mu(x) \in \mathcal{OMf}(n)$.
(5) If $\{f_{\lambda^r}(x)\}$ *is a sequence of ordered median functions that pointwise converges to a function* f*, then* $f \in \mathcal{OMf}(n)$.
(6) 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 in general not an* $\mathcal{OMf}$.
(7) Let $p < n-1$ *and* $x^p = (x_1, \ldots, x_p)$, $x^{\setminus p} = (x_{p+1}, \ldots, x_r)$. *If* $f_\lambda(x) \in \mathcal{OMf}(n)$ *then* $f_{\lambda^p}(x^p) + f_{\lambda^{\setminus p}}(x^{\setminus p}) \lesseqgtr f_\lambda(x)$.
(8) Every ordered median function $\mathcal{OMf}(n)$ *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 recall 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 a 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:

$$\min_{x \in \mathbb{R}^n} F(x) = \langle \lambda, sort_n((w_1\gamma(x - a_1), \ldots, w_n\gamma(x - a_n))) \rangle. \tag{10.4}$$

Note that the problem is well-defined even if ties occur. In that case any order of the tied positions gives the same value.

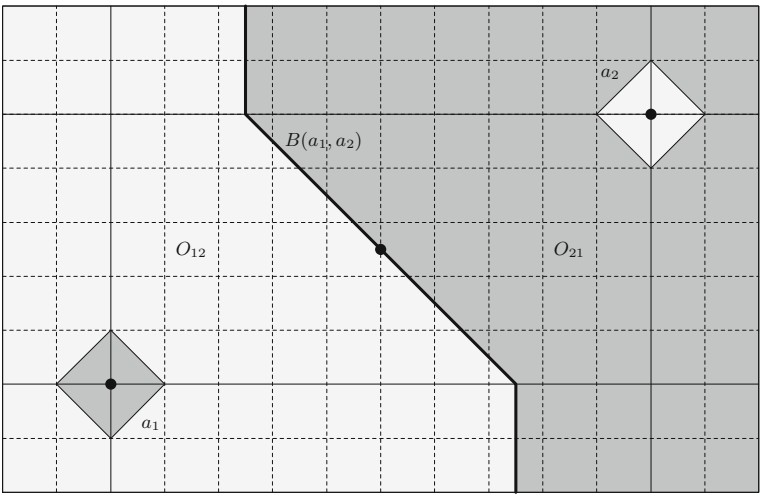


Fig. 10.1 Two regions where the function of Example 10.1 has different linear representation

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 as gauge 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,

$$F(a_1) = 100 \times 0 + 1 \times 15 = 15$$
$$F(a_2) = 100 \times 0 + 1 \times 15 = 15$$
$$F(\frac{1}{2}(a_1 + a_2)) = 100 \times 7.5 + 1 \times 7.5 = 757.5.$$

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 $Ext(B) = \{e_g : g = 1, \ldots, \mathscr{G}\}$. 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 $\mathbb{R}^n$ 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 $O(n^2\mathscr{G}^2)$ where $\mathscr{G}$ 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

$$O_\sigma = \{x \in \mathbb{R}^n : w_{\sigma_1}\gamma(x - a_{\sigma_1}) \leq \ldots \leq w_{\sigma_n}\gamma(x - a_{\sigma_n})\}.$$

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 $O(n^4\mathscr{G}^2)$, 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 in the interior of 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 doable in the plane for polyhedral gauges, where one can easily derive an appealing geometrical algorithm to solve these problems. 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 $O(n^4\mathscr{G}^2)$. Next, one needs to evaluate the objective function in each vertex of the subdivision. Each evaluation can be done in $O(n\mathscr{G}\log n\mathscr{G})$. This results in an algorithm that solves the problem in the plane with a complexity of $O(n^5\mathscr{G}^3\log n\mathscr{G})$.

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:

$$\begin{array}{lll} \text{minimize} & \sum_{i=1}^{n} \lambda_i z_{\sigma_i} & \\ \text{subject to} & w_i \langle e_g^0, x - a_i\rangle \le z_i, \; e_g^o \in B^o, i = 1, 2, \ldots, n & (P_\sigma) \\ & z_{\sigma_i} \le z_{\sigma_{i+1}} \qquad i = 1, 2, \ldots, n-1 & \end{array}$$

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

$$w_i \langle e_{g_i}^o, X^* - a_i\rangle = z_i, i = 1, 2, \ldots, n \text{ , for some } g_i,$$

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 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 of the next algorithm 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 the global convexity property of the objective function, 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 for the convex case. Alternatively, one could simply resort to general randomized subgradient descent algorithms which, under mild conditions (see Ruszczynski and Syski 1986) will converge to the global optimal solution due to the finiteness of the linearity regions of these problems.

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. For this, 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.

$$\text{minimize} \sum_{j=1}^{n} \lambda_j \theta_j \tag{10.5}$$

$$\text{subject to } (1 - z_{ij})M + \theta_j \geq w_i \langle e_g^0, x - a_i \rangle, e_g^o \in B^o, \ i, j = 1, 2, \ldots, n \tag{10.6}$$

$$\sum_{i=1}^{n} z_{ij} = 1, \quad j = 1, \ldots, n \tag{10.7}$$

$$\sum_{j=1}^{n} z_{ij} = 1, \quad i = 1, \ldots, n \tag{10.8}$$

$$\theta_j \leq \theta_{j+1}, \quad j = 1, \ldots, n-1 \tag{10.9}$$

$$\theta_j \geq 0, \quad j = 1, \ldots, n \tag{10.10}$$

$$z_{ij} \in \{0, 1\}, \quad i, j = 1, \ldots, n \tag{10.11}$$

$$x \in \mathbb{R}^d. \tag{10.12}$$

Constraints (10.7) and (10.8) define a permutation by placing a single distance to a facility at each position 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 an 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. (2016), 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 nonspecialized 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 a 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, Kalcsics et al. 2015, 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 problem 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 $(u_1, \ldots, u_n)$), $x \in \mathbf{K}$ 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 $\tilde{f}_j(x) := \|x - a_j\|_\tau$.

- $\mathbf{K} := \{x \in \mathbb{R}^d : g_j(x) \geq 0,\ j = 1, \ldots, \ell\} \subset \mathbb{R}^d$ 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 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 lifts 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.

$$\overline{\rho}_\lambda = \text{minimize} \sum_{j=1}^{m} \lambda_j \sum_{i=1}^{m} f_i(u) w_{ij} := p_\lambda(x, u, v, w) \tag{10.13}$$

$$\text{subject to} \sum_{j=1}^{m} w_{ij} = 1, i = 1, \ldots, m, \tag{10.14}$$

$$\sum_{i=1}^{m} w_{ij} = 1, j = 1, \ldots, m, \tag{10.15}$$

$$\sum_{i=1}^{m} w_{ij} f_i(u) \leq \sum_{i=1}^{m} w_{ij+1} f_i(u), j = 1, \ldots, m-1, \tag{10.16}$$

$$w_{ij}^2 - w_{ij} = 0, i, j = 1, \ldots, m, \tag{10.17}$$

$$v_{k\ell}^{2s} = (x_\ell - a_{k\ell})^{2r}, \ k = 1, \ldots, n, \ \ell = 1, \ldots, d, \tag{10.18}$$

$$u_k^r = (\sum_{\ell=1}^{d} v_{k\ell})^s, \ k = 1, \ldots, n, \tag{10.19}$$

$$\sum_{j=1}^{m} w_{ij}^2 \leq 1,\ i = 1, \ldots, m, \tag{10.20}$$

$$\sum_{j=1}^{d} v_{ij}^2 \leq \bar{M}^{2\tau},\ i = 1, \ldots, n, \tag{10.21}$$

$$w_{ij} \in \mathbb{R},\ i, j = 1, \ldots, m, \tag{10.22}$$

$$v_{k\ell} \geq 0, u_k \geq 0, k = 1, \ldots, n, \ell = 1, \ldots, d, \tag{10.23}$$

$$x \in \mathbf{K}. \tag{10.24}$$

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 $f_{(1)}(u) \leq \cdots \leq f_{(m)}(u)$. 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:

$$
\begin{aligned}
\text{minimize} \quad & \sum_{j=n-k+1}^{n} \sum_{i=1}^{n} u_i w_{ij} & \\
\text{subject to} \quad & \sum_{i=1}^{n} w_{ij} = 1, & j = 1, \ldots, n, \\
& \sum_{i=1}^{n} w_{ij} = 1, & j = 1, \ldots, n, \\
& \sum_{i=1}^{n} w_{ij} u_i \leq \sum_{i=1}^{n} w_{ij+1} u_i, & j = 1, \ldots, n-1 \\
& w_{ij}^2 - w_{ij} = 0, & i, j = 1, \ldots, n, \\
& v_{k\ell}^2 = (x_\ell - a_{k\ell})^6, & k = 1, \ldots, n, \ \ell = 1, \ldots, 2, \\
& u_k^3 = (\sum_{\ell=1}^{d} v_{k\ell}), & k = 1, \ldots, n, \\
& \sum_{j=1}^{n} w_{ij}^2 \leq 1, & i = 1, \ldots, n, \\
& \sum_{j=1}^{2} v_{ij}^2 \leq \bar{M}^6, & i = 1, \ldots, n, \\
& w_{ij} \in \mathbb{R}, & i, j = 1, \ldots, m, \\
& v_{k\ell} \geq 0, u_k \geq 0, & k = 1, \ldots, n, \ell = 1, \ldots, d, \\
& x \in \mathbf{K}
\end{aligned}
$$

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 conclusion, $\rho_\lambda = \overline{\rho}_\lambda$. Moreover, if $\mathbf{K} \subset \mathbb{R}^d$ satisfies the Archimedean property then $\overline{\mathbf{K}} \subset \mathbb{R}^{d+m^2+n(d+1)}$ also satisfies the 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 $g = \sum_{\gamma \in \mathbb{N}^d} g_\gamma x^\gamma \in \mathbb{R}[x]$, 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 $\nu_j := \lceil (\deg h_j)/2 \rceil$, where $\{g_1, \ldots, g_\ell\}$, and $\{h_1, \ldots, h_{3m+m^2+n(d+3)}\}$ are the polynomial constraints that define $\mathbf{K}$ and $\overline{\mathbf{K}} \setminus \mathbf{K}$ in (10.13)–(10.24), respectively. For

$$r \geq r_0 := \max\{\max_{k=1,\ldots,\ell} \xi_k, \max_{j=0,\ldots,3m+m^2+n(d+3)} \nu_j\},$$

we introduce the hierarchy of semidefinite programs:

$$\begin{array}{lll} \text{minimize}_\mathbf{y} & \mathrm{L}_\mathbf{y}(p_\lambda) & \\ \text{subject to} & \mathrm{M}_r(\mathbf{y}) \succeq 0, & \\ & \mathrm{M}_{r-\xi_k}(g_k, \mathbf{y}) \succeq 0, & k = 1, \ldots, \ell, \\ & \mathrm{M}_{r-\nu_j}(h_j, \mathbf{y}) \succeq 0, & j = 1, \ldots, 3m + m^2 + n(d+3), \end{array} \qquad (\overline{\mathbf{Q}}_r)$$

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 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.

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 to a nonnegative weight. A point x on an edge $e = \{i, j\}$ is defined as a pair $x = (e, t),\ t \in [0, 1]$, with

$$d(v_k, x) := d(x, v_k) := \min\{d(v_k, v_i) + t\ell(e), d(v_k, v_j) + (1 - t)\ell(e)\}. \tag{10.26}$$

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.

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

10.4.1 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 to be 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

$$f_\lambda(d(x)) := \sum_{i=1}^{n} \lambda_i d_{(i)}(x) \quad \text{with} \quad \lambda = (\lambda_1, \ldots, \lambda_n) \geq 0\,, \tag{10.27}$$

and

$$M(\lambda) := \min_{x \in P(G)} f_\lambda(d(x)). \tag{10.28}$$

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 3 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).

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

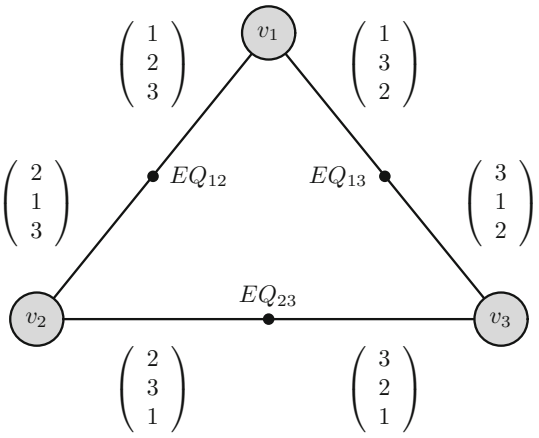


Fig. 10.2 A 3-node network with $EQ = \{EQ_{12}, EQ_{13}, EQ_{23}, v_1, v_2, v_3\}$ and the geometrical subdivision

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

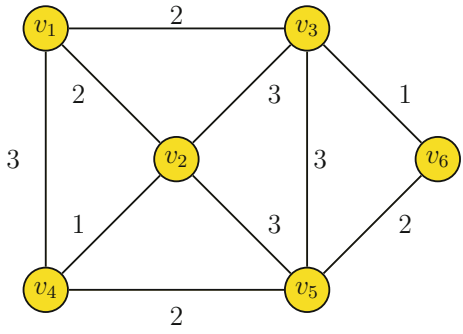


Fig. 10.3 A 6-node network used in Example 10.3 where the numbers on the edges represent their length

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}		$\frac{1}{2}$		$\frac{2}{3}$			$\frac{1}{9}$			$\frac{2}{3}$

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 choices for λ. To describe the solution set we use the notation EQ_{kl}^{ij} 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, i.e., 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 $d(v, X_p) := \min_{i=1,\ldots,p} d(v, x_i)$ for all $v \in V$; $d(X_p) := (w_1 d(v_1, X_p), \ldots, w_n d(v_n, X_p))$ 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

$$d_\lambda(d(X_p)) = \sum_{i=1}^{n} \lambda_i d_{(i)}(X_p) = \text{minimize}\{\sum_{i=1}^{n} \lambda_i d_{\pi(i)}(X_p) : \pi \in \Pi(\{1, \ldots, n\})\}.$$

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 $d(v, x(s)) = s$, $s \in [0, l(e)]$.

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	$\lambda = (0, 0, 0, 0, 0, 1)$	$EQ_{46}^{23}, EQ_{46}^{35}, EQ_{34}^{56}$	5
2-Centra	$\lambda = (0, 0, 0, 0, \frac{1}{2}, \frac{1}{2})$	$[EQ_{35}^{23}, EQ_{56}^{23}]$, $[EQ_{36}^{35}, EQ_{14}^{35}]$, $[EQ_{14}^{56}, EQ_{13}^{56}]$	5
3-Centra	$\lambda = (0, 0, 0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	EQ_{26}^{23}	$\frac{40}{9}$
Median	$\lambda = (1, 1, 1, 1, 1, 1)$	$EQ_{16}^{23} = v_3$	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_{34}^{56}, $0 \leq \hat{\lambda} \leq \frac{36}{43}$, v_3 otherwise	$-\frac{17}{12}\hat{\lambda} + 5$, $-5\hat{\lambda} + 8$
Noname	$\lambda = (1, 1, 0, 0, 1, 1)$	$EQ_{14}^{23}, EQ_{12}^{56}$	13

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

$$d_{\pi(j)}(X_p(s)) = \min\{d(v_{\pi(j)}, x_1), \ldots, \min\{d(v_{\pi(j)}, a) + s, d(v_{\pi(j)}, b) + \ell(e) - s\}, \ldots, d(v_{\pi(j)}, x_n)\}$$

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 one vertex of V instead of x_i. 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 single facility problem (independent of the structure of λ). It may 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, see Puerto et al. (2018) for an updated review. 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 (different) 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 (the same) 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 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.

Fig. 10.4 Network of Example 10.4 where the dots, the ticks and the small ticks are the nodes, the equilibrium points and the elements of Y, respectively. Observe that in this case there are no pairs in T

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

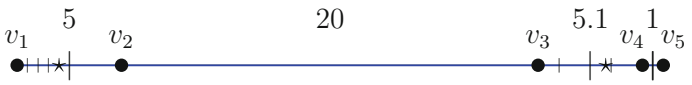


Fig. 10.5 Network of Example 10.5 where the dots, the ticks, the small ticks and the stars are the nodes, the equilibrium points, the elements of Y and T, respectively. By domination and symmetry arguments not all the candidates are necessary and therefore, they are not depicted

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		

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 $\ell(\{1,2\}) = 5, \ell(\{2,3\}) = 20, \ell(\{3,4\}) = 5.1, \ell(\{4,5\}) = 1$. 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 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 do not need 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-

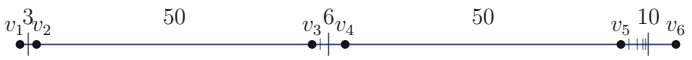


Fig. 10.6 Network of Example 10.6, using the same notation as in Fig. 10.4

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 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 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.

For the case of locating extensive facilities on the line, in Rozanov and Tamir (2018), it is proved a nestedness property (given any two facility lengths t_1, t_2, $0 \leq t_1 < t_2$, there is an optimal solution with length t_1 which lies within some optimal solution with length t_2). In addition, in Schnepper (2017), Schnepper et al. (2019), it is analyzed the p-k-max problem on networks, a particular case of the ordered median problem. The reader is referred to Puerto et al. (2018) for an updated review of results on location of extensive facilities on networks.

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

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 problem 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), Labbé et al. (2017), Deleplanque et al. (2018), 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:

$$x_{ij}^k = \begin{cases} 1, & \text{if client } i \text{ is supplied by server } j \text{ and is the } k\text{-th} \\ & \text{cheapest cost allocation} \\ 0, & \text{otherwise,} \end{cases} \qquad i, j, k = 1, \ldots, n,$$

$$y_j = \begin{cases} 1, & \text{if the server at } j \text{ is open} \\ 0, & \text{otherwise,} \end{cases} \qquad j = 1, \ldots, n.$$

Hence, the formulation of the model is:

$$\text{minimize} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \lambda_k c_{ij} x_{ij}^k \tag{10.33}$$

$$\text{subject to} \sum_{j=1}^{n} \sum_{k=1}^{n} x_{ij}^k = 1, \quad i = 1, \ldots, n \tag{10.34}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij}^k = 1, \quad k = 1, \ldots, n \tag{10.35}$$

$$\sum_{i=1}^{n} \sum_{k=1}^{n} a_i x_{ij}^k \leq b_j y_j, \quad j = 1, \ldots, n, \tag{10.36}$$

$$\sum_{j=1}^{n} y_j = p, \tag{10.37}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} x_{ij}^k \leq \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} x_{ij}^{k+1}, \quad k = 1, \ldots, n-1. \tag{10.38}$$

$$x_{ij}^k \in \{0, 1\}, \quad i, j, k = 1, \ldots, n; \tag{10.39}$$

$$y_j \in \{0, 1\}, \quad j = 1, \ldots, n\ . \tag{10.40}$$

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 capacity constraints 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 H 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: $c_{(0)} := 0 < c_{(1)} < c_{(2)} < \cdots < c_{(H)} := \max_{1\leq i,j\leq n}\{c_{ij}\}$.

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 ($j = 1,\ldots,n$ and $k = 1,\ldots,H$):

$$u_{jk} := \begin{cases} 1, \text{ if the } j\text{–th smallest allocation cost is at least } c_{(k)}, \\ 0, \text{ otherwise.} \end{cases} \tag{10.41}$$

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 $\sum_{j=1}^{n}\sum_{k=1}^{H}\lambda_j \cdot (c_{(k)} - c_{(k-1)}) \cdot u_{jk}$.

First of all, we need to impose the following group of sorting constraints on the u_{jk}–variables: $u_{j+1,k} \geq u_{jk}$, $j = 1,\ldots,n-1$; $k = 1,\ldots,H$. 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 and capacities are satisfied. For these reasons we introduce: (1) the variables x_{ij} (binary allocation) :

$$x_{ij} = \begin{cases} 1, & \text{if the client } i \text{ is allocated to server } j \\ 0, & \text{otherwise} \end{cases} \tag{10.42}$$

and (2) the 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

$$\text{minimize} \sum_{j=1}^{n}\sum_{k=1}^{H}\lambda_j(c_{(k)}-c_{(k-1)})u_{jk} \tag{10.43}$$

$$\text{subject to} \sum_{j=1}^{n} x_{ij}=1, \quad i=1,\ldots,n \tag{10.44}$$

$$\sum_{i=1}^{n} a_i x_{ij} \le b_j y_j, \quad j=1,\ldots,n, \tag{10.45}$$

$$x_{ij} \le y_j \quad i,j=1,\ldots,n \tag{10.46}$$

$$\sum_{j=1}^{n} y_j = p \tag{10.47}$$

$$\sum_{j=1}^{n} u_{jk} = \sum_{i=1}^{n} \sum_{\substack{j=1\ldots,n \\ c_{ij}\ge c_{(k)}}} x_{ij}, \quad k=1,\ldots,H \tag{10.48}$$

$$u_{j+1k} \ge u_{jk}, \quad j=1,\ldots,n-1;\ k=1,\ldots,H \tag{10.49}$$

$$u_{jk} \in \{0,1\}, \quad j=1,\ldots,n;\ k=1,\ldots,H \tag{10.50}$$

$$x_{ij}, y_j \in \{0,1\}, \quad i,j=1,\ldots,n; \tag{10.51}$$

Since the proposed formulation contains $O(nH)$ binary variables and $O(nH)$ constraints, fast solution times for larger problem instances, using standard software tools, are very unlikely. In this sense, 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 $u_{jk} \ge u_{jk+1}, \quad j=1,\ldots,n, \quad k=1,\ldots,H-1$. 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: $\sum_{i\in S}\sum_{k:c_{ik}\geq c_{(j)}} x_{ik} \leq \sum_{i=n-r+1}^{n} u_{ij}, \quad \forall\ S \subseteq \{1,\ldots,n\}, \quad |S| = r, \quad r = 1,\ldots,n,\ j = 1,\ldots,H$. Note that there is 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: $\sum_{j:c_{ij}\geq c_{(k)}} x_{ij} + \sum_{j:c_{ij}<c_{(k)}} y_j \geq 1,\ i = 1,\ldots,n$.

In addition, we mention the staircase inequalities introduced by Labbé et al. (2017), where several new formulations for the Uncapacitated Discrete Ordered Median Problem (DOMP) based on its similarity with some scheduling problems are presented (some of them with a considerably smaller number of constraints).

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 and Cut algorithm to solve faster and hence 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 valid inequalities, as for instance in Puerto et al. (2011), leading to solve larger problem sizes of capacitated discrete ordered median problems.

Finally, we would like to mention that two ad-hoc solution procedures have been developed for the uncapacitated DOMP, the first one based on a parallelized Lagrangian relaxation approach, see Redondo et al. (2016) and the second one is a Branch-Price-and-Cut procedure, see Deleplanque et al. (2018). These two approaches could also be adapted to tackle the capacitated version of this problem.

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 solution 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

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	1939.3	2092.6
Mt	1.7	9.9	4.5	12.3	20.2	66.6	102	135.9	59.6	63.5	198.3	2644.6	427.6	3600.8	3600.4
An	47	10.2	51.6	12.2	191	1557.8	344	1055.8	607.4	124.4	2357	31,512.6	716.4	23,586.4	42,081.6
Mn	231	31	127	39	451	4077	767	4865	1647	485	4221	129,917	1848	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	1786.9	338.9	260.2	2162.7	1977.3	1416.2	2578.6
Mt	16.7	13.9	4.3	81.4	34.4	629.9	130.6	3599.7	3599.3	568	713	3600	3599.9	3600	3600.4
An	17.4	391	65	605.6	1558	21,804	685.8	51,162.4	82,292.4	3052.4	7549	68,845.4	16,741.2	22,036.8	41,283.6
Mn	37	925	123	1189	3467	47,542	1391	211,405	167,386	4465	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	2336.9	676.1	1155.4	2222
Mt	18.5	10.6	9.2	22.7	115.3	191.7	162	209.2	3012.1	394.9	1051.8	3600	814.2	2265	3599.7
An	20.6	101.4	32.2	140.6	1933.8	3527.2	1662.2	3552.8	44,945	2563.4	14,941.6	71,892.6	5663.4	24,109.4	29,005.6
Mn	81	487	127	493	8047	12,769	5623	9626	171,469	7745	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
Trimmed mean															
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	4487.6	998.2	225	2540.6	6100.2
Mn	8	1	13	187	335	2235	406	1191	27	29	19,803	3223	753	3806	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

ordered median location problems. Moreover, the extensive list of references that have been included may result in an interesting source, for expert readers, to carry out a deeper study of this topic.

Acknowledgements The authors were partially supported by projects MTM2016-74983-C2-01/02-R (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 (2014) Revisiting several problems and algorithms in continuous location with l_p norms. Comput Optim Appl 58:563–595

Blanco V, Puerto J, Ben-Ali SEH (2016) Continuous multifacility ordered median location problems. Eur J Oper Res 250(1):56–64

Blanco V, Puerto J, Salmerón, R (2018) A general framework for locating hyperplanes to fitting set of points. Comput Oper Res 95:172–193

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

Deleplanque S, Labbé M, Ponce D, Puerto J (2019) An extended version of a branch-price-and-cut procedure for the discrete ordered median problem. Informs J Comput. https://doi.org/10.1287/ijoc.2019.0915

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 ed. 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
Kalcsics J, Nickel S, Puerto J, Rodríguez-Chía AM (2015) Several 2-facility location problems on networks with equity objectives. Networks 65(1):1–9
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
Labbé M, Ponce D, Puerto J (2017) A comparative study of formulations and solution methods for the discrete ordered p-median problem. Comput Oper Res 78:230–242
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
Martínez-Merino LI, Albareda-Sambola M, Rodríguez-Chía AM (2017) The probabilistic p-center problem: planning service for potential customers. Eur J Oper Res 262:509–520
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, Dresden, OR 2000, September 9–12, 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 Optimiz Theory App 126:657–683
Okabe A, Boots B, Sugihara K (1992) Spatial tessellations: concepts and applications of Voronoĭ diagrams. In: 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, September 5–7, 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 234:61–76

Puerto J, Ricca F, Scozzari A (2018) Extensive facility location problems on networks: an updated review. TOP 26(2):187–226

Redondo JL, Marín A, Ortigosa PM (2016) A parallelized Lagrangean relaxation approach for the discrete ordered median problem. Ann Oper Res 246(1–2):253–272

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

Rozanov M, Tamir A (2018) The nestedness property of location problems on the line. TOP 26:257–282

Ruszczynski A, Syski W (1986) On convergence of the stochastic subgradient method with on-line stepsize rules. J Math Anal Appl 114(2):512–527

Schnepper T (2017) Location problems with k-max functions-modelling and analysing outliers in center problems. In: PhD dissertation, Universität Wuppertal, Germany

Schnepper T, Klamroth K, Stiglmayr M, Puerto J (2019) Exact algorithms for handling outliers in center location problems on networks using k-max functions. Eur J Oper Res 273(2):441–451

Schöbel A, Scholz D (2010) The big cube small cube solution method for multidimensional facility location problems. Comput Oper Res 37:115–122

Turner L, Hamacher HW (2011) On universal shortest paths. In: Operations research proceedings 2010, pp 313–318

Turner L, Ehrgott M, Hamacher HW (2015) On the generality of the greedy algorithm for solving matroid base problems. Discrete Appl Math 195:114–128

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 This chapter covers different aspects related with facility location problems involving time-dependent parameters. The emphasis is put on problems defined over a multi-period finite planning horizon. An overview of continuous and network problems is presented, although most of the chapter focuses 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 time-invariant one. Current challenges and future trends on the topic are discussed.

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 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.

In order to capture predictable variations in the parameters of a facility location problem, we often have to consider a so-called dynamic or time-dependent model.

S. Nickel
Institute for Operations Research and Research Center for Information Technology (FZI), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
e-mail: stefan.nickel@kit.edu

F. Saldanha-da-Gama (✉)
Departamento de Estatística e Investigação Operacional e Centro de Matemática, Aplicações Fundamentais e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal
e-mail: fsgama@ciencias.ulisboa.pt

G. Laporte et al. (eds.), *Location Science*,
https://doi.org/10.1007/978-3-030-32177-2_11

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 to (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 over 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 (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 there are 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.

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 devised by Wesolowsky (1973) is the following:

$$\text{Minimize} \quad \sum_{t \in T} \sum_{j \in J_t} c_{tj}(x_t, y_t) + \sum_{t=2}^{|T|} f_t z_t \tag{11.1}$$

$$\text{subject to} \quad z_t = 0 \text{ if } d_{t-1,t} = 0;\ z_t = 1, \text{otherwise}, \quad t \in T \tag{11.2}$$

$$z_t \in \{0, 1\}, \quad t \in T. \tag{11.3}$$

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 $t \in T \setminus \{1\}$. 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 investigating 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 is defined considering the optimal location for the facility in the different periods. These locations define the possible sates for all periods (stages). Incomplete dynamic programming is then applied. The method was later converted into an exact one by Sweeney and Tatham (1976) who enlarged 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) who 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 $w_j(.)$ as in Drezner and Wesolowsky (1991). At the beginning of each time period $t \in T$, exactly one facility is to be installed. The decision variables represent 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 two-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 a general network, Mesa (1991) studied 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 that was closer to what could be refereed to as an extension of the p-median problem to a multi-period setting. In that problem, α_t represents the number of points that must be located in each period $t \in T$. Such values must satisfy $\sum_{t \in T} \alpha_t = p$. 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 authors define the exact location x_t as the distance between the median and one of the extreme vertices of the edge that contains the median; $d_t(v_j, x_t)$ is the length of the shortest path between v_j and x_t in period $t \in T$; $g(t)$ is a function representing the unit cost for relocating the facility in the end of period t moving it from location x_t in that period to location x_{t+1} in period $t+1$ ($t \in T, x_{|T|+1} = x_{|T|}$). Hakimi et al. (1999) also proved that the vertex-optimality property holds for the 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 minimum weight of a 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 node x' in X_t and for every node x'' in X_{t+1} the weight of the edge (x', x'') is set equal to $d_t(x', x'')$. 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 has 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 one-center problem on a network was proposed also by Hakimi et al. (1999). The model is the following (the notation is the same as above):

$$\underset{x_1, x_2, \ldots, x_{|T|}}{\text{Minimize}} \quad \sum_{t \in T} \max_{j \in V} \left\{ w_{jt} d_t(v_j, x_t) + g(t) d_t(x_t, x_{t+1}) \right\}. \tag{11.11}$$

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. Assume a set of nodes J 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 every period. The problem of deciding the best location for the facilities in each period, 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 of 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.

The above model still has little "multi-period flavor" because it can be decoupled, leading to $|T|$ single-period problems. Nevertheless, it represents a good starting point for what we discuss next. In fact, a more interesting problem emerges if we account for opening and closing costs for the facilities. This was first done by Wesolowsky and Truscott (1975), who proposed the following model for the extended problem:

$$\text{Minimize} \quad \sum_{t\in T}\sum_{i\in I}\sum_{j\in J} c_{ijt}x_{ijt} + \sum_{t\in T}\sum_{i\in I} g_{it}z'_{it} + \sum_{t\in T}\sum_{i\in I} h_{it}z''_{it} \tag{11.17}$$

$$\text{subject to} \quad (11.13)\text{–}(11.16)$$

$$\sum_{i\in I} z'_{it} \leq m_t, \qquad t \in T \tag{11.18}$$

$$x_{iit} - x_{ii,t-1} + z''_{i,t-1} - z'_{it} = 0, \qquad t \in T \setminus \{1\},\ i \in I \tag{11.19}$$

$$z'_{it}, z''_{it} \in \{0, 1\}, \qquad t \in T,\ i \in I. \tag{11.20}$$

In this model, facilities can be opened (closed) at the beginning (end) of a time period; m_t is the maximum number of facilities that can be opened in each period $t \in T$. 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. The parameters g_{it} and h_{it} ($i \in I, t \in T$) denote the opening and closing costs, respectively. Wesolowsky and Truscott (1975) solved the above problem using dynamic programming. However, the dimension of the state space is exponential in the number of potential locations for the facilities and thus the procedure can only be applied to small instances.

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

$$z'_{it} \geq x_{iit} - x_{ii,t-1}, \quad t \in T,\ i \in I, \tag{11.22}$$

with $x_{ii0} = 0$, $i \in I$. For this problem, Galvão and Santibañez-Gonzalez (1992) proposed two Lagrangian 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 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. We discuss this aspect by considering another well-known problem: the uncapacitated facility location problem (UFLP)—see Chap. 4. 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 can be formulated as follows:

$$\text{Minimize} \quad \sum_{t \in T} \sum_{i \in I} f_{it} y_{it} + \sum_{t \in T} \sum_{i \in I} \sum_{j \in J} c_{ijt} x_{ijt} \tag{11.23}$$

$$\text{subject to} \quad \sum_{i \in I} x_{ijt} = 1, \quad t \in T,\ j \in J \tag{11.24}$$

$$\sum_{j \in J} x_{ijt} \leq n y_{it}, \quad t \in T,\ i \in I \tag{11.25}$$

$$x_{ijt} \geq 0, \quad t \in T,\ i \in I,\ j \in J \tag{11.26}$$

$$y_{it} \in \{0, 1\}, \quad t \in T,\ i \in I. \tag{11.27}$$

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 was operating in some past period and then closed). 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:

$$\sum_{t \in T} \sum_{i \in I} g_{it} y_{it} \left(1 - y_{i,t-1}\right) + \sum_{t \in T} \sum_{i \in I} f_{it} y_{it} + \sum_{t \in T} \sum_{i \in I} \sum_{j \in J} c_{ijt} x_{ijt}, \tag{11.28}$$

with $y_{i0} = 0$, $i \in I$. Warszawski (1973) proposed a dynamic programming algorithm for 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 Lagrangian 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 suggested by Canel and Khumawala (1997), who 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 introduced a profit maximization objective. The problem was formulated 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 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 $t \in T$. We will see below that constraints (11.30) had been proposed before but for another problem. Canel and Khumawala (1997) developed a branch-and-bound procedure for this problem, adapting the algorithm proposed by Khumawala (1972).

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 applied to obtain lower and upper bounds. The gap is closed by 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 more than once during the planning horizon. This may make sense for seasonal facilities, such as warehouses that can often be rented for short time intervals but it 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 open (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:

Minimize (11.23)

subject to (11.24)–(11.27)

$$y_{it} \leq y_{i,t-1}, \quad t = 2, \ldots, |T|,\ i \in I^c \tag{11.39}$$

$$y_{it} \geq y_{i,t-1}, \quad t = 2, \ldots, |T|,\ i \in I^o. \tag{11.40}$$

The above model contains the pure phase-in problem (facilities can only be opened) and the pure phase-out problem (facilities can only be closed) as particular cases. In fact, Roodman and Schwarz (1977) extended the work presented by Roodman and Schwarz (1975) in which a pure phase-out problem had been considered.

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 used by Canel and Khumawala (1997).

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) reformulated model (11.23)–(11.27), (11.39) and (11.40). The 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 binary variable z_{it}, is 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 the facility 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 models (11.23)–(11.27), (11.39), (11.40) and (11.41)–(11.45) are equivalent. 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 exploited by Van Roy and Erlenkotter (1982). By multiplying constraints (11.43) by -1 we obtain the

following dual of the linear relaxation of model (11.41)–(11.45):

$$\text{Maximize} \quad \sum_{t\in T}\sum_{j\in J} v_{jt} \tag{11.46}$$

$$\text{subject to} \quad v_{jt} - w_{ijt} \le c_{ijt}, \quad t \in T,\ i \in I,\ j \in J \tag{11.47}$$

$$\sum_{j\in J}\sum_{\tau\in T_{it}} w_{ij\tau} \le F_{it}, \quad t \in T,\ i \in I \tag{11.48}$$

$$w_{ijt} \ge 0, \quad t \in T,\ i \in I,\ j \in J. \tag{11.49}$$

The dual variables v_{jt} and w_{ijt} ($t \in T$, $i \in I$, $j \in J$) are associated with constraints (11.42) and (11.43), respectively (with the latter 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) become:

$$v_{jt}\left(\sum_{i\in I} x_{ijt} - 1\right) = 0 \qquad t \in T,\ j \in J$$

$$w_{ijt}\left(\sum_{\tau\in \overline{T}_{it}} z_{i\tau} - x_{ijt}\right) = 0, \quad t \in T,\ i \in I,\ j \in J$$

$$x_{ijt}\left(v_{jt} - c_{ijt} - w_{ijt}\right) = 0, \quad t \in T,\ i \in I,\ j \in J$$

$$z_{it} S_{it} = 0, \qquad t \in T,\ i \in I,$$

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 a trivial dual feasible solution ($v_{jt} = \min_{i \in I}\{c_{ijt}\}$, $t \in T$, $j \in J$) 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) developed 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 miss one important feature found in many applications: capacity constraints. Suppose that the capacity of a facility located at $i \in I$ is limited and denote it by Q_i. Capacity constraints can be easily embedded in model (11.23), (11.24)–(11.27), (11.39), (11.40) by replacing (11.25) with

$$\sum_{j \in J} d_{jt} x_{ijt} \leq Q_i \, y_{it}, \quad i \in I, \; t \in T. \tag{11.51}$$

Castro et al. (2017) studied the pure phase-in version of this problem (facilities cannot be closed) and enriched the model in two ways: first, by considering the limits $p_1, \ldots, p_{|T|}$ above introduced representing the maximum number of facilities that can be operating in each period $t \in T$; second, by assuming that a service level below 100% is acceptable given that an opportunity cost is paid.

Let us denote by d_{jt} the demand of customer $j \in J$ in period $t \in T$, by o_{jt} the cost per unit of demand of customer $j \in J$ in period $t \in T$ that is not supplied and by v_{jt} the proportion of the demand of customer $j \in J$ in period $t \in T$ that is not supplied (or that is outsourced). The problem studied by Castro et al. (2017) is the following:

$$\text{Minimize} \quad \sum_{t \in T} \sum_{i \in I} f_{it} y_{it} + \sum_{t \in T} \sum_{j \in J} \left(\sum_{i \in I} c_{ijt} x_{ijt} + o_{jt} d_{jt} v_{jt} \right) \tag{11.52}$$

$$\text{subject to} \quad \sum_{i \in I} x_{ijt} + v_{jt} = 1, \quad t \in T, \; j \in J \tag{11.53}$$

$$\sum_{i \in I} y_{it} \leq p_t, \quad t \in T \tag{11.54}$$

(11.26), (11.27), (11.40), (11.51)

$$v_{jt} \geq 0, \quad j \in J, \ t \in T. \tag{11.55}$$

This problem contains as particular case the "standard" multi-period pure phase-in location problem as described above, as well as the well-known capacitated facility location problem. Nevertheless, it covers features not included in those two cases. For the above problem, the authors developed a Benders decomposition. By using a specialized interior-point method for solving the Benders subproblems, the authors were able to solve instances of a size never attempted for capacitated static and multi-period facility location problems (up to 200 locations and one million customers).

The inclusion of capacity constraints in model (11.41)–(11.45) can be accomplished by replacing (11.43) with

$$\sum_{j \in J} d_{jt} x_{ijt} \leq Q_i \sum_{\tau \in \overline{T}_{it}} z_{i\tau}, \quad t \in T, \ i \in I. \tag{11.56}$$

The resulting model was adopted 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

$$\sum_{t \in T} \sum_{i \in I} R_{kit} z_{it} \leq r_k, \quad k \in K. \tag{11.57}$$

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, states that the operational capacity must be at least equal to the total demand. In the multi-period case, these constraints are written as

$$\sum_{i \in I} \left(Q_i \sum_{\tau \in \overline{T}_{it}} z_{i\tau} \right) \geq \sum_{j \in J} d_{jt}, \quad t \in T, \tag{11.58}$$

which can be easily accommodated in (11.57).

For the linear relaxation of model (11.41)–(11.45), (11.56) and (11.57), 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.

When considering capacity constraints in multi-period facility location models, we can envisage the possibility of making adjustments in the capacity of the facilities throughout the planning horizon. This aspect was also taken into account by Van Roy and Erlenkotter (1982) who considered exogenous time-dependent capacities Q_{it} ($i \in I, t \in T$). Nevertheless, for some applications this may still be insufficient because no connection is established between the capacities in different periods.

The problem of planning for the capacity expansion of existing facilities was very popular 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. 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 considered for representing the facility types that can be located at i. Denote by c_{ijpt} the cost of 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 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:

$$\text{Minimize} \quad \sum_{t\in T}\sum_{i\in I}\sum_{p\in P_i} f_{ipt}\, y_{ipt} + \sum_{t\in T}\sum_{i\in I}\sum_{j\in J}\sum_{p\in P_i} c_{ijpt}\, x_{ijpt} \tag{11.59}$$

$$\text{subject to} \quad \sum_{i\in I}\sum_{p\in P_i} x_{ijpt} = 1, \quad t \in T,\ j \in J \tag{11.60}$$

$$\sum_{j \in J} d_{jt} x_{ijpt} \leq n_{ip0} Q_p + \sum_{\tau=1}^{t} Q_p y_{ip\tau}, \quad t \in T,\ i \in I,\ p \in P_i \tag{11.61}$$

$$x_{ijpt} \geq 0 \quad t \in T,\ i \in I,\ j \in J,\ p \in P_i \tag{11.62}$$

$$y_{ipt} \in \{0, 1\}, \quad t \in T,\ i \in I,\ p \in P_i. \tag{11.63}$$

The coefficients c_{ijpt} may include the transportation costs between facilities and customers as well as handling costs at the facilities. Shulman (1991) proposed a Lagrangian relaxation based procedure for obtaining lower and upper bounds for the problem. Constraints (11.60) 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 where 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 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. That 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 re-opened, and once installed, a facility must remain open until the end of the planning horizon. Hinojosa et al. (2000) proposed a Lagrangian 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 Lagrangian relaxation based procedure was also developed.

A multi-period discrete facility location problem was 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 is enhanced via polyhedral analysis.

Albareda-Sambola et al. (2009) have 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 must 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 Lagrangian relaxation based procedure was proposed in that paper for obtaining lower and upper bounds. A particular case of this problem was studied by Albareda-Sambola et al. (2010), assuming that each customer requires service only in a subset of periods. 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 applied to this problem, which were compared computationally.

Correia and Melo (2016, 2017) focus on multi-period facility location problems with so-called demand satisfaction delay. In both works, two types of customers are considered: those who impose a strict timing for demand satisfaction and those whose demand can be fulfilled with delay. In both works existing facilities at the beginning of the planning horizon may be closed and new facilities may be opened in potential locations considered for that purpose. For new facilities, the capacity level at which they will operate is also a decision to make. In the first paper, such level is chosen from a finite set of possibilities previously identified for each location. In the second case, modular capacities are considered (with a maximum number of modules allowed at each location). Several optimization models are proposed and compared both theoretically (in terms of the lower bounds provided by linear relaxation) and computationally.

Jena et al. (2015a) studied a so-called logging camp location problem. This is a multi-period facility location problem with multiple commodities and modular capacities for the facilities. The real application underlying the problem calls for the possibility of making partial closing or reopening of facilities throughout the planning horizon. To the best or our knowledge, this is the first work considering such a possibility. Additionally, due to the nature of the facilities involved in the problem, the total demands assigned to a certain location is rounded up to the next integer value—the so-called round up capacity constraints are introduced. These are capacity constraints that instead of considering explicitly the total demand assigned to some location consider binary variables that help account for the number of (discrete) units of each commodity supplied from a facility of a certain size located in some location at some time period. Such capacity constrains are useful when, for instance, a facility cannot produce any arbitrary amount of a product but only modular sized packages of products. Jena et al. (2015a) introduce and compare several optimization models for the problem. This problem would be later complemented by the introduction of a stronger formulation and a hybrid heuristic that first applies a Lagrangian relaxation and afterwards constructs a restricted MIP problem using the previously obtained Lagrangian solutions (Jena et al. 2016).

Jena et al. (2015b) investigate the single-commodity version of the problem introduced in Jena et al. (2015a). Nevertheless, modular capacities are considered as well as a very general cost structure. In particular, the costs for capacity changes are given by a matrix that is an input to the problem. The new problem extends two cases of practical relevance: first, facility closing and reopening; second, capacity expansion and reduction. Different models are introduced and compared. This work would be later extended to the multi-commodity case (Jena et al. 2017) and the methodology introduced in Jena et al. (2016) successfully adapted to the extended problem.

Other contributions to the study of multi-period facility location problems include the work by Escudero and Pizarro Romero (2017), who consider a pure phase-in multi-period facility location problem for unit demand customers (demand is satisfied by means of a service that is provided and not by some quantity of a commodity). The set of customers is partitioned into different categories. In each period, some customers need to be served while some others do not. A service level of 100% is not imposed. Two decisions must be made in each period: new facilities to set operating and the assignment of customers to available operating. The costs involved in the problem include setup and maintenance for the facilities, assignment of customers to facilities, interaction costs incurred when customers of any two categories are assigned to the same facility in the same period, and penalty costs for unsatisfied service requests. A MILP formulation is derived for the problem and so-called fix-and-relax procedure is proposed for finding high quality feasible solutions to the problem. The proposed algorithm is a matheuristic introduced by Dillenberger et al. (1994) for general mixed 0–1 deterministic optimization problems.

11.6 The Value of the Multi-Period Solution

Multi-period modeling frameworks like those described in the previous sections, involve one extra dimension in the decision space: the time. The models tend to be of large scale and therefore more difficult to tackle, even for instances of moderate size. Hence, 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 time-invariant 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 again consider 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, i.e., the value of the multi-period solution is not necessarily 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, remaining 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 by simply 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 should remain unchanged during 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 (11.56). 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 papers within the context of facility location, in which the relevance of using a multi-period modeling framework is measured, are those by Alumur et al. (2012) and Marín et al. (2018).

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.

Nowadays, one can find much work focusing on facility location problems arising in the context of logistics systems. 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). The strong relation between facility location and logistics network design is also made clear by Melo et al. (2009).

Another aspect of relevance in many applications regards the uncertain nature of the data underlying the problems. Aghezzaf (2005) tackle a multi-period facility location problem under uncertainty through a robust optimization modeling framework. Multi-period stochastic facility location problems were investigated by Nickel et al. (2012), Albareda-Sambola et al. (2013), and Marín et al. (2018). These works show that embedding 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 that 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) studied multi-criteria, multi-period facility location problems in the context of logistics network design.

Most of the content in this chapter constitutes a basis for investigating more complex real-world problems. In fact, several models presented in the previous sections have already been extended to problems arising in other areas such as Hub Location (Chap. 12), Location-routing and Location-arc routing (Chap. 15), and Supply Chain Management and Logistics (Chap. 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, namely those just mentioned above. These correspond to research directions worth exploring namely for better tackling 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 Marketing 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

Castro J, Nasini S, Saldanha-da-Gama F (2017) A cutting-plane approach for large-scale capacitated multi-period facility location using a specialized interior-point method. Math Program A 163:411–444

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

Correia I, Melo T (2016) Multi-period capacitated facility location under delayed demand satisfaction. Eur J Oper Res 255:729–746

Correia I, Melo T (2017) A multi-period facility location problem with modular capacity adjustments and flexible demand fulfillment. Comput Ind Eng 110:307–321

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

Dillenberger C, Escudero LF, Wollensak A, Zhang W (1994) On practical resource allocation for production planning and scheduling with period overlapping setups. Eur J Oper Res 75:275–286

Drezner Z (1995) Dynamic facility location: the progressive p-median problem. Loc Sci 3:1–7

Drezner Z, Wesolowsky GO (1991) Facility location when demand is time dependent. Nav Res Log 38:763–777

Erlenkotter D (1981) A comparative study of approaches to dynamic location problems. Eur J Oper Res 6:133–143
Escudero LF, Pizarro Romero C (2017) On solving a large scale problem on facility location and customer assignment with interaction costs along a tie horizon. Top 25:601–622
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. Loc 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 SD, Cordeau J-F, Gendron B, (2015) Dynamic facility location with generalized modular capacities. Transp Sci 49:484–499
Jena SD, Cordeau J-F, Gendron B, (2015) Lagrangian heuristics for large-scale dynamic facility location with generalized modular capacities. INFORMS J Comput 29:388–404
Jena SD, Cordeau J-F, Gendron B, (2015) Modeling and solving a logging camp location problem. Ann Oper Res 232:151–177
Jena SD, Cordeau J-F, Gendron B, (2015) Solving a dynamic facility location problem with partial closing and reopening. Comput Oper Res 67:143–154
Khumawala BM (1972) An efficient branch and bound algorithm for the warehouse location problem. Manage 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
Marín A, Martínez-Merino LI, Rodríguez-Chía AM, Saldanha-da-Gama F (2018) Multi-period stochastic covering location problems: modeling framework and solution approach. Eur J Oper Res 268:432–449
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). PhD Thesis. Faculty of Science, University of Lisbon, Portugal

Saldanha-da-Gama F, Captivo ME (1998) A heuristic approach for the discrete dynamic location problem. Loc Sci 6:211–223

Scott AJ (1971) Dynamic location-allocation systems: some basic planning strategies. Environ Plann 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. Manage Sci 22:748–758

Van Roy T, Erlenkotter D (1982) A dual-based procedure for dynamic facility location. Manage Sci 28:1091–1105

Warszawski A (1973) Multi-dimensional location problems. Oper Res Qt 24:165–179

Wesolowsky GO (1973) Dynamic facility location. Manage Sci 19:1241–1248

Wesolowsky GO, Truscott WG (1975) The multi-period location-allocation problem with relocation of facilities. Manage 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 and Morton O'Kelly

Abstract *Hub Location Problems* (HLPs) lie at the heart of network design planning in transportation and telecommunication systems. They constitute 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.

12.1 Introduction

Transportation, telecommunications and computer networks frequently employ hub-and-spoke architectures efficiently to 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 are then combined with other flows having different origins but the same destination. This helps reduce setup costs, centralize commodity handling and sorting operations,

I. Contreras (✉)
Logistics and Transportation (CIRRELT), Concordia University and Interuniversity Research Centre on Enterprise Networks, Montreal, QC, Canada
e-mail: icontrer@encs.concordia.ca

M. O'Kelly
Department of Geography, The Ohio State University, Columbus, OH, USA
e-mail: okelly.1@osu.edu

G. Laporte et al. (eds.), *Location Science*,
https://doi.org/10.1007/978-3-030-32177-2_12

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.

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 or water. Hub facilities correspond to sorting centers or transportation terminals in which one or more transportation modes interact. Consolidation of flows at hubs enables 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 setup costs for hub facilities and communication links, motivate the use of hub-and-spoke architectures.

The study of HLPs began with the 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 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) concentrates 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 scheme 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, and provides insights on their modeling implications. We point out how these assumptions simplify network design decisions by 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 (Campbell and O'Kelly 2012), but rather our personal treatment on some of the most interesting research on this field. 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), Wieberneit (2008), and Saito et al. (2009) 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 their 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 lies in 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 the facilities and flows thus originate at demand nodes and their destination are the facilities. 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 cost is minimized by assigning each demand node to its nearest (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 which is more difficult. All demand flows from the same origin or to the same 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: (1) service demand is associated with flows between O/D pairs, (2) hub facilities are intermediate nodes in the O/D paths which act as transshipment or consolidation points, (3) there is a benefit (or requirement) of routing flows via hubs, (4) 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 setup cost for locating a hub, whereas for each $e \in E$, g_e denotes the fixed setup 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 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 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 setup 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: (1) *hub arcs* connecting two hubs, (2) *access arcs* connecting non-hub nodes and hubs, and (3) *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.

Four common assumptions underlie most HLPs:

1. Flows have to be routed via a set of hubs.
2. Access arcs and bridge arcs have no setup 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 setup 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 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 $\tau < \delta$. We note that in the literature, the path (i, k, m, j) is sometimes written with the alternative order (i, j, k, m).

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

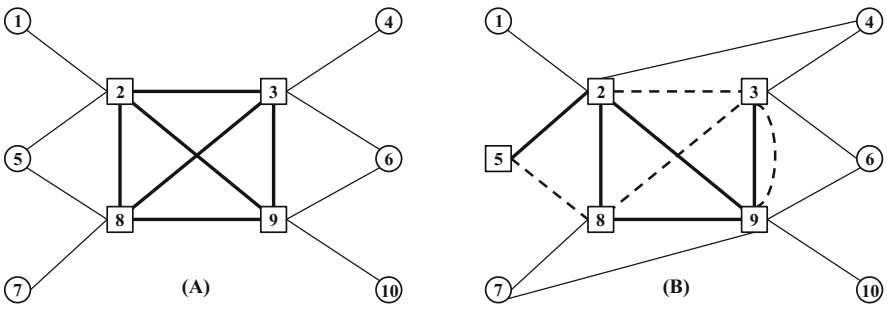


Fig. 12.1 Solution network of a hub node location problem (**a**) and a hub arc location problem (**b**)

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 imposed. 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 setup 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

$$c(S, R) = \sum_{i \in R} c_i; \qquad g(S, R) = \sum_{e \in S} g_e; \qquad h(S, R) = \sum_{i,j \in N} h^{ij}(S) = \sum_{i,j \in N} \min_{(k,m) \in S} F_{ijkm},$$

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 setup costs for setting hub nodes, $g(S, R)$ is the total setup 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 the problem of finding 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 split.

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 setup 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* O'Kelly (1987), Campbell (1996) 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* O'Kelly (1992), Campbell (1994b) consider that the number of hubs to locate is not known a priori, but a fixed setup 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* Campbell (1994b), Kara and Tansel (2000) are minimax problems that focus on the minimization of a maximum service or cost measure between O/D pairs. Some of these measures are: (1) the maximum flow cost (or travel time) of all O/D pairs, (2) the maximum flow cost (or travel time) of all arcs of the hub network, and (3) the maximum flow cost (or travel time) associated with an access arc.
- *Hub Covering Problems* Campbell (1994b), Kara and Tansel (2003) impose a maximum threshold value on the service level (travel time) and focus on the minimization of the setup 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: (1) the length of the path (i, k, m, j) is within a specified value, (2) the length of each arc in the path (i, k, m, j) does not exceed a specified value, or (3) 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 with 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, 2016) study 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 HLPs that integrate both cost and service objectives. Alibeyg et al. (2016, 2018) introduce hub location problems with profit-oriented objectives that measure the tradeoff between the revenue derived from served commodities and the overall network design and flow costs.

12.3 Formulating Hub Location Problems

One of the major modeling challenges in HLPs is due to the fact 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 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 (2008), Wagner (2008a), Ernst et al. (2009), 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 of formulating HLPs with single assignments is to consider them as facility location problems with additional quadratic costs associated with the interaction between hub facilities. For each pair $i, k \in N$, we define location-allocation variables z_{ik}, equal to one if and only if node i is assigned to hub k. 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 $D_i = \sum_{j \in N} W_{ji}$. The first term of the objective function represents the total setup cost of the hub facilities, whereas the second and third terms 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. 2). 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):

$$\text{minimize} \quad \sum_{k \in N} f_k z_{kk} + \sum_{i,j,k,m \in N} F_{ijkm} x_{ijkm}$$

$$\text{subject to} \quad (12.4)\text{–}(12.6)$$

$$\sum_{m \in N} x_{ijkm} = z_{ik} \qquad i, j, k \in N \tag{12.7}$$

$$\sum_{k \in N} x_{ijkm} = z_{jm} \qquad i, j, m \in N \tag{12.8}$$

$$x_{ijkm} \geq 0 \qquad i, j, k, m \in N. \tag{12.9}$$

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 at a node j assigned to hub m from some node i. There is no need to state explicitly the integrality on the x_{ijkm} variables given that constraints (12.7)–(12.8), in combination with (12.6), ensure that for each node pair $i, j \in N$ exactly one variable x_{ijkm} equals to one and the rest of them to zero. 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

$$\text{minimize} \quad \sum_{k \in N} f_k Z_k + \sum_{i,k \in N} (\chi O_i + \delta D_i) d_{ik} z_{ik} + \sum_{k,m \in N} \alpha d_{km} y_{km}$$

$$\text{subject to} \quad (12.4)\text{–}(12.6)$$

$$y_{km} \geq \sum_{(i,j) \in K} W_{ij} \left(z_{ik} + z_{jm} - 1 \right) \qquad k, m \in N, K \subseteq N \times N \tag{12.10}$$

$$y_{km} \geq 0 \qquad k, m \in N. \tag{12.11}$$

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):

$$\text{minimize} \quad \sum_{k \in N} f_k z_{kk} + \sum_{i,k \in N} (\chi O_i + \delta D_i) d_{ik} z_{ik} + \sum_{i,k,m \in N} \alpha d_{km} Y_{ikm}$$

subject to (12.4)–(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, 2017) 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).

Hence we 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 exists 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. Similar to 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. When solving this formulation, it may be possible to find an optimal solution in which, for a subset of node pairs $i, j \in N$, more than one x_{ije} variable is strictly positive (i.e., two or more paths are used to route flow between i and j). This would imply that there is more than one OD path with the same route cost. In that case, one can recover an integer solution by arbitrarily selecting one of such paths for each node pair $i, j \in N$ with multiple paths and making the associated x_{ije} variable equal to one and the others equal to zero. This path-based 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_{ijr} to the r-th element according to that ordering. That is, $F_{ije_1} \leq F_{ije_2} \leq \cdots \leq F_{ije_{|E|}} \leq F_{e_{ij|E|+1}}$, where $F_{e_{ij|E|+1}} = F_{ije^*}$ is the cost for the

fictitious edge e^* such that (1) $F_{ije^*} > \max_{e \in E} F_{ije}$, for all $i, j \in N$; and (2) $\sum_{i,j \in N} F_{ije*} > \max_{e \in E}(f_e + \sum_{i,j \in N} F_{ije})$. 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):

$$\text{minimize} \quad \sum_{k \in N} f_k Z_k + \sum_{i,j \in N} \eta_{ij}$$

$$\text{subject to} \quad \eta_{ij} \geq F_{ije_r} + \sum_{e \in E} (F_{ije} - F_{ije_r})^- y_e \quad r = 1, \ldots, |E| + 1, \; i, j \in N \tag{12.18}$$

$$y_e \leq z_k \qquad e = (k, m) \in E \tag{12.19}$$

$$y_e \leq z_m \qquad e = (k, m) \in E \tag{12.20}$$

$$y_e, z_i \in \{0, 1\} \qquad e \in E, i \in N, \tag{12.21}$$

where η_{ij} are continuous decision variables used to evaluate the flow cost of O/D pair (i, j) and $(x)^- = \min\{0, x\}$. 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):

$$\text{minimize} \quad \sum_{k \in N} f_k Z_k + \sum_{i,k \in N} \chi d_{ik} Z_{ik} + \sum_{i,k,m \in N} \alpha d_{km} Y_{ikm} + \sum_{ijm} \delta d_{jm} X_{ijm}$$

$$\text{subject to} \quad (12.17)\text{–}(12.13)$$

$$\sum_{k \in N} Z_{ik} = O_i \qquad i \in N \tag{12.22}$$

$$\sum_{m} X_{ijm} = W_{ij} \qquad i, j \in N \tag{12.23}$$

$$Z_{ik} + \sum_{m \in N} Y_{ikm} = \sum_{m \in N} Y_{imk} + \sum_{j} X_{ijm} \qquad i, k \in N \tag{12.24}$$

$$Z_{ik}, X_{ijm} \geq 0 \qquad i, j, m \in N. \tag{12.25}$$

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; O'Kelly et al. 2015a). 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), Tanash et al. (2017), O'Kelly et al. (2015a), Miranda et al. (2017), and Martins de Sá et al. (2018a,b), among others, study the design of incomplete hub networks in which no network structure other than connectivity is imposed on the backbone network. Miranda et al. (2017) also consider a variant in which hop constraints are used to limit the number of arcs in OD paths. Other works study models that do not consider a complete backbone network but rather, a particular topological structure. Figure 12.2 shows some examples of different hub network structures.

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)

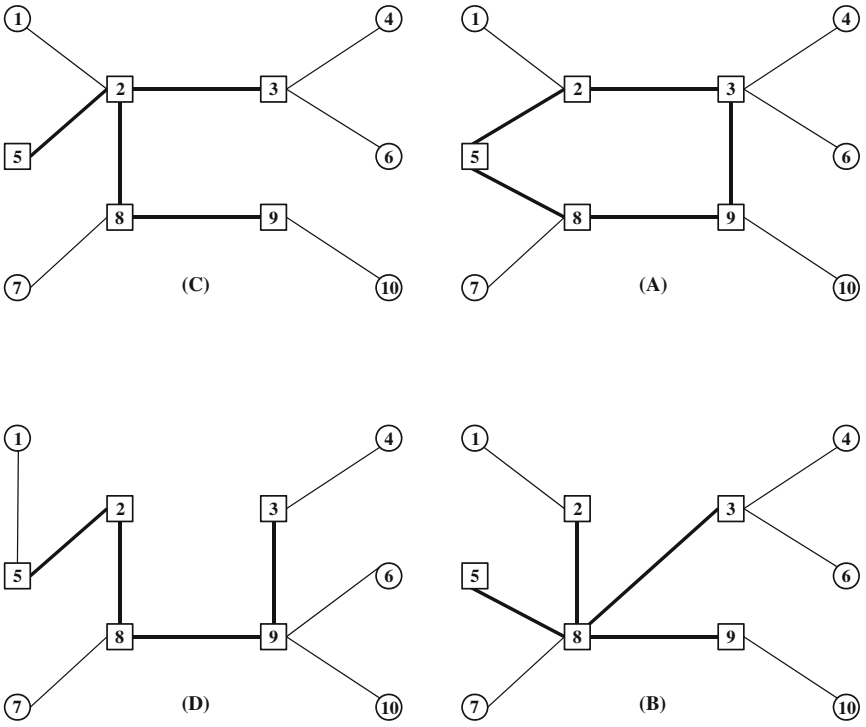


Fig. 12.2 Structure of a cycle-star (**a**), star-star (**b**), tree-star (**c**), and line-star (**d**) hub network

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. (2014b) 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. (2017) focus on the design of cycle-star networks in which the hubs are connected by means of a cycle. O'Kelly et al. (2015a) analyze the role of setup costs for link activation decisions in the design of hub networks. The proposed model allows particular versions of hub networks to emerge from the cost structure, rather than assuming a predefined network structure.

Some papers focus on the design of more complex access networks that are no longer determined by a single or multiple assignment pattern of O/D nodes to hubs. Figure 12.3 depicts some examples of various access network structures. 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). 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

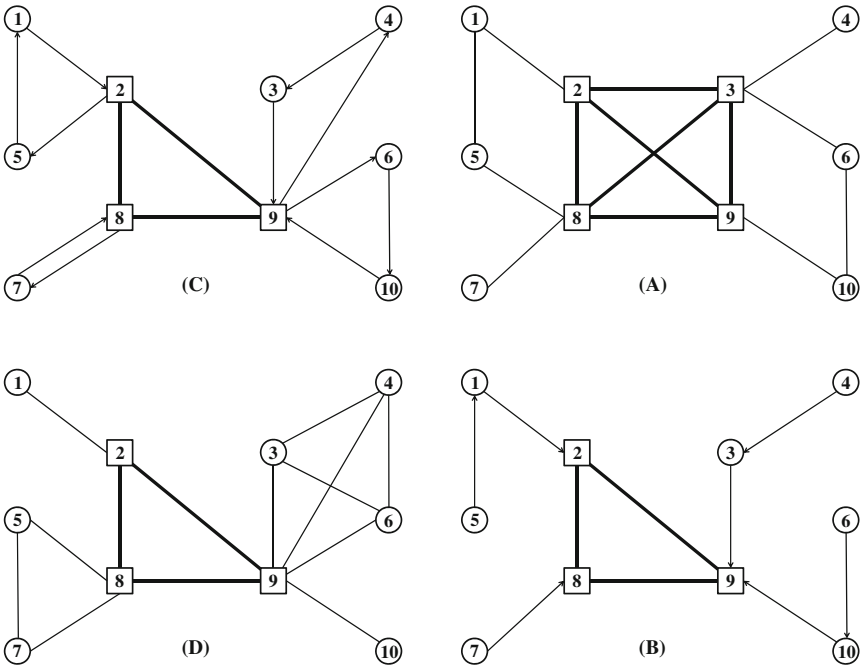


Fig. 12.3 Access network with direct connections (**a**), multi-stops (**b**), tours (**c**), and complete subgraphs (**d**)

access networks are fully interconnected. Finally, we refer to Chap. 14 for references considering hub network topologies arranged in a hierarchical structure.

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 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 economies of scale 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. In contrast 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. Cunha and Silva (2007) and Lüer-Villagra and Marianov (2019) consider an alternative linear flow cost function in which a threshold for switching cost lines is used. In this case, the flow-independent discount rate applies only when the amount of flow on a link exceeds the threshold.

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 consists of a fixed flow-independent setup 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. O'Kelly et al. (2015a) study a similar uncapacitated problem in which fixed and variables flow costs for arcs are incorporated to provide a flow-dependent cost rate. Yaman and Carello (2005), Tanash et al. (2017), and Hoff et al. (2017) study modular hub location problems with single assignments in which a stepwise function is used to model flow-dependent costs on hub arcs. Contrary to fixed-charge cost functions, stepwise functions do not consider a variable cost component and are frequently used to model transportation costs in vehicle routing and pick-up and delivery 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 on 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), Boland et al. (2004), and Puerto et al. (2016) 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, 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.

Alumur et al. (2016) introduced models with single and multiple assignments in which capacities at hub nodes can be gradually expanded over a finite planning horizon. Serper and Alumur (2016) presented a more comprehensive capacitated model in which capacities have to be determined in both hubs and arcs of the network. Given that alternative transportation modes and different types of vehicles are considered, the design of the hub network is done by explicitly determining the number of vehicles of each type to operate on each link of the network.

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 studied 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 setup 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 setup 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 setup costs and demand flows and are modeled as two-stage minimax regret programs with recourse. Correia et al. (2018) introduce a more general two-stage stochastic multi-period capacitated hub location problem in which uncertainty is assumed for the demands. The first-stage decisions deal with the location of the hubs over the planning horizon and their initial installed capacity. The second-stage decisions concern the assignment of non-hubs to hubs, the routing of flows, and the capacity expansion for existing hubs.

Merakli and Yaman (2016) introduce robust uncapacitated p-hub median problems with multiple assignments under polyhedral demand uncertainty. They employ a hose model and a hybrid model to characterize demand uncertainty. The former assumes that the only available information is an upper bound on the total flow adjacent to each node, while the latter incorporates in addition lower and upper bounds on each OD flow. Merakli and Yaman (2017) extends the hose uncertainty model to a more challenging scenario in which capacities at hubs are considered, impacting the feasibility of solutions. Zetina et al. (2017) present robust counterparts for UHLPMAs in which the level of conservatism is controlled with a budget of uncertainty. The proposed models incorporate both independently and jointly demand and flow costs uncertainties when the only available information is an interval of uncertainty. The considered objectives aim at a minimizing the sum of the hub setup costs and of demand flow costs in the worst-case scenario.

Martins de Sá et al. (2018a) study a robust counterpart of an incomplete hub location problem with multiple assignments in which link activation decisions are taken into account. The model considers uncertainty in setup costs for hub nodes and hub arcs as well as demand and uses a budget of uncertainty to control the level of conservatism. Martins de Sá et al. (2018b) address another robust incomplete hub location problem in which service time constraints for each demand flow are incorporated. In this case, the uncertainty is related to travel times between nodes and the goal is to obtain cost-effective solutions with a high probability of being feasible with respect to the service time constraints.

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. Aziz et al. (2018) consider the design of hub networks

under stochastic demand and congestion. Hubs are modeled as spatially distributed $M/G/1$ queues and congestion is captured using the expected queue lengths at hubs.

An important uncertainty aspect neglected until recently is the reliability of hub networks. Kim and O'Kelly (2009) present 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. (2015), Aziz et al. (2016), and Rostami et al. (2018) 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 (one at a time) by using backup hubs and alternative routes for demand flows. The objective of these models is to minimize the total expected flow cost considering both the regular and the disruptive situation. Tran et al. (2016) assume that more than one hub can simultaneously fail, each of which can fail with a site-specific probability. Ramamoorthy et al. (2018) present multiple allocation hub interdiction and hub protection problems. In the hub interdiction problem, the goal is to determine a set of r critical hubs from an existing set of p hubs such that when interdicted results in the maximum post-interdiction flow cost. In the hub protection problem, the decision maker seeks to fortify a set of u hubs from an existing set of p hubs against interdiction. These models lead to complex bi-level and tri-level optimization problems which are known to be extremely difficult to solve.

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) study 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 et al. (2015) presents another multi-period hub location model arising in the design of public transportation networks in which the full interconnection assumption is relaxed and thus, additional hub arc selection are considered. Alumur et al. (2016) study multi-period models with single and multiple

assignments in which capacities at hub facilities can be gradually expanded over a planning horizon.

Another important feature in some applications is the presence of strategic decisions related to the choice for mode of transportation. Most HLPs consider that only one mode of transportation is available and hence there is 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 setup 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. Serper and Alumur (2016) present capacitated models considering alternative transportation modes and different types of vehicles. The models select an optimal number of vehicles of each type to operate on each link of the network. Dukkanci and Kara (2017) study a hub covering problem with service time constraints. They propose a hierarchical multimodal hub network structure in which different types of vehicles can be used in each layer.

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 flows) 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 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 captured on a particular path is modeled through 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. Lüer-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 derived from captured flows, minus the a fixed and variable costs. Customer preferences on selected firm and route are modeled using a logit model. Mahmutogullari and Kara (2016) propose other competitive models in which two decision-makers sequentially determine the location of their hubs and then customers choose one firm with respect to provided service levels. The goal of each firm is to maximize its own market share. O'Kelly et al. (2015b) introduce a model with price-sensitive demands. It considers three different service levels for routing

flow between OD pairs that use either two-hub OD paths, on-hub OD paths or direct connections. The authors model the problem as an economic equilibrium problem that maximizes a nonlinear concave utility function, minus the flow cost and setup cost for the location of the hubs.

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 polynomially 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 hence, 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. 2017).

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 yields solutions for very large-scale instances with up to 1000 nodes and $p = 20$ within reasonable CPU times. The best results for the UHLPSA seem to be obtained using the learning-based probabilistic tabu search recently designed by Guan et al. (2018). This heuristic has the best performance when compared with other heuristics, 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. (2017) 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. (2014b) propose an adaptive large neighborhood search and GRASP algorithms to design hub networks with multiple hub lines. Tran et al. (2016) develop a parallel tabu search to solve reliable hub location problems. Hoff et al. (2017) present a metaheuristic based on adaptive memory programming and path-relinking to solve a capacitated modular hub location problem.

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.

Lagrangian 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 10 time periods. Alibeyg et al. (2018) develop an exact BB algorithm that uses a LR to solve hub location problems with profits involving up to 100 nodes.

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, 2014b), HLPs arising in public transportation networks (Gelareh and Nickel 2011), liner shipping applications (Gelareh and Nickel 2011; Gelareh and Pisinger 2011), and incomplete hub networks (Miranda et al. 2017; Martins de Sá et al. 2018a,b).

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, 2017) 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. (2017) 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. Meier and Clausen (2018) present a novel linearization technique together with a cutting plane algorithm to solve uncapacitated and capacitated hub location problems with single assignments. This linearization, requiring only two-index variables, is applicable in the case of Euclidean data and can be used to solve instances with up to 200 nodes.

Column generation (CG) is the method that has received the least attention in the hub location literature. Thomadsen and Larsen (2007) present a branch-and-price method for solving a HLP with fully interconnected access networks. Contreras et al. (2011d) develop an 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. Rothenbcher et al. (2016) propose an exact branch-and-price-and-cut algorithm for the service network design and hub location problem. It uses a path-based formulation solved via column generation as a bounding procedure at the nodes of the tree. It also uses several families of valid inequalities to strengthen the LP bounds.

12.6 Conclusions

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, 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 among 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 have been used within a cutting plane framework to solve challenging hub location variants. Over the past few years, promising two-index (integer linear) 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, which 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 full 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 represent more accurately 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 represent 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

Alibeyg A, Contreras I, Fernández E (2016) Hub network design with profits, Transport Res E-Log 96:40–59

Alibeyg A, Contreras I, Fernández E (2018) Exact solution of hub network design problems with profits, Eur J Oper Res 266:57–71

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, 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. Transport Res E-Log 48:1107–1120

Alumur SA, Nickel S, Saldanha-da-Gama F, Secerdin Y (2016) Multi-period hub network design problems with modular capacities. Ann Oper Res 246:289–312

An Y, Zhang Y, Zeng B (2015) The reliable hub-and-spoke design problem: models and algorithms. Transp Res B Methodol 77:103–122

Asgari N, Zanjirani Farahani R, Goh M (2013) Network design approach for hub ports-shipping companies competition and cooperation. Transp Res A Policy Pract 48:1–18

Aykin T (1988) On the location of hub facilities. Transport 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. Transport Sci 3:201–221

Aziz N, Chauhan S, Vidyarthi N (2016) The impact of hub failure in hub-and-spoke networks: mathematical formulations and solution techniques. Comput Oper Res 65:174–188

Aziz N, Vidyarthi N, Chauhan S (2018) Modelling and analysis of hub-and-spoke networks under stochastic demand and congestion. Ann Oper Res 264:1–2

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 Regional 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, Luna HP (2009) Benders decomposition for hub location problems with economies of scale. Transport Sci 43:86–97
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, 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, 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
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 Locat Anal 6:31–43
Campbell JF (1994b) Integer programming formulations of discrete hub location problems. Eur J Oper Res 72:387–405
Campbell JF (1996) Hub location and the p-hub median problem. Oper Res 44:923–935
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. Transport 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, 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, 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, Díaz JA, Fernández E (2009a) Lagrangean relaxation for the capacitated hub location problem with single assignment. OR Spectr 31:483–505
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. Transport Sci 45:18–32
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, 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, Cordeau J-F, Laporte G (2012) Exact solution of large-scale hub location problems with multiple capacity levels. Transport Sci 46:439–459
Contreras I, Tanash M, Vidyarthi N (2017) Exact and heuristic approaches for the cycle hub location problem Ann Oper Res 258:655–677
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
Correia I, Nickel S, Saldanha-da-Gama F (2018) A stochastic multi-period capacitated multiple allocation hub location problem: formulation and inequalities. Omega 74:122–134
Costa MG, Captivo ME, Climaco J (2008) Capacitated single allocation hub location problem—a bi-criteria approach. Comput Oper Res 35:3671–3695
Cunha CB, Silva MR (2007) A genetic algorithm for the problem of configuring a hub-and-spoke network for a LTL trucking company in Brazil. Eur J Oper Res 179:747–758
Dukkanci O, Kara BY (2017) Routing and scheduling decisions in the hierarchical hub location problem. Comput Oper Res 85:45–57
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. Loc 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, Nickel S (2011) Hub location in transportation networks. Transport Res E-Log 47:1092–1111
Gelareh S, Pisinger D (2011) Fleet deployment, network design and hub location of liner shipping companies. Transport Res E-Log 47:947–964
Gelareh S, Nickel S, Pisinger D (2010) Liner shipping hub network design in a competitive environment. Transport Res E-Log 46:991–1004
Gelareh S, Monemi RN, Nickel S (2015) Multi-period hub location problems in transportation. Transport Res E-Log 75:67–94
Gendron B, Crainic TG, Frangioni A (1999) Multicommodity capacitated network design. In: Sansó B, Soriano P (eds) Telecommunications Network planning. Kluwer, Norwell, 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. Transport Res E-Log 41:567–583
Guan J, Lin G, Feng H-B (2018) A learning-based probabilistic tabu search for the uncapacitated single allocation hub location problem. Comput Oper Res 98:1–12
Hamacher HW, Labbé M, Nickel S, Sonneborn T (2004) Adapting polyhedral properties from facility to hub location problems. Discrete Appl Math 145:104–116
Hoff A, Peiró J, Corberán A, Martí, R (2017) Heuristics for the capacitated modular hub location problem. Comput Oper Res 86:94–109
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. DUV, Weisbaden, 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
Lüer-Villagra A, Marianov V (2013) A competitive hub location and pricing problem. Eur J Oper Res 231:734–744

Lüer-Villagra A, Eiselt HA, Marianov V (2019) A single allocation p-hub median problem with general piecewise-linear costs in arcs. Comput Ind Eng 128:477–491
Mahmutogullari AI, Kara BY (2016) Hub location under competition. Eur J Oper Res 250:214–225
Marianov V, Serra D (2003) Location models for airline hubs behaving as M/D/c queues. Comput Oper Res 30:983–1003
Marianov V, Serra D, ReVelle, CS (1999) Location of hubs in a competitive environment. Eur J Oper Res 114:363–371
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, 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
Martins de Sá E, Contreras I, Cordeau J-F, de Camargo RS, de Miranda R (2015a) The hub line location problem. Transport Sci 9:500–518
Martins de Sá E, Contreras I, Cordeau J-F (2015b) Exact and heuristic algorithms for the design of hub networks with multiple lines. Eur J Oper Res 246:186–198
Martins de Sá E, Morabito R, de Camargo RS (2018a) Benders decomposition applied to a robust multiple allocation incomplete hub location problem. Comput Oper Res 89:31–50
Martins de Sá E, Morabito R, de Camargo RS (2018b) Efficient Benders decomposition algorithms for the robust multiple allocation incomplete hub location problem with service time requirements. Expert Syst Appl 93:50–61
Meier JF, Clausen U (2018) Solving single allocation hub location problems on Euclidean Data. Transport Sci 52:1141–1155
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
Merakli M, Yaman H (2016) Robust intermodal hub location under polyhedral demand uncertainty. Transp Res B Methodol 86:66–85
Merakli M, Yaman H (2017) A capacitated hub location problem under hose demand uncertainty. Comput Oper Res 88:58–70
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
Miranda G, de Camargo RS, O'Kelly ME, Campbell JF (2017) Formulations and decomposition methods for the incomplete hub location problem with and without hop-constraints. Appl Math Model 51:274–301
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. Transport 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. Transport Res D 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, Campbell JF, de Camargo RS, Miranda G (2015a) Multiple allocation hub location model with fixed arc costs. Geogr Anal 47:73–96
O'Kelly ME, Luna PL, de Camargo RS, Miranda G (2015b) Hub location problems with price sensitive demands. Netw Spat Econ 15:917–945

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
Puerto J, Ramos AB, Rodriguez-Chia AM, Sanchez-Gil MC (2016) Ordered median hub location problems with capacity constraints. Transport Res C 70:142–156
Racunica I, Wynter L (2005) Optimal location of intermodal freight hubs. Transp Res B Methodol 39:453–477
Ramamoorthy P, Jayaswal S, Sinha A, Vidyarthi N (2018) Multiple allocation hub interdiction and protection problems: model formulations and solution approaches. Eur J Oper Res 270:230–245
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 JJ, Yaman H (2014) A branch-and-cut algorithm for the hub location and routing problem. Comput Oper Res 50:161–174.
Rothenbcher A-K, Drexl M, Irnich S (2016) Branch-and-price-and-cut for a service network design and hub location problem, Eur J Oper Res 255:935–947
Rostami B, Kämmerling N, Buchheim C, Clausen U (2018) Reliable single allocation hub location problem under hub breakdowns. Comput Oper Res 96:15–29
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, Fukushima M (2001) Stackelberg hub location problem. J Oper Res Soc Jpn 44:390–405
Sasaki M, Campbell JF, Ernst AT, Krishnamoorthy M (2009) Hub arc location with competition. Technical report NANZAN-TR-2009-02
Serper EZ, Alumur SA (2016) The design of capacitated intermodal hub networks with different vehicle types. Transp Res B Methodol 86:51–65
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
Tanash M, Contreras I, Vidyarthi N (2017) An exact algorithm for the modular hub location problem with single assignments. Comput Oper Res 85:32–44
Thomadsen T, Larsen J (2007) A hub location problem with fully interconnected backbone and access networks. Comput Oper Res 34:2520–2531
Tran TH, O'Hanley JR, Scaparra MP (2016) Reliable hub network design: formulation and solution techniques. Trans Sci 51:358–375

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 Spect 30:77–112
Yaman H (2008) Star p-hub median problem with modular arc capacities. Comput Oper Res 35:3009–3019
Yaman H, Carello G (2005) Solving the hub location problem with modular link capacities. Comput Oper Res 32:3227–3245
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
Zetina C, Contreras I, Cordeau J-F, Nikbakhsh E (2017) Robust uncapacitated hub location. Transp Res B Methodol 106:393–410

Chapter 13
Hierarchical Facility Location Problems

Ivan Contreras and Camilo Ortiz-Astorquiza

Abstract *Hierarchical facility location problems* (HFLPs) are an important class of problems arising in numerous contexts such as in the design of health care, telecommunications and transportation systems. HFLPs deal with the location of interacting facilities at different levels of a hierarchical system. This chapter describes the distinguishing features and main areas of application of HFLPs and provides a comprehensive classification scheme based on several attributes. It also presents a concise overview of four classes of HFLPs that have received the most attention in the literature: multi-level facility location, median and covering hierarchical facility location, multi-echelon location-routing problems, and hierarchical hub location problems. For these classes of HFLPs, we highlight their main characteristics and point out to some of the integer programming formulations and efficient algorithms that have been developed.

13.1 Introduction

Health care, telecommunications, and transportation systems are examples where hierarchical structures arise having different types of interacting facilities that collectively provide services or products to a set of customers. There are three key features of such hierarchical systems which are crucial for their design. The first one is that the different types of facilities in the system are characterized by the services they provide. The second one is that there exists either an inherent hierarchy given by the nature of the system or a ranking mechanism that allows the complete ordering of the different types of facilities into levels, each of which contains only

I. Contreras (✉)
Concordia University and Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT), Montreal, QC, Canada
e-mail: icontrer@encs.concordia.ca

C. Ortiz-Astorquiza
Departamento de Matemáticas, Pontificia Universidad Javeriana, Bogotá, Colombia
e-mail: camiloortiz@javeriana.edu.co

G. Laporte et al. (eds.), *Location Science*,
https://doi.org/10.1007/978-3-030-32177-2_13

facilities of the same type. The third feature is that there is an implicit or explicit relationship between facilities among them.

This chapter focuses on a general class of discrete location problems arising in the design of hierarchical systems. Given the wide variety of applications in which a hierarchical structure is present, several variants of this class of problems have been studied, usually under different names. Depending on the application, as well as authors' discipline and fields of expertise, the terms *hierarchical*, *multi-level*, *multi-echelon*, *multi-stage*, *multi-tier*, and *multi-layer* have been used to refer to different facility location problems with an underlying hierarchical structure. Therefore, the main goal of this chapter is to present a unified view of such problems that fit under one umbrella: *hierarchical facility location*. Broadly speaking, given a set of customers that demand one or multiple services and a set of facilities of k different types associated with the various services, *hierarchical facility location problems* (HFLPs) consist of selecting a set of facilities to open so that each customer receives the requested service(s) from facilities of one or multiple types, while optimizing an objective function.

HFLPs are closely related to classical facility location problems such as p-median, uncapacitated facility location, p-center, and covering problems. In these problems, there is the implicit assumption that all customers request one and the same type of service, which is offered by any candidate facility. In contrast, HFLPs extend these problems to deal with more realistic situations in which facilities are arranged in a hierarchical structure determined by the bundle of services that each of them provides or by other natural order inherited from the considered system. In this sense, classical facility location problems are single-level variants of the more general class of HFLPs.

Applications of HFLPs are abundant. These include supply chains and production-distribution systems, where manufacturing facilities, warehouses, distribution centers, and retail stores interact to provide cost-effective production, storage and transportation services to customers. Health care systems are another common example, where patients seeking health related services travel to different facilities such as local clinics, community and regional hospitals, each of them providing a variety of services. Other examples of applications of HFLPs arise in solid waste management, education systems, emergency medical services, air freight and passenger travel, postal delivery, telecommunication networks, and urban transportation planning.

The study of hierarchical systems in location science has its origins in the area of health care planning with the works by Schultz (1970) and Dökmeci (1973, 1977), for continuous location models, and those by Calvo and Marks (1973), Schilling et al. (1979), and Moore and ReVelle (1982), for discrete models. Hierarchical discrete location models were initially studied in the context of production-distribution systems by Kaufam et al. (1977), where the so-called plant and warehouse location problem was introduced. Narula (1984) provided the first classification scheme for HFLPs based on the relationship between the different types of facilities and the flow pattern. Narula (1986) and Church and Eaton (1987) are early reviews on HFLPs focusing on median and covering-based models, respectively. Şahin and

Süral (2007) provide a more comprehensive classification scheme and review of the growing literature on HFLPs before 2007. Daskin (2013) discusses the basic notions of hierarchical facilities and applications and describes some of the fundamental median and covering HFLPs. Zanjirani Farahani et al. (2014) review applications, models and solution algorithms mainly for median and covering-based HFLPs. Ortiz-Astorquiza et al. (2018) present a comprehensive review of a special case of HFLPs, denoted multi-level facility location problems, and propose a unified framework to classify them based on the types of strategic and tactical decisions involved.

This chapter is organized as follows. In Sect. 13.2 we first discuss the fundamentals of discrete HFLPs, including their distinguishing features and key concepts. We also provide a comprehensive classification scheme based on several attributes such as nature of customer demand, service availability, flow pattern, and decisions and objectives. A review of classical and more recent applications areas is then given in Sect. 13.3. Finally, in Sect. 13.4 we provide a concise overview of four classes of HFLPs that have received the most attention in the literature: *multi-level facility location*, *median* and *covering hierarchical facility location*, *multi-echelon location-routing*, and *hierarchical hub location*. In particular, we highlight their main characteristics and point out to the some integer programming formulations and efficient algorithms to solve some variants of them. We note that our intention is neither to pose a disjoint partition of the set of HFLPs nor to offer an exhaustive survey on the topic but rather to clarify the most relevant features of HFLPs.

13.2 Fundamentals

The distinguishing features and key concepts arising in hierarchical systems need to be introduced in order to have a better understanding of the structure and inherent complexity of modeling and solving HFLPs. We next discuss the following aspects: (1) nature of customer demand, (2) service availability, (3) flow pattern and spatial configuration, and (4) decisions and objectives.

13.2.1 Nature of Demand

When referring to the nature of demand in the context of a hierarchical system we must identify who the customers are, which services they may request, and how these services will be offered by the facilities.

The most common setting consists of a set of customers represented as demand points in a graph with service given *at* or *from* a facility. Depending on the context, customers may correspond to patients, retail stores, patrons, students, geographical zones, etcetera. These *customers* can have one or more types of demand. In other words, each customer may have different service requirements to be offered by

facilities. For instance, a patient that requires a specialized test offered at regional hospitals has a type of demand different from one whose requirement is an annual control visit offered at local clinics. Other patients could request both diagnostic and out-patient surgery services offered at regional hospitals. Note that in this context, service is given at facilities and thus customers have to travel to them to receive service. We denote as *single demand* and *multiple demand* customers to distinguish cases in which one or more types of demand are requested by customers, respectively.

In some other situations, customers may request several types of service as well as multiple products. For example, in the fashion industry, retail stores sell hundreds of products produced in dozens of manufacturing facilities. Customers (i.e. retail stores) continuously request production, transportation and storage services of various products such as clothes, shoes, accessories, etcetera. Note that in this context, service is provided by facilities and thus products are eventually delivered to the customers. We note that in this context, each customer demands both products and services, and these may be different from one customer to the other. For instance, one customer may request production and transportation services for clothes and shoes whereas another customer may request production, transportation, and storage services only for accessories. We denote as *single product* and *multiple product* customers to differentiate cases in which one or more types of product are requested by customers, respectively.

Another class of customers arising in hierarchical systems are those in which service demand corresponds to the movement of commodities, people, or information between an origin and a destination point. Each origin/destination (O/D) pair represents one or more customers requesting transportation or communication services between two specific points. Facilities provide such services by acting as transshipment and consolidation points in the paths of many O/D pairs. For instance, in express package delivery, a package is usually picked-up directly at its origin by a small vehicle and then moved to a branch office in order to be consolidated with other packages. It is then moved forward using larger vehicles to regional and possible central hub facilities where the package is sorted and rerouted to its destination.

Except from the hierarchical hub location models discussed in Sect. 13.4.4, in this chapter we focus on models in which customers are represented as demand points.

13.2.2 Service Availability

The nature of the customers demand is intimately related to the services that facilities can offer. This has been referred to as *service availability* or *service varieties* criterion in classifications of HFLPs (see, Narula 1986; Şahin and Süral 2007).

According to Narula (1986), a *successively inclusive facility hierarchy* is one in which a facility at level r provides all services offered by a facility at level $r-1$ plus one or more additional services. A *successively exclusive facility hierarchy* is one in which the set of services offered by a facility at level r are not offered by any other facility at levels $q \neq r$. We note that there is a third category, denoted as *mixed facility hierarchy*, in which the set of services offered by a facility at level r has a non-empty intersection with the set of services offered by facilities at other levels $q \neq r$, without necessarily offering all services of lower-level facilities. Şahin and Süral (2007) refer to the first category as *nested* and the last two as *non-nested*.

Additional categories can be considered in more complex settings in which services offered by the facilities also depend on geographical considerations. For example, one can refer to *locally inclusive service hierarchies*. These make use of a measure such as distance to determine whether or not a service is offered to a particular customer. For instance, suppose that facility of type r located at node i offers services 1 through r to demands from node i, but does not offer services 1 to $r-1$ to demands from nodes $j \neq i$. That is, only service r is available for demands whose origin is not i. For example, a regional hospital located in a given district may provide basic care to patients living is such district but not to other patients living in other districts. However, it may provide out-patient surgical services to any patient when needed. In contrast, *globally inclusive service hierarchies* consider that a facility at level r located at node i can offer services 1 to r to all customers requesting any of such services. Some examples of problems in which these service hierarchies arise can be found in Tien et al. (1983), Mirchandani (1987), and Daskin (2013).

13.2.3 Flow Pattern and Spatial Configuration

The *flow pattern* refers to the way in which network flows are routed through the various levels of a hierarchical system. Şahin and Süral (2007) propose two possible patterns: *single-flow* or *multi-flow*. In a single-flow pattern, flow can start at a demand point at the customers level and visit facilities in each of the levels until it reaches the highest level k. Similarly, flow can start in a facility at the highest level k and pass through all levels until it arrives to its demand point at the customer levels. On the other hand, in a multi-flow pattern, flow can start at a demand point and visit facilities in a subset of levels, possibly skipping some levels. Alternatively, the flow can start at a facility in any level r and pass through a subset of levels until arriving to its demand point. A third type of pattern, denoted as *bidirectional-flow*, arises whenever demand flow is routed in both directions. That is, in a bidirectional-flow pattern, flow starts at the customer level and visit some or all higher-levels and then is routed back to lower-levels until it arrives to a demand point at the customer level.

In some applications, regardless of the considered flow pattern, a portion of customer demand served at a given level is eventually referred for additional services to a higher level. This is denoted as *referral* systems. Alternatively, when

referrals are not considered between levels, this can be denoted as *non-referral* systems.

Şahin and Süral (2007) describes another relevant concept, denoted as coherency, which relates to the *spatial configuration* of levels in the hierarchy. A *coherent* system is one in which all demand flow entering to a particular facility at a lower-level is assigned to exactly one facility at a higher-level. Some authors refer to this concept as single assignment due to its resemblance with single-sourcing in single-level discrete location problems. A *non-coherent* system is one in which facilities at lower-levels can send demand flows to more than one facility at a higher level.

13.2.4 Decisions and Objectives

We now discuss the types of decisions and objectives commonly involved in HFLPs. For this purpose, we introduce the main notation used throughout this chapter. Let $G = (V, E)$ be a graph with node set $V = I \cup J$ and edge set E. The set I corresponds to the sites of potential facilities and J to the customers. We consider facilities of types 1 to k. An HFLP involves some of the following decisions.

- **Design Decisions: Facility Location and Edge Activation** The *location decisions* determine where to open the facilities. Given an underlying network G, facilities may be located at both the nodes or the edges of the network. Here we focus on discrete location problems, where it is assumed that facilities can only be located at the nodes of G. The *network design decisions* select the edges to be activated. These edges are used to provide transportation or communication services between demand points and facilities, and between facilities of the same or different levels. We concentrate on those problems where the facility location decisions are non-trivial.
- **Tactical Decisions: Allocation and Routing** The *allocation decisions* determine which facilities will be used to serve each demand point. In FLPs, two types of allocation strategies have been considered. In single allocation, each customer is assigned to exactly one facility, whereas in multiple allocation each customer is allowed to be assigned to more than one facility, if beneficial. The *routing decisions* indicate the routes (or paths) on G that will be used to satisfy the customer demands. We use the term route to indicate the sequence of edges used to send flows between pairs of nodes. These types of decisions commonly appear in network flow problems which have been widely studied (Ahuja et al. 1993). In the general case of an HFLP the term routing could also be used in the sense of location-routing models (see Chap. 15) where tours or paths between nodes of the same level in the hierarchy are considered. Finally, observe that the network design and routing decisions are interrelated, since the edges that can be used in the paths are determined by the network design decisions.

Both of the above types of decisions are directly related to the fixed and variable costs. For example, when a node $i \in I$ is selected to locate a facility, a setup cost f_i

may be incurred. However, note that these costs in the context of HFLPs typically depend on the type of facility. Analogously, when an edge $\{a, b\} \in E$ is activated a setup cost h_{ab} may be paid. On the other hand, the tactical decisions are affected by variable costs. A common example are transportation costs which are generally related to the distances between the nodes. Assuming that customers are at level 0, transportation costs (or distances) $c_{i_r i_s}$, for r and s in $\{0, \cdots, k\}$ are variable since they typically depend on the flow passing through the corresponding edge $\{i_r, i_s\}$ which is dependent on customers demands.

In any case, we note that in an HFLP there must be location decisions involved for one or more types of facilities. Depending on the application, network design and routing decisions may be explicitly considered or not, that is, the activation of edges and flow patterns are not necessarily non-trivial decisions in this context. These types of decisions will help us define differentiating features for families of HFLPs in the following sections.

Analogously to single-level facility location, HFLPs can be classified based on the type of objective function.

- *fixed-charge models:* consider that the number of facilities to locate at each level is not known a priori, but a fixed setup cost f_i for each facility at each level is considered. The objective is to minimize the sum of facilities fixed costs and of demand-weighted distance.
- *median models:* consist of selecting a set of facilities to open, such that no more than a given number of facilities is opened with the objective of minimizing the total demand-weighted distance (or transportation cost). In this case, the maximum number of facilities to open can be given by type r, as p_r, or in total. Moreover, it can be generalized to one or multiple budget constraints that limit either the total setup cost incurred in locating facilities at each level or for all of the facilities.
- *coverage-based models:* assume that a customer is covered if its demand point is within a specified distance of facilities offering the requested service(s). *Set-covering* models assume that all service demand must be covered and the goal is to minimize the setup cost for the facilities. *Maximum covering* models consider that the number of facilities to locate is given as an input and the objective is to maximize the total number of demands of all types that are covered. Some variants of maximum covering models assume that each demand point can only be considered as being covered if all the requested services at such point are satisfied.

13.2.5 Classification Scheme

The classification scheme takes into account the attributes mentioned above, namely *flow pattern and spatial configuration*, *service availability*, *nature of customers demand* and *decisions and objective*. The classification criteria for HFLPs is

Table 13.1 Classification criteria

Criterion	Description
Nature of the demand	Single demand/multiple demand, single product/multiple product
Flow pattern	Single-flow/multi-flow/bidirectional-flow
Service availability	Nested/non-nested/locally inclusive/globally inclusive
Spatial configuration	Coherent (single assignment)/non-coherent
Decisions	Network design/tactical/network design and tactical
Objective function	Median/fixed-charge/covering

summarized in Table 13.1. This classification generalizes the scheme proposed by Şahin and Süral (2007) which in turn extends that of Narula (1986). Note that additional attributes may be considered such as capacities at facilities and edges as well as the interaction between facilities of the same type or between customers. However, we focus on those that we consider to have the most impact on the hierarchical structure.

13.3 Applications

We next review some of the most relevant areas of application of HFLPs. These range from heath care and production-distribution systems, which are arguably among the oldest and most studied hierarchical systems in location science, to telecommunication and transportation systems which have given rise to new variants of HFLPs.

13.3.1 Health Care Systems

Given the wide variety of services that health care systems must provide within a specific region, these services are naturally governed by a hierarchical structure. Although the number of levels may vary by country and region, three-level systems are commonly found. Patients in a geographical region (i.e. neighborhood, district, or county) requesting a variety of services are modeled with a demand point at level 0. Local clinics (level 1) may provide basic care and diagnostic services. Community hospitals (level 2) could offer, in addition to basic care and diagnostic services, other services such as out-patient surgery and specialized clinical tests. Regional hospitals provide a wide variety of in-patient services and may or may not provide basic care and diagnostic services. Successively inclusive facility and mixed facility hierarchies as well as multi-flow patterns are predominant in health care systems. It is also somehow common to observe systems with a locally inclusive service hierarchy, specially in countries with a public health care system (Rahman and Smith 2000; Smith et al. 2013). In this case, a regional hospital may provide

basic care and diagnostic services but only to patients living in the proximity of the hospital. The interaction between facilities of different levels arises when patients are referred to community or regional hospitals after being diagnosed at a local clinic. Thus, it is important for health authorities to jointly determine the location of various health care facilities and how these will be interconnected to provide the best possible service to patients. We refer to Ahmadi-Javid et al. (2017) for a recent survey on the topic and to Chap. 23 for other facility location problems in health care.

13.3.2 Production-Distribution Systems

The design of production-distribution systems plays a central role in supply chain management. In particular, for companies that produce and deliver their goods. In such cases, a variety of products are produced in manufacturing facilities (level 3) and shipped to distributions centers (level 2), in which products are sorted, consolidated, and rerouted to regional warehouses (level 1) in order to be stored for a period of time (days, weeks, months) before being finally distributed to retail stores (level 0) for sale. This is a common example of a single-flow pattern on a multiple product environment. The interaction between facilities arises naturally due to the routing of products through the supply chain (from the highest level to its lowest level). Sometimes a multi-flow pattern may arise whenever products are directly shipped from manufacturing facilities to regional warehouses or retail stores. Both coherent and non-coherent structures have been reported in the literature (see, for instance Şahin and Süral 2007; Gendron et al. 2016). Locational and network design decision arising in supply chain management have been extensively studied in the literature. We refer to Melo et al. (2006) and Chap. 16 for an in-depth discussion of facility location problems in the context of supply chain and logistics.

13.3.3 Telecommunications Systems

Telecommunication networks are frequently built with a hierarchical structure having two or three levels. A classical example is the so-called hub-and-spoke architecture used in various distributed data networks. In those cases, service demand corresponds to electronic data transmissions between O/D pairs that are routed over a variety of links in the access-level and backbone-level networks. Hub facilities correspond to electronic equipment such as concentrators, multiplexors, and routers. An example of a three-level hub network arises in an intra-local access transport area network architecture (see, Wu et al. 1988; Yaman 2009). O/D nodes are the central offices and each central office is served by a regional hub. A group of central offices served by the same hub is referred to as a cluster. Each hub is then connected to a central hub (gateway). A group of clusters served by the same

central hub is a sector. Central hubs are connected by a complete fiber network. Communication services between O/D pairs are offered by using O/D paths in which data transmissions first visit the regional hub of its cluster, then one or two central hubs, and finally the regional hub of the cluster where the destination point belongs to. That is, demand flows move from lower-levels to higher-levels and back to lower-levels of the hierarchy (i.e. a bidirectional-flow pattern). Catanzaro et al. (2011) provide another example of a two-level hub network arising in the deployment of an Internet routing protocol called Intermediate System-Intermediate System (ISIS). We refer to Chap. 12 for a description of hub location problems arising in the design of hub-and-spoke networks and to Fortz (2015) for an overview of other location problems arising in telecommunications.

13.3.4 Urban Transportation Systems

Hierarchical structures of interacting facilities have recently appeared in the area of city logistics, in which consolidation activities can take place at different levels of an urban supply chain (Mancini et al. 2014; Savelsbergh and Van Woensel 2016). Many logistics companies deliver goods destined for an urban area by using long-haul transportation vehicles that arrive at consolidation facilities (level 2). These facilities are referred to as urban consolidation centers and are usually located in the boundaries of the urban zone. Commodities are then unloaded, sorted, consolidated, and loaded into smaller vehicles which are then routed to other intermediate logistics facilities (level 1), usually referred as cross-dock satellites. From these facilities commodities are shipped to retail stores (level 0) using different vehicle fleets to avoid the presence of large vehicles in the city center. Contrary to urban consolidation centers, cross-dock satellites can be located within the urban zone, even in dense populated areas. These may correspond to basic rendezvous points such as parking lots, rail stations or bus exchanges, where commodities are transferred from one vehicle to another. Cross-dock satellites can also be small warehousing facilities with limited storage capabilities. In any case, transshipment of flows is done in a highly synchronized fashion. The use of these consolidation-distribution strategies can follow a single-flow or multi-flow pattern and all flows are clearly routed form top to bottom.

13.3.5 Air Transportation Systems

Hub-and-spoke architectures are also widely used in air freight and passenger travel systems. In the case of the airline industry, global alliances and mergers have given rise to complex global air transportation systems (Adler and Smilowitz 2007; Bernardes Real et al. 2018). Some alliances operate extensive three-level hub networks, in which local airports (level 0), regional hubs (level 1), and international

gateways (level 3) interact among them to route millions of passengers each year. Regional hubs allow passengers to make connections along their routes, while gateways are necessary for connecting continents and for performing immigration, customs and security checks. Passengers traveling within the same continent or geographical region are routed via regional hubs, whereas transcontinental passengers are frequently routed via a combination of regional hubs and gateways to reach their destinations. Local airports are connected to one or more regional hubs and maybe a gateway. International gateways are connected (indirectly or directly) to all regional hubs in its geographical region. All international gateways are interconnected across continents. Similar to telecommunications systems, demand flow follows a bidirectional-flow pattern.

13.3.6 Cargo and Postal Delivery Systems

Other transportation systems, such as cargo and postal delivery, employ a hierarchical structure in their facilities and a mix of air and ground transportation services. In the case of postal services, customers deliver mail or small parcels at post boxes (level 0) in a city. At branch offices (level 1) customers can deposit mail and obtain other services such as buying stamps and envelopes, among other things. Postal flow is then routed to central post offices (level 2) to be sorted and rerouted to other central post offices and branch offices for delivery.

A similar situation arises in the case of cargo delivery systems where branch offices (level 1) collect and distribute cargoes from/to customers (level 0) directly using small trucks. Operations centers (level 2) collect and distribute cargoes on different geographical regions, which are connected with a central hub facility (level 3). This means that all flow must pass through a central hub facility. In some applications (see, for instance, Dukkanci and Kara 2017), operation centers and a central hub facility are connected with a set of tours performed by airplanes.

In both cargo and postal delivery systems the time aspect plays a major role. Thus, it is usually integrated in the design of the system with location and link activation decisions. In particular, these hierarchical networks must be designed in such as way that transportation services between O/D demand points can be performed within a predefined service time limit (i.e. same-day or next-day delivery). Operational scheduling decisions, such as release times at branch offices and operations centers need to be taken into account while designing the network to ensure demand flows can be delivered on time (Yaman et al. 2012).

13.4 Families of Hierarchical Facility Location Problems

Several variants of HFLPs have been studied under various names. We recall that when referring to different types of facilities numerous terms have been used such as level, layer, echelon, stage, tier, among others. We next discuss four classes

of HFLPs: multi-level facility location, median and covering hierarchical facility location, multi-echelon location-routing, and hierarchical hub location. The first two families are those HFLPs that were first studied since the early 1970's and also the ones that have received the most attention in the literature. On the other hand, the last two families of HFLPs have emerged over the last two decades and fewer references can be found. We would like to clarify that these four classes of HFLPs do not constitute a partition of the general field of HFLPs. That is, there could also be some overlapping between them and also some HFLPs may not necessarily belong to any of these classes of problems.

13.4.1 Multi-Level Facility Location Problems

Multi-level facility location problems (MLFLPs) are typically found in the context of supply chains and production-distribution systems. There are two main distinguishing features underlying most MLFLPs: there exist an inherent hierarchy given by the nature of the system and a successively exclusive facility hierarchy is usually considered. For example, in the case of production-distribution systems, the products need to be first produced in order to be shipped to regional warehouses for temporary storage. Once the products are requested with a given due date, they are routed to retail stores. In this case, the hierarchy of the different types of facilities is implicitly given by the nature of the system, i.e one cannot store a product which has not yet been produced. Moreover, production services offered at manufacturing facilities are not available at warehouses. Similarly, warehousing services such as sorting, labeling and consolidation operations as well as storage space are usually not available at manufacturing facilities.

Another distinguishing feature of MLFLPs is that non-trivial facility location decisions are taken at every level of the hierarchy, simultaneously. Other problems involve two or more types of facilities but only in one of them is the selection of facilities considered (see for instance Sect. 13.4.3). Moreover, in a MLFLP, there is no direct interaction between customers, and no horizontal interactions between facilities of the same level. Typically, the edges between facilities of different types are defined sequentially. Thus, a sequence of exactly one opened facility at each level is required which corresponds to what we called a single-flow pattern. Nevertheless, some problems with multi-flow patterns could also be considered as MLFLPs when demand flows are allowed to skip levels in the hierarchy (i.e some services may not be requested by some customers). Most of these multi-flow pattern problems can be modeled as single-flow-patterns by simply adding dummy nodes in the corresponding missing levels, at the expense of increasing the instance size. Also, when flow directions are considered, the flow between levels of an MLFLP must go in one direction and there ought to be only one type of arc available. Some HFLPs, especially those that arise in the framework of waste management systems, consider bidirectional-flows or more than one type of arc (see, for instance Barros et al. 1998; Mitropoulos et al. 2009). In terms of the coherency criterion both cases

have been studied for MLFLPs. Finally, we note that fixed-charge and median-based objective functions are more common for this class of problems.

As an example of an MIP formulation for an uncapacitated MLFLP that extends the *uncapacitated facility location problem* (UFLP) and the *p-median problem* (*p*-MP), we consider the following so-called path-based formulation (Tcha and Lee 1984; Aardal et al. 1999). Let $G = (I \cup J, E)$ be a graph with vertex set $I \cup J$ partitioned into $k+1$ levels, where J represents the set of customers, I is partitioned into $\{I_1, \cdots, I_k\}$, corresponding to the sets of potential facilities at levels 1 to k, and E is the set of edges. Let S be the set of all possible simple paths having exactly one node from each level, starting from some node $i_1 \in I_1$, finishing at some node $i_k \in I_k$. Also, consider c_{js} to be the cost associated with the allocation of customer $j \in J$ to the sequence of facilities in path $s \in S$. Now, let $p = (p_1, \cdots, p_k)$ be a vector of positive integers corresponding to the maximum number of facilities that can be opened at each level, and let f_{i_r} be the non-negative fixed cost associated with opening facility i_r at level r. We define the binary variables x_{js} equal to one if and only if customer $j \in J$ is assigned to path $s = i_1, \cdots, i_k \in S$. Also, we define the binary variables y_{i_r} equal to one if and only if facility i_r of level r is open. The formulation is the following:

$$\text{minimize} \quad \sum_{j \in J} \sum_{s \in S} c_{js} x_{js} + \sum_{r=1}^{k} \sum_{i_r \in I_r} f_{i_r} y_{i_r} \tag{13.1}$$

$$\text{subject to} \quad \sum_{s \in S} x_{js} = 1 \quad \forall \ j \in J \tag{13.2}$$

$$\sum_{s \in S : i_r \in s} x_{js} \leq y_{i_r} \quad \forall \ j \in J, \ \ i_r \in I_r, \ \ r = 1, \cdots, k \tag{13.3}$$

$$\sum_{i_r \in I_r} y_{i_r} \leq p_r \quad r = 1, \cdots, k \tag{13.4}$$

$$x_{js} \geq 0 \quad \forall \ j \in J, \ \ s \in S \tag{13.5}$$

$$y_{i_r} \in \{0, 1\} \quad \forall \ i_r \in I_r, \ \ r = 1, \cdots, k. \tag{13.6}$$

The objective (13.1) is to minimize the sum of the assignment costs and the setup cost for opening facilities at different levels. Constraints (13.2) ensure that exactly one path is assigned to every customer, while constraints (13.3) are the linking constraints which ensure that if a path is assigned to a customer, then all the facilities in such path must be open. Constraints (13.4) are the cardinality restrictions. Finally, note that the variables x_{js} can be relaxed from binary to continuous variables, as for the UFLP (see Chap. 4).

Some properties and characteristics of classical FLPs have been extended for the more general case of MLFLPs. For instance, Aardal et al. (1996) showed that all non-trivial facet defining inequalities for the UFLP also define facets for the two-level uncapacitated facility location problem. Aardal et al. (1999), Bumn and Kern (2001), and Zhang (2006) use ideas previously developed for the UFLP,

such as dual ascent and adjustment techniques (Erlenkotter 1978), in order to develop approximation algorithms for the multi-level UFLP. In this context, it is important to note that most research efforts towards the development of algorithms for MLFLPs have focused on heuristics. In particular, we can differentiate two main research streams in this field: heuristics without a performance guarantee, and ρ-approximation algorithms i.e., polynomial-time heuristics that yield a feasible solution with an objective function value lying within a factor of ρ of the optimal value. Most of the work has focused on the latter stream. A more recent example is the work of Krishnaswamy and Sviridenko (2016) who presented inapproximability results for the multi-level UFLP and showed that in the general case, the two-level UFLP is computationally harder than the single-level UFLP.

Most of the early works on MLFLPs introduced exact algorithms for different variants of the problem. For example, Kaufam et al. (1977) presented a branch-and-bound method that extended from the single-level case. Barros and Labbé (1994a) introduced a general version of an MLFLP including design and tactical decisions and developed a branch-and-bound procedure using the corresponding upper and lower bounds obtained from different Lagrangian relaxations of two formulations, and those obtained from an extension of the greedy heuristic proposed for the UFLP. More recently (Gendron et al. 2016; Ortiz-Astorquiza et al. 2019), developed efficient exact methods for MLFLPs based on Lagrangian relaxation and Benders decomposition, respectively.

As mentioned before, most of the techniques used to solve MLFLPs are especially modified from successful algorithms developed for single-level FLPs. One very important property in discrete optimization that has led to the development of algorithms for FLPs is submodularity. This property somehow resembles convexity for continuous functions on set functions. For the single-level case (Cornuéjols et al. 1977), presented worst-case bounds for greedy and local improvement heuristics for the maximization version of an FLP which includes as special cases the UFLP and the p-MP (see Chap. 4 for details on supermodularity and supermodular reformulations for the minimization version of the UFLP). Some of the first articles discussing MLFLPs assumed that the submodularity property extends directly from the single-level cases (Ro and Tcha 1984; Tcha and Lee 1984). Later (Barros and Labbé 1994b), showed that the set function associated with the natural combinatorial representation of the multi-level UFLP does not satisfy submodularity. However, other equivalent combinatorial optimization problems modeling the multi-level UFLP have an objective function that actually satisfies submodularity, as was shown in Ortiz-Astorquiza et al. (2015). This observation allowed to provide sufficient conditions to extend the results on worst-case bounds for greedy heuristics and submodular reformulations of single-level UFLP and pMP to MLFLPs (Ortiz-Astorquiza et al. 2017).

Similarly, the case of having capacities in the facilities has also received important attention. From the early works, we note that of Aardal (1992), who presented an MILP formulation for the two-level capacitated FLP and a polyhedral study. Aardal (1998) later introduced a reformulation along with computational results. Marín and Pelegrín (1999) compared two-index and a three-index formulations for

the development of an exact algorithm for the two-level capacitated FLP based on Lagrangian relaxations. As for the uncapacitated case (Bumn and Kern 2001; Ageev 2002; Du et al. 2009), developed ρ-approximation algorithms for capacitated MLFLPs with values of ρ equal to 12, 9 and $k+2+\sqrt{k^2+2k+5}+\epsilon$, respectively. Finally, multi-period (or dynamic) extensions of MLFLPs have also been studied in Hinojosa et al. (2000, 2008). For more details on classification, models, properties and solution methods for MLFLPs we refer to Ortiz-Astorquiza et al. (2018).

13.4.2 Median and Covering Hierarchical Location Problems

HFLPs that are referred to as *median-based hierarchical location problems* (MHLPs) and *covering-based hierarchical location problems* (CHLPs) in the literature are those which are frequently found in the context of health care systems, educational systems, and emergency medical services. One of the main distinguishing features of MHLPs and CHLPs is that either a successively inclusive facility hierarchy or a mixed facility hierarchy is considered. Similar to MLFPLs, there exist an inherent hierarchy given by the nature of the system. For instance, regional hospitals are clearly in a higher level of the hierarchy as compared to community hospitals and local clinics. The service availability may be successively inclusive or not but the hierarchy of the hospitals is implicitly given by the level of urgency or criticality of the offered services.

The objective function has been one of the most important factors when categorizing FLPs in general. The modeling and solution structures might change drastically when different objectives are considered (e.g. p-median, p-center, fixed charged and covering). One stream of research has focused on MHLPs in which a median objective is considered. An example of a nested multiple demand MFLPs can be formulated as follows. Let d_j^s denote the demand of service s at node j and c_{ji} denote the cost associated with the allocation of customer j to facility i. Additionally, consider the decision variables y_{ir} equal to one if and only if facility of type r is located at node i and variables x_{is}^j equal to one if and only if demand of service s at node j is satisfied with facility at node i. Then, we obtain

$$\text{minimize} \quad \sum_{i\in I}\sum_{j\in J}\sum_{s=1}^{k} d_j^s c_{ji} x_{is}^j \tag{13.7}$$

$$\text{subject to} \quad \sum_{i\in I} x_{is}^j = 1 \qquad \forall j \in J \;\; s = 1,\cdots,k \tag{13.8}$$

$$\sum_{i\in I} y_{ir} \le p_r \qquad r = 1,\cdots,k \tag{13.9}$$

$$x_{is}^j \le \textstyle\sum_{l=s}^{k} y_{il} \qquad \forall \;\; j \in J,\; i \in I \;\; s = 1,\cdots,k \tag{13.10}$$

$$x_{is}^{j} \in \{0,1\} \qquad \forall i \in I \;\; j \in J, \;\; s = 1, \cdots, k \tag{13.11}$$

$$y_{ir} \in \{0,1\} \qquad \forall \; i \in I, \;\; r = 1, \cdots, k. \tag{13.12}$$

The objective (13.7) is to minimize the total assignment cost. Constraints (13.8) ensure that the demand of each service of every node is met. Constraints (13.9) limit the number of open facilities of each type, while inequalities (13.10) are linking constraints which in this case define a successively (globally) inclusive HFLP. Note that replacing constraints (13.10) with $x_{is}^{j} \leq y_{is}$ modifies the problem into a non-nested HFLP. Then, we would have a successively exclusive formulation. Some of the early works on the subject are precisely those that identified the differences between a successively exclusive, successively inclusive or locally inclusive HFLP (Tien et al. 1983; Mirchandani 1987).

Later Weaver and Church (1991), formulated a nested MHLP with two types of facilities minimizing an objective function similar to that of Narula and Ogbu (1985). They proposed a Lagrangian procedure and a primal exchange substitution heuristic. Other examples where hierarchical p-median models have been studied are those of Galvão et al. (2002), Yasenovskiy and Hodgson (2007) and Hodgson and Jacobsen (2009). Also (Serra and ReVelle 1993; Alminyana et al. 1998), present solution methods for a nested and coherent hierarchical structure combining two p-median problems referred to as the pq-median problem.

On the other hand, when referring to covering objectives, an important notion is that of a demand point being *covered* (see Chaps. 3 and 5). In the context of HFLPs the three most common types of covering found in single-level FLPs have also been studied, namely the hierarchical extensions of the set covering location problem, the p-center problem, and the maximum covering problem (Toregas et al. 1971; Church and ReVelle 1974). Note that in these cases because we are considering different types of facilities, it is more intuitive to talk about different types of demand. However, defining a critical distance is also more challenging than in the single-level case. For example, one may be interested in covering a demand point by each type of facility. Another case would be, for instance, when each customer must be covered by a first level facility, first level facilities in turn are covered by second level facilities and so on. In the context of health care systems for example, this is referred to as bottom-up referral system (Church and Eaton 1987; Gerrard and Church 1994). In such cases the facilities are typically service-nested and thus a second level facility can also cover customers. Therefore, covering type objectives are more commonly found with multiple demand type of customers.

For example, let α_{ir}^{js} be a parameter whose value is equal to one if and only if demand at node j of service s can be covered by facility of type r from node i. Also, let z_{js} be the binary variables equal to one if and only if demand of service s

from node j is covered. A hierarchical maximum coverage location problem can be formulated as

$$\text{maximize} \quad \sum_{j \in J} \sum_{s=1}^{k} d_j^s z_{js} \tag{13.13}$$

$$\text{subject to} \quad \sum_{i \in I} y_{ir} \leq p_r \qquad r = 1, \cdots, k \tag{13.14}$$

$$z_{js} \leq \sum_{i \in I} \sum_{r=1}^{k} \alpha_{ir}^{js} y_{ir} \quad \forall \ j \in J, \ \ s = 1, \cdots, k \tag{13.15}$$

$$0 \leq z_{js} \leq 1 \qquad \forall \ j \in J, \ \ s = 1, \cdots, k \tag{13.16}$$

$$y_{ir} \in \{0, 1\} \qquad \forall \ i \in I, \ \ r = 1, \cdots, k. \tag{13.17}$$

The objective (13.13) is to maximize the sum of covered demand of each type of service. Constraints (13.14) limit the number of open facilities of each type, whereas constraints (13.15) ensure that demand at each node j for each service s is considered to be covered if and only if there exist at least one open facility which can provide such service to that node. The integrality restrictions on the z_{js} variables can be relaxed due to the sense of the objective function and constraints (13.17).

Note that in this formulation we may count as covered some demand points that are only partially covered. That is, customers with multiple demands which are covered for only some services still add value to the objective function. Another case is to impose that only complete covered customers add value to the objective function of total coverage.

Given the applicability of CHLPs several different variants have been presented focusing on case studies and analysis of solutions. One of the first works on CHLPs is that of Moore and ReVelle (1982) who proposed an IP formulation for a hierarchical problem with two types of facilities. Later (Gerrard and Church 1994), discussed and compared three additional CHLPs to the one proposed by Moore and ReVelle (1982). Marianov and Serra (1998, 2001) studied a CHLP in the context of congested systems. Espejo et al. (2003) developed dual based heuristics using Lagrangian relaxation to solve instances of a CHLP with two types of facilities. More recently (Lee and Lee 2010), proposed tabu-based heuristics for a generalization of the model introduced by Moore and ReVelle (1982). For more examples, formulations and solution methods on this family of HFLPs we refer the reader to Şahin and Süral (2007), Daskin (2013) and references therein.

13.4.3 Multi-Echelon Location-Routing Problems

The term *echelon* is generally associated with distribution networks where products are transported between each pair of levels. Such pairs are called echelons (Aiken 1985; Gao and Robinson 1992). Multi-echelon FLPs are thus very similar to

MLFLPs. In fact, many studies use both terms indistinctly. However, we note two main differences. The first one is that although all of the multi-echelon problems involve a multi-level environment, not all of them require facility location decisions at every level as in an MLFLP. For example, in one of the early works on the topic Geoffrion (1974) studied two-echelon FLPs in which facilities to be opened are only selected at one of the levels. This is partially because the predominant decisions are made at the echelons, and these typically involve routing decisions. Routing decisions in both senses, i.e., the flow between types of facilities and customers as well as the routes or tours in the same level of the hierarchy. Indeed, the second differentiating feature lies precisely in these routing patterns. MLFLPs are concerned with problems where facility, and sometimes network design decisions, are predominant with no routing decisions between nodes of the same level involved. The paper of Cuda et al. (2015) on two-echelon routing problems reviews a more general class of problems in which locational decisions are optional at all levels. That is, they include problems that may have no location decisions involved.

Another term that is generally related to echelons is the word *tier*, which has mainly been used in the context of freight transportation systems and city logistics (Crainic et al. 2009; Mancini et al. 2014). These HFLPs typically involve vehicle routing decisions extending those FLPs studied in Chap. 15. The term *stage* has also been used in this context and is possibly the most elusive one when trying to associate it to something in particular. In some references (e.g. Marín 2007) the term stage is used when referring to what we denote as levels. However, in other papers it has been used in the sense of what we identified as echelons (e.g. Klose 1999).

One example of a route-based formulation (Cuda et al. 2015) for a two-echelon capacitated location-routing problem is as follows. Let T^1 be the set of routes where each $t \in T^1$ starts from a facility of level 1 (e.g. warehouses) and visits one or several customers. Similarly, define the set of routes T^2 for the second echelon starting at a facility of the second level (e.g. plants) and visiting a group of warehouses. The binary parameters α_{ti_1} and β_{tj} indicate whether facility i_1 or customer j are in route t or not. Finally, given a route $t \in T^1$, let $d_t = \sum_{j \in t} d_j$ be the total demand for customers visited. Additionally to the fixed costs for setting up facilities f_i, consider the fixed costs paid for each vehicle used in each echelon g_r and the cost per route b_t. Now, let the binary decision variables y_{i_r} equal to one if and only if facility $i_r \in I_r$ of level $r = 1, 2$ is opened. Also, let variables x_t equal to one if and only if the route $t \in T^1 \cup T^2$ is in the solution, and let w_{ti_1} be a flow variable for route $t \in T^2$ that must be delivered to facility i_1. Then we have the following MIP formulation:

$$\text{minimize} \quad \sum_{r=1}^{2} \left(\sum_{i_r \in I_r} f_{i_r} y_{i_r} + \sum_{t \in T^r} (g_r + b_t) x_t \right) \tag{13.18}$$

$$\text{subject to} \quad \sum_{i_1 \in I_1} w_{ti_1} \leq c^2 x_t \qquad \forall t \in T^2 \tag{13.19}$$

$$\sum_{t \in T^2 : i_2 \in t} \sum_{i_1 \in I_1} w_{ti_1} \leq q_2 y_{i_2} \qquad \forall i_2 \in I_2 \tag{13.20}$$

$$\sum_{t \in T^2} w_{ti_1} \leq q_1 y_{i_1} \qquad \forall i_1 \in I_1 \tag{13.21}$$

$$\sum_{t \in T^2} \alpha_{ti_1} x_t = y_{i_1} \qquad \forall i_1 \in I_1 \tag{13.22}$$

$$\sum_{t \in T^2} w_{ti_1} = \sum_{t \in T^1 : i_1 \in t} d_t x_t \qquad \forall i_1 \in I_1 \tag{13.23}$$

$$\sum_{t \in T^1} \beta_{tj} x_t = 1 \qquad \forall j \in J \tag{13.24}$$

$$y_{i_r} \in \{0, 1\} \qquad \forall\ i_r \in I_r,\ \ r = 1, 2 \tag{13.25}$$

$$x_t \in \{0, 1\} \qquad \forall\ t \in T^1 \cup T^2 \tag{13.26}$$

$$w_{ti_1} \geq 0 \qquad \forall\ t \in T^2\ \ i_1 \in I_1, \tag{13.27}$$

where c^r is the capacity of the vehicles in echelon r and q_r is the capacity of facilities in level r. Constraints (13.19) ensure that if a route $t \in T^2$ is selected, then the total load delivered to all the satellites visited in that route must not exceed the vehicle capacity. Constraints (13.20) and (13.21) correspond to the capacity limits for each opened facility at both levels. Constraints (13.22) impose that if the first level facility is opened then it must be visited by one vehicle. Constraints (13.23) represent flow balance equations for each first level facility and constraints (13.24) ensure that each customer is served by one vehicle.

One of the first papers in this family of HFLPs is the one of Jacobsen and Madsen (1980) in which the problem considers location decisions at only one level of the hierarchy. More recent articles have focused on two-echelon location-routing problems having location decisions at both levels. For example, Boccia et al. (2010) and Contardo et al. (2012) present various formulations and solution methods extended from location-routing and vehicle routing problems. In general, as in the more global view of HFLPs, most of the related papers propose heuristic algorithms. In this particular family of HFLPs, perhaps one of the few exceptions where a specialized exact solution algorithm is developed is the branch-and-cut proposed in Contardo et al. (2012).

13.4.4 Hierarchical Hub Location Problems

All previously described HFLPs consider that customers, regardless if they are single/multiple demand or single/multiple product, can be represented with demand points. We now turn our attention to a different class of problems in which service demands correspond to the routing of commodities between O/D pairs over a hierarchical network. We refer to this class of HFLPs as *hierarchical hub location problems* (HHLPs). HHLPs arise in the design of hub-and-spoke networks

in telecommunications, air transportation, cargo, and postal delivery systems. Even though standard hub-and-spoke networks involving two-levels (access and backbone levels) have already a hierarchical structure, in this section we limit our study to hub networks involving two (or more) levels and in which location decisions arise. We refer to Chap. 12 for an in-depth analysis of hub-and-spoke networks involving locational decisions in one level.

An important feature of HHLPs is that, given the nature of their demand, most of these problems consider a bidirectional-flow pattern. That is, demand flow originates in an O/D node (level 0) and is routed to the highest level via one or more facilities of each intermediate level and then is routed back to an O/D node in the lowest level visiting once more one or more facilities of each level. Another feature of HHLPs is that facilities of the same level are connected and thus, additional link activation decisions are usually present (unless full interconnection between them is assumed). Yet another interesting characteristic of HHLPs is that in some situations the hierarchy of the different types of hub facilities is given by the nature of the system, such as the case of telecommunication networks in which the role of the electronic equipment determines the sequence in which the flow must be routed. However, in some other situations the hierarchy is given by a ranking mechanism that allows the ordering of the facilities into levels. For example, in the case of passenger airline networks there exist multiple levels of hubs. According to the Federal Aviation Administration (FAA), air traffic hubs are classified based on the percentage of total passengers enplaned in the area into one of four types of hubs: large hubs, medium hubs, small hubs, and non-hubs (Shaw 1993). This means that the hierarchy of facilities (hub airports), is determined by such metric.

We now provide a brief overview of the HHLPs that have been studied in the last decade. In the context of telecommunications networks (Yaman 2009), studies the problem of designing a three-level hub network, where the top layer consists of a complete network connecting the central hubs, and the second and third layers are unions of star networks connecting the remaining hubs to central hubs and the O/D nodes to hubs, respectively. The objective is to minimize the total routing cost while taking into account a cardinality constraint on the number of open hubs at each level. The author also studies an extension incorporating the same delivery time restriction for all O/D pairs. Yaman and Elloumi (2012) focus on two variants of HHLPs with covering-based objectives: the star p-hub center problem and the star p-median problem. These problems consist of locating p hubs (level 1) and connect them at central hub (level 2) via a star topology. Each O/D node is assigned to one hub in level 1. The objective of the former problem is to minimize the length of the longest path between O/D pairs. The objective of the latter is to minimize the total routing cost, while taking into consideration the service quality in terms of the length of paths between pairs of O/D nodes.

In the context of cargo delivery systems (Alumur et al. 2012), introduce a multimodal HHLP in which a three-level hub network is considered. O/D nodes are connected to ground hubs (level 1), which in turn are connected to airport hubs (level 2). A central airport hub (level 3) is connected with a star topology to all airport hubs. Depending on the O/D pairs, demand flow may visit ground hubs, or

a combination of ground and airport hubs. The objective is to minimize the sum of setup costs for link activation decisions and routing costs, while taking into account service time constraints on O/D paths. Dukkanci and Kara (2017) study a variant of the previous problem in which a ring-star-star topology is assumed. In particular, the airport hubs are connected with the central airport hub with a set of rings (or routes) performed by the aircrafts. The objective is to minimize the setup cost for the activation of the links between airport hubs while taking into account service time constraints on O/D paths.

For the case of air transportation systems (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. The authors develop a game theoretic approach in which merger and location decisions are considered. Bernardes Real et al. (2018) introduce a more comprehensive model in which gateways (level 2) and regional hubs (level 1) need to be located on a tree-level hub network. Backbone networks are no longer assumed to be complete. Unlike previous models, local and global flows are differentiated as the structure of OD paths associated with each type of flow is different. In particular, global flows can only leave or enter a geographic region via a gateway hub, while local flows can only use domestic hubs within their region.

Zhong et al. (2018) present a HHLPs arising in the design of public transport systems in which a three-level hub network is considered. O/D nodes correspond to traffic districts (level 0) in urban and rural areas. Hubs in central towns (level 1) provide service to rural areas, which are connected to urban public transport hubs (level 2) located inside the city or on the rural-urban boundary. These urban transportation networks are used to satisfy demand generated by urban and rural residents moving into and out of the city each day. The authors consider a fixed charge objective that minimizes the sum of setup costs for the installation of hubs at both levels and transportation costs.

13.5 Conclusions

In this chapter we presented a unified view of hierarchical facility location problems. They constitute a general class of discrete location problems arising in the design of hierarchical systems, in which different types of facilities interact to collectively provide services or products to a set of customers. We discussed the fundamentals of these problems, including their distinguishing features and key concepts. We also provided a comprehensive classification scheme that combines and extends previous schemes. We provided a review on applications areas, focusing on classical and new applications that have recently emerged. We also presented a concise overview of four classes of HFLPs that have received the most attention in the literature: multi-level facility location, median and covering hierarchical facility location, multi-echelon location-routing, and hierarchical hub location.

Although a substantial progress has been done by researches and practitioners in the area of hierarchical facility location, there is still significant work to be done. Identifying new areas of application will give rise to more realistic and complex models capable of capturing features of real-life. For instance, the recent work of Smith et al. (2017) focusing on the location of IV/AIDS test laboratories in South Africa and of Teixeira et al. (2019) dealing with the location of courts of justice in Portugal, provide good examples of the innovative applications of HFLPs. Moreover, although some recent progress has been done in the solution of HFLPs of realistic size (Ortiz-Astorquiza et al. 2019), sophisticated solution algorithms capable of exploiting the network flow structure of hierarchical models still need to be investigated to solve large-scale instances of more realistic variants. Other aspects of HFLPs that have received limited attention include the uncertainty in demand and travel times, as well as the multi-period nature of decision problems involving strategic decisions. Finally, another aspect that has been rarely discussed in HFLPs is that of incorporating into the decision making process the allocation of services to facilities and to exogenously determine the number of levels in the hierarchy (see, Narasimhan and Pirkul 1992).

References

Aardal K (1992) On the solution of one and two-level capacitated facility location problems by the cutting plane approach. Louvain-la-Neuve, Belgium: Université Catholique de Louvain PhD Thesis

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

Aardal K, Labbé M, Leung J, Queyranne M (1996) On the two-level uncapacitated facility location problem. INFORMS J Comput 8:289–301

Aardal K, Chudak FA, Shmoys DB (1999) A 3-approximation algorithm for the *k*-level uncapacitated facility location problem. Inf Process Lett 72:161–167

Adler N, Smilowitz K (2007) Hub-and-spoke network alliances and mergers: price-location competition in the airline industry. Transp Res B-Method 41:394–409

Ageev, A (2002). Improved approximation algorithms for multilevel facility location problems. Oper Res Let 30:327–332

Ahmadi-Javid A, Seyedi P, Siddhartha SS (2017) A survey of healthcare facility location. Comput Oper Res 79:223–263

Ahuja RK, Magnanti TL, Orlin JB (1993) Network flows: theory, algorithms, and applications. Prentice Hall, New Jersey

Aikens CH (1985) Facility location models for distribution planning. Eur J Oper Res 22:263–279

Alminyana M, Borras F, Pastor JT (1998) A new directed branching heuristic for the pq-median problem. Locat Sci 6:1–23.

Alumur SA, Yaman H, Kara BY (2012) Hierarchical multimodal hub location problem with time-definite deliveries. Transp Res E-Log 48:1107–1120

Barros AI, Labbé M (1994a) A general model for the uncapacitated facility and depot location problem. Locat Sci 2:173–191

Barros AI, Labbé M (1994b) The multi-level uncapacitated facility location problem is not submodular. Eur J Oper Res 72:607–609

Barros AI, Dekker R, Scholten V (1998) A two-level network for recycling sand : a case study. Eur J Oper Res 110:199–214

Bernardes Real L, O'Kelly M, de Miranda G, de Camargo R (2018) The gateway hub location problem. J Air Transp Manag 73:95–112

Boccia M, Crainic TG, Sforza A, Sterle C (2010) A metaheuristic for a two echelon location-routing problem. In: Festa P (ed). Experimental algorithms: lecture notes in computer science, vol 6049. Springer, Berlin, pp 288–301

Bumb A, Kern W (2001) A simple dual ascent algorithm for the multilevel facility location problem. In: Goemans M, Jansen K, Rolim JDP, Trevisan L (eds) Approximation, randomization, and combinatorial optimization: algorithms and techniques. Springer, Berlin, pp 55–63

Calvo AB, Marks DH (1973) Location of health care facilities: an analytical approach. Socio Econ Plan Sci 7:407–422

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

Church R, Eaton DJ (1987) Hierarchical location analysis using covering objectives. In: Gosh A, Rushton G (eds) Spatial analysis and location-allocation models. Van Nostrand Reinhold, New York, pp 163–185

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

Contardo C, Hemmelmayr V, Crainic TG (2012) Lower and upper bounds for the two-echelon capacitated location routing problem. Comput Oper Res 39:3215–3228

Cornuéjols G, Fisher ML, Nemhauser GL (1977) Location of bank accounts to optimize float: an analytic study of exact and approximate algorithms. Manag Sci 23:789–810

Crainic TG, Ricciardi N, Storchi G (2009). Models for evaluating and planning city logistics systems. Transp Sci 43:432–454

Cuda R, Guastaroba G, Speranza MG (2015) A survey on two-echelon routing problems. Comput Oper Res 55:185–199

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

Dökmeci VF (1973) An optimization model for a hierarchical spatial system. J Reg Sci 13:439–451

Dökmeci VF (1977) A quantitative model to plan regional health facility systems. Manag Sci 24:411–419

Du, D Wang, X Xu, D (2009) An approximation algorithm for the k-level capacitated facility location problem. J Comb Opt 20:361–368

Dukkanci O, Kara BY (2017) Routing and scheduling decisions in the hierarchical hub location problem. Comput Oper Res 85:45–57

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

Espejo LGA, Galvão RD, Boffey B (2003) Dual-based heuristics for a hierarchical covering location problem. Comput Oper Res 30:165–180

Fortz B (2015) Location problems in telecommunications. In: Laporte G, Nickel S, Saldanha da Gama, F(eds) Location science. Springer, Berlin, pp 537–554

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

Gao L-L, Robinson EPJ (1992) A dual-based optimization procedure for the two-echelon uncapacitated facility location problem. Nav Res Logist 39:191–212

Gendron B, Khuong P-V, Semet F (2016). A Lagrangian-based branch-and-bound algorithm for the two-level uncapacitated facility location problem with single-assignment constraints. Transp Sci 50:1286–1299

Geoffrion, AM (1974) Multicommodity distribution system design by Benders decomposition. Manag Sci 20:822–844

Gerrard RA, Church RL (1994) A generalized approach to modeling the hierarchical maximal covering location problem with referral. Pap Reg Sci 73:425–453

Hinojosa Y, Puerto J, Fernández F (2000) A multiperiod two-echelon multi-commodity 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
Hodgson MJ, Jacobsen SK (2009) A hierarchical location–allocation model with travel based on expected referral distances. Ann Oper Res 167:271–286
Jacobsen SK, Madsen OBG (1980) A comparative study of heuristics for a two-level routing-location problem. Eur J Oper Res 5:378–87
Kaufman L, Eede M, Hansen P (1977) A plant and warehouse location problem. Oper Res Quart 28:547–554
Klose A (1999) An LP-based heuristic for two-stage capacitated facility location problems. J Oper Res Soc 50:157–166
Krishnaswamy R, Sviridenko M (2016) Inapproximability of the multilevel uncapacitated facility location problem. ACM Trans Algorithms 13:1–26
Lee JM, Lee YH (2010) Tabu based heuristics for the generalized hierarchical covering location problem. Comput Ind Eng 58:638–645
Mancini S, Gonzalez-Feliu J, Crainic TG (2014) Planning and optimization methods for advanced urban logistics systems at tactical level. In: Gonzalez-Feliu J, Semet F, Routhier J-L (Eds.), Sustainable urban logistics: concepts, methods and information systems. Springer, Berlin, pp. 145–164
Marianov V, Serra D (1998) Probabilistic, maximal covering location-allocation models for congested systems. J Reg Sci 38:401–424
Marianov V, Serra D (2001) Hierarchical location-allocation models for congested systems. Eur J Oper Res 135:195–208
Marín A (2007) Lower bounds for the two-stage uncapacitated facility location problem. Eur J Oper Res 179:1126–1142
Marín, A Pelegrín, B (1999) Applying Lagrangian relaxation to the resolution of two-stage location problems. Ann Oper Res 86:179–198
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
Mirchandani P (1987) Generalized hierarchical facility locations. Transp Sci 21:123–125
Mitropoulos P, Giannikos I, Mitropoulos I (2009) Exact and heuristic approaches for the locational planning of an integrated solid waste management system. Operat Res 9:329–347
Moore GC, ReVelle CS (1982) The hierarchical service location problem. Manag Sci 28:775–780
Narasimhan S, Pirkul H (1992) Hierarchical concentrator location problem. Comput Commun 15:185–91
Narula SC (1984) Hierarchical location-allocation problems: a classification scheme. Eur J Oper Res 15:93–99
Narula SC (1986) Hierarchical location problems. Ann Oper Res 6:257–272
Narula SC, Ogbu UI (1985) Lagrangean relaxation and decomposition in an uncapacitated 2-hierarchical location-allocation problem. Comput Oper Res 12:169–180
Ortiz-Astorquiza C, Contreras I, Laporte G (2015) The multi-level facility location problem as the maximization of a submodular set function. Eur J Oper Res 247:1013–1016
Ortiz-Astorquiza C, Contreras I, Laporte G (2017) Formulations and approximation algorithms for multi-level uncapacitated facility location. INFORMS J Comput 29:767–779
Ortiz-Astorquiza C, Contreras I, Laporte G (2018) Multi-level facility location problems. Eur J Oper Res 267:791–805
Ortiz-Astorquiza C, Contreras I, Laporte G (2019) An exact algorithm for multilevel uncapacitated facility location. Transp Sci 53:1085–1106
Rahman SU, Smith DK (2000) Use of location-allocation models in health service development planning in developing nations. Eur J Oper Res 123:437–452
Ro H-B, Tcha D-W (1984) A branch and bound algorithm for the two-level uncapacitated facility location problem with some side constraints. Eur J Oper Res 18:349–358

Savelsbergh M, Van Woensel T (2016) City logistics: challenges and opportunities. Transp Sci 50:579–590

Şahin G, Süral H (2007) A review of hierarchical facility location models. Comput Oper Res 34:2310–2331

Schilling D, Elzinga DJ, Cohon J, Church RL, ReVelle CS (1979) The team/fleet models for simultaneous facility and equipment siting. Transp Sci 13:163–175

Schultz CP (1970) The logic of health care facility planning. Socio Econ Plan Sci 4:383–393

Serra D, ReVelle CS (1993) The *pq*-median problem: location and districting of hierarchical facilities. Locat Sci 1:299–312

Shaw SL (1993) Hub structures of major US passenger airlines. J Transp Geogr 1:47–58

Smith HK, Harper PR, Potts CN (2013) Bicriteria efficiency/equity hierarchical location models for public service application. J Oper Res Soc 64:500–512

Smith H, Cakebread D, Battarra M, Shelbourne B, Cassim N, Coetzee L (2017) Location of a hierarchy of HIV/AIDS test laboratories in an inbound hub network: case study in South Africa. J Oper Res Soc 68:1068–1081

Tcha D-W, Lee B (1984) A branch-and-bound algorithm for the multi-level uncapacitated facility location problem. Eur J Oper Res 18:35–43

Teixeira JC, Bigotte JF, Repolho HM, Antunes AP (2019) Location of courts of justice: the making of the new judiciary map of Portugal. Eur J Oper Res 272:608–620

Tien JM, El-Tell K, Simons, GR (1983) Improved formulations to the hierarchical health facility location-allocation problem. IEEE Trans Syst Man Cybern 13:1128–1132

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

Weaver JR, Church RL (1991) The nested hierarchical median facility location model. INFOR Inf Sys Oper Res 29:100–115

Wu TH, Kolar DJ, Cardwell RH (1988) Survivable network architectures for broad-band fiber optic networks: model and performance comparison. J Lightwave Technol 6:1698–1709

Yaman H (2009) The hierarchical hub median problem with single assignment. Transp Res B-Method 43:643–658

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, Karasan OE, Kara BY (2012) Release time scheduling and hub location for next-day delivery. Oper Res 60:906–917

Yasenovskiy VS, Hodgson MJ (2007) Hierarchical location-allocation with spatial choice interaction modeling. Ann Assoc Am Geogr 97:496–511

Zanjirani Farahani R, Hekmatfar M, Fahimnia B, Kazemzadeh N (2014) Hierarchical facility location problem: models, classifications, techniques, and applications. Comput Ind Eng 68:104–117

Zhang J (2006) Approximating the two-level facility location problem via a quasi-greedy approach. Math Program 108:159–176

Zhong W, Juan Z, Zong F, Su H (2018) Hierarchical hub location model and hybrid algorithm for integration of urban and rural public transport. Int J Distrib Sens Netw 14:1–14

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 time under special consideration of customer behavior. After developing a general framework for customers' decision making, the main results are cast within this framework. The conclusion outlines a number of areas, in which existing models can be refined and made more realistic.

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 uniformly distributed. (The often-mentioned "ice cream vendors on the beach" were

H. A. Eiselt
Faculty of Business Administration, University of New Brunswick, Fredericton, NB, Canada
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, USA
e-mail: tdrezner@fullerton.edu

G. Laporte et al. (eds.), *Location Science*,
https://doi.org/10.1007/978-3-030-32177-2_14

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 order to follow the argument, first consider a graphical representation of Hotelling's

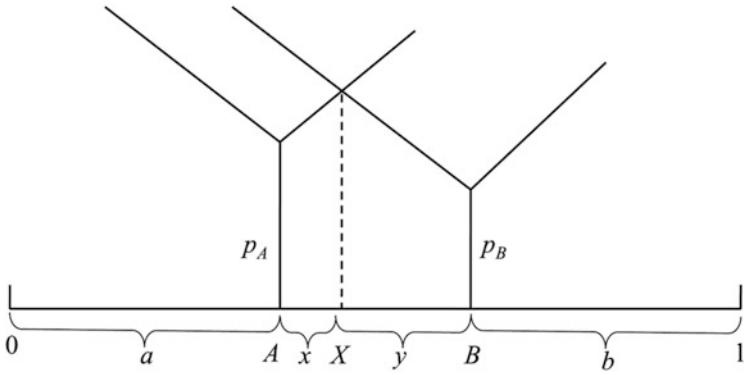


Fig. 14.1 Hotelling's duopoly on a linear market

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 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.

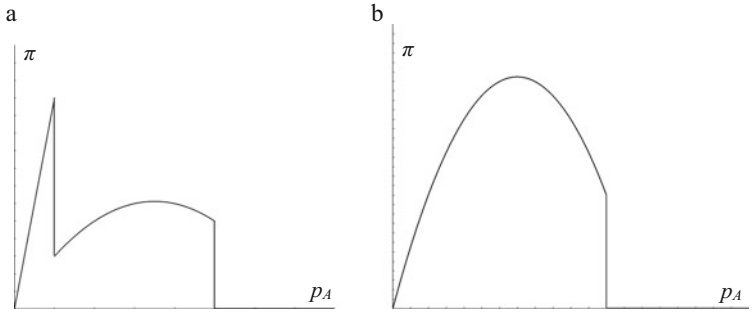


Fig. 14.2 Competitor A's profit functions with linear transportation costs. (a) A and B are close to each other and (b) A and B are far apart

The obvious question is whether or not there exists 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 show 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 exists a point, at which it is sufficiently high so that firm B cuts out firm A, and thus A's profit drops to zero. Note 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). Fig. 14.3a, b show 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

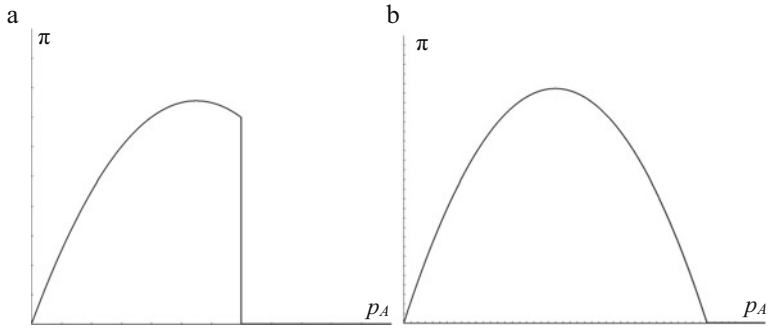


Fig. 14.3 Competitor A's profit functions with quadratic transportation costs. (a) A and B are close to each other and (b) A and B are far apart

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

$$\begin{aligned} TC &= t\left[\int_{\Phi=0}^{A}(A-\Phi)\,d\Phi+\int_{\Phi=A}^{X}(\Phi-A)\,d\Phi+\int_{\Phi=X}^{B}(B-\Phi)\,d\Phi+\int_{\Phi=B}^{1}(\Phi-B)\,d\Phi\right] \\ &= t\left[3A^2/4+3B^2/4-AB/2-B+\tfrac{1}{2}\right] \end{aligned}$$

Partial differentiation $\frac{\partial TC}{\partial A}=0$ and $\frac{\partial TC}{\partial B}=0$ results in the optimal points $A=¼$ and $B=¾$, 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 $(A, B) = (¼, ¾)$ 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. Kress and Pesch (2012) surveyed the subject, but concentrate on problems on networks, while Drezner (2014) surveys problems in the plane. Similar to the aforementioned contribution by Eiselt et al. (1993), Ashtiani (2016) first outlines some of the main characteristics of competitive

location problems and then reviews individual papers published in 2000–2014. While Karakitsiou and Migdalas (2017) survey competitive location problems with respect to Nash equilibria, Aras and Küçükaydın (2017) review contributions that focus on von Stackelberg solutions. Finally, Marianov and Eiselt (2016) investigate existing competitive location models with respect to the tendencies of facilities to agglomerate or disperse.

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 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 Hakimi (1964) on network location properties, in which he proved that in some classes of 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. 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 act or react. Consider the simple case of pure location competition. In such a 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 asymmetry 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 has a Nash equilibrium. Applied to our type of problem, players may choose a "first location, then price" strategy (see, e.g., Anderson and 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. Note 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. We would like to point out, though, that there are a number of instances, in which allocation and customer choice models are quite similar. If a firm delivers goods to customers, it may ship from the facility closest to the customer. Similarly, the same customer, in case he purchases the good from a facility and transports it home, may also choose the closest facility. The main difference between the two cases is that in the former,

transportation costs appear explicitly in the firm's objective function, whereas they do not in the latter, where proximity enters in the form of which facility is chosen by a customer, but not in the form of transportation costs. 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 located at site i (for simplicity, we will refer to "customer i") associates with goods or services 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 one when it is sold through specialty retail outlets.

The simplest (deterministic) utility function is

$$\text{UD1a} : u_{ij} = -td_{ij},$$

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. (1999a, b).

An extension is the utility function

$$\text{UD1b} : u_{ij} = -p_j - td_{ij}.$$

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

$$\text{UD1} : u_{ij} = R_i - p_j - td_{ij},$$

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 when 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 square 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. Such a utility function can be written as

$$\text{UD1c} : u_{ij} = R_i - p_j - td_{ij} - W_i,$$

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

$$\text{UD2a} : u_{ij} = \frac{A_j}{d_{ij}^{\lambda}}$$

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

$$\text{UD2b} : u_{ij} = A_j e^{-\beta d_{ij}}$$

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

$$\text{UP1} : u_{ij} = R_i - p_j - td_{ij} + \varepsilon_i \mu,$$

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

$$\text{UP2} : u_{ijk},$$

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 defined below.

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

$$\text{AD1} : \text{winner} - \text{take} - \text{all},$$

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

$$\text{AD2} : \text{proportional allocation}.$$

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 $u_{ij} / \sum_k u_{ik}$. 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,

$$\text{AP1a} : p_{ij} = \frac{A_j / d_{ij}^{\lambda}}{\sum_k A_k / d_{ik}^{\lambda}}$$

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

$$\text{AP1} : p_{ij} = \frac{\prod_k u_{ijk}^{\alpha_k}}{\sum_j \prod_\ell u_{ij\ell}^{\alpha_\ell}}.$$

Following the arguments of McFadden (1974), the use of the probabilistic utility function UP1 leads to the demand allocation rule

$$\text{AP2} : \quad p_{ij} = \frac{e^{(R_i - p_j - t d_{ij})/\mu}}{\sum_k e^{(R_i - p_k - t d_{ik})/\mu}}.$$

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

$$\text{Q1} : \text{fixed},$$

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

$$\text{Q2}: q_{ij} = f\left(p_j + td_{ij}, u_{ij}\right),$$

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 respectively 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 his 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 choice rules outlined in the previous section. Each subsection will examine one customer choice rule, given a specific space in which customers and facilities are (going to be) located in, and the type of solutions that are investigated, *viz.*, followed by results in the literature regarding Nash equilibria, and von Stackelberg solutions. To avoid too much fragmentation, we will list those contributions that deal with some discrete space under the header "plane."

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 ¼ and ¾ 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, 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. The global equilibrium is reached when Voronoi cells form a regular 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. The 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 facility, and the follower locates $r > 1$ 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. Following the asymmetry of objectives already mentioned by Eiselt and Marianov (2017), Gentile et al. (2018) consider three scenarios, each with a specific combination of objectives by leader and followers, in a discrete space. Pelegrín et al. (2015) explore the effects of tie-

breaking rules in customer choice on the solutions, while Santos-Peñate et al. (2017) suggest a heuristic for the solution of the centroid problem. Seyhan et al. (2018) consider the leader-follower problem in a discrete space and, in order to make the reaction function of the follower more tractable, suggest a greedy heuristic for that purpose. Xue et al. (2017) have the follower maximize his revenue, but allow the total demand to increase in case additional facilities locate. Zhang et al. (2016) study the usual leader-follower model, but include the possibility of disruptions of service. Finally in this category, a number of authors study competitive hub location problems, as there are Sasaki et al. (2014), Mahmutogullari and Kara (2015), Niknamfar et al. (2017), and Ghaffarinasab et al. (2018).

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 for the cases where 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 that 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. The author introduces leader and follower, respectively, but does not make the connection to von Stackelberg's work. The paper proves that on a tree network, the 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, 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. Benati and Laporte (1994) devise a tabu search algorithm for the solution of these difficult problems. 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 Pérez and Pelegrín (2003) follow the analysis of Eiselt (1992) 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 author 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. The work by Serra et al. (1999a, b) discusses the MAXCAP problem with different rules for the location of the entering firm. The rules, both of which belong to the class of proportion models, are based on different assumption concerning customer behavior.

Serra et al. (1999a, b) 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 is provided; the rule checks viability first and then locates and reallocates demand; if any store is not viable at this point, the one with least demand is deleted. This process is repeated until it converges. To solve the problem, heuristic concentration is the method of choice.

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. Marianov et al. (2004) extend these results to facilities with waiting lines. Ruiz-Hernández et al. (2017) discuss the case of delocation, "i.e., the possibility of optimally closing facilities. The usual customer choice is applied, except with some degree of loyalty. The authors investigate whether or not the first mover advantage occurs, and also study Nash equilibria.

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 to this 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 for 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 that, 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 ¼ and ¾, 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 ½.

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 (2016). 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 Bertrand 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 exist, and without coordination, the result

of maximum differentiation is invalidated. Eaton (1972) follows Smithies (1941) by considering a model that 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. Eiselt and Marianov (2017) determine the line between existence and nonexistence of locational Nash equilibria for location problems with asymmetries, such as those with different transportation costs, different production costs, and those that have different objective functions. The reference also investigates von Stackelberg solutions for these problems. While the work by Eiselt and Marianov (2017) focuses on asymmetric competitive location models, Colombo (2016) investigates equilibria on a linear market in the presence of three cities given Cournot (i.e., quantity) and Bertrand (i.e., price) competition.

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 Jr. 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. Firms 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 shows 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 et al. (1979) result. While this result shows maximum differentiation in one direction, it has minimum differentiation in the other. Lederer and Hurter Jr. (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 are 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. The model by Fernández et al. (2014) is the usual two-phase "first location, then price" game with delivered pricing, for which the authors demonstrate that a price equilibrium exists, which reduces the game to a pure location game. The paper then describes a branch-and-bound approach for small to medium problems and a heuristic for larger problems. Rohaninejad et al. (2017) present two models, in which firms maximize profits, and minimize the maximum deviation from the highest possible profit, respectively. Computational evidence is provided.

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. The paper by Berglund and Kwon (2014) has a von Stackelberg firm competing with Cournot-Nash firms given capacities at the facilities. Equilibrium results are presented and the computational method of choice is a simulated annealing heuristic.

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 a 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. In Guo and Lai's (2017) simultaneous location-and-price game on a linear market, firms face a non-uniform distribution of demand. Another feature is the inclusion of an online e-tailer. The long run will see the brick-and-mortar retailers more densely agglomerated than without the online competition, and they will serve with urban population, while the e-tailer will specialize in the rural population.

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 so, 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 UD1, Plane, von Stackelberg Solution

Panin et al. (2014) uses price discrimination in a sequential "first location, then price" game. Customers in their model have reservation prices and firms are assumed to have budget constraints. While the Phase 1 location competition uses the standard leader-follower concept, the Phase 2 pricing game searches for Nash equilibria. The work formulates the problem as a bilevel optimization problem and devises heuristic algorithms of the "alternating" type to solve the problem. Customers in Kononov et al. (2018) have a budget and finite demand. The study concentrates on complexity results and solvable cases.

14.4.14 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.

14.4.15 UD2a, Plane, von Stackelberg Solution

This special field has been very active in the last few years. Earlier work by 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. The contributions by Fernández et al. (2017a, b) both investigate the effects of different choice rules have on locational patterns. In particular, they "rediscover" Hakimi's (1983) binary and partially binary choice rules and solve the resulting problems with branch-and-bound methods and heuristics, respectively. Hendrix (2016) includes different costs for leader and follower as he determines optimal locations and qualities. It turns out that there is no equilibrium in qualities, so that a von Stackelberg solution for qualities is determined. Qi et al. (2017) apply the usual leader-follower concept, but will serve only customers, if they are within a prespecified distance from the facility. The work by Bagherinejad and Niknam (2018) follows similar lines in that the competitors do not just locate their facilities, but choose qualities of the facilities as well. The model under consideration allows the closing of facilities. Similarly, the contribution by Arrondo et al. (2014) choose locations and qualities and investigates exact and heuristic solution techniques. The papers by Rahmani (2016) and Sadjadi et al. (2016) both allow adjustments of a facility's attractiveness in addition to its location. The former contribution relaxes part of the problem and uses an exact algorithm to solve the problem, while the latter work uses a methheuristic, which is then applied to some real data.

14.4.16 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 than 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.

14.4.17 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.

14.4.18 UD2b, Plane, von Stackelberg Solution

Drezner et al. (2015) discuss a model, in which facilities attract customers that are located within a "sphere of influence." Given that the follower will react by maximizing its market share, the leader's objective is to maximize his own market share after the follower has reacted. A summary of leader-follower models in the plane is provided by Drezner and Drezner (2017). Levanova and Gnusarev (2018) consider the follower's problem, in which the follower has a limited budget, which

can be used to locate a facility with a given attractiveness. The authors develop an ant colony algorithm (the main piece of this paper), which they use to solve randomly generated instances of the problem.

14.4.19 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 with travel time, and add waiting time as a competitive factor. Shan et al. (2017) consider the follower's location-pricing game with mill pricing and a budget that limits the construction of stores. The lower-level pricing game represents a Nash equilibrium. The proposed algorithm for the follower problem is tabu search, and a numerical example concludes the paper.

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.

14.4.20 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.

14.4.21 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 pseudoconcave, the game possesses 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. The thesis by Tuan (2017) considers the follower problem in a discrete setting and evaluates different probabilistic choice rules.

14.4.22 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.

14.4.23 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. Čvokić et al. (2016) considers leader and follower, who locate their respective hubs. Both firms are profit maximizers. The problem is formulated, and the follower part of the formulation is solved by way of an "alternate" heuristic. Kress and Pesch (2016) also consider the follower's problem. Their formulation of the problem includes conditions for a price equilibrium. The authors then state conditions for the existence of a price equilibrium, followed by *NP*-hardness results, and a method to compute equilibrium prices, and some computational experiments.

14.4.24 UP2, Plane, von Stackelberg Solution

Drezner et al. (2002) discuss a medianoid problem in the plane, in which customers' choices are modeled in probabilistic fashion 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.

14.4.25 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.

14.5 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, which 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 or 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, see, e.g., Applegate et al. (2007), or a traveling purchaser problem, as described in Laporte et al. (2003). Planning multi-purpose shopping trips has been shown to foster the agglomeration of facilities; see, e.g., Marianov et al. (2018).

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 expected 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, Internet 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. This webrooming effect, i.e., the practice of obtaining information online and then shopping locally, counteracts the effects of showrooming, at least to some extent.

A different aspect that appears to be very promising deals not with the development of more realistic models, but with their visualization, which may provide insight and increase acceptability by decision makers. A good survey of the use of geographical information systems in location analysis is provided in Chap. 19 of this volume.

Acknowledgments This paper was in part supported by grants from the Institute Complex Engineering Systems, through grant CONICYT PIA FB 0816) and FONDECYT 1160025. This support is gratefully acknowledged. The authors would also like to thank a 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, Neven DJ (1989) Market efficiency with combinable products. Eur Econ Rev 33:707–719

Anderson SP, Neven DJ (1991) Cournot competition yields spatial agglomeration. Int Econ Rev 32:793–808

Anderson SP, de Palma A (1992) Spatial equilibrium with footloose firms. J Reg Sci 32(3):309–320

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, NJ

Aras N, Küçükaydın H (2017) Bilevel models on the competitive facility location problem. In: Mallozzi L, D'Amato E, Pardalos P (eds) Spatial interaction models. Springer optimization and its applications 118. Springer, Cham, pp 1–19

Arrondo AG, Fernández J, Redondo JL, Ortigosa PM (2014) An approach for solving competitive location problems with variable demand using multicore systems. Optim Lett 8:555–567

Ashtiani MG (2016) Competitive location: a state-of-art review. Int J Ind Eng Comput 7:1–18

Bagherinejad J, Niknam A (2018) Solving the competitive facility location problem considering the reactions of competitor with a hybrid algorithm including Tabu search and exact method. J Ind Eng Int 14:171–183

Beckmann MJ (1972) Spatial Cournot oligopoly. Pap 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

Benati S, Laporte G (1994) Tabu search algorithms for the ($r|X_p$) medianoid and the (r|p) centroid problems. Locat Sci 2(4):193–204

Berglund PG, Kwon C (2014) Solving a location problem of a Stackelberg firm competing with Cournot-Nash firms. Netw Spat Econ 14:117–132

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, New York, 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. Manag Sci 36(2):175–199

Colombo S (2016) A model of three cities: the locations of two firms with different types of competition. Int Reg Sci Rev 39(4):386–416

Cremer H, Marchand M, Thisse J-F (1991) Mixed oligopoly with differentiatied products. Int J Ind Organ 9:43–53

Čvokić DD, Kochetov YA, Plyasunov AV (2016) A leader-follower hub location problem under fixed markups. In: Kochetov Y, Khachay M, Beresnev V, Nurminski E, Pardalos P (eds) Discrete optimization and operations research. Lecture notes in computer science, vol 9869. Springer, Cham, pp 350–363

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. Manag 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. A chapter. In: Drezner Z (ed) 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(114):1–12

Drezner T, Drezner Z (1997) Replacing continuous demand with discrete demand in a competitive location model. Nav Res Logist 44:81–95

Drezner T, Drezner Z (2017) Leader-follower models in facility location. In: Mallozzi L, d'Amato E, Pardalos PM (eds) Spatial interaction models. Springer, Cham, pp 73–104

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

Drezner T, Drezner Z, Kalczynski P (2015) A leader-follower model for discrete competitive facility location. Comput Oper Res 64:51–59

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. Chapter 7. In: Eiselt HA, Marianov V (eds) Foundations of location analysis. Springer, New York, pp 139–162
Eiselt HA, Bhadury J (1998) Reachability of locational Nash equilibria. OR-Spektrum 20:101–107
Eiselt HA, Laporte G (1991) Locational equilibrium of two facilities on a tree. Recherche Opérationnelle/Operations Research 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) In: Drezner Z (ed) Objectives in location problems. Pp. 151–180 in facility location: a survey of applications and methods. Springer, New York
Eiselt HA, Marianov V (2017) Asymmetries in competitive location models on the line. In: Mallozzi L, D'Amato E, Pardalos PM (eds) Spatial interaction models: facility location using game theory. Springer, Cham, pp 105–128
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
Fernández J, Salhi S, Tóth BG (2014) Location equilibria for a continuous competitive facility location problem under delivered pricing. Comput Oper Res 41:185–195
Fernández P, Pelegrín B, Lančinskas A, Žilinskas J (2017a) New heuristic algorithms for discrete competitive location problems with binary and partially binary customer behavior. Comput Oper Res 79:12–18
Fernández J, Tóth GB, Redondo JL, Ortigosa PM, Arrondo AG (2017b) A planar single-facility competitive location and design problem under the multi-deterministic choice rule. Comput Oper Res 78:305–315
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
Gentile J, Pessoa A, Poss M, Roboredo MC (2018) Integer programming formulations for three sequential discrete competitive location problems with foresight. Eur J Oper Res 265(3):872–881
Ghaffarinasab N, Motallebzadeh A, Jabarzadeh Y, Kara BY (2018) Efficient simulated annealing based solution approaches to the competitive single and multiple allocation hub location problems. Comput Oper Res 90:173–192
Ghosh A, Buchanan B (1988) Multiple outlets in a duopoly: a first entry paradox. Geogr 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 51(3):309–324
Guo W-C, Lai F-C (2017) Prices, locations and welfare when an online retailer competes with heterogeneous brick-and-mortar retailers. J Ind Econ 65(2):439–468
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. Pap 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
Hendrix EMT (2016) On competition in a Stackelberg location-design model with deterministic supplier choice. Ann Oper Res 246:19–30
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
Karakitsiou A, Migdalas A (2017) Nash type games in competitive facilities location. Int J Decis Support Syst 2(1–3):4–12
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
Kononov AV, Panin AA, Plyasunov AV (2018) A new model of competitive location and pricing with the uniform split of the demand. In: Eremeev A, Khachay M, Kochetov Y, Pardalos P (eds) Optimization problems and their applications. Communications in Computer and Information Science, vol 871. Springer, Cham, pp 16–28
Kress D, Pesch E (2012) Sequential competitive location on networks. Eur J Oper Res 217:483–499
Kress D, Pesch E (2016) Competitive location and pricing on networks with random utilities. Netw Spat Econ 16(3):837–863
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
Laporte G, Riera-Ledesma J, Salazar-González JJ (2003) A branch-and-cut algorithm for the undirected traveling purchaser problem. Oper Res 51(6):940–951
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
Levanova T, Gnusarev A (2018) Ant colony optimization for competitive facility location problem with elastic demand. In: Belim S et al. (eds.) OPTA-SCL 2018, Omsk. Available online at http://ceur-ws.org/Vol-2098/paper21.pdf. Last Accessed 1 Aug 2019
Liou JJH (2009) A novel decision rules approach for customer relationship management of the airline market. Expert Syst Appl 36:4374–4381
Lösch A (1954) The economics of location, 2nd rev. edn. Yale University Press, New Haven
Mahmutogullari AI, Kara BY (2015) Hub location under competition. Eur J Oper Res 250(1):214–225
Marianov V, Eiselt HA (2016) On agglomeration in competitive location models. Ann Oper Res 246:31–55

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
Marianov V, Eiselt HA, Lüer-Villagra A (2018) Effects of multipurpose shopping trips on retail store location. Eur J Oper Res 269(2):782–792
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 Algebraic Discret Methods 4:253–161
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 Mark 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
Niknamfar AH, Niaki STA, Niaki SAA (2017) Opposition-based learning for competitive hub location: a bi-objective biogeography-based optimization algorithm. Knowl-Based Syst 128: 1–19
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 Plan A19:1067–1082
Osborne MJ, Pitchik C (1986) The nature of equilibrium in a location model. Int Econ Rev 27(1):223–237
Osborne MJ, Pitchik C (1987) Equilibrium in Hotelling's model of spatial competition. Econometrica 55(4):911–922
Panin AA, Pashchekno MG, Plyasunov AV (2014) Bilevel competitive facility location and pricing problems. Autom Remote Control 75(4):715–727
Pelegrín B, Fernández P, Suárez R, García MD (2006) Single facility location on a network under mill and delivered pricing. IMA J Manag Math 17(4):373–385
Pelegrín B, Fernández P, García Pérez MD (2015) On tie breaking in competitive location under binary customer behavior. Omega 52:156–167
Penn M, Kariv O (1989) Competitive location in trees: parts I and II. In: Working paper 89-MSC-029. Faculty of Commerce and Business Administration, University of British Columbia, Vancouver
Plastria F (1992) On destination optimality in asymmetric distance Fermat-weber problems. Ann Oper Res 40:369–355
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
Qi M, Xia M, Zhang Y, Miao L (2017) Competitive facility location problem with foresight considering service distance limitations. Comput Ind Eng 112:483–491

Rahmani A (2016) Competitive facility location problem with attractiveness adjustment of the follower on the closed supply chain. Cogent Math 3:1–19. Available at https://www.cogentoa.com/article/10.1080/23311835.2016.1189375.pdf. Last accessed on 1/8/2019
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
Rohaninejad M, Navidi H, Nouri BV, Kamranrad R (2017) A new approach to cooperative competition in facility location problems: mathematical formulations and an approximation algorithm. Comput Oper Res 83:45–53
Rothschild R (1979) The effect of sequential entry on choice of location. Eur Econ Rev 12:227–241
Ruiz-Hernández D, Elizalde J, Delgado-Gómez D (2017) Cournot-Stackelberg games in competitive delocation. Ann Oper Res 256(1):149–170
Sadjadi SJ, Ashtiani MG, Ramezanian R, Makui A (2016) A firefly algorithm for solving competitive location-design problem: a case study. J Ind Eng Int 12:517–527
Santos-Peñate DR, Campos-Rodríguez C, Moreno-Pérez JA (2017) A Matheuristic to solve a competitive location problem. In: Quesada-Arencibia A, Rodríguez-Rodríguez JC, Moreno-Díaz R, Moreno-Díaz R Jr (eds) Computer-aided systems theory. Ramon Llull's Ars magna. Springer, Cham, pp 80–81
Sasaki M, Campbell JF, Krishnamoorthy M, Ernst AT (2014) A Stackelberg hub arc location model for a competitive environment. Comput Oper Res 47:27–41
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, Eiselt HA, Laporte G, ReVelle CS (1999a) Market capture models under various customer choice rules. Environ Plan B Plan Design 26:741–750
Serra D, ReVelle C, Rosing K (1999b) Surviving in a competitive spatial market: the threshold capture model. J Reg Sci 39(4):637–652
Seyhan TH, Snyder LV, Zhang Y (2018) A new heuristic formulation for a competitive maximal covering location problem. Transp Sci 52(5):1035–1296
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
Shan W, Yan Q, Chen C, Zhang M, Yao B, Fu X (2017) Optimization of competitive facility location for chain stores. Ann Oper Res 273(1):187–205. see, e.g., https://link.springer.com/article/10.1007/s10479-017-2579-z. Last Retrieved 01 Aug 2019
Shigehiro S, Shiode S, Hiroaki I, Teraoka Y (1995) A competitive facility location problem. In: Fushimi M, Tone K (eds) Proceedings of APORS '94. World Scientific, 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. Expert Syst Appl 21(3):157–168
Spoerhase J, Wirth H-C (2008) (r|p) centroid 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
Tuan L-A (2017) A data-driven approach for the maximum capture problem in competitive facility location. Minh Hoang Ha (supervisor). Available at http://ds.libol.fpt.edu.vn/bitstream/123456789/2423/1/TuanLA_Thesis.pdf. Last Retrieved 01 Sept 2019
von Stackelberg H (1943) Grundlagen der theoretischen Volkswirtschaftslehre (translated as: The theory of the market economy). W. Hodge & Co. Ltd., London, p 1952
Xue Z, Gao Y, Yin J (2017) Competitive facility location problem with exotic products. In: 4th international conference on industrial engineering and applications ICIEA. IEEE Press, Nagoya, pp 173–177
Younies H, Eiselt HA (2011) Sequential location models. Chapter 8. In: Eiselt HA, Marianov V (eds) Foundations of location analysis. Springer, Berlin
Zhang ZJ (1995) Price-matching policy and the principle of minimum differentiation. J Ind Econ 43:287–299
Zhang Y, Snyder LV, Ralphs TK, Xue Z (2016) The competitive facility location problem under disruption risks. Transp Res Part E Logist Transp Rev 93:453–473

Chapter 15
Location-Routing and Location-Arc Routing

Maria Albareda-Sambola and Jessica Rodríguez-Pereira

Abstract This chapter overviews the most relevant contributions on location-routing problems. Although there exist several 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.

15.1 Introduction

Combined location-routing problems (LRPs) are location problems in which 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

M. Albareda-Sambola (✉)
Universitat Politècnica de Catalunya, BarcelonaTech, Barcelona, Spain
e-mail: maria.albareda@upc.edu

J. Rodríguez-Pereira
HEC Montréal, Montréal, QC, Canada
e-mail: Jessica.Rodriguez@cirrelt.ca

G. Laporte et al. (eds.), *Location Science*,
https://doi.org/10.1007/978-3-030-32177-2_15

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 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, 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 available works 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).

In the last years, several LRP extensions and variants have been considered. To mention just a few some, authors have considered problems with time windows (Farham et al 2018), heterogeneous fleets (Koç et al 2016), uncertain data (Caunhye et al 2016) or environmental effects (Koç et al 2016b). Other works concerning LRP extensions are surveyed in Drexl and Schneider (2015).

Given the little attention that LARPs have received, this chapter focuses 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 Sects. 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.

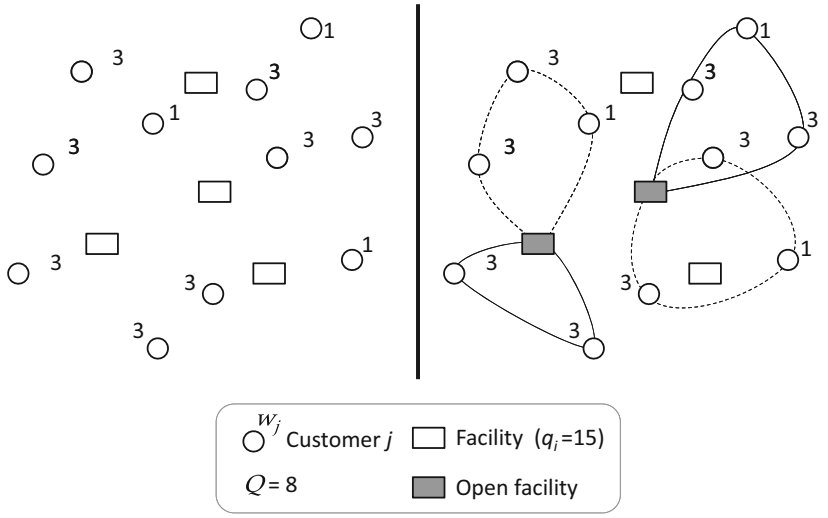


Fig. 15.1 Example of an LRP solution

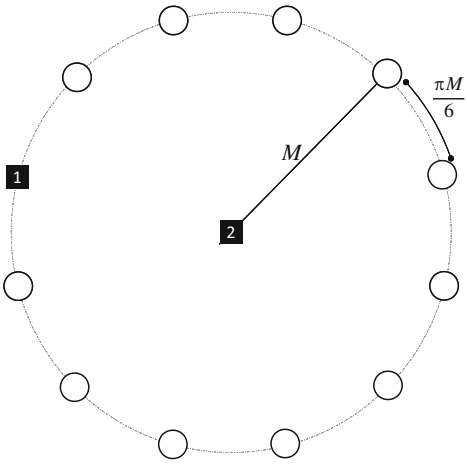


Fig. 15.2 Influence of facility location on the routing costs

Further considerations and characteristics of the main elements of the problem (number of facilities to locate, types of customers, size and characteristics of the 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 $2M + (10\pi M)/6 \simeq 7.24M$. 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.

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, especially if they are tightly constrained. 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:

$$\text{(LRP1)} \quad \text{minimize} \sum_{i \in I} f_i y_i + \sum_{k \in K} \sum_{(i,j) \in E} \ell_{ij} z_{ijk} \tag{15.1}$$

$$\text{subject to} \sum_{k \in K} \sum_{i \in V} z_{ijk} = 1 \qquad j \in J \tag{15.2}$$

$$\sum_{j \in J} w_j \sum_{i \in V} z_{ijk} \leq Q \qquad k \in K \tag{15.3}$$

$$\sum_{j \in J} w_j x_{ij} - q_i y_i \leq 0 \qquad i \in I \tag{15.4}$$

$$\sum_{k \in K} \sum_{i \in S} \sum_{j \in V \setminus S} z_{ijk} \geq 1 \qquad I \subseteq S \subset V \tag{15.5}$$

$$\sum_{j \in V} z_{ijk} - \sum_{j \in V} z_{jik} = 0 \qquad k \in K, i \in V \tag{15.6}$$

$$\sum_{i \in I} \sum_{j \in J} x_{ijk} \leq 1 \qquad k \in K \tag{15.7}$$

$$\sum_{t \in J} z_{itk} + \sum_{t \in V} z_{jtk} - x_{ij} \leq 1 \qquad i \in I, j \in J, k \in K \tag{15.8}$$

$$y_i \in \{0, 1\} \qquad i \in I \tag{15.9}$$

$$x_{ij} \in \{0, 1\} \qquad i \in I, j \in J \tag{15.10}$$

$$z_{ijk} \in \{0, 1\} \qquad (i, j) \in E, k \in K. \tag{15.11}$$

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 because they have an exponential number of connectivity constraints and because they contain $O(|V|^3)$ 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). Additionally, 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.

$$\text{(LRP3) minimize} \sum_{i\in I} f_i y_i + \sum_{r\in\Gamma} \ell_r \lambda_r + \sum_{\{i,j\}\in E_{IJ}} 2\ell_{ij} z^2_{ij} \tag{15.22}$$

$$\text{subject to} \sum_{r\in\Gamma}\sum_{i\in V} a_{ijr}\lambda_r + \sum_{i\in I} 2z^2_{ij} = 2 \qquad j \in J \tag{15.23}$$

$$\sum_{r\in\Gamma_i}\sum_{\{j,s\}\in E} (w_j + w_s) a_{jsr}\lambda_r + \sum_{j\in J} 2w_j z^2_{ij} \le 2q_i y_i \qquad i \in I \tag{15.24}$$

$$y_i \in \{0,1\} \qquad i \in I \tag{15.25}$$

$$z^2_{ij} \in \{0,1\} \qquad \{i,j\} \in E \tag{15.26}$$

$$\lambda_r \in \{0,1\} \qquad r \in \Gamma. \tag{15.27}$$

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) with their following stronger variant, the strengthened degree constraints:

$$\sum_{r\in\Gamma}\sum_{k:\{j,k\}\in E} a_{jkr}\lambda_r + \sum_{i\in I} z^2_{ij} \ge 1 \quad j \in J. \tag{15.28}$$

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.

y-Strengthened Capacity Cuts (y-SCC)
For $S \subset J$, and $r \in \Gamma$, 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).

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:

$$\sum_{i\in I\setminus H}\sum_{r\in\Gamma_i}\xi_{rS}\lambda_r+\sum_{i\in I\setminus H}\sum_{j\in S\setminus S'}z_{ij}^2\geq\kappa_1(S,H\setminus\{i\})+y_i\Big(\kappa_1(S,H)-\kappa_1(S,H\setminus\{i\})\Big).\tag{15.30}$$

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.

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 indeed 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 $S = \cup_{s=1}^{4} S_s$,

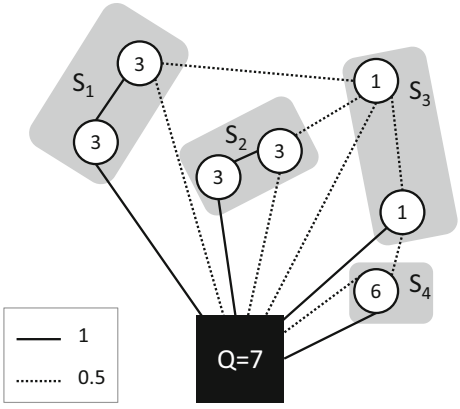


Fig. 15.3 Example of unsatisfied FrCI

$w(S) = 20$ and $Q = 7$, so that $\kappa_1(S) = 3$. Therefore, 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 $z-$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 Escobar et al (2014) 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 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) taking an average CPU time of about 4 min, although this average is about 10 min for the most demanding set of instances.

- *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; Escobar et al 2014; Ferreira and de Queiroz 2018). Actually, this is the usual category for the most recent works, since they tend to yield better results. Indeed, the results obtained in Ferreira and de Queiroz (2018) are superior to those of previous heuristics in terms of solution quality, although at a high computational cost, whereas Escobar et al (2014) provides an excellent tradeoff between solution quality and computing time.

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 was proposed before the recent work Rodríguez-Pereira (2017), and only heuristic algorithms for different variants could 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.

Rodríguez-Pereira (2017), after studying the multi-depot rural postman problem as an intermediate step, propose two alternative formulations for different uncapacitated LARPs which are solved through branch-and-cut algorithms. The first one is a natural formulation with flow variables with three-indices. These indices associate each variable with an edge and the facility where the route traversing it starts. The second one uses twice-indexed variables; now, indices are associated with edges but not with facilities, which requires a new set of constraints to guarantee that the routes are consistent and return to the original depot. The second formulation allowed to solve instances with almost 200 nodes, over 300 edges and between 100 and about 200 required edges in small CPU times. These results can be found in Fernández et al. (2019).

Bearing in mind the evolution of the formulations for the capacitated arc routing problem (CARP), one might expect set partitioning formulations to yield 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 on 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 developed and compared several heuristics combining tabu search, variable neighborhood search, and GRASP for which they also tested several 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 1 min, 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 multiperiod LARP extension where inventories are considered is addressed in Riquelme-Rodríguez et al (2016). The work is motivated by a road watering application in open-pit mines and inventories are used to model road dust retention. Depots are located to provide service for the whose time horizon, whereas different routes must be designed for the different time periods.

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 rise 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

al Ajdad SMHM, Torabi SA, Salhi S (2012) A hierarchical algorithm for the planar single-facility location routing problem. Eur J Oper Res 39:461–470

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. Transport 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

Caunhye A, Zhang Y, Li M, Nie X (2016) A location-routing model for prepositioning and distributing emergency supplies. Transport Res E-Log 90:161–176

Contardo C, Cordeau J-F, Gendron B (2013) A computational comparison of flow formulations for the capacitated location-routing problem. Discrete 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

Drexl M, Schneider M (2015) A survey of variants and extensions of the location-routing problem. Eur J Oper Res 241:283–308

Escobar JW, Linfati R, Toth P (2013) A two-phase hybrid heuristic algorithm for the capacitated location-routing problem. Comput Oper Res 40:70–79

Escobar JW, Linfati R, Baldoquin MG, Toth P (2014) A granular variable tabu neighborhood search for the capacitated location-routing problem. Transp Res B Methodol 67:344–356

Farham MS, Süral H, Iyigun C (2018) A column generation approach for the location-routing problem with time windows. Comput Oper Res 90:249–263
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
Fernández E, Laporte G, Rodríguez-Pereira J (2019) Exact solution of several families of location-arc routing problems. Transp Sci 53:1213–1499
Ferreira K, de Queiroz T (2018) Two effective simulated annealing algorithms for the location-routing problem. Appl Soft Comput 70:389–422
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. Transport 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
Koç Ç, Bektaş T, Jabali O, Laporte G (2016) The fleet size and mix location-routing problem with time windows: Formulations and a heuristic algorithm. Eur J Oper Res 248:33–51
Koç Ç, Bektaş T, Jabali O, Laporte G (2016) The impact of depot location, fleet composition and routing on emissions in city logistics. Transp Res B Methodol 84:81–102
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
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 Methodol 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. Transport 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
Riquelme-Rodríguez JP, Gamache M, Langevin A (2016) Location arc routing problem with inventory constraints. Comput Oper Res 76:84–94
Rodríguez-Pereira J (2017) New models and algorithms for several families of arc routing problems. Ph.D. Thesis, Statistics and Operations Research Department. Technical University of Catalunya, http://hdl.handle.net/10803/461412

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, Rand GK (1989) Effect of ignoring routes when locating depots. Eur J Oper Res 39:150–156

Salhi S, Nagy G (2009) Local improvement in planar facility location using vehicle routing. Ann Oper Res 167:287–296

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

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 Logistics in Supply Chain Management

Iris Heckmann and Stefan Nickel

Abstract Location decisions play a key role in strategic logistics and supply chain management. In this chapter, we place the emphasis on the interaction of logistics activities and long-term supply chain decision-making on location logistics models. We cover modeling formulations of logistics core activities related to different industrial supply chain settings. In particular, we relate current challenges in supply chain management and their implications on relevant logistics activities. Finally, new research directions and areas of interest are provided.

16.1 Introduction

Since the 1960s many models developed in the context of location theory incorporate logistics aspects. For this reason they are also applicable for logistics and supply chain problems (see for example Melo et al. 2008). However, these inclusions have not always been systematic. In this chapter, we approach location decisions by starting from a logistics point of view and problem description. In particular we discuss logistics settings and their suitability for location models.

It is worth-noting that the terminology and the definitions in logistics are not as consistent and unified as in operations research. Many terms are used in practice before they are introduced into the academic literature. Therefore, we sometimes give our own or refine existing definitions. Whenever a specific reference is useful, we provide it. Nevertheless, we can directly list some sources where definitions and terms in logistics can be found: CSCMP (2013), Zijm et al. (2019) and Web Finance Inc. (2019).

I. Heckmann (✉)
Camelot ITLab GmbH, Köln, Germany
e-mail: ihec@camelot-itlab.com

S. Nickel
KIT Karlsruhe Institute of Technology, Karlsruhe, Germany
e-mail: stefan.nickel@kit.edu

G. Laporte et al. (eds.), *Location Science*,
https://doi.org/10.1007/978-3-030-32177-2_16

Location decisions in an industrial context imply the opening, the closing or the positioning of facilities. While the first and the second type of decisions focus on whether or not to open or close facilities such as production sites, distribution centers, or warehouses, positioning decisions refer to the location of suppliers, customers or facilities of similar or successive functions among each other. Those decisions have to be made whenever companies need to expand their capacities because they enter new markets or grow into new product segments. The ultimate reason for making these decisions, however, arises from the fact that facilities are not autonomous entities, but they have to interact with each other as well as with their environment. Due to this interaction, facility location problems are often cast as network design problems.

The activities that take place within a set of facilities include, for example, the shipment of raw material or finished goods from suppliers to production sites or from production sites to storage facilities or end-customers. The manufacturing or production, the storage and the handling of raw material and finished goods, take place within one facility. Nevertheless, they have to be coordinated among several locations. Generally, these activities are referred to as *logistics* and more precisely described as procurement and distribution, production or manufacturing, transportation, storage and handling, respectively (CSCMP 2013; Zijm et al. 2019; Web Finance Inc. 2019). Logistics activities that take place at a single location such as materials handling, forklift transportation and inventory management are referred to as *site logistics* or *on-site-logistics* (Logistik-Lexikon 2019). We define logistics activities that interact with other locations or that have to be coordinated among several locations as *location logistics*.

Facility location and allocation represent a core link between supply chain and logistics management. In the supply chain management literature it is also often referred to as supply chain network design. When considering a single location instead of a set of interacting locations that have to be coordinated, location selection is often referred to as *plant location*.

In order to leverage the efficiency of the resulting set of facilities, e.g. respect capacities, costs and availabilities, activities are subject to an overall logistics management, which is part of modern supply chain management.

We follow the Council of Supply Chain Management Professionals (CSCMP 2013) that defines Logistics Management as

> that part of supply chain management that plans, implements, and controls the efficient, effective forward and reverses flow and storage of goods, services and related information between the point of origin and the point of consumption in order to meet customers' requirements.

For a definition of supply chain management we refer to CSCMP (2013) and for an in-depth discussion on the topic we refer to the review papers by Lummus and Vokurka (1999) and Mentzer et al. (2001). It is important to note that supply chain management differs from logistics management by important aspects: In addition to the planning and management of logistics activities supply chain management includes coordination and collaboration of business partners as well as integration of major business functions and business processes.

Table 16.1 Common terms used in supply chain management

Geographic	Granularity	Modeling	Management
Site	Site logistics	Plant location	Site management
(Supply chain) Locations	Location logistics	Facility location and allocation/supply chain network design	Logistics management
Supply chain	Supply chain logistics		Supply chain management

Due to the increased complexity of today's businesses supply chains should be called supply networks. For the remainder of this chapter we use the term supply chain and supply network interchangeably.

The terms used to refer to geographical entities, type of logistics granularities, strategic location selection modeling frameworks, and management paradigms are summarized in Table 16.1.

In this chapter, we discuss the interaction of logistics activities and challenges for supply chain management as well as the consequences when building a facility location model. The focus is on modeling aspects rather than on solution methods. Therefore we only consider literature relevant for such aspects.

The remainder of the chapter is organized as follows. Section 16.2 introduces logistics activities and their inclusion in location models. Section 16.3 provides a first integrated location model capturing relevant logistics aspects. In Sect. 16.4, some challenges of modern supply chain management are discussed and a mapping between each such challenges and the corresponding logistics activity is presented. Section 16.5 discusses extensions of the first integrated location model with respect to logistics activities and relevant challenges for supply chain management. Finally, in Sect. 16.6 further research directions are discussed.

16.2 From Logistics to Location Models

An adequate model for a facility location problem emerging in the context of logistics systems calls for a clear understanding of the fact that logistics activities and processes affect location decisions. Consequently, we must answer to some major questions prior to modeling and analyzing a problem, namely:

- Which logistics activities are to be considered?
- Which logistics activities must be integrated in a model?
- Which modeling paradigm is the most adequate given the nature of the underlying data?

We start this section by briefly discussing the aforementioned questions. Next, we present logistics elements for a facility location model in the context of supply chain management. We offer models and discuss the importance of each element.The last Paragraph is dedicated to the presentation of a first integrated location logistics model.

16.2.1 Why Logistics Matters in Location Modeling

Historically, researchers have focused relatively early on the design of distribution systems (Geoffrion and Powers 1995), but missed to consider logistics processes as interacting functions over the whole supply chain (Melo et al. 2009) as well as to analyze the importance of logistics activities for location models.

Somehow, it seems to be an unwritten rule that strategic decision making only considers those activities and processes that are either associated with high investments or not flexible enough to change when new circumstances demand for modifications. In the context of logistics, Daskin et al. (2005) among others, discussed how decisions on transportation and inventory may change within a short- to mid-term time frame, when relevant characteristics of the underlying supply chain indicate the necessity of such modifications. Production quantities can be modified in a mid-term time horizon, when material shortages or customers demands make it necessary. However, decisions on production capacities are typically fixed for longer time periods and they are less flexible. Consequently, they are considered in strategic decision making. The investments associated with the installation of new production plants are usually high compared to those of transportation or inventory. It seems natural, though, that investments on production facilities are included in strategic location models. In fact, the relocation of a production plant due to changes in customer demands, transportation costs, or component prices is hardly acceptable (Daskin et al. 2005). Moreover, the relocation of production facilities is often expensive and nearly impossible except in the long-term. Finally, modern distribution centers containing highly technologized–thus expensive–material handling equipment or transportation hubs such as airports are difficult or even impossible to relocate (Daskin et al. 2005). General aspects of logistics planning with time dependent decisions are discussed in Dunke et al. (2018).

The main conclusion to be derived from this discussion is that making location decisions ignoring primary logistics activities like production or distribution, may result in excessive costs incurred throughout the lifetime of the facilities supporting the logistics system. Inefficiencies and excessive costs, however, may be a consequence of other aspects. For instance, transportation costs may raise or labor costs may evolve differently from what was expected. Additionally, inventory holding costs may increase due to unexpected changes of interest or exchange rates. Overall, a logistics planning ignoring relevant logistics activities may lead to bad location decisions. In fact, apart from production, the location decisions made for a logistics network carry out all logistics activities in one way or another. Decisive for facility location modeling, however, is the way logistics activities are taken into account.

The logistics tasks of a facility in a supply chain can be manifold. It can be a raw material plant, a production plant, a warehouse, a transshipment center, a hub, or even a retailer. Despite its major logistics function each location often fulfills a number of additional logistics activities, which need to be respected and sometimes integrated in location models (Cordeau et al. 2006). Before formulating a location decision model, it is necessary to analyze the industrial setting in which

the underlying supply chain is or will be operated as well as the business objectives the supply chain is exposed to. Sometimes it is not necessary to integrate all existing logistics activities—at least not in every detail.

Consider as an example a set of production sites engaged in the chemical industry. In this case, raw material and finished products are often stored in silos, whose capacities can vary over time since whenever a silo becomes empty it can be used for another product. However, when a silo is not empty it can only be used for the product it is already filled with. It is very complex to model this type of inventory management. Nevertheless, this may not be relevant if decision makers conclude, that silo capacities are not determinant for an opening, closing or positioning decision. Silos may be assumed to be at any production site with the necessary capacity. In other words, a decision maker might decide to leave the inventory management aspects out of the location model.

This motivates another important aspect when modeling location logistics for supply chain management: the appropriate way for modeling logistics activities. Facilities as elements of the supply chain are often globally dispersed with separated data bases and different logistics operation modes. This complicates the availability, accuracy, and thus the reliability of information and data needed for building a facility location model. Additionally, globally spread facilities are exposed to numerous environmental, cultural and infrastructural uncertainties that provoke changes in information that often is assumed to be deterministic. In order to avoid that a set of efficient sites suddenly becomes inefficient, uncertainty influencing logistics activities should be taken into account in advance. The nature and type of data uncertainty is however in itself uncertain and decisively affects the modeling paradigm that should be considered. Uncertainty in data can be tackled using different tools such as stochastic programming, chance-constrained programming, or robust optimization (see Chap. 8). The paradigm to consider strongly depends on the nature of the uncertainty.

16.2.2 Building Blocks of Logistics

From the discussion presented so far we conclude that the traditional hierarchical planning sequence starting with the strategic decisions, then tactical and finally operational may lead to low quality, conflicting or even infeasible decisions. The challenge lies in the integration of the three planning levels in order to find feasible and good decisions for logistics execution.

Integrated facility location problems may turn into large-scale complex optimization problems that call for sophisticated solution methods. In the light of location problems, a common approach to overcome such difficulties is to split larger problems into smaller sub-problems (Stadtler 2008). Unfortunately, such approaches may lead to sub-optimal solutions. However, while technology is further developing and new solution techniques for nonlinear and large-scale linear math-

ematical models evolve, increasingly larger integrated planning problems become more tractable (Zanjirani Farahani et al. 2015).

Next we take a close look on prominent logistics activities, namely: procurement (or inbound logistics), production (or assembly or manufacturing), inventory (and handling), routing and distribution (or transport), as well as layout.

In what follows we assume that we have a finite planning horizon divided into several time periods. Additionally, we consider a set of customers whose demand (known for all periods) is to be supplied throughout the planning horizon. We consider the possibility of having a service level below 100%. This may be due to high costs associated to some demand satisfaction, shortage of production capacity or service times impossible to fulfill. In an optimization model, unsatisfied demand is often accounted for by introducing a penalty in the objective function. Finally, we note the multi-commodity nature of many logistics systems. Hence, production capacity and resource availability must be balanced across the different products or commodities.

16.2.2.1 Procurement

Procurement or inbound logistics, is an activity that focuses on the acquisition of goods needed for production, assembly or manufacturing. Typically the amount acquired from suppliers as well as related variable costs and fixed costs describe the procurement activity. Nevertheless, before procurement activities can even begin, strategic decisions such as the selection of suppliers based on their solvency as well as quality and availability of goods have to be made.

Supplier selection represents often, by itself, a decision to make. However, some qualitative aspects should also be integrated in models tailored for location logistics in supply chain management. Solvency and product quality can be integrated by including supplier-dependent penalty costs or reward terms in the objective function. Nevertheless, the availability of goods is often captured via a capacity constraint that limits the amount that a location can purchase from a specific supplier.

From a supply chain management perspective, the type of product and the company-specific logistics requirements are important aspects to analyze up-front, because they can have an impact when modeling the aforementioned aspects. For instance, if the products involved in a supply chain require a sparse bill-of-materials (BOM), or if only a few suppliers exist, it becomes relevant to consider supplier shortages in a model. Accordingly, attention should be given to product type, technology knowledge, available capacity, initial investment required, and specific logistics requirements before integrating procurement relevant formulations in the location model (Simchi-Levi et al. 2007).

Although capturing procurement is recognized as a vital element in supply chain management (Kraljic 1983), it is rarely present in the facility location literature (see Melo et al. 2009; Zijm et al. 2019).

16.2.2.2 Production

Production activities transform one or several materials or components into one or several products. They include the production from raw materials as well as the assembly of several products to one final product. Note that we consider production as part of logistics without the special aspects of production technology. Similarly to procurement, production activities are described by an amount produced as well as related variable and fixed costs. Often specific limits for the production capacities are given. In addition, the production process itself can be further described by consumption factors provided by the BOM. They represent the amount of materials needed for the production of one unit of a product. Resource capacities such as those induced by production lines in a discrete production setting or capacities of converters in a continuous setting and occasionally surplus capacity provide a more detailed description of the necessary production infrastructure. Typically, a capacity constraint has to be considered limiting the production. For further reading we refer to Esmaeilian et al. (2016).

16.2.2.3 Inventory

The main functionality of storing materials, components, semi-finished or finished products is the decoupling of precedent or successive logistics activities such as sourcing, production and distribution facilitating the planning of such activities. During the decoupling period, material, goods and products have to be stored at production sites, warehouses, or distribution centers resulting in inventory costs. The consumption of stored products is generally formulated as inventory balancing constraints. In the light of industrial (and even civil or public) supply chain management, inventory models have to include decisions on safety stocks, re-order points, turnovers, and service levels. A relevant issue when developing a model for supporting decision making, is to describe centralized and decentralized inventory systems, to capture lead times or safety stocks, and to integrate multi-layer supply chains in a multi-period setting.

For a deeper discussion of logistics activities related to inventory as well as model formulations tailored for location-inventory problems, we refer to Melo et al. (2009) and Zanjirani Farahani et al. (2015).

16.2.2.4 Routing and Distribution

Routing and distribution—transportation in general—can take place between all entities within a supply chain. Material and products are transported from one location to another in distinct time periods and at certain costs. Besides distance, the level of transportation costs depend on the type of product and on the transportation mode. In the facility location literature, most often trucks or airplanes are considered. In the particular case of road transportation, two possibilities exist: full-

truck-load (FTL) and less-than-full-truck-load (LTL). Decision-makers often favor FTL. However, when delivery becomes urgent (production may stop or customer service level is at risk) LTL may be necessary. At an operational level (e.g. short-term decisions) discounts for larger volumes play an important role. In this case, the cost curve is often non-linear (concave). However, in a strategic setting, a linear approximation is in most cases sufficient. The shipping from and the entrance of transported products at a facility is generally formulated using balancing constraints.

While transportation is a concept describing the movement of goods in general, distribution refers to the allocation of material or goods to the end user of material or goods and routing refers to the determination of the optimal path to serve a group of customers. Routing and distribution decisions have been extensively discussed in location theory because they integrate two major decisions: location and routing. For more details we refer to Nagy and Salhi (2007) as well as to Chap. 15.

16.3 A Basic Integrated Logistics Location Model

Following the aforementioned logistics activities we introduce a basic integrated logistics location model, $BILL$, as a mixed-integer linear program. The $BILL$ model considers capacities of different logistics activities as well as multiple products. It assumes that there is an underlying planning horizon divided into several time periods. Additionally, several general non-hierarchical levels are considered. The model includes decisions about location, procurement and production, inventory and distribution as well as customer demand fulfillment. It takes into account costs for procurement and production, inventory (stock-level and stock-turnover), installation and closing of facilities, transportation and non-fulfillment of customer demand. The overall objective of the model is to minimize the total cost. All entities of a supply chain—whether they belong to the same organization or not—can be divided into so-called *selectable* and *non-selectable* facilities (see e.g. Melo et al. 2006). Selectable facilities are those that may have their status changed. Non-selectable facilities cannot have their status changed.

The mathematical formulation is presented in Sect. 16.3.2 and captures the aforementioned features. The required notation is first introduced in Sect. 16.3.1.

16.3.1 Notation

Table 16.2 presents the sets used in the $BILL$ model.

Table 16.3 introduces the parameters related to both tactical logistics activities and strategic location decisions. Besides the demand requirements, we need input for capacity resources. Each product consumes a certain share of the overall resource capacities. Similarly, handling capacities are taken into account. we assume that resource capacities can be expanded at additional costs. Extra handling capacity

Table 16.2 Sets used for the $BILL$ model

Set	Description
L	Locations
S	Selectable locations
S^o	Selectable facilities that can be opened
S^c	Existing selectable facilities that can be closed
T	Time periods
P	Products
R^p	Production resources
R^h	Handling resources

Table 16.3 General parameters for the $BILL$ model

Symbol	Description
d_{ipt}	Demand of location $i \in L$ for product $p \in P$ in period $t \in T$
a_{iqp}	Number of units of product $q \in P$ required to produce one unit of product $p \in P$ $(q \neq p)$ at facility $i \in L$
μ_{irp}	Amount of resource $r \in R^p$ required to produce one unit of product $p \in P$ in facility $i \in L$
$\lambda^{in}_{irp}, \lambda^{out}_{irp}$	Amount of resource $r \in R^h$ required to handle one unit of product $p \in P$ upon its entrance at and its shipment from facility $i \in L$, respectively
K_{rt}	Initial capacity of resource $r \in R^p \cup R^h$ in period $t \in T$
K^e_{rt}	Maximum capacity expansion of resource $r \in R^p \cup R^h$ in period $t \in T$

can be made available through overtime work or outsourcing (e.g. via external service providers). Additional storage or production capacities can be acquired by purchasing or leasing additional space or production lines.

There are three different ways of modeling the relationship between facilities and resources. In a *one-to-many* relationship, the same resource is used at several facilities. This is the case, for instance when a production manager is responsible for several production lines in different facilities. A *one-to-one* relationship represents the situation where the same resource is used by all the products of a facility. Typical examples include a foiling machine or a storage place. In a *many-to-one* relationship, several resources are used at the same facility. A set of resources can be product-specific and a different set of resources can be used for multiple products. The former is the case, for instance when a machine is dedicated to a particular product; the latter refers for example to a production manager or a picking system.

In Table 16.4 cost parameters are introduced. Finally, Table 16.5 presents the decision variables of the problem.

16.3.2 The BILL Model

The objective function to be minimized includes the total cost for procurement and production, distribution, inventory, penalty for unsatisfied demand, opening for new

Table 16.4 Cost parameters for the $BILL$ model

Symbol	Description
OC_{it}	Fixed cost for opening a facility in location $i \in S^o$ at the beginning of period $t \in T$. This parameter includes the operation costs until the end of the planning horizon
CC_{it}	Fixed costs for closing a facility in location $i \in S^c$ at the end of period $t \in T$. This parameter includes the operation costs until the end of t
XC_{ipt}	Unit penalty cost for not serving demand of facility $i \in L$ for product $p \in P$ in period $t \in T$
BC_{ipt}	Unit cost for buying/procuring product $p \in P$ at facility $i \in L$ from an external source in period $t \in T$
PC_{ipt}	Unit cost for producing product $p \in P$ at facility $i \in L$ in period $t \in T$
HC_{ipt}	Unit cost for holding/storing product $p \in P$ at facility $i \in L$ in period $t \in T$
TC_{ijpt}	Unit cost for shipping product $p \in P$ from facility $i \in L$ to facility $j \in L$ in period $t \in T$
EC_{rt}	Unit cost of expanding resource $r \in R^p$ or handling resource $r \in R^p \cup R^h$ in period $t \in T$

Table 16.5 Decision variables for the $BILL$ model

Symbol	Description
y_{it}	Binary variable equal to 1 if facility $i \in S^o$ is opened at the beginning of period $t \in T$ and 0 otherwise
y_{it}	Binary variable equal to 1 if facility $i \in S^c$ is closed at the end of period $t \in T \setminus \|T\|$ and 0 otherwise
$y_{i\|T\|}$	Binary variable equal to 1 if facility $i \in S^c$ is kept open during the entire planning horizon, 0 otherwise
φ_{ipt}	Quantity of unsatisfied demand of location $i \in L$ for product $p \in P$ in period $t \in T$
b_{ipt}	Quantity of product $p \in P$ procured from facility $i \in L$ from an external source in period $t \in T$
X_{ipt}	Quantity of product $p \in P$ produced at facility $i \in L$ in period $t \in T$
h_{ipt}	Quantity of product $p \in P$ stored at facility $i \in L$ in period $t \in T$
x_{ijpt}	Quantity of product $p \in P$ shipped from facility $i \in L$ to facility $j \in L$ in period $t \in T$
w_{rt}	Extra capacity to acquire of production resource $r \in R^p$ or handling resource $r \in R^h$ in period $t \in T$

facilities and removal of existing ones.

$$\min \sum_{t\in T}\sum_{i\in L}\sum_{p\in P}(BC_{ipt}b_{ipt} + PC_{ipt}X_{ipt}) +$$

$$\sum_{t\in T}\sum_{i,j\in L, i\neq j}\sum_{p\in P} TC_{ijpt}x_{ijpt} +$$

$$\sum_{t\in T}\sum_{r\in R^h\cup R^p} EC_{rt}w_{rt} +$$

$$\sum_{t\in T}\sum_{i\in L}\sum_{p\in P} HC_{ipt}h_{ipt} + \tag{16.1}$$

$$\sum_{t\in T}\sum_{i\in L}\sum_{p\in P} XC_{ipt}\varphi_{ipt} +$$

$$\sum_{t\in T}\sum_{i\in S^o} OC_{it}y_{it} + \sum_{t\in T}\sum_{i\in S^c} CC_{it}y_{it}$$

The flow conservation constraints balance incoming amounts with the outgoing amounts of each logistics activity, production and procurement, transportation, inventory and demand. They can be written as follows:

$$b_{ipt} + \sum_{j\in L, i\neq j} x_{jipt} + X_{ipt} + h_{ipt-1} =$$

$$\sum_{j\in L, i\neq j} x_{ijpt} + \sum_{q\in P} a_{iqp}X_{iqt} + h_{ipt} + d_{ipt} - \varphi_{ipt} \quad i \in L, p \in P, t \in T \tag{16.2}$$

Capacity constraints are necessary for limiting the resources consumption of different logistics activities, namely for production and handling as well as their expansions. Mathematically we can write:

$$\sum_{i\in L}\sum_{p\in P} \mu_{irp}X_{ipt} \leq K_{rt} + w_{rt} \quad r \in R^p, t \in T \tag{16.3}$$

$$\sum_{p\in P}\left(\sum_{i,j\in L, i\neq j}(\lambda^{in}_{jrp} + \lambda^{out}_{jrp})x_{ijpt} + \sum_{i\in L}\lambda^{in}_{irp}b_{ipt}\right) \leq K_{rt} + w_{rt} \quad r \in R^h, t \in T \tag{16.4}$$

$$0 \leq w_{rt} \leq K^e_{rt} \quad r \in R^p \cup R^h, t \in T \tag{16.5}$$

The selectable facilities can have their status changed at most once during the planning horizon. Formally we have:

$$\sum_{t\in T} y_{it} \leq 1 \quad i \in S^o \tag{16.6}$$

$$\sum_{t\in T} y_{it} = 1 \quad i \in S^c \tag{16.7}$$

Furthermore, for $i \in S$ and $t \in T$ we define:

$$T^t_i = \begin{cases} \{1, \ldots, t\}, & \text{if } i \in S^o. \\ \{t, \ldots, T\}, & \text{if } i \in S^c. \end{cases} \tag{16.8}$$

This helps writing constraints ensuring that the logistics activities are limited by their capacities but in those facilities that are operating:

$$b_{ipt} \leq M \sum_{\tau \in T_i^t} y_{i\tau} \qquad i \in L, p \in P, t \in T \tag{16.9}$$

$$X_{ipt} \leq M \sum_{\tau \in T_i^t} y_{i\tau} \qquad i \in L, p \in P, t \in T \tag{16.10}$$

$$h_{ipt} \leq M \sum_{\tau \in T_i^t} y_{i\tau} \qquad i \in L, p \in P, t \in T \tag{16.11}$$

$$x_{ijpt} \leq M \sum_{\tau \in T_i^t} y_{i\tau} \qquad i, j \in L, p \in P, t \in T \tag{16.12}$$

$$x_{jipt} \leq M \sum_{\tau \in T_i^t} y_{i\tau} \qquad i \in L, j \in L \setminus \{S\}, p \in P, t \in T \tag{16.13}$$

The model is concluded by the domain constraints:

$$h_{ip0} = 0 \qquad i \in L, p \in P \tag{16.14}$$

$$b_{ipt}, h_{ipt}, h_{ipt} \geq 0 \qquad i \in L, p \in P, t \in T \tag{16.15}$$

$$0 \leq \varphi_{ipt} \leq d_{it} \qquad i \in L, p \in P, t \in T \tag{16.16}$$

$$x_{ijpt} \geq 0 \qquad i, j \in L, p \in P, t \in T \tag{16.17}$$

$$y_{it} \in \{0, 1\} \qquad i \in L, t \in T \tag{16.18}$$

Computationally, the above problem is NP-hard since it generalizes the capacitated plant location problem (see Chap. 4). Nevertheless, the existing literature shows that it can be tackled within an acceptable CPU time using a general purpose solver for small- and medium-sized instances. For larger instance we may have to resort to heuristic algorithms (see Melo et al. 2008, 2012, 2014).

16.4 Challenges in Industrial Logistics

The management of logistics activities operates in an environment that is usually set by corporate supply chain strategies. The latter follow business strategies that nowadays are influenced by upcoming new information and production technologies, new business opportunities, and new political as well as environmental changes. Consequently, supply chain management has become a major strategic issue for every company involved in the efficient processing of value creation—be it through products or services. Trends in the economy and society resulting from computerization, increased complexity and uncertainty of trade flows, increased competition. These facts together with the need for sustainable developments, has resulted in major big structural as well as organizational effects on supply chain designs (Eskandarpour et al. 2015).

It turns out that currently the major challenges in supply chain management are sustainability, uncertainty and the digital transformation of the supply chain (Garcia

and You 2015; Kache and Seuring 2017). The aim of this section is to discuss these major research streams. It is not the goal in this chapter to discuss in detail every obstacle that hinders efficient supply chain management in general and location logistics in specific.

16.4.1 Sustainability

One of the current trends and challenges in supply chain management is the design and operation of sustainable supply chains. In this context, three dimensions can be considered: economic aspects, environmental (green) performance and social responsibility (Eskandarpour et al. 2015). The increasing interest in sustainable development has pushed supply chains to be sustainable as well: Nowadays, they have to be socio-political aware, ecologically sensitive, and green.

Until some time ago, repair and container logistics stood in the foreground when it came to plan and manage a supply chain. More recently, reverse logistics and reusable logistics have started playing a greater role due to the increase in customer expectations.

We do not go further into that topic since there is a complete chapter in this book devoted to green logistics (see Chap. 20). Nevertheless, in the following sections we provide another model related to sustainability.

16.4.2 Uncertainty, Risk and Disaster Events

Decision-making in industrial supply chain management calls for information about future developments (e.g. demand and lead-time forecast, spot prices for transportation and inventory). A major concern for the achievement of any business goal, including that of a supply chain system or a logistics task, is the treatment of uncertainty. Usually a decision maker has a certain amount of information about future developments. Customers demands for example most often slightly deviate from the initial outlook. In an industrial context, modern supply chains have evolved into transnational systems and since then they are often caught in a crossfire of influences (e.g., political, environmental) that are hardly predictable. Additionally, in the presence of the continuously increasing fierce competition for customers and their profitable satisfaction, supply chain management needs to account for numerous optimization criteria and different information sources that are all subject to uncertainty. This evolution has led to a wider range of uncertainty to be dealt with. The lack of a good uncertainty management becomes visible when unexpected incidents interfere with the normal operation of the supply chain. For instance, natural disasters such as earthquakes, can destroy production facilities or roads, and impede the possibility to satisfy customer's needs as promised. Similarly, effects are triggered by socio-economic or socio-political turmoils. Unpredictable and slightly

aggravating deviations, e.g. lead time increase, exchange rate fluctuations or oil price variability, may also affect supply chain's goal achievement.

Unknown deviations, supply chain disruptions and disasters as well as the supply chain risk impede the availability of resources, the realization of the plan, and ultimately, the satisfaction of demand. For an in depth discussion of the different concepts we refer to Heckmann et al. (2015) and Heckmann (2016). In order to anticipate these perils, supply chain models need to be endowed with the information about uncertain developments. Different types of models, capturing both different types of decisions and uncertainty, exist (Melo et al. 2009).

The consideration of uncertainty, risk, or disasters that have the potential to impede a supply chain goal achievement is carried out within different research streams. One such stream emerges in the context of facility location and focuses on disaster prevention and management (see Chap. 22). For general uncertainty extensions the reader is referred to Chap. 8. Instead of going into detail concerning these extensions we concentrate in Sect. 16.5.2 on capturing and quantifying supply chain risk in facility location models.

16.4.3 Digital Supply Chain Transformation and Supply Chain Integration

Contemporary supply chains evolved into highly stretched and interdependent systems (Christopher 2016). The variety of products, suppliers and customers, who constantly emerge with new and demanding expectations, has increased tremendously. The possibility to integrate logistics as well as other supply chain related activities has reached its limits—as stated at the annual meeting of the World Economic Forum (WEF) by global chief executives WEF (2017). Influences of Industry 4.0 and IoT on supply chain planning are starting to be considered in scientific papers. See for example Manavalan and Jayakrishna (2019), Müller et al. (2019) and references therein. The new aspects emerging increase immensely the complexity of the systems and limit most of the originally laid-out infrastructures. Accordingly, the WEF asks for new forms of structural and organizational agility that offers better supply chain visibility. Instruments for automated data identification (Auto-ID/RFID) and the intelligent integration of systems, assemblies, and sensors into higher-level value networks, allow to continuously acquire and process data. In turn, this provides data and information for the decision making process on different scales: online, operational, tactical and strategic. Note, however, that these technologies could not yet be leveraged to the fullest possible extent. Once this is accomplished, supply chain integration will also change.

The best way for integrating a network is still an ongoing discussion. For instance, it can be done by acquiring new supply chain entities, activities or products (e.g. through direct acquisitions or joint ventures). Alternatively, in the case of many enterprises, outsourcing emerges as a possibility to consider. Since this discussion

is still evolving many concepts and methodological approaches are still to be adequately framed.

Network integration approaches including outsourcing and joint ventures are very specific and depend on the circumstances as well as the current environment. Nevertheless, we can find several authors discussing these aspects such as Babazadeh et al. (2013), Johansson and Olhager (2018), Wilhelm et al. (2013) and Dou and Sarkis (2010).

16.5 Modeling Formulations for Industrial Location Decisions

There is no one-to-one solution, in terms of modeling formulation capturing the emerging challenges faced by supply chain management. However, there are facility location models available that address some well-framed sub-problems in this context. In Table 16.6 we present some of these challenges and some related aspects.

In the following we give two examples for location models addressing each one of the challenges in sustainability and uncertainty. Of course we are not able to provide in a book chapter all the details, but we decided to state always a complete model, which can be used in courses or for learning by the example. For a deeper understanding we cite the respective references.

16.5.1 Reverse Logistics

Reverse logistics and closed-loop supply chain have become a major area of supply chain management. Contrary to forward or traditional logistics which considers material flows from upstream to downstream of a value chain, reverse logistics refers to all operations related to the reuse of products.

According to Srivastava (2007) most often the model formulation relies on single economic objectives and miss to explicitly address environmental and social dimensions. The resolution of this mismatch can lead to sustainable supply chains. In this section we revisit a general facility location logistics model for reverse

Table 16.6 Some challenges faced by supply chain management and related topics

Sustainability	Uncertainty	Digital transformation
Reverse logistics	Interdiction and fortification	Collaboration
Supply sourcing	Supply chain risk	Network integration
Carbon footprint	Multi-period decision making	“Infinite” labor
Green supply chain	Multiple-criteria decision making	Organizational agility

logistics. This is a model fist introduced by Alumur et al. (2012) (see also Alumur et al. 2015).

Reverse logistics focuses on one of the first and still important objectives of sustainable supply chains: waste disposal. Additionally, it also includes what we can call return logistics and repair logistics as well as container and returnable container logistics (pallets, lattice boxes, small load carriers and reusable containers).

Following the Council of Supply Chain Management Professionals, reverse logistics is the process of moving goods from their typical final destination for the purpose of capturing value, or proper disposal (CSCMP 2013).

Before discussing an optimization model for reverse logistics ($RLND$) we introduce some notation. We make use of notation introduced in the context of the $BILL$ model presented in Sect. 16.3.1. Note, that the latter is introduced as a multi-period model and the $RLND$ model presented below as a single-period one.

We consider multiple products which include used, inspected, repaired or refurbished products, components, or raw materials. In order to take different states into account (inspected, repaired, refurbished, etc.), different product states need to be defined.

A recovery option describes an activity that transfers a product from one state to another. It includes all options related to real-life reverse logistics networks such as returns, recalls, repair, refurbishment, and recycle as well as non-recovery alternatives such as inspection, disassembly, repackaging for restock or resale, selling to suppliers, to the secondary market or to external (re)manufacturing facilities, and disposal. The latter alternative is operated by third-party logistics providers, which are external and therefore represent non-selectable facilities (see Alumur et al. 2015). Table 16.7 introduces the sets underlying the $RLND$ model.

Table 16.8 describes the parameters underlying the model. We highlight, in particular, parameters that represent the reverse BOM structure. For example, a damaged product can be converted into a repaired product through the recovery option repair. Another possibility is to have a used product 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 secondary market, or to external (re)manufacturing facilities. Some

Table 16.7 Sets considered for the $RLND$ model is addition to those already presented for the *BILL* model

Set	Description
R	Recovery options
I_r	Selectable facilities with recovery option $r \in R$
E_r	Existing facilities with recovery option $r \in R$
N_r	Potential locations for installing recovery option $r \in R$
J_r	Non-selectable location with recovery option $r \in R$ (secondary market, disposal)
L	All locations, $L = \cup_{r \in R} (I_r \cup J_r)$

Table 16.8 New general parameters used for the $RLND$ model

Symbols	Description
g_{ip}	Amount of product $p \in P$ generated at location $i \in L$
a_{rqp}	Number of units of product $q \in P$ required to produce one unit of product $p \in P$ $(q \neq p)$ using recovery option $r \in R$
K_{ri}	Capacity of recovery option $r \in R$ at location $i \in L$
K^e_{ri}	Maximum increase in capacity for recovery option $r \in R$ at location $i \in I_r$
RT_{rp}	Target amount of products $p \in P$ with recovery option $r \in R$

Table 16.9 New cost parameters used for the $RLND$ model

Symbols	Description
RE_{rip}	Revenue from recovering one unit of product $p \in P$ with recovery option $r \in R$ at location $i \in L$
RC_{rip}	Cost of recovering one unit of product $p \in P$ with recovery option $r \in R$ at location $i \in L$
FC_{ri}	Fixed setup cost of establishing recovery option $r \in R$ at location $i \in N_r$
CC_{ri}	Fixed cost of closing recovery option $r \in R$ at existing facility $i \in E_r$
OC_{ri}	Fixed cost of operating recovery option $r \in R$ at location $i \in L$
EC_{ri}	Unit cost of expanding capacity of recovery option $r \in R$ at location $i \in I_r$

Table 16.10 New decision variables used for the reverse logistics model

	Description
y_{ri}	Binary variable equal to 1 if recovery option $r \in R$ is operated at the selectable facility $i \in I_r$ and 0 otherwise
v_{rip}	Amount of product $p \in P$ recovered with recovery option $r \in R$ or collected at location $i \in L$
w_{ri}	Extra capacity established for recovery option $r \in R$ at location $i \in I_r$

recovery options may also incur costs as in the case of product disposal (see Alumur et al. 2015).

Table 16.9 introduces the cost parameters for the $RLND$ Model.

In Table 16.10 we present the decision variables. While in the $BILL$ model the decision variable y refers to the opening or closing of a location, in the $RLND$ model it refers to the selection of a recovery option. Similarly, decision variable w defines extra capacity for a production resource in the $BILL$ model and it defines extra capacity for the recovery option in the $RLND$ model.

The $RLND$ model can be formulated as a MILP. Its objective function (16.20) maximizes the total profit, which sums up the revenues of various recovery options and subtracts the costs involved in the system.

$$\begin{aligned}\max \; & \sum_{r\in R}\sum_{i\in L}\sum_{p\in P} RE_{rip}v_{rip} \\ & -\sum_{r\in R}\sum_{i\in L}\sum_{p\in P} RC_{rip}v_{rip} - \sum_{r\in R}\sum_{i\in N_r} FC_{ri}y_{ri} \\ & -\sum_{r\in R}\sum_{i\in E_r} CC_{ri}(1-y_{ri}) - \sum_{r\in R}\sum_{i\in I_r} OC_{ri}y_{ri} \\ & -\sum_{r\in R}\sum_{j\in J_r} OC_{rj} \\ & -\sum_{i\in L}\sum_{j\in L\setminus\{i\}}\sum_{p\in P} TC_{ijp}x_{jip} - \sum_{r\in R}\sum_{i\in I_r} EC_{ri}w_{ri}\end{aligned} \tag{16.19}$$

The flow balance equalities (16.20) relate incoming flows like products shipped to a location and the amount of product obtained after processing at a location with outgoing flows like products recovered at a location and products shipped to other locations. The recovery target for each product category and recovery option should be achieved due to constraint (16.21). Inequalities (16.22)–(16.24) restrict capacities. The former guarantees that the amount of recovered products at selectable facilities does not exceed the available capacity. Inequality (16.23) formulates a similar conditions for non-selectable facilities. Constraints (16.24) limit the level of capacity expansions at selectable facilities. Constraints (16.25) and (16.26) ensure that products can only be shipped from operating facilities. Conditions (16.27)–(16.29) set the domains of the decision variables.

$$\text{s.t.} \quad g_{ip} + \sum_{r\in R}\sum_{q\in P} a_{rqp}v_{riq} + \sum_{j\in L\setminus\{i\}} x_{jip} = \sum_{r\in R} v_{rip} + \sum_{j\in L\setminus\{i\}} x_{ijp} \quad i\in L, p\in P \tag{16.20}$$

$$\sum_{i\in L} v_{rip} \geq RT_{rp} \quad r\in R, p\in P \tag{16.21}$$

$$\sum_{p\in P} v_{rip} \leq K_{ri}y_{ri} + w_{ri} \quad r\in R, i\in I_r \tag{16.22}$$

$$\sum_{p\in P} v_{rjp} \leq K_{ri} \quad r\in R, i\in J_r \tag{16.23}$$

$$0 \leq w_{ri} \leq K^e_{ri}y_{ri} \quad r\in R, i\in I_r \tag{16.24}$$

$$0 \leq x_{ijp} \leq \mathscr{M} \sum_{r \in R} y_{ri} \quad i \cup_{r \in R} I_r, j \in L \setminus \{i\}, p \in P \tag{16.25}$$

$$0 \leq x_{jip} \leq \mathscr{M} \sum_{r \in R} y_{ri} \quad j \in L \setminus \{i\}, i \cup_{r \in R} I_r, p \in P \tag{16.26}$$

$$x_{ijp} \geq 0 \quad i, j \in \cup_{r \in R} J_r (i \neq j), p \in P \tag{16.27}$$

$$v_{rip} \geq 0 \quad r \in R, i \in L, p \in P \tag{16.28}$$

$$y_{ri} \in 0, 1 \quad r \in R, i \in I_r \tag{16.29}$$

Again, this problem contains as a special case the CFLP. For more details and solution approaches concerning this and related problems we refer the reader to Alumur et al. (2012), Alshamsi and Diabat (2015), Chen et al. (2015), Govindan et al (2015), Khatami et al. (2015).

16.5.2 Supply Chain Risk

While uncertainty definitely is an important topic also in reverse logistics, we show in this section how to explicitly model uncertainty in a location model by addressing the notion of supply chain risk.

Over the last decade supply chain risk became increasingly relevant, although the notion of risk or more precisely supply chain risk was not clearly defined. An extensive literature review on the topic concluded that supply chain risk can be defined by three elementary characteristics, namely: risk objective, risk exposition, and risk attitude (Heckmann et al. 2015). A risk-aware capacitated plant location model ($CPLP^{Risk}$) aims at overcoming systematic definitional inconsistencies and offers a risk-aware location formulation founded on the general capacitated plant location problem ($CPLP$) (Heckmann 2016).

If uncertainty can be captured by a joint CDF, a model incorporating uncertainty and risk can often be formulated as a two-stage stochastic program (see Chap. 8). The decisions consist of first stage and recourse decisions. Initially, the opening and capacity extension decisions are made for each facility, while minimizing the expected costs of the consequences of these decisions. When uncertain parameters are disclosed, the recourse or second-stage decisions lean on, improve or correct the decisions made at the first stage. The selection of the type of expansion level for every period depicts the second stage decision. It follows that the overall objective function minimizes the costs of the first plus the expected costs of the second stage decision. In what follows we assume that uncertainty can be captured by a finite number of scenarios each of which occurring with some probability that we also assume to be known in advance.

Table 16.11 introduces the sets underlying the $CPLP^{Risk}$ model.

Table 16.12 contains the deterministic and stochastic parameters underlying the model.

Table 16.13 presents the cost parameters.

In Table 16.14 we present the decision variables, which are similar to those introduced in the context of the *BILL* model.

Table 16.11 Sets used for the $CPLP^{Risk}$ model

Set symbol	Description
I	Facilities
J	Customers
T	Time periods
H	Expansion levels
S	Scenarios

Table 16.12 General deterministic and stochastic parameters for the $CPLP^{Risk}$ model

Symbol	Description
d_{jts}	Demand of customer j in period t under scenario s
β^o	Level of targeted service level
K_i	Capacity of facility i
K_h^e	Extra capacity of expansion level h
γ_{its}	Relative capacity reduction within facility i in time period t and scenario s
π_s	Probability associated with scenario s

Table 16.13 Cost parameters for the $CPLP^{Risk}$ model

Symbol	Description
OC_i	Fixed cost of opening a facility in location $i \in I$
EC_i^o	Fixed cost of installing optional extra-capacity at facility i
TC_{ij}	Unit transportation cost between facility i and customer j
R_j	Unit revenue provided by customer j
XC	Unit penalty cost for not reaching target service level
EC_h	Unit cost of extra-capacity of expansion level h

Table 16.14 Decision variables for the $CPLP^{Risk}$ model

Symbol	Description
y_i	Binary variable equal to 1 iff facility i is opened
z_i	Binary variable equal to 1 iff expansion options are installed at facility i
x_{ijts}	Amount transported from facility i to customer j in time period t under scenario s
φ_{jts}	Unsatisfied demand of customer j in time period t under scenario s
ω_{iths}	Binary variable equal to 1 iff in scenario s expansion level h is installed at facility i at time period t
β_s	Service level in scenario s
Δ_s	Service level reduction in scenario s

The $CPLP^{Risk}$ model can be formulated as a MILP. Its objective function (16.30) minimizes the total costs, which sums up the costs related to the first-stage decision and costs associated to the recourse decision which are offset or decreased by the revenue.

$$\min \sum_{i\in I}\left(OC_i y_i + EC_i^o z_i\right) + \tag{16.30}$$

$$\sum_{s\in S}\pi_s\left(XC\,\Delta_s + \sum_{i\in I}\sum_{h\in H} EC_h K_h^e \sum_{t\in T}\omega_{iths} + \sum_{t\in T}\sum_{i\in I}\sum_{j\in J}\left(TC_{ij} - R_j\right) d_{jts} x_{ijts}\right)$$

$$\text{s.t.} \quad \sum_{i\in I} d_{jts} x_{ijts} + \varphi_{jts} = d_{jts} \quad j \in J, t \in T, s \in S \tag{16.31}$$

$$\sum_{j\in J} d_{jts} x_{ijts} \le \gamma_{its} K_i y_i + \sum_{h\in H} K_h^e \omega_{iths} \quad i \in I, t \in T, s \in S \tag{16.32}$$

$$\sum_{h\in H} \omega_{iths} \le z_i \quad i \in I, t \in T, s \in S \tag{16.33}$$

$$z_i \le y_i \quad i \in I \tag{16.34}$$

$$\beta_s = 1 - \frac{\sum_j \sum_t \varphi_{jts}}{\sum_j \sum_t d_{jts}} \quad s \in S \tag{16.35}$$

$$\Delta_s = \beta^o - \beta_s \quad s \in S \tag{16.36}$$

$$0 \le \Delta_s \le 1 \quad s \in S \tag{16.37}$$

$$x_{ijts} \ge 0 \quad i \in I, j \in J, t \in T, s \in S \tag{16.38}$$

$$\varphi_{its} \ge 0 \quad i \in I, t \in T, s \in S \tag{16.39}$$

$$z_i \in \{0, 1\} \quad i \in I \tag{16.40}$$

$$y_i \in \{0, 1\} \quad i \in I \tag{16.41}$$

$$\omega_{iths} \in \{0, 1\} \quad i \in I, t \in T, h \in H, s \in S \tag{16.42}$$

Demand constraint (16.31) equalizes demand fulfillment and unsatisfied demand with customer demand. Capacity constraints (16.32) restrict the ratio of demand fulfillment of each facility to the available capacity at the facility considered. Facility-related capacity sums to the reduced capacity and the capacity extension units. For each time period and facility only one extension level is allowed to be executed, constraint (16.33), if and only if a capacity extension option has been allotted to the facility, constraint (16.34). The amount of service level deterioration is calculated by constraints (16.35)–(16.37). Additionally, variables are limited to appropriately accomplish the aforementioned requirements by constraints (16.38)–(16.42).

Equivalent to the $RLND$ model this problem contains as a special case the CFLP. For more details and solution approaches concerning the $CPLP^{Risk}$ model we refer the reader to Heckmann (2016) and Heckmann et al. (2016).

16.6 Conclusions

In this chapter we have put an emphasis on the importance of logistics activities, their strong presence in supply chain management and their necessary integration into location modeling. Although several models and approaches have been published addressing logistics activities in location problems, the focus is mostly on some technical details missing the holistic point of view of logistics. Summing up the insights of this chapter, a location modeler has three main tasks to accomplish in order to adequately address these hurdles:

- to identify logistics activities that are relevant to the underlying industrial supply chain problem,
- to integrate relevant logistics activities in location decision models,
- to frame industrial challenges to smaller and well-defined problems of location modeling.

We presented a basic integrated logistics location ($BILL$) model that captures logistics activities for location decision making. In addition to a good integration of relevant logistics activities into location models, location modelers are confronted with several emerging challenges in the context of supply chain management which nowadays especially demand for effective location decisions. Three main challenges were discussed. We introduced two location models that address some well-framed sub-problems of the aforementioned supply chain challenges. In accordance to the discussion presented, we offered some further well-framed sub-problems to specific supply chain challenges in Table 16.15.

The need for future research directions emerges from the discussion within this chapter. In addition to new modeling approaches that integrate logistics activities and align to current challenges in supply chain management, we want to put an emphasis on dovetailing location decision with production structure. Operational production decisions are modeled through the inclusion of the BOM in the *BILL* model. Structural production decisions are modeled through facility layout decisions and should also be included in a holistic location logistics point of view. However,

Table 16.15 Supply chain challenges, corresponding location models and related chapters in this book

Challenge	Location Model	Reference
Uncertainty in supply chains	Facility location under uncertainty	See Chap. 8
	Location models with multiple-criteria	See Chap. 9
Transformation of supply chains	Multi-period facility location	See Chap. 11
Disaster events	Location problems under disaster events	See Chap. 22

we could not find models integrating strategic in-house decisions (layout) and inter-facility decisions (Supply Chain Design). Nevertheless, this might be an interesting research direction. We therefore recommend the interested reader to have a look at current reviews on facility layout, such as those by Briskorn and Dienstknecht (2017) and Anjos and Vieira (2017). However, the current view of facility layout problems is rather limited and models miss to include the general logistics perspective. Industrial supply chains continue to evolve demanding decision makers to adopt and to apply sophisticated decision support systems. This implies that locators as well have to follow closely the developments in industrial supply chains.

References

Alshamsi A, Diabat A (2015) A reverse logistics network design. J Manuf Syst 37:589–598
Alumur SA, Nickel S, Saldanha-da-Gama F, Verter V (2012) Multi-period reverse logistics network design. Eur J Oper Res 220(1):67–78
Alumur SA, Kara BY, Melo MT (2015) Location and logistics. In: Laporte G, Nickel S, Saldanha da Gama F (eds) Location science, Chapter 16. Springer, Berlin, pp 419–443
Anjos MF, Vieira MV (2017) Mathematical optimization approaches for facility layout problems: the state-of-the-art and future research directions. Eur J Oper Res 261(1):1–16
Babazadeh R, Razmi J, Ghodsi R (2013) Facility location in responsive and flexible supply chain network design (SCND) considering outsourcing. Int J Oper Res 17(3):295–310
Briskorn D, Dienstknecht M (2017) Survey of quantitative methods in construction. Comput Oper Res 92:194–207
Chen YT, Chan FTS, Chung SH (2015) An integrated closed-loop supply chain model with location allocation problem and product recycling decisions. Int J Prod Res 53(10):3120–3140
Christopher M (2016) Logistics & supply chain management. Pearson, London
Cordeau J-F, Pasin F, Solomon MM (2006) An integrated model for logistics network design. Ann Oper Res 144:59–82
CSCMP (2013) Logistics management. In: Supply chain management: terms and glossary. The council of supply chain management professionals. https://cscmp.org/. May 2019
Daskin MS, Snyder LV, Berger RT (2005) Facility location in supply chain design. In: Langevin A, Riopel D (eds) Logistics systems: design and optimization. Springer, Boston, pp 39–65
Dou Y, Sarkis J (2010) A joint location and outsourcing sustainability analysis for a strategic offshoring decision. Int J Prod Res 48(2):567–592
Dunke F, Heckmann I, Nickel S, Saldanha-da-Gama F (2018) Time traps in supply chains: Is optimal still good enough? Eur J Oper Res 264(3):813–829
Eskandarpour M, Dejax P, Miemczyk J, Péton O (2015) Sustainable supply chain network design: an optimization-oriented review. Omega 54:11–32
Esmaeilian B, Behdad S, Wang B (2016) The evolution and future of manufacturing: a review. J Manuf Syst 39:79–100.
Garcia DJ, You, F (2015) Supply chain design and optimization: challenges and opportunities. Comput Aided Chem Eng (281):153–170
Geoffrion AM, Powers RF (1995) Twenty years of strategic distribution system design: an evolutionary perspective. Interfaces 25:105–127
Govindan K, Soleimani H, Kannan D (2015) Reverse logistics and closed-loop supply chain: a comprehensive review to explore the future. Eur J Oper Res 240(3):603–626
Heckmann I (2016) Towards supply chain risk analytics. Springer Gabler, Wiesbaden
Heckmann I, Comes T, Nickel S (2015) A critical review on supply chain risk–definition, measure and modeling. Omega 52:119–132

Heckmann I, Nickel S, Saldanha-da-Gama F (2016) The risk-aware multi-period capacitated plant location problem (CPLP-Risk). In: International conference on information systems in supply chain, Bordeaux, pp 1–25

Johansson M, Olhager J (2018) Comparing offshoring and backshoring: the role of manufacturing site location factors and their impact on post-relocation performance. Int J Prod Econ 205:37–46

Kache F, Seuring S (2017) Challenges and opportunities of digital information at the intersection of big data analytics and supply chain management. Int J Oper Prod Man 37(1):10–36

Khatami M, Mahootchi M, Zanjirani Farahani R (2015) Benders' decomposition for concurrent redesign of forward and closed-loop supply chain network with demand and return uncertainties. Transport Res E-Log 79:1–21

Kraljic P (1983) Purchasing must become supply management. Harvard Bus Rev 61(5):109–117

Logistik-Lexikon (2019) https://www.logistik-lexikon.de/lexikon/liste/. May 2019

Lummus RR, Vokurka RJ (1999) Defining supply chain management: a historical perspective and practical guidelines. Ind Manage Data Syst 99(1):11–17

Manavalan E, Jayakrishna K (2019) A review of Internet of Things (IoT) embedded sustainable supply chain for industry 4.0 requirements. Comput Ind Eng 127:925–953

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 – symposium of the Collaborative Research Center 559 "Modelling of Large Logistics Networks". Verlag Praxiswissen, 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(2):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(1):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

Mentzer JT, DeWitt W, Keebler JS, Min S, Nix NW, Smith CD, Zacharia ZG (2001) Defining supply chain management. J Bus Logist 22(2):1–25

Müller F, Jaeger D, Hanewinkel M (2019) Digitization in wood supply – a review on how Industry 4.0 will change the forest value chain. Comput Electron Agric 162:206–218

Nagy G, Salhi S (2007) Location-routing: issues, models and methods. Eur J Oper Res 177(2):649–672

Simchi-Levi D, Kaminsky P, Simchi-Levi E (2007) Designing and managing the supply chain – concepts, strategies and case studies. McGraw-Hill/Irwin, New York

Srivastava S (2007) Green supply-chain management: a state-of-the-art literature review. Int J Manag Rev 9(1):53–80

Stadtler H (2008) Supply chain management—an overview. In: Stadtler H, Kilger C (eds) Supply chain management and advanced planning, vol. 528. Springer, Berlin, pp 9–36

Web Finance Inc. (2019) List of logistics definitions – business dictionary. http://www.businessdictionary.com/topic/logistics/definitions/. May 2019

WEF (2017) Impact of the fourth industrial revolution on supply chains. World economic forum. Available online https://www.weforum.org/whitepapers/impact-of-the-fourth-industrial-revolution-on-supply-chains. May 2019

Wilhelm W, Han X, Lee C (2013) Computational comparison of two formulations for dynamic supply chain reconfiguration with capacity expansion and contraction. Comput Oper Res 40(10):2340–2356

Zanjirani Farahani R, Rashidi Bajgan H, Fahimnia B, Kaviani M (2015) Location-inventory problem in supply chains: a modelling review. Int J Prod Res 53(12):3769–3788

Zijm H, Klumpp M, Regattieri A, Heragu S (2019) Operations, logistics and supply chain management. Springer, Berlin

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. We specifically outline the advances and new approaches that have been developed since the previous edition of this volume.

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 a *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
e-mail: berman@rotman.utoronto.ca; krass@rotman.utoronto.ca

G. Laporte et al. (eds.), *Location Science*,
https://doi.org/10.1007/978-3-030-32177-2_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 will 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 these assumptions 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; we refer the reader specifically to Chap. 8 for an in-depth discussion of the issues outlined below.

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 and related models are discussed in Chap. 22. Such 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., 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. We refer to Snyder (2006) as well as to Chaps. 8 and 22 in this volume for a review of reliability and related models.

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 we are interested in as they 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, repositioning servers between facilities, re-routing a sever 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, theoretically, more 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 8 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.

Several interesting new ideas have been introduced to SLCIS models since the previous edition of this volume. These are highlighted in the present version.

The remainder of this chapter is organized as follows. We start by describing the main model components in Sect. 17.2. 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

In this section we specify the key model components that allow us to identify the main classes of SLCIS models. These classes and the relevant solution approaches will be described in the following sections. 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 the 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 looser 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 customer nodes generate "identical" demands (possibly, within certain priority classes), i.e., that the streams of demand are indistinguishable with respect to the originating node 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 given, 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 and Brimberg et al. 1997), or $I \subset J \cup A$ for a network $G = (J, A)$, 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 enforced by

$$\mu_i \leq \mu^{\max} y_i, \quad i \in I, \tag{17.2}$$

where $\mu^{\max}$ 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 $\mu_i \in [0, \mu^{\max}]$, 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.

We note that even in settings where μ is a continuous decision variable, it is sometimes useful to discretize it. This is because, as will be seen shortly, μ appears in many non-linear expressions for service levels and waiting times; discretization is a common trick used to linearize the corresponding expressions—this idea was first explored in Vidyarthi and Jayaswal (2014). When discretization is used, it is

assumed that the capacity μ_i of each facility i must satisfy

$$\mu_i \in \{\mu^1, \ldots, \mu^L\},$$

where $\mu^l, l = 1, \ldots, L$ represent a discrete set of options for service levels. Defining binary decision variables z_{il} which take on a value of 1 is $\mu_i = \mu^l$ and 0 otherwise, we can now write:

$$\mu_i = \sum_{l=1}^{L} z_{il}\mu^l, \ i \in I \tag{17.3}$$

$$\sum_{l=1}^{L} z_{il} = 1, \ \ i \in I, \tag{17.4}$$

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) $\mathscr{F}_S(w)$, 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 the *demand allocation* decision variables, specifying what portion of demand from customer node $j \in J$ is directed to facility $i \in I$. 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

$$\sum_{i \in I} x_{ij} \leq 1, \ j \in J \tag{17.5}$$

$$x_{ij} \leq y_i, \ i \in I, \ j \in J \tag{17.6}$$

$$x_{ij} \in \{0, 1\}, \ i \in I, \ j \in J \tag{17.7}$$

These constraints are quite standard in location models: (17.5) ensures that at most 100% of customer demand from j is allocated to the facilities, (17.6) prevents allocating a customer to an unopened facility. Constraint (17.7) enforces the binary assumption for the allocation variables x_{ij}, with the value of 1 if the demand stream generated by customer node j is directed to facility i, and 0 otherwise.

The integrality of x_{ij} reflects the "single sourcing" assumption made in most SLCIS models, requiring each customer point to be assigned to at most one facility. An alternative is to allow "multi-sourcing", in which case x_{ij} is allowed to be continuous, by replacing (17.7) with its linear relaxation. We also note that constraints (17.5)–(17.7) 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 *lost demand*, i.e. demand that could have been captured but was lost, usually due to insufficient system capacity. 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

$$\Lambda_i = \sum_{j \in N_i} \lambda_j = \sum_{j \in J} \lambda_j x_{ij}. \tag{17.8}$$

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., a $M/G/1$ (or $M/G/\kappa_i$) queue with service rate μ_i.

System stability (i.e., ensuring that queue lengths are finite) requires that

$$\Lambda_i \leq \mu_i, i \in I, \tag{17.9}$$

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 be also 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

$$\Lambda_i \geq \Lambda^{min} y_i, i \in I. \tag{17.10}$$

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: 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 relies on 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 is more useful to differentiate models with dynamic vs. static 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 λ_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. Following the common modeling practice, all costs below are assumed to be per unit time.

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 $TC(d), d \geq 0$, representing the cost of traveling distance d is assumed to be non-decreasing and non-negative. This yields the System Travel Cost per time unit of

$$STC = \sum_{j \in J} \sum_{i \in I} TC(d(i, j))\lambda_j x_{ij}, \tag{17.11}$$

where we assume that constraint (17.6) 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

$$STC = \beta \sum_{j \in J} \sum_{i \in I} d(i, j)\lambda_j x_{ij}, \tag{17.12}$$

where $\beta > 0$ is a parameter relating the travel cost to other terms in the objective function (the meaning of this parameter is discussed in Sect. 17.3). We will use this linear form in place of (17.11) from this point on.

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*:

$$\sum_{i \in I} d(i, j)x_{ij} \leq R \quad \text{for all } j \in J, \tag{17.13}$$

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 (17.12) or (17.13); 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, in particular of SLCIS models. 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 $\bar{L}_i$, 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):

$$\overline{W} = \overline{W}^q + \frac{1}{\mu} = \frac{1+\gamma^2}{2}\frac{\rho}{1-\rho}\frac{1}{\mu} + \frac{1}{\mu} \tag{17.14}$$

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, $\frac{\rho}{1-\rho}$ 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

$$\overline{W}^q = [\text{Variability Factor}]\left[\frac{\text{Prob system busy}}{\text{Prob system free}}\right][\text{Scale Effect}]. \tag{17.15}$$

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?

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:

$$\overline{W} = \overline{W}^q + \frac{1}{\mu^0} = \frac{1+\gamma^2}{2}\frac{\rho^{\sqrt{2(\kappa+1)}-1}}{1-\rho}\frac{1}{\kappa\mu^0} + \frac{1}{\mu^0}, \tag{17.16}$$

which is very similar to (17.14): focusing on the expression for $\overline{W}^q$, we see that the only difference is that ρ in the numerator of (17.14) is replaced with $\rho^{\sqrt{2(\kappa+1)}-1}$ in (17.16). 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 (17.15) 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 former system will process customers κ times faster than the latter, and the 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 (17.14) 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 $i \in I$:

$$\overline{W}_i = \overline{W}_i^q + \frac{1}{\mu_i} = \frac{(1+\gamma^2)\Lambda_i}{2\mu_i(\mu_i - \Lambda_i)} + \frac{1}{\mu_i} \tag{17.17}$$

with Λ_i given by (17.8). We assume that $\overline{W}_i = 0$ if there is no facility at i.

One important question is how to treat the term γ^2 in the preceding expression. The "traditional" approach, adopted by all models described in the previous edition of the current text, has been to treat γ^2 as an intrinsic model parameter, rather than a decision variable, i.e., to 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., in $M/M/1$ 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: 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 not only Λ_i, but also the variability of service times γ_i. Nevertheless, the treatment of this parameter as exogenous, rather than a decision variable is quite common in SLCIS models; moreover its value is typically assumed to be identical at all facilities, which is reflected in our usage of γ without a subscript.

Several recent papers have relaxed the assumption that γ^2 is a fixed model parameter. One approach is to assume a one-to-one relationship between coefficient of variation of service times γ_i and service capacity μ_i at facility i, replacing γ^2 with $\gamma^2(\mu_i)$ in the previous expression. This idea is explored in Ahmadi-Javid et al. (2018), where γ_i is assumed to be a linear function of μ_i.

If the discretization of service times described by (17.3) and (17.4) is used, a very general relationship between μ and γ can be modeled. Recall that this approach assumes there are L discrete choices of capacity level. It is quite natural to assume that each choice $l \in L$ defines a pair (μ^l, γ^l) (in fact, two different choices could have identical capacity but different variability values). The coefficient of variation at facility i can now be written as

$$\gamma_i = \sum_{l=1}^{L} z_{il}\gamma^l, \tag{17.18}$$

where the decision variables $z_{il}, i \in I, l \in \{1, \ldots, L\}$ represent the choice of capacity level, as before. Now, for each fixed arrival rate Λ_i and capacity level l at facility i we can pre-compute the values of $\overline{W}_i^l(\Lambda_i)$ and write

$$\overline{W}_i(\Lambda_i) = \sum_{l=1}^{L} \overline{W}_i^l(\Lambda_i) z_{il},$$

which is linear in the decision variable. If, in addition we assume that Λ_i is discrete (which is natural in many contexts), we can further simplify the previous expression, while allowing for different coefficients of variation at different facilities (at the cost,

of course, of the approximation inherent in the discretization approach). Variations of this approach are used in Ahmadi-Javid and Hoisenpour (2018), Azizi et al. (2017), and Schön and Saini (2018).

Another observation regarding (17.17) is 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 (17.9) hold. These properties are exploited in several SLCIS models that follow.

Let $WC(w)$ represent the "waiting cost", i.e. the cost incurred by customers waiting w units of time in the system (here, and hereafter, we assume that waits include service times; 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

$$SWC = \sum_{j\in J}\sum_{i\in I} WC(\overline{W}_i)x_{ij}. \tag{17.19}$$

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 "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 not servicing a customer.

There are two potential issues with using (17.19) 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):

$$\overline{W}_i^q \le EW, \quad i \in I, \tag{17.20}$$

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

$$P(W_i^q > T) \leq \alpha_T, \quad i \in I, \tag{17.21}$$

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 (17.20) and (17.21) above are examples of *Service level Constraints* (SCs) that are quite common in SLCIS models. Since (17.20) refers to the expected behavior of the system, while (17.21) 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 (17.17) into (17.20), 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 (17.21) 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 (17.21) is satisfied whenever

$$G_S(\frac{V(T, \alpha_T)}{\mu_i})(\Lambda_i - 1) \leq V(T, \alpha_T), \tag{17.22}$$

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 (17.22) 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 (17.22) 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

$$\mu_i \geq \bar{\mu}(\Lambda_i), \tag{17.23}$$

where $\bar{\mu}(\Lambda_i)$ is computed (perhaps numerically) from (17.20), (17.21), or (17.22). 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

$$\Lambda_i \leq \bar{\Lambda}(\mu_i), \tag{17.24}$$

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.

$$\bar{\mu}(\Lambda) \approx \Lambda + Q\sqrt{\Lambda} \tag{17.25}$$

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 approximation for Q is provided by

$$\sqrt{2}Q \approx \sqrt{\gamma^2 + 1}P(W > 0).$$

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

$$\mu_i \geq \bar{\mu}(\Lambda_i) \approx \Lambda_i + \left[\sqrt{\frac{\gamma^2 + 1}{2}}\alpha_0\right]\sqrt{\Lambda_i},\ i \in I. \tag{17.26}$$

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 (17.26).

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:

$$SFC = \sum_{i \in I} FC_i y_i + \sum_{i \in I} VC(\mu_i) \tag{17.27}$$

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 now be expressed as

$$SR = r \sum_{i \in I} \Lambda_i = r \sum_{j \in J} \lambda_j \sum_{i \in I} x_{ij}. \tag{17.28}$$

In principle, parameter r can be treated as a decision variable—the price charged by the decision-maker for service. However, in the 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 (17.28), (17.12), (17.19), and (17.27). 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 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 to the service offered by the facilities. 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 on 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 (17.5)–(17.7). One may

Table 17.1 Model classification by customer response

	Demand allocation	
	Decision-maker	Customer
Inelastic demand	NR	AR
Elastic demand	DR	FR

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 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 (17.5)–(17.7) 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. 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 facility this customer actually visits. Since $\sum_j x_{ij} \leq 1, x_{ij} \in \{0, 1\}$, this leads to

$$u_j = \sum_{i \in I} U(d(i, j), \overline{W}_i) x_{ij}, \tag{17.29}$$

Note that this definition remains valid even when the single-sourcing assumption is relaxed. In this case, $x_{ij} \in [0, 1]$ 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

$$U^L(d, w) = -(\tau_d d + \tau_w w), \tag{17.30}$$

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 (17.30) 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

$$U^D(d, w) = -\tau_d d, \tag{17.31}$$

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 (though 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 (17.31) 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

$$U^E(d, w) = exp(-\tau_d d - \tau_w w), \tag{17.32}$$

which is quite similar to (17.30) 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—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 Stackelberg 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, i.e. allocates their demand to the utility-maximizing facility. Unfortunately, as we will see shortly, this may lead to situations where no equilibrium solution (i.e., set of choices for all customers) exists.

One fundamental issue is the implicit assumption that faced with the same set of alternatives (here, set of open facilities and processing capacities) customers always make the same choice. There is a rich body of research in marketing and economics that suggests that this may not be the case. A related question is how well can the customers measure their own utility? After all, if the utility function includes waiting times, a stochastic element is automatically present in measuring $U(d, w)$. Other stochastic elements, including uncertainties about travel times or even the non-waiting time aspects of the quality of the service interaction at the facility may also be present. Game Theory and Marketing literature have defined two notions of utility: deterministic and stochastic, with the associated large bodies of research. SLCIS literature have also adopted these two different notions of utility, leading to distinct classes of models.

As discussed below, in order to ensure the existence of equilibrium in deterministic utility models one has to allow for fractional choice, where the customers allocate their purchases among many (possibly all) facilities. Thus, the random choice element naturally enters in the deterministic utility setting, with the allocation vector derived from the equilibrium conditions. This set of models is discussed next.

An alternative approach, discussed in Sect. 17.3.2.4 is to assume a Proportional Allocation (PA) mechanism, where customers allocate their demand among the available facilities proportionally to the utility derived from each facility. The main advantage of this approach is that the allocation vector is specified from the start in closed form, leading to a simpler structure. Moreover, if one assumes a stochastic utility setting together with some additional assumptions, the (PA) mechanism naturally arises, providing additional axiomatic justification to this model class.

17.3.2.1 Customer Reaction Models with Deterministic Utility 1: Models with Allocation-Only Reaction (AR)

Here we assume that, once the facility locations and service capacities are determined by the decision-maker, the customer allocates their demand so as to maximize their deterministic utility function $U(d, w)$. Moreover, AR models assume that the demand rate of each customer node is fixed *a priori*, with $\lambda_j = \lambda_j^{\max}$ for all $j \in J$. For concreteness, we will assume the linear specification of the utility function

$U^L(d, w)$ given by (17.30), though much of the discussion extends to alternative specifications as well.

Even in this relatively simple setting 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, e.g., Nagurney (1999).

Consider first the original "single-sourcing" assumption, i.e. that a customer will only patronize a single facility. Utility maximization implies that if $x_{ij} = 1$ for some $i \in I$ and $j \in J$, then

$$U^L\left(d(i, j), \overline{W}_i\right) \geq U^L\left(d(k, j), \overline{W}_k\right) \quad \text{for all } k \in I \text{ with } y_k = 1,$$

which, assuming for simplicity that $\tau_w = \tau_d = 1$ in (17.30), 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 (17.8) and $\overline{W}_i$ by (17.17), this leads to the following equilibrium conditions that must be satisfied by allocations x_{ij}:

$$d(i, j) + \overline{W}_i \leq [d(k, j) + \overline{W}_k] y_k + M(1 - x_{ij}), \quad i, k \in I, j \in J \tag{17.33}$$

$$\overline{W}_i = \frac{(1 + \gamma^2)\Lambda_i}{2\mu_i(\mu_i - \Lambda_i)} + \frac{y_i}{\mu_i + M(1 - y_i)}, \quad i \in I \tag{17.34}$$

$$\Lambda_i = \sum_{j \in J} \lambda_j^{\max} x_{ij}, \quad j \in J \tag{17.35}$$

$$\sum_{i \in I} x_{ij} \leq 1, \quad j \in J \tag{17.36}$$

$$x_{ij} \leq y_i, \quad i \in I, j \in J \tag{17.37}$$

$$x_{ij} \in \{0, 1\}, \tag{17.38}$$

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 (17.36) 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 (17.33)–(17.38) 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 (17.33)–(17.38). 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 (17.38) with its linear relaxation, the equilibrium-defining inequality (17.33) has to be adjusted as follows.

Recall the definition of u_j given by (17.29), 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

$$d(i,j) + \overline{W}_i \begin{cases} = -u_j^* \text{ if } x_{ij} > 0; \\ \geq -u_j^* \text{ if } x_{ij} = 0 \end{cases} \tag{17.39}$$

(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 (17.33) with the following complementarity conditions:

$$d(i,j) + \overline{W}_i \geq v_j, \ \ j \in J, i \in I \tag{17.40}$$

$$(d(i,j) + \overline{W}_i - v_j)x_{ij} = 0, \ \ j \in J, i \in I \tag{17.41}$$

$$v_j \geq 0. \tag{17.42}$$

Note that for a feasible solution we must have $v_j = -u_j^*$, indicating that the new decision variable represents the equilibrium "disutility" for customers at $j \in J$. We will refer to a solution of the system (17.34)–(17.42) 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, they extend 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, as illustrated in the following example, there may be multiple allocation vectors satisfying the equilibrium conditions for a particular fully served subset of customer nodes.

Example 17.2 Consider adding a second identical customer node j' to the system in Example 17.1. 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 in Brandeau and Chiu (1994):

Theorem 17.2 *For any two customer flow equilibria under which the same subset $J' \subset J$ is fully served, the values of $\Lambda_i \;\; i \in I$ (total demand seen at each facility) and $u_j, \;\; j \in J$ (equilibrium utility of each customer) 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 u_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 (17.34)–(17.42). Even for fixed facility locations and capacities, solving the customer flow equilibrium conditions is far from easy. While the system is a linear complementarity problem with respect to variables v_j, $\overline{W}_i$ and x_{ij}, the waiting time is, in general, non-linear with respect to the capacity decision μ_i, resulting in a non-linear complementarity problem, which is often computationally challenging.

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. For example, assuming $M/M/1$ allows for a variable substitution $\mu_i = \lambda_i + 1/\overline{W}_i$, where the waiting times, rather than capacities, are used as decision variables. This turns the equilibrium conditions into a linear complementarity problem, making the system much more solvable. Zhang et al. (2010) were able to compute equilibria for such a system with $|J| \approx 500$ and $|I| \approx 40$ (note that their model also had elastic demands, which likely increased computational complexity). However, computing the equilibrium is only a subproblem of an SLCIS model. Thus embedding even a simplified computation in an overall exact optimization procedure is very computationally challenging. Hence both of the papers cited above resort to search heuristics for the upper level (location and capacity allocation decisions).

An interesting recent development was presented in Aboolian et al. (2016) who show that for the $M/M/1$ system traffic equilibrium constraints can be linearized through the introduction of additional binary variables $z_{ij} = 1$ if $x_{ij} > 0$ (i.e. customer j makes some use of facility i) and $z_{ij} = 0$ if $x_{ij} = 0$. It is not clear if this approach can be extended to non-$M/M/1$ settings.

In view of the difficulties involved in using the customer flow equilibrium approach above, it is natural to think of model simplifications. We mention two 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, non-existence 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 non-existence 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 (17.33)–(17.38) 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 (17.31). 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 (17.33) with

$$d(i,j) \le d(k,j)y_k + M(1-x_{ij}), \quad i,k \in I, j \in J, \tag{17.43}$$

which leads to significant simplifications (obviously, single-sourcing assumption can be retained here as well).

Another alternative, which bypasses some of the difficulties discussed above, is to use stochastic utility model, which is discussed in Sect. 17.3.2.4.

17.3.2.2 Customer Reaction Models with Deterministic Utility 2: Models with Demand-Only Reaction (DR)

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 (17.29), 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: 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, the demand from customer j that is captured in response to the offered travel distance of $d(i(j), j)$ and waiting time $\overline{W}_{i(j)}$ is given by the composition of the decay function and the utility functions by:

$$\lambda_j(d(i(j), j), \overline{W}_{i(j)}) = \lambda_j^{\max} h(U(d(i(j), j), \overline{W}_{i(j)})), \quad j \in J. \tag{17.44}$$

One example of a functional forms that satisfy the required assumptions is the identity function $h(u) = u$ together with the exponential utility U^E given by (17.32), leading to the popular "exponential decay" demand specification:

$$\lambda_j(d(i(j), j), \overline{W}_{i(j)}) = \lambda_j^{\max} \exp(-\tau_d d(i(j), j) - \tau_w \overline{W}_{i(j)}), \quad j \in J. \tag{17.45}$$

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 (17.44) 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, (17.44) 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 $j \in J$ assigned to facility i (i.e., with $i(j) = i$) for each open facility (i.e., $y_i = 1$). 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 a 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 (17.44) without the need for explicit

constraints (17.9). Moreover, even though customer assignments are "dictated" by the decision-maker through the specification of x_{ij}, assigning a customer to a more distant or more congested facility leads to a lower demand λ_j, with the resulting loss of revenue. Thus, the model assures that the objectives of the decision-maker and customers are aligned, while avoiding the complexities of full traffic equilibrium treatment (another way to interpret the DR model is that the hard constraint requiring each customer to be assigned to their utility-maximizing facility is replaced with a soft constraint, allowing violations of such assignments at a cost). In fact, Aboolian et al. (2012) report (based on computational experiments) that at optimum all customers are almost always assigned to their utility-maximizing facility, though rare exceptions 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 Customer Reaction Models with Deterministic Utility 3: Full Response Models (FR)

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 (17.40)–(17.42), as well as the elastic demand equilibrium (17.44). 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. One approach to deal with this complexity is by using the DR model as an approximation—as noted above, computational experiments suggest that optimal solutions to DR and AR models often coincide. Another approach, which is becoming more popular, is to use an alternative specification of demand allocation vectors described in the following section.

17.3.2.4 Proportional Allocation (PA) Models

As discussed above, the PA modelling framework 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 shares of this market.

The simplified structure, where customer demand allocations appear in closed form and can be analyzed for additional insights, together with several attractive mathematical properties have attracted significant recent interest to this model class, with several new approaches appearing since the first edition of this book.

These models have their theoretical origins in the MCI model of Cooper and Nakanishi (1988). As discussed below, they are also closely linked to stochastic utility theory. 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 Chap. 14, as well as the 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. This mechanism specifies the solution to the non-cooperative game played between customers once the decision-maker's strategy is specified as follows: for customers at $j \in J$ and (open) facility at $i \in I$, the demand allocations are given by

$$x_{ij} = \frac{U(d(i,j), \overline{W}_i) y_i}{\sum_{k \in I} U(d(k,j), \overline{W}_k) y_k}, \tag{17.46}$$

where the numerator represents the utility derived from facility i by customers at j, 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 additional term $\sum_{k\in C} U(d(k,j),\overline{W}_k)$ 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.

Note that with the specification (17.46), it is easy to see that $\sum_{k\in I} x_{kj} = 1$ for all customers j, implying that all of customer's visits will be captured by the open facilities. In case where none of the open facilities provide adequate service (e.g., all are too far away to be considered), this may be unrealistic. A common modification is the inclusion of "outside option", i.e., the option for the consumer not to use the service offered by the facilities at all. Suppose the utility of this option for customers at j is given by U_{0j}. Then by adding this term to the denominator of the expression above we obtain

$$x_{ij} = \frac{U(d(i,j),\overline{W}_i)y_i}{U_{0j} + \sum_{k\in I} U(d(k,j),\overline{W}_k)y_k}, \tag{17.47}$$

where the outside option is modeled as a pre-existing competitive facility providing utility constant U_{0j}. Observe that in this case $\sum_{k\in I} x_{ij} < 1$.

In both cases, the demand allocations are fractional, and the demand rate from j attracted by facility i is given by $\lambda_j x_{ij}$. For deterministic utility models we drew a distinction between FR and AR models depending on whether λ_j is elastic or not. A similar distinction can, in principle, be drawn for PA models, with $\lambda_j = \lambda_j^{max}$ for AR models and λ_j being elastic with respect total utility derived by customer j from all facilities: $U_j = \sum_i U(d(i,j),\overline{W}_i)x_{ij}$. While PA-FR models of this type have been considered in deterministic location literature (see, e.g., Aboolian et al. 2007, 2012), we are not aware of any SLCIS models of this type. Thus, all current PA models follow the AR assumption that available customer demand at each node is equal to λ_j^{max}.

Note, however, that when specification (17.47) is used, the resulting model automatically retains some aspects of elastic demand. This because the total captured demand from customers j is given by $\lambda_j^{max}(1 - \frac{1}{U_{0j}+U_j(I)})$, where $U_j(I) = \sum_{k\in I} U(d(k,j),\overline{W}_k)y_k$ is the total utility derived by customers at j from the service offered by all open facilities. Thus, as the value of offered service declines, the amount of captured demand declines as well—exhibiting similar behavior as when the demand is specified explicitly. The fact that this elasticity of demand is represented by a single model parameter U_{0j} makes the model (as well as the parameter estimation) simpler, accounting for the popularity of this representation. On the other hand, it should be obvious that explicit demand specification via (17.45) provides much more modeling flexibility.

To complete the specification of the proportional allocation model one needs to select a particular utility function. The popular Multinomial Logit (MNL)

specification (McFadden 1974) employs exponential utilities, leading to

$$x_{ij} = \frac{\exp(-\tau_d d(i,j) - \tau_w \overline{W}_i) y_i}{U_{0j} + \sum_{k \in I} \exp(-\tau_d d(k,j) - \tau_w \overline{W}_k) y_k}, \tag{17.48}$$

where weights τ_d, τ_w, as well as the outside option parameter U_{0j} can be estimated from the available consumer demand allocation data using the MNL methodology.

Two interesting observations can be made with respect to the MNL model. First, it can be derived axiomatically from the stochastic utility theory. The following discussion is based on McFadden (2005)—please refer there for further details. If one assumes that customer utility is given by

$$U_{ij}^s = U^L(d(i,j), \overline{W}_i) + \epsilon_{ij},$$

where $U^L(d,w)$ is the linear utility function given by (17.30) and ϵ_{ij} is a Gumbel random variable, then under further assumption that Independence of Irrelevant Alternatives holds, Eq. (17.48) can be shown to be a unique equilibrium demand allocation vector. This important result, due to McFadden (1974), provides a link between stochastic utility and proportional allocation models. Indeed, the (MNL) model is extremely popular in econometrics and marketing literature, being the dominant model in brand choice and related fields. On the other hand, Independence of Irrelevant Alternatives assumption is routinely observed to be broken, leading to many generalizations of stochastic utility models; see McFadden (2005) for further discussion.

The second observation for the (MNL) model is that, under very mild conditions, the user equilibrium conditions (17.33) can be regarded as the limiting case of the (MNL) model above. Assume that the weights τ_d, τ_w are scaled by same parameter θ. It is shown in Fisk (1980) that the (MNL) allocation (17.48) approaches the user equilibrium solution (17.39) as $\theta \to \infty$. This result holds as long as the waiting times at the facility are continuous and non-decreasing in the total demand seen by the facility. Thus, the (MNL) model can be viewed as a proper generalization of the user equilibrium model with exponential utilities. This, together with its attractive analytical properties described below, accounts for the popularity of this model in some of the recent SLCIS papers.

The key advantage of the proportional allocation approach is that the values of x_{ij} are directly computable from (17.46) or (17.48) 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 the 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, (17.46) together with waiting time specification (17.17) and facility-level demand specification (17.8) form a system of non-linear equations. A solution to this system represents the equilibrium demand allocations and waiting times. The issues of existence and uniqueness of the equilibrium 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 (17.46) 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 (17.46) will naturally reach an equilibrium. They derive sufficient conditions for both uniqueness and stability.

Theorem 17.6 *For proportional allocation models the equilibrium is unique and stable*

Some of the key results stated above also extend to PA models of FR type (i.e., elastic demand), though sometimes certain additional conditions are required. However, as noted earlier, no SLCIS models of this type have been described in the literature (though AR models with outside option partially fill this gap).

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:

maximize $Z =$

$$r \sum_{j \in J} \lambda_j \sum_{i \in I} x_{ij} \tag{17.49}$$

$$- \sum_{j \in J} \sum_{i \in I} \beta d(i, j) \lambda_j x_{ij} \tag{17.50}$$

$$- \sum_{j \in J} \sum_{i \in I} WC(\overline{W}_i) x_{ij} \tag{17.51}$$

$$- \sum_{i \in I} FC_i y_i - \sum_{i \in I} VC(\mu_i) \tag{17.52}$$

$$\overline{W}_i = \frac{(1+\gamma^2)\Lambda_i}{2\mu_i(\mu_i - \Lambda_i)} + \frac{y_i}{\mu_i + M(1-y_i)}, \qquad i \in I \tag{17.53}$$

$$[\ \lambda_j \text{ specification for DR and FR models }] \tag{17.54}$$

$$[\ x_{ij} \text{ specification for AR, FR, and PA models }] \tag{17.55}$$

$$[\text{ Coverage Constraints }] \tag{17.56}$$

$$[\text{ SC Constraints }] \tag{17.57}$$

$$\sum_{i \in I} y_i \leq m \tag{17.58}$$

$$\Lambda_i = \sum_{j \in J} \lambda_j x_{ij}, \qquad i \in I \tag{17.59}$$

$$\sum_{i \in I} x_{ij} \leq 1, \qquad j \in J \tag{17.60}$$

$$x_{ij} \leq y_i, \qquad i \in I, j \in J \tag{17.61}$$

$$\mu_i \geq \Lambda_i \qquad i \in I, j \in J \tag{17.62}$$

$$x_{ij} \geq 0;\ \mu_i \geq 0;\ y_i \in \{0, 1\}. \tag{17.63}$$

The objective function (17.49)–(17.52) represents the total profit which includes the revenue, travel, congestion, and facility fixed and capacity costs, respectively. Constraints (17.53) 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 (17.54) 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 (17.55) are only relevant for user-choice models of AR and FR type. Constraints (17.58)–(17.62) 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 (17.54)–(17.57) 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 five common types of models: Coverage-Type (CT), Service-Objective (SO), Balanced-objective (BO), Explicit Customer Response (ECR), and Proportional Allocation (PA) models. These are described in more detail in the following sections. The relevant references are summarized on Tables 17.2–17.6. 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 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 corresponding terms in the objective function. The last two columns briefly describe the solution approach and any additional comments.

17.5.1 Coverage-Type (CT) Models

These models, listed on Table 17.2, 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 general coverage constraints (17.56), for instance constraints (17.13). The CT models 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).

CT 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

Table 17.2 Coverage-type (CT) models

Reference	Cust. resp.	Utility function	Queueing model	Flexible # Facil.?	Flexible proc. rate μ?	Coverage constr./ Lost demand	Service constr.	Obj: revenue (captured demand)	Obj: travel time	Obj: congest. cost	Obj: facility fixed cost	Obj: server variable cost	Solution approach	Additional comments
Baron et al. (2008)	AR	Distance	GI/G/N, GI/G/1	Yes	Yes	Yes	Yes				Yes	Yes, General concave	Decompose the problem into several simpler sub-problems. Developed heuristic based on the equitable facility configurations	Both single and multiple server models considered
Berman et al. (2006)	AR	Distance	M/M/1/c	Yes	No	Yes	Yes: % blocked calls				Yes, Min Total # of facilities		Variety of heuristics including tabu search and random adaptive search	Demand is lost due to coverage and congestion constraints
Kakhki and 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	No testing or comp. results

(continued)

Table 17.2 (continued)

Reference	Cust. resp.	Utility function	Queueing model	Flexible # Facil.?	Flexible proc. rate μ?	Coverage constr./ Lost demand	Service constr.	Obj: revenue (captured demand)	Obj: travel time	Obj: congest. cost	Obj: facility fixed cost	Obj: server variable cost	Solution approach	Additional comments
Marianov and Serra (1998)	NR	N/A	M/M/1, M/M/K	No	No	Yes	Yes	Yes, Max covered demand					Exact: Linearized the SLC, leading to a linear MIP	
Yang (2018)	AR and NR	Based on fac. service capacity, not waits	M/M/k	Yes	Flexible number of servers k	Yes	Yes: on prob. of long waits	Yes, Max covered demand			Budget constraint on total fixed and variable cost		SLC approx-imated by an upper bound on demand seen by each facility, where the latter is approxi-mated separately	Both NR and AR models considered

Table 17.3 Service-objective (SO) models

Reference	Cust. resp.	Utility function	Queueing model	Flexible # Facil.?	Flexible proc. rate μ?	Coverage constr./ Lost demand	Service constr.	Obj: revenue (captured demand)	Obj: travel time	Obj: congest. cost	Obj: facility fixed cost	Obj: server variable cost	Solution approach	Additional comments
Aboolian et al. (2009)	AR	Distance	M/M/k	No	Yes. Total # of servers bounded	No	No		Minmax	Minmax			Meta-heuristics (genetic, tabu)	
Berman and 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	
Boffey et al. (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	
Drezner and Drezner (2011)	AR/Prop. Alloc.	Distance, exp	M/M/k	No	No	No	No		Yes	Yes			Heuristic (descent, tabu search, simulated annealing, genetic)	

(continued)

Table 17.3 (continued)

Reference	Cust. resp.	Utility function	Queueing model	Flexible # Facil.?	Flexible proc. rate μ?	Coverage constr./ Lost demand	Service constr.	Obj: revenue (captured demand)	Obj: travel time	Obj: congest. cost	Obj: facility fixed cost	Obj: server variable cost	Solution approach	Additional comments
Hamaguchi and Nakade (2010)	AR	Distance	M/G/1	No	No	No	No		Ignored	Max prob $W < \tau$			Heuristic (greedy + tabu), service times computed exactly via Laplace transform	Maximize probability that waiting time is below t
Marianov and Serra (2011)	NR	N/A	M/M/c/K	No	No	No	Demand loss due to block-ages		Yes	Yes			Ant colony heuristic	Bi-objective; (1) Travel cost, (2) Congestion cost (with a coefficient for the number of customers in the system)
Marianov et al. (2009)	NR	N/A	M/Er/K/c	No	No	No	Yes: # blocked		Yes				Similar to Boffey et al. (2010) with SLC estimated via Erlang queues	
Wang et al. (2002)	AR	Distance	M/M/1	No	No (fixed mu)	No	Yes: max uti-lization rate bounded		Yes	Yes			Greedy, tabu search and Lagrangian relaxation heuristics	

Table 17.4 Balanced-objective (BO) models

Reference	Cust. resp.	Utility function	Queueing model	Flexible # Facil.?	Flexible proc. rate μ?	Coverage constr./ Lost demand	Service constr.	Obj: revenue (captured demand)	Obj: travel time	Obj: congest. cost	Obj: facility fixed cost	Obj: server variable cost	Solution approach	Additional comments
Hoisenpour and Ahmadi-Javid (2016)	NR	N/A	M/M/1 with interrupt.	Yes	Yes	No	No	Yes (fixed price)	Yes	Yes	Yes	Yes	Exact algorithm based on Lagrangian relaxation	Random service interruptions
Aboolian et al. (2008)	AR	Distance	M/M/k	Yes	Yes	No	No		Yes	Yes	Yes	Yes	Exact algorithm and heuristics	
Aboolian et al. (2018)	NR	N/A	M/M/1	Yes	Yes	No	Prob. of late delivery (Mod 1), penalty on late delivery per incident (Mod 2), or per unit time (Mod 3)		Yes	Yes	Yes	Yes	Semi-exact algorithm using tangent line approximation method	Capacity discretized and used to linearize the model
Abouee-Mehrizi et al. (2011)	AR/ Prop. Alloc.	Exp	M/M/1/ balking	No	Yes	No		Yes: demand loss due to balking			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 limited room capacity for waiting

(continued)

Table 17.4 (continued)

Reference	Cust. resp.	Utility function	Queueing model	Flexible # Facil.?	Flexible proc. rate μ?	Coverage constr./ Lost demand	Service constr.	Obj: revenue (captured demand)	Obj: travel time	Obj: congest. cost	Obj: facility fixed cost	Obj: server variable cost	Solution approach	Additional comments
Ahmadi-Javid and Hoisen-pour (2018)	NR	N/A	M/G/1	Yes	Yes	No	No		Yes	Yes	Yes	Yes	Exact algorithm based on Conic programming (standard CPLEX cuts)	Capacity discretized. Coefficient of variation part of capacity decision
Ahmadi-Javid et al. (2018)	NR	N/A	M/G/1	Yes	Yes	No	No		Yes	Yes	Yes	Yes	Exact algorithm based on Conic programming (both standard CPLEX cuts and additional valid cuts used)	Coeff of variation a function of service rate
Azizi et al. (2017)	NR	N/A	M/G/1	Yes	Yes	No	No		Yes	Yes	Yes	Yes	Hub and Spoke network. Exact solution via MIP	Capacity discretized. Coefficient of variation part of capacity decision
Castillo et al. (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	
Elhedhli (2006)	NR	N/A	M/M/1	Yes	Yes	No	No		Yes	Yes	Yes		Exact: linearization approach which eliminates capacity variables and replaces non-linear term in the	

													objective with a family of linear constraints; used column generation	
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 generation	
Marianov and Rios (2000)	NR	N/A	M/M/1	Yes	No	No	Yes: prob queue below a threshold		Link constr. cost		Yes		Exact: linearized the SLC, then solved MIP	Application to the location of ATM switches
Pasandideh and 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		Yes (Obj 1)	Yes (Obj 1)		Obj 2: Min Ave % idle time per facility	Genetic heuristic	Bi-objective: (1) total waiting time, (2) total % idle at the facilities
Vidyarthi and Jayaswal (2014)	NR	N/A	M/G/1	Yes	Yes	No	No		Yes	Yes	Yes		Exact: linearized "nasty" term in obj function, leading to a convex problem with exp number of constraints. Similar to Elhedhli (2006)	
Wang et al. (2004)	AR	Distance	M/M/1	Yes	Yes	No	Yes: max utilization rate bounded		Yes	Yes	Yes	Yes	Greedy-type heuristics; shown to be optimal for some of the models considered	Present several different models, but most general one is of "social optimum" type

Table 17.5 Explicit customer response (ECR) models

Reference	Cust. resp.	Utility function	Queueing model	Flexible # Facil.?	Flexible proc. rate μ?	Coverage constr./ Lost demand	Service constr.	Obj: revenue (captured demand)	Obj: travel time	Obj: congest. cost	Obj: facility fixed cost	Obj: server variable cost	Solution approach	Additional comments
Aboolian et al. (2012)	DR	Exp. Demand/ lin. Utility	M/M/1, M/M/k	Yes	Yes	No	Yes: wait	Yes				Yes	Exact algorithm and heuristics	Explicit response. Max profit including a feedback loop between customer demand and congestion. Exogenous price
Aboolian et al. (2016)	FR	Exp. Demand/lin. Utility	M/M/1	Yes	Yes	No	No	Yes			Constraint specifying upper and lower capacity bound for each open facility		Exact algorithm and heuristics	Explicit response. Max profit (exogenous price). Linearization of traffic equil.
Berman and Drezner (2006)	DR	Distance	M/M/1	No	No	No	Yes: exp. wait times	Yes					Single facility: exact O(n3); Multi-facility: NLIP, heuristic algorithms (ascent algorithm, tabu, simulated annealing)	Explicit response
Berman and Kaplan (1987)	DR	Linear	M/M/1	No: m = 1	No	No	No	Yes					Exact algorithm (1-facility)	Explicit response. Single facility setting. Exog. price

Baron et al. (2007)	AR/ search	Distance	M/M/1/c	No	No	No	Demand loss to blockage, search costs	Demand lost to block-ages, search					Heuristics combined with an iterative calibration scheme to estimate the expected demand rate faced by the facilities	Explicit response. Comes closer to mobile server models due to dynamic search behavior by customers
Berman et al. (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	Explicit response. Demand elastic in travel, wait, and endogenous price
Tavakkoli-Moghaddam et al. (2009)	FR	Distance and price; linear	M/M/k/K	Yes	Yes (flex. Number of servers)	No	No (Service optimized by separate objectives)	Yes for one of the objec-tives		Indirectly: one objective is waiting time		Indirectly: one objective is ide time	Metaheuristics based on vibration damping optimization	Multiple objectives. Blockages not accounted for in customer utility of waiting times. Endogenous prices
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	Explicit response. Considers several models, including FR models with traffic equilibrium conditions. Endog. price
Zhang et al. (2009)	FR	Linear	M/M/1	Yes: min workload per facility	No	No	No	Yes					Location allocation heuristic (that includes an equilibrium facility and client allocation sets)	Explicit response. Max total participation (captured demand)
Zhang et al. (2010)	FR	Linear	M/M/k	Yes: min workload per facility	Yes	No	No	Yes					Bi-Level optimization model; lower level solved by variational inequalities and upper level heuristically	Explicit response. Max total participation (captured demand)

Table 17.6 Proportional allocation (PA) models

Reference	Cust. resp.	Utility function	Queueing model	Flexible # Facil.?	Flexible proc. rate μ?	Coverage constr./ Lost demand	Service constr.	Obj: revenue (captured demand)	Obj: travel time	Obj: congest. cost	Obj: facility fixed cost	Obj: server variable cost	Solution approach	Additional comments
Dan and Marcotte (2017)	AR	Linear w/ blockage penalty/ MNL	M/M/1/c	Yes	Yes	No	No	Demand lost due to blockage			Budget constraint on total fixed and variable cost		Bi-level optimization model. Semi-exact algorithm based on piecewise linear approximation of lower level. Heuristic	Corrects model of Marianov et al. (2008)
Marianov et al. (2008)	AR	Linear/ MNL	M/M/1/c	No	No	No	No	Demand lost due to blockage					Greedy-based Heuristics	Max captured demand (at own facilities); blockages do not affect customer utilities
Rabieyan and Seif-barghy (2010)	AR	Distance	M/M/1/ balking	No	No	Constraint on idle rate at facilities	No	Demand loss due to balking					Genetic, simulated annealing, and k-opt-type heuristic. The latter outperforms by up to 43% for larger instances	Max total benefit that incorporates travel distance and accounts for balking
Schön and Saini (2018)	AR	Linear	M/G/1	Yes	No	No	No	Yes	Yes		Yes	Yes	Exact MIP + Heuristic. Both capacity and promised service level are discretized. Various forms of PA modeled (incl. MNL)	Exogenous price
Zhang et al. (2012)	Model 1: PA/AR, Model 2: ECR/FR	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: PA (MNL based on shortest time), model 2: ECR (shortest time)

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 (17.31), 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 an implicit stepwise utility function, with customers using service only if coverage and service level constraints are met.

The typical formulation maximizes the objective consisting of (17.49) with revenue $r = 1$, reflecting the maximization of captured demand, subject to constraints (17.56)–(17.61). 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 (17.23), requiring adequate service capacity at each facility, or (17.24), 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 CT (and some 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, (17.24) 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 CT 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.

We note an important result from Baron et al. (2008), who studied a very general version of the CT 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 decided, the optimal capacities μ_i can be computed through a separate optimization model.

This result provides an important insight for CT models: when the goal is to ensure "satisfactory" service experience, the optimal design should equalize loads on 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).

While traditional applications of CT models (with or without congestion) is in emergency services, an interesting new theme is the location of recharging stations for electrical vehicles. Due to limited battery range, coverage constraints are crucial. On the other hand, user choice behavior must be taken into account as well. An AR-type SLCIS model with these features is developed in Yang (2018), where each station is modeled as an $M/M/K$ queue, with the number of stations and the number of servers at each station being decision variables. A service constraint limiting the probability of long waits is assumed. Users select facilities based on travel distance and capacities, not waits, which eschews the issues related to traffic equilibria (but the assumption does seem questionable). Due to non-equal capacities at the facilities and non-linearities inherent in the $M/M/K$ system, a heuristic approximation is developed to linearize the SC constraints.

17.5.2 Service-Objective (SO) Models

These models, listed on Table 17.3, seek to design a system that optimizes "customer service" using limited resources. 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 (17.52). "Customer service" is typically defined as the combination of travel and congestion costs; thus the objective function typically includes terms (17.50) and (17.51). Since the congestion cost term (17.51) only measures the aggregate congestion, some authors (e.g., Boffey et al. 2010; Marianov et al. 2009; Marianov and Serra 2011; Wang et al. 2002) impose service level constraints to ensure that congestion is controlled at each facility. SO 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 constraint (17.60) is specified as equality.

The models of this class are either of NR 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 (17.32) 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 SO models is quite similar to that of CT 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 a 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.

SO 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 (BO) Models

These 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 facility-related costs. They are listed on Table 17.4.

One may view BO models as seeking to achieve a "social optimum"; the objective functions in these models are similar to social welfare functions in economics, with the resulting models being similar to the "first best" models. 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 one exception appears to be Aboolian et al. (2018) where service constraint is present in one of the three proposed models.

The objective function typically includes the "customer-borne" cost terms (17.50)–(17.51) representing travel and congestion costs, as well as the "operator-borne" facility costs (17.52). Since most models do not assume any demand losses,

the revenue term (17.49) is not included; the exception being Abouee-Mehrizi et al. (2011), who model revenue losses due to balking and thus optimize the net profit. Two of the models in Aboolian et al. (2018) include penalty terms for late deliveries (i.e., delayed service), where the penalty is charged per instance or per amount of delay.

Most models in this class assume relatively simple queuing systems at the facilities with the two recent exceptions being Hoisenpour and Ahmadi-Javid (2016) who study a system with random service interruptions, and Azizi et al. (2017) who assume $M/G/1$-based hub-and-spoke system.

Other distinguishing features of most BO models are typically 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 (third) term of the objective function (17.51). 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

$$L_i(\Lambda_i, \mu_i) = \frac{\Lambda_i}{\mu_i - \Lambda_i}. \tag{17.64}$$

All costs are assumed to be linear and uniform (i.e., identical for all facilities), leading to the following objective function:

$$\text{minimize } Z = \beta \sum_{j\in J} \sum_{i\in I} d(i,j)\lambda_j x_{ij} + WC \sum_{i\in I} L_i(\Lambda_i, \mu_i) + FC \sum_{i\in I} y_i + VC \sum_{i\in I} \mu_i, \tag{17.65}$$

where WC, FC, VC are the waiting cost, fixed cost and variable cost parameters respectively. This function is minimized subject to constraints (17.58), (17.60) specified as equality, as well as (17.59), (17.61) and (17.62).

Note 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 (17.25) 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 (17.65) and performing some

algebraic manipulations and noting that for NR models the total customer demand is an exogenous parameter, allows us to re-state the objective function as

$$\text{mininize } Z = \beta \sum_{j\in J}\sum_{i\in I} d(i,j)\lambda_j x_{ij} + 2\sqrt{WC \cdot VC}\sum_{i\in I}\sqrt{\sum_{j\in J}\lambda_j x_{ij}} + FC\sum_{i\in I} y_i,$$

subject to constraints (17.58), (17.61), and (17.60) specified as equality; the variables Λ_i and μ_i are no longer required.

This is a MIP with a single concave 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 tangent-line approximation (TLA) of the concave term, is presented in Aboolian et al. (2007). The TLA leads to an $\epsilon-$optimal solution, where the maximum relative error from the exact solution is bounded by ϵ, with the value of this parameter set by the user (the smaller the ϵ, the higher the computational effort required; $\epsilon = 10\%, 5\%, 1\%$ are typical choices). Recently, Hoisenpour and Ahmadi-Javid (2016) apply Lagrangian Relaxation to a model with random service interruptions at the facilities.

It should be noted that in view of the discussion preceding (17.25), 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 BO models is based on capacity discretization ideas described earlier. The following discussion follows Elhedhli (2006). Once again we start with the model whose objective function is given by (17.65) and assume an $M/M/1$ queue at each facility. Assume the processing capacity 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 (17.64) as

$$\Lambda_i = \frac{L_i}{1+L_i}\sum_{h=1}^{H}\mu^h z_{ih},\ i \in I, \tag{17.66}$$

where z_{ih}, as defined in (17.3 and 17.4) is a binary decision variable taking the value of 1 if $\mu_i = \mu^h$ and 0 otherwise. 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 (17.66) 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. A similar approach is applied to hub-and-spoke SLCIS system in Azizi et al. (2017).

The capacity discretization approach with the resulting MIP with concave objective naturally lands itself to the TLA methodology mentioned above. This approach is applied, with promising computational results, to a set of balanced objective models with explicit (per occurrence or per delay length) penalties on service delays in Aboolian et al. (2018).

An interesting recent development in MIP literature is the efficient treatment of conic functions (particularly conic constraints)—see Atamtürk and Vishnu Narayanan (2011) for a general treatment and Atamtürk et al. (2012) for an application to a location-inventory problem. Some standard solvers, e.g., CPLEX, now provide automatic treatment of conic inequalities. The resulting methodology has seen recent applications in the SLCIS literature as well. Ahmadi-Javid and Hoisenpour (2018) consider a BO model with $M/G/1$ queues at the facilities, where capacity is discretized and each choice leads to a certain μ_i, γ_i pair. The initial MIP with non-linear objective is re-formulated as a conic program with a linear objective and conic constraint to which CPLEX solver can be directly applied. A further development along this lines is presented in Ahmadi-Javid et al. (2018) where instead of using discretization, an affine relationship is assumed between the coefficient of variation γ_i and facility capacity μ_i. Once again an original non-linear MIP is recast as a conic program, but in addition to now-standard CPLEX treatment, a number of additional valid cuts are developed. The latter lead to a strong improvement in computational efficiency.

In summary, the simpler structure of BO SLCIS 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 (ECR) Models

ECR models specify an "explicit" customer response mechanism, i.e., they are of types AR, DR, or FR. These models are listed on Table 17.5. 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 (17.49), and may also include the facility cost terms (17.52), 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 ECR 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 determine a good set of open facilities, often using meta-heuristics. As in the case of BO 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 ECR 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 (17.32) for transparency, with the elastic demand given by (17.45). 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:

$$\text{maximize } Z = r \sum_{i \in I} \Lambda_i - FC \sum_{i \in I} y_i - VC \sum_{i \in I} \mu_i \tag{17.67}$$

$$s.t. \quad \overline{W}_i = \frac{y_i}{\mu_i - \Lambda_i} \qquad i \in I \tag{17.68}$$

$$\Lambda_i = \sum_{j \in J} \lambda_{ij}^{\max} \exp(-\tau_w \overline{W}_i) x_{ij} \qquad i \in I \tag{17.69}$$

(17.60), (17.61).

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 (17.62) are omitted, since 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:

$$\Lambda_i^{\max} = \sum_{j \in J} \lambda_{ij}^{\max} x_{ij}. \tag{17.70}$$

For each facility i we now solve the following univariate "capacity optimization" model:

$$\begin{aligned} &\text{maximize } r\Lambda_i - VC\mu_i \\ s.t. \quad &\Lambda_i = \Lambda_i^{\max} \exp(-\tau_w \frac{\Lambda_i}{\mu_i - \Lambda_i}) \\ &\mu_i \geq 0. \end{aligned}$$

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 (17.67)–(17.69) we obtain

$$\begin{aligned} \text{maximize } Z = &r \sum_{i \in I} \Lambda(\Lambda_i^{\max}) - FC \sum_{i \in I} y_i - VC \sum_{i \in I} \mu(\Lambda_i^{\max}) \\ &(17.60), (17.61), (17.70), \end{aligned}$$

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 ECR 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 computational results presented in Aboolian et al. (2012) suggest 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.

A notable recent advance for ECR models was made in Aboolian et al. (2016). They assumed $M/M/1$ system with the fixed costs and budget constraint replaced by the requirement that any open facility must have the capacity of at least μ^{min} and at most μ^{max} (a reasonable assumption in case of public service facilities). As described earlier, using waiting times W_j in place of capacities μ_j as decision variables and adding additional binary variables z_{ij} to represent whether customer i makes any use of facility j, they derive an MIP with the only non-linearity limited to $1/W_j$ terms. Since this is convex in W_j, the TLA methodology can be used to obtain a linear MIP which is ϵ-optimal for the original problem. They were able to solve fairly large problem instances (up to 900 customer nodes and up to 40 potential locations) to within (at most) 0.1% of optimality. However, as noted earlier, the approach may be quite fragile with respect to the $M/M/1$ assumption.

17.5.5 Proportional Allocation (PA) Models

As discussed earlier, these models incorporate explicit customer response to the service offered by the decision-maker; however the form of this response (allocation of customer's demand amongst the facilities) is pre-specified via Eq. (17.47). In the first edition of this volume these models were classified under the ECR type. However, with several interesting recent developments, these models now merit a separate category; they are listed on Table 17.6.

There are well-established methods for linearizing the fractional market share equation (17.47) when customer decisions are decoupled. However, as observed in Sect. 17.3.2.4, when customer's utility includes waiting time (or another measure of congestion at the facilities), the decisions become coupled and (17.47) defines a system of non-linear equations that make the resulting SLCIS computationally very challenging.

The $M/M/1$ system offers significant simplifications since it is possible to treat the waiting time, rather than capacity, as the decision variable. Zhang et al. (2012)

uses this approach to linearize the customer-level problem in their Model 1, while optimizing the decision-maker's level via heuristics.

A more general approach (at the cost of discretizing some key decisions) is developed in Schön and Saini (2018). For an $M/G/1$ system they use capacity discretization (which also allows them to model coefficients of variation as part of decision variables). In addition, they discretize offered service levels, i.e., wait times, at the facilities. All non-linearities in the model, such as both the numerator and denominator in (17.47), can now be discretized, and thus linearized through the introduction of additional integer variables. The resulting model is quite general—it can incorporate a variety of utility functions, as well as revenue and cost terms in the objective—is formulated as a linear MIP. However, the formulation is very large, and thus even relatively small instances cannot be solved to optimality by CPLEX. This leads to the development of several heuristic approaches.

A different approach, heavily rooted in economics literature, is taken by Dan and Marcotte (2017). Their starting point is the model of Marianov et al. (2008), the first published SLCIS model with PA mechanism. The facilities are modeled as limited buffer $M/M/1/b$ queues where b is the buffer size; customers are blocked from entering the facility when the queue size reaches b. The objective is to locate m facilities to maximize total captured demand, where customers have an option to choose either new or pre-existing "competitive" facilities. The model employs linear utilities (17.30) and MNL structure (17.48). A metaheuristic procedure, combining GRASP and Tabu Search, is proposed.

Dan and Marcotte (2017) point out and correct several deficiencies in this model: (1) the "captured demand" does not account for demand lost to blockages, (2) customer's utility function does not account for dis-utility due to blockages, leading to a perverse situation where a customer who obtains service after experiencing some waiting time has a lower utility than a customer who traveled the same distance but was then blocked from joining the queue, (3) the capacity μ_j was assumed to be identical at all facilities and was treated as an exogenous parameter. In addition, the new model of Dan and Marcotte (2017) introduces a budget constraint:

$$\sum_i (FC y_i + VC \mu_i) \leq B,$$

where B is the available budget, and other notation is consistent with the general model in Sect. 17.4. Note that the capacity decision is treated as a continuous variable (though the buffer size b is treated as an exogenous parameter with an identical value for all facilities).

The problem is first formulated as a bilevel model, with the upper level (leader) specifying the facility locations and capacities, with the objective of maximizing captured demand (both the objective and the constraints are linear), while the lower level (follower) allocating customer demand according to MNL mechanism and constraints relating wait times and blockage probabilities. In this initial form, the lower level is a fixed point equation, rather than an optimization problem. However, using the standard results from Fisk (1980), the lower level is converted to an

non-linear optimization problem, whose objective is shown to be convex. Next, a "semi-exact" solution procedure is developed, based on a similar procedure in Gilbert et al. (2015), using the following steps: (1) the lower-level objective is approximated with piecewise linear function and re-cast as an LP, (2) the optimality (i.e. duality) conditions for the LP are added as complementarity constraints to the upper level, resulting in a single-level integer program with complementarity constraints, (3) finally, similarly to Aboolian et al. (2016), the complementarity conditions are linearized through the addition of binary decision variables, resulting in a linear MIP. The resulting model yields an approximate solution to the original model due to the piece-wise linear approximation in step (1), however this approximation can be made arbitrarily precise by increasing the number of segments, hence the "semi-exact" nature of the algorithm. It should be noted that the resulting model tends to be quite large even when the original instance is of relatively small size, leading to computational difficulties. Thus a heuristic approach is proposed as well.

While these results may be quite fragile with respect to the $M/M/1$ assumption, they do indicate that capacity discretization is not the only way to approach PA-type models. They also point out that many methods developed in the transportation economics literature may be applicable to SLCIS models as well.

We finish the previous two sections with an important message from Zhang et al. (2012). In much of the literature, the difference between deterministic utility optimization of Sect. 17.3.2 and the proportional allocation is considered mainly on theoretical grounds, focusing on the difference between utility specifications, choice axioms, etc. Theoretical arguments can be made in favor of either approach. However, as shown in Zhang et al. (2012), these different mechanisms for modeling customer response may lead to very different optimal facility network designs, with wide-ranging implications: for example, it is shown that if PA choice model is assumed, while customers are actually following the utility optimization model (or vice-versa), many of the facilities will be over/ under-used, resulting in very different congestion patterns and network performance than what is predicted by the model. Thus, the choice of customer reaction model must be made based on empirical evidence of customer behavior in a given setting, rather than theoretical arguments for one or the other 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 systematized the models along two dimensions: by customer response and demand elasticity (leading to our NR/AR/DR/FR types), and by the key structural elements of the models (leading to our CT/SO/BO/ECR/PA model classes), 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 with respect to customer behavior and key model objectives, and with regards to realizing what key difficulties may arise for a given model type. We are pleased to note that several papers that were published after the first edition of this volume have adopted this classification.

We also hope that the proposed systematization will motivate the authors to ensure internal consistency of implicit assumptions in their models. This should help to avoid models where customer utilities are affected by travel times, but not waiting times, or by waiting times but not by blockages, etc. Of course, such simplifications may be necessary to make the model computationally tractable, but they should be explicitly spelled out and discussed.

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.

Some important strides towards deriving exact or semi-exact solution algorithms for models with realistic customer response mechanisms have been made since the first edition and are described above. These include (1) leveraging capacity discretization to incorporate variability of service times as endogenous parameter of the model, and also to develop clever linearization schemes; (2) adapting advances in conic programming to SLCIS models, and (3) pushing the boundary on the PA-type models. However, 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. The importance of basing modeling choices on empirical evidence of customer behavior must also be emphasized.

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

Aboolian R, Berman O, Verter V (2016) Maximal accessibility network design in the public sector. Transp Sci 50(1):336–347

Aboolian R, Berman O, Wang J (2018) Responsive make-to-order supply chain network design. Working Paper

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

Ahmadi-Javid A, Hoseinpour P (2018), Convexification of queueing formulas by mixed-integer second-order cone programming: an application to a discrete location problem with congestion. Working Paper. arXiv:1710.05794

Ahmadi-Javid A, Berman O, Hoseinpour P (2018) Location and capacity planning of facilities with general service-time distributions using conic optimization. Working Paper, arXiv:1809.00080

Ashtiani H, Magnanti T (1981) Equilibria on a congested transportation network. SIAM J Algebra Discr Methods 2:213–226

Atamtürk A, Narayanan V (2011) Lifting for conic mixed-integer programming. Math Program 126(2):351–363

Atamtürk A, Berenguer G, Shen Z-J (2012) A conic integer programming approach to stochastic joint location-inventory problems. Oper Res 60(2):366–381

Azizi N, Vidyarthi N, Chauhan S (2017) Modelling and analysis of hub-and-spoke networks under stochastic demand and congestion. Ann Oper Res 264:1–40

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. Manuf Serv Oper Manag 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, 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, Berlin, 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
Cooper L, Nakinishi M (1988) Market share analysis. Kluwer Academic Publishers, Boston
Dan T, Marcotte P (2019) Competitive facility location with selfish users and queues. Oper Res https://doi.org/10.1287/opre.2018.1781
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. Manuf Serv Oper Manag 8:92–97
Fisk C (1980) Some developments in equilibrium traffic assignment. Transp Res B-Meth 14(3):243–255
Gilbert M, Marcotte P, Savard G (2015) A numerical study of the logit network pricing problem. Transp Sci 49(3): 709–719
Gross D, Harris C (1985) Fundamentals of queueing theory, 2nd edn. John Wiley and Sons, 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
Hoseinpour P, Ahmadi-Javid A (2016) A profit-maximization location-capacity model for designing a service system with risk of service interruptions. Transp Res E-Log 96:113–134
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. In: 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 M (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, Rios M, Icaza MJ (2008) Facility location for market capture when users rank facilities by shorter travel and waiting times. Europ J Oper Res 191(1):32–44
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, New York

McFadden DL (2005) Revealed stochastic preference: a synthesis. Econ Theory 26(2), 245–264
Nagurney A (1999) Network economics: a variational inequality approach. Kulwer Academic Publishers, 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
Schön C, Saini P (2018) Market-oriented service network design when demand is sensitive to congestion. Transp Sci https://doi.org/10.1287/trsc.2017.0797
Shen Z-J (2005) Multi-commodity supply chain design problem. IIE Trans 37:753–762
Snyder LV (2006) Facility location under uncertainty: a review. IIE Trans 38:537–554
Suzuki A, Drezner Z (2009) The minimum equitable radius location problem with continuous demand. Eur J Oper Res 195:17–30
Tavakkoli-Moghaddam R, Vazifeh-Noshafagh S, Taleizadeh A, Hajipour V, Mahmoudi A (2009) Pricing and location decisions in multi-objective facility location problem with *M/M/m/k* queuing systems. Engr Opt 49(1):136–160
Tong D (2011) Optimal Pricing and Capacity Planning in Operations Managemnet. Ph.D. Thesis, University of Toronto
Vidyarthi N, Jayaswal S (2014) Efficient solution of a class of location-allocation problems with stochastic demand and congestion. Comp Oper Res 48:20–30
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 Log 51:138–152
Whitt W (1992) Understanding the efficiency of multi-server service systems. Manag Sci 38:708–723
Yang W (2018) A user-choice model for locating congested fast charging stations. Transp Res E-log 110:189–213
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
Zhang Y, Berman O, Verter V (2012) The impact of client choice on preventitive healthcare network design. OR Spektrum 34(2):349–370

Chapter 18
Aggregation in Location

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.

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), Eiselt and Marianov (2011), 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 corresponding problems

R. L. Francis
Industrial and Systems Engineering Department, 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*,
https://doi.org/10.1007/978-3-030-32177-2_18

may be NP-hard to optimize (Kariv and Hakimi 1979). For problems as diverse as those including the location of branch banks (Chelst et al. 1998), 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, 3, 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 $q << n$.

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 $D(S,V) \equiv (D(S,v_j): j = 1,...n)$, $D(S,V') \equiv (D(S,v_j'): j = 1,...n) \in R_+^n$. 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, $g(U) = w_1 u_1 + \ldots + w_n u_n$, and $g(U) = \max\{w_1 u_1, \ldots, w_n u_n\}$ 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 $w_j = 1$: $j = 1,...n$). 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

$$f(S:V) = \Sigma\ \left\{w_j\ D\left(S, v_j\right) : j \in J\right\}.$$

Let $\omega_1 = \Sigma\{w_j: j \in J_1\}$, $\omega_2 = \Sigma\{w_j: j \in J_2\}$. We then have $f(S:V') = \Sigma\{w_j\ D(S,v_j'): j \in J\} = \Sigma\{w_j\ D(S,\gamma_1): j \in J_1\} + \Sigma\{w_j\ D(S,\gamma_2): j \in J_2\} = \omega_1\ D(S,\gamma_1) + \omega_2\ D(S,\gamma_2)$.

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 $\omega = \Sigma\{w_j: j \in J\}$, and $\gamma = \Sigma\{(w_j/\omega)\ v_j: j \in J\}$ 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 $\ell\boldsymbol{p}$ 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

$$f(S:V) = \max\left\{w_j\ D\left(S, v_j\right) : j \in J\right\}.$$

Let $w_1^+ = \max\{w_j: j \in J_1\}$, $w_2^+ = \max\{w_j: j \in J_2\}$. We then have $f(S:V') = \max\{w_j\ D(S,v_j'): j \in J\} = \max\{\max\{w_j\ D(S,v_j'): j \in J_1\}, \max\{w_j\ D(S,v_j'): j \in J_2\}\} = \max\{\max\{w_j\ D(S,\gamma_1): j \in J_1\}, \max\{w_j\ D(S,\gamma_2): j \in J_2\}\} = \max\{w_1^+\ D(S,\gamma_1), w_2^+\ D(S,\gamma_2)\}$. 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 $D(S,v_j) \le r_j$, $j \in J$, $S \subset \Omega$, 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 $D(S,\gamma_1) \le \rho_1$, $D(S,\gamma_2) \le \rho_2$, $S \subset \Omega$. 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 have no 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. 2004b). 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 (2266 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 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 6600 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 3900 of the parcels seemed possibly usable for DRC sites, as they had public or commercial buildings whose total usable footage exceeded 2000 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 $S = \{s_1,...,s_k\}$, 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 6600 rows and 3900 columns. The size of this model was beyond the resources

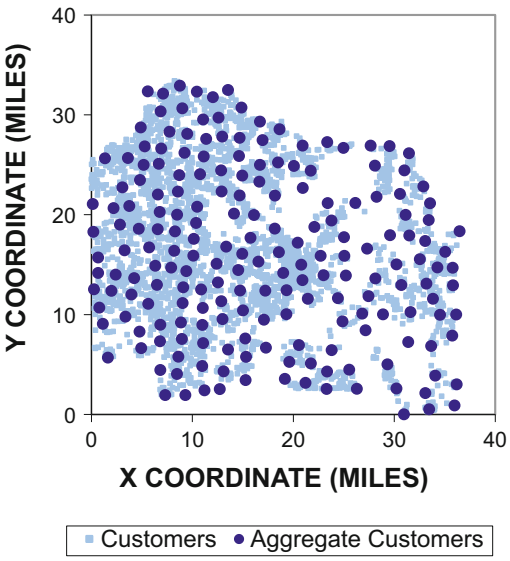


Fig. 18.1 Plot of demand points and aggregate demand points

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 that replaces v_j; the v_j are distinct while the v_j' are not. Choosing the v_j' to keep this error bound small keeps the covering error small. Note, if there are n distinct v_j', that $\max\{d(v_j,v_j'): j = 1, \ldots, n\}$ 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 6600 to 198 and the number of potential DRC sites from 3900 to 162.

The team's 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

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			b		
	Idealized			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

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 $d(v',v) \leq b$. Thus for any center s that covers v', $d(s,v) \leq d(s,v') + d(v',v) \leq r + b$. 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).

Note that $\max\{d(v_i,v_i'): \text{all unchosen DPs } v_j\} \leq b$ 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 11 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 DRCs respectively; see Fig. 18.2 for the case of three 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

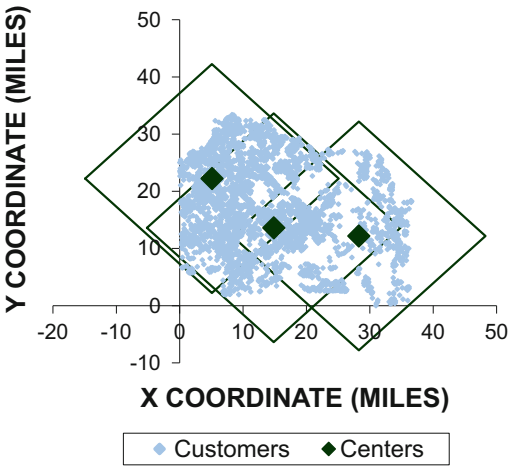


Fig. 18.2 The set of points within 20 miles of three Disaster Recovery Centers (DRCs)

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 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 to not 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. Some of these error measures are discussed in NaimiSadigh and Fallah (2009), as well as in Irawan and Salhi (2015a). All of the error measures in Table 18.2 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-DP distances and other error measures, including the *distance difference error*. For the PMM,

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	$d(v_j', v_j)$, $j \in J$
2	Distance difference error	$D(S,v_j') - D(S,v_j)$, $j \in J$, all S
3	DP error, PMM	$e_j(S) = w_j\,[D(S,v_j') - D(S,v_j)]$, $j \in J$, all S
4	Total DP error, PMM	$e(S) = \Sigma\{e_j(S){:}\ j \in J\}$, all S
5	ABC error for PMM: $J_1, \ldots, J_q$ is a partition of $J = \{1, \ldots, n\}$; $\omega_i \equiv \Sigma\{w_j{:}\ j \in J_i\}$ for $i = 1, \ldots, q$	$abc_i(S) = \omega_i D(S,\gamma_i) - \Sigma\{w_j D(S,v_j){:}\ j \in J_i\}$, 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	A number eb with $ae(S) \le eb$ for all S
	Ratio error bounds (when $f(S{:}V)$, $f(S{:}V') > 0$)	$\lvert f(S{:}V')/f(S{:}V) - 1\rvert \le eb/f(S{:}V)$, $\lvert f(S{:}V)/f(S{:}V') - 1\rvert \le eb/f(S{:}V')$ for all S
10	Location error	a measure, $diff(S', S_*)$, of the "difference" between p-servers S' and S_*

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 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 $\sum \{w_j D(S,v_j)\colon j \in J_i\}$ is replaced in the aggregate model by $\sum\{w_j D(S, \gamma_i)\colon j \in J_i\} = \omega_i D(S,\gamma_i)$, where $\omega_i \equiv \sum \{w_j\colon j \in J_i\}$. 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 $\sum\{w_j D(S,v_j)\colon j \in J_i\}$ 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 $\sum\{w_j D(S,v_j)\colon j \in J_i\}$. 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 $abc_i(S) = \sum \{e_j(S)\colon j \in J_i\}$ 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., -3000) 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 $S \subset \Omega$:

$$\left| \frac{f\left(S : V'\right)}{f\left(S : V\right)} - 1 \right| \leq \frac{eb}{f\left(S : V\right)} \iff \left| \frac{f\left(S : V\right)}{f\left(S : V'\right)} - 1 \right| \leq \frac{eb}{f\left(S : V'\right)}.$$

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 $|f(S:V') - F(S:V)| \leq eb$ and $|f(S:V') - f(S:V)| \leq eb$ 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 $d(v_j, v_j') = d(b_j, v_j')$ 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 $|f(S':V') - f(S_*:V)| \leq eb$, then $|f(S':V) - f(S_*:V)| \leq 2eb$. Supposing $f(S':V) > 0$, we thus have $|1 - f(S_*:V)/f(S':V)| \leq 2eb/f(S':V)$. 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 $p \geq q$, 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 $p << q$.

Table 18.3 Various types of optimality errors for any location model *f(S:V)*. Ideal error measures are zero

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	$f(S_*:V) - f(S':V)$

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 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 $p \times p$ matrix $C = (c_{ij})$ with $c_{ij} = d(x_i,y_j)$. Define an assignment (permutation matrix) to be any 0/1 p x 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, and 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 $|1 - f(S_*:V)/f(S':V)| \leq 2\ eb/f(S':V)$ 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. Remove space The second approach introduces an additional source of error, since a heuristic is used, but may possibly result in a better solution. Promising examples of this second approach include the work of Avella et al. (2012), Irawan and Salhi (2015b), and Irawan et al. (2014). 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, $|f(S:V') - f(S:V)| \leq eb$.

Consider now obtaining error bounds for the PMM and PCM, say eb_{pmm} and eb_{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 $S \subset \Omega$:

$$-d\left(v_j', v_j\right) \leq D\left(S, v_j'\right) - D\left(S, v_j\right) \leq d\left(v_j', v_j\right) \iff \left| D\left(S, v_j'\right) - D\left(S, v_j\right) \right| \leq d\left(v_j' v_j\right). \tag{18.1}$$

The p-median and the p-center models have the following error bounds respectively:

$$eb_{pmm} = \Sigma \left\{ w_j \, d\left(v_j{}', v_j\right) : j \in J \right\}, eb_{pcm} = \max \left\{ w_j \, d\left(v_j{}', v_j\right) : j \in J \right\}.$$

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

$$eb_{pmm} = \Sigma \left\{ w_j \, D\left(\Gamma, v_j\right) : j \in J \right\}, eb_{pcm} = \max \left\{ w_j \, D\left(\Gamma, v_j\right) : j \in J \right\}.$$

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 $D(S,v_j) \leq r_j$ is equivalent to $D(S,v_j)/r_j \leq 1$, from (18.1) we obtain.

$$\left| D\left(S, v_j{}'\right)/r_j - D\left(S, v_j\right)/r_j \right| \leq d\left(v_j{}', v_j\right)/\mathrm{r_j}, \text{ for all } j \in J \text{ and all S} \subset \Omega. \tag{18.2}$$

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 $f_j(S) \leq r_j$, $j \in J$, $S \subset \Omega$. 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 $f_j'(S) \leq r_j$, $j \in J$, $S \subset \Omega$. If we now define functions $f(S)$ and $f'(S)$ by $f(S) \equiv \max\{(1/r_j)\, f_j(S): j \in J\}$, $f'(S) \equiv \max\{(1/r_j)\, f_j'(S): j \in J\}$, then the constraints for the two models are equivalent to $f(S) \leq 1$ and $f'(S) \leq 1$ 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) for example, that if $f_j'(S)$ and $f_j(S)$ have error bound b_j ($= d(v_j', v_j)/r_j$ for the CLM) 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 $f'(S) \leq 1\text{-}eb \Rightarrow f(S) \leq 1$; $f(S) \leq 1 \Rightarrow f'(S) \leq 1 + eb$. Thus, the constraint $f'(S) \leq 1\text{-}eb$ gives a restriction of the original constraint, while $f'(S) \leq 1 + eb$ 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 $eb << 1$ (which is clearly desirable), feasibility conclusions about one model thus allow us to draw feasibility or "near-feasibility" conclusions about the other model.

Following Francis et al. (2004c), Table 16.4 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. (2004c) 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 1000 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 $\max\{w_j\, d(v_j', v_j): j \in J\}$ 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). Although most of the error bound results were developed for the case of discrete demand, Francis and Lowe (2014) study the relationship between error bounds for the discrete and the continuous demand cases.

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	$D(S, v_j) \leq r_j$, $j \in J$, all S
3	Aggregate constraints	$D(S, v_j') \leq r_j$, $j \in J$, all S
4	Restrictions of both original and aggregate constraints	$D(S, v_j') \leq r_j - \delta_j$, $j \in J$, all $S \iff$ $D(S, \gamma_i) \leq \beta_i$, $i = 1, \ldots, q$, all S
5	Relaxations of both original and aggregate constraints	$D(S, v_j') \leq r_j + \delta_j$, $j \in J$, all $S \iff$ $D(S, \gamma_i) \leq \rho_i$, $i = 1, \ldots, q$, all S

18.6 Conclusions

For location problems with many 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 (1978) 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, Efrat A, Sharir M (1999) Vertical decomposition of shallow levels in 3-dimensional arrangements and its applications. SIAM J Comput 29:912–953

Agarwal PK, Varadarajan KR (1999) Approximation algorithms for bipartite and nonbipartite matchings in the plane. In: 10th ACM-SIAM symposium on discrete algorithms (SODA). ACM, New York, pp 805–814

Ahuja RK, Magnanti TL, Orlin JB (1993) Exercise 12.23. In: Network flows: theory, algorithms, and applications. Prentice Hall, Englewood Cliffs, p 505
Avella P, Boccia M, Salerno S, Vasilyev I (2012) An aggregation heuristic for large scale p-median problem. Comput Oper Res 39:1625–1632
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 (1998) Issues and decision aids for designing branch networks. J Retail Bank X: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
Eiselt HA, Marianov V (2011) Foundations of location analysis. Springer, New York
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, Lowe TJ (2014) Comparative error bound theory for three location models: continuous demand vs discrete demand. TOP 22:144–169
Francis RL, Lowe TJ, Rayco MB, Tamir A (2003) Exploiting self-canceling demand point aggregation error for some planar rectilinear median problems. Nav Res Log 50:614–637
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, 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, McGinnis LF, White JA (1992) Facility layout and location: an analytical approach, 2nd edn. Prentice-Hall, Englewood Cliffs
Francis RL, Rayco MB, Lowe TJ, Tamir A (2009) Aggregation error for location models: survey and analysis. Ann Oper Res 167:171–208

Francis RL, White JA (1974) Facility layout and location: an analytical approach. Prentice-Hall, Englewood Cliffs, p 324. problem 7.25
Gavrilіouk EO (2003) Aggregation in hub location models. M Sc thesis, Department of Mathematics, Clemson University, Clemson, SC
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. MIT Press, Cambridge, NJ
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
Irawan C, Salhi S (2015a) Aggregation and non aggregation techniques for large facility location problems - a survey. Yugosl J Oper Res 25:313–341
Irawan C, Salhi S (2015b) Solving large p-median problems by a multistage hybrid approach iusing demand points aggregation and variable neighborhood search. J Glob Optim 63:537–554
Irawan C, Salhi S, Scaparra M (2014) An adaptive multiphase approach for large unconditional and conditional p-median problems. Eur J Oper Res 237:590–605
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, 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, 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
NaimiSadigh A, Fallah H (2009) Demand point aggregation analysis for location models. Chap. 22. In: Farahani R, Hekmatfar M (eds) Contributions to management science. Physica-Verlag, Heidelberg
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. IEEE, Los Alamitos, pp 320–331

Part III
Applications

Chapter 19
Location and Geographic Information Systems

Burcin Bozkaya, Giuseppe Bruno, and Ioannis Giannikos

Abstract Since their early stages of development, Geographic Information Systems (GIS) were utilized in a variety of ways to support analytical models in the field of location science. The interactions between the two disciplines soon became so evident that one can safely argue that GIS and location science are influencing each other in multiple ways. The rapid technological advances in the field of GIS and the ever increasing availability of geographically referenced information create even more possibilities for interconnections between GIS and location science. This chapter highlights these new possibilities and the new directions that emerge within academic research as well as in practical applications. We also attempt to point out further possibilities that may emerge in the future for linking these two disciplines.

19.1 Introduction

The realization that it is vital to take into account the locations where certain events or phenomena took place in order to fully analyze them, is not new. In fact, one of the earliest documented examples of spatial analysis is the study by Picquet in 1832 in which he represented the 48 districts of the city of Paris by half-tone color gradient according to the percentage of deaths by cholera per 1000 inhabitants. Perhaps the most celebrated application in the nineteenth century was the representation by John Snow in 1854 of a cholera outbreak in London using points to represent the locations of some individual cases. It was this representation which played a major role in

B. Bozkaya (✉)
Sabanci University, Istanbul, Turkey
e-mail: bbozkaya@sabanciuniv.edu

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*,
https://doi.org/10.1007/978-3-030-32177-2_19

the effort by the authorities to identify the source of the disease, a contaminated water pump, whose handle was disconnected, thus terminating the outbreak. The term Geographic Information Systems (GIS) initially appeared following the work by Roger Tomlinson and his colleagues who developed a digital natural resources inventory system for Canada in the 1960s. This system provided capabilities for measurement, digitizing, scanning and overlay, thus enabling the spatial analysis of stored data. The rapid advances in ICT opened the way for the development of modern GIS which have now become accessible to a wide variety of users ranging from large corporations to individuals.

Academics and practitioners dealing with location science models were quick to recognize the possibilities for interaction between location science and GIS. Initially, GIS were loosely coupled with location models and were mainly used for handling data and visualizing results. As GIS developed further, the two domains inevitably came closer together in the sense that each of them has influenced the other in a variety of ways. Apart from suitability analysis and data generation, the linkages include topics such as formulation of new models, analysis of uncertainty and error propagation and development of new solution methods for location science problems. These aspects were discussed in the earlier version of this chapter that was an attempt to evaluate how these linkages had evolved over time in comparison to the earlier reviews by Church (1999, 2002) and Murray (2010). In the last few years significant technological developments have certainly occurred such as the proliferation of smartphones and wearable devices, the massive availability of location data made accessible through platforms such as Google Earth or OpenStreetMap. In this new version of the chapter we intend to highlight the new directions that have been initiated both in terms of research and in terms of applications as a result of these transformations. In this work we will also avoid making extensive references to specific software tools mainly because we still expect that GIS technology will continue to evolve and will most probably make any reference to particular packages obsolete.

The rest of the chapter is organized as follows. In the following section we give an overview of Geographic Information Systems and their functionality, as it has advanced over the recent period. We then present the main ingredients of location analysis models and discuss the most common classes of problems where GIS have been successfully utilized. Finally, we comment on some recent applications where GIS have been employed to address realistic problems in location science and finish with some conclusions and directions for further research.

19.2 Overview of GIS

Generally speaking, GIS are information systems that integrate, store, edit, analyze, share and display geographic information as well as non-spatial information for supporting decision making. Over the years, the term GIS came to indicate a

technology as well as a tool or a way of data acquisition, management, manipulation, analysis and display.

Typically, GIS store information as a collection of thematic layers that are linked together by geography. In practical terms, GIS combine *spatial data*, that is in some way referenced to locations on the earth and *attribute data* that provides additional information about each of the spatial features and is typically represented in tabular format. As GIS technology developed, other types of data such as image or multimedia also became relevant. Documentation of GIS datasets is known as *metadata* and may concern the coordinate system, the date and time when the data was created or 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. 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. Each of these two spatial data models is characterized by certain advantages and disadvantages (see Church and Murray 2009).

Image data may also be used to store remotely sensed imagery, such as satellite scenes or aerial photos and it is typically stored in a variety of formats (e.g., .TIFF, .PNG, .JIF, etc.). Most GIS software packages allow the inputs 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. 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), 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).

Since their early stages of development all GIS tools offered a set of basic functions which characterized their primary purposes and objectives. This set of functions included the *management*, *transformation*, *analysis* and *visual presentation* of spatially referenced information. For a more detailed description of the

functionality of early GIS, see Chang (2018) and https://gisgeography.com/history-of-gis/.

The rapid advances in technology resulted in an almost simultaneous increase in GIS functional requirements. As a result, the basic functions mentioned above were rapidly enhanced by incorporating an ever increasing number of capabilities. For instance, off the shelf as well as open source GIS employ a series of machine learning algorithms in order to predict, cluster or classify spatial data.

The development which probably caused the most dramatic changes in GIS functionality is the fact that nowadays millions of users are sending and receiving information from mobile devices such as smartphones or tablets all over the world. The rapid shift towards a digitized society has totally changed the way people access and utilize location data. Whereas GIS technology and location data were previously available only to experts, the digital transformation of modern societies has made it possible for small enterprises as well as individuals to require access to location data anytime and on any device. GIS developers were quick to adapt to the new reality. As a result, Web based GIS applications are becoming increasingly popular as they offer a much broader range of functions and capabilities. In turn, this trend leads to the need for *standardization*, *interoperability* and *sharing of data*. Along these lines the Open Geospatial Consortium (OGC) initiated the Open Web Service (OWS) program based on service-oriented architectures and web service and proposed several geospatial specifications to support geospatial data sharing and interoperation, such as Web Map Service (WMS), Web Feature Service (WFS), and Web Processing Service (WPS) (see Verma et al. 2012). WMS offers the ability to produce maps rather than access specific data holdings and generates spatially referenced maps dynamically. WFS defines the interfaces for the access and manipulation of geographical features and elements through Geography Markup Language (GML) whereas WPS provides standardized interfaces to facilitate publishing, discovering and binding geospatial services that enable spatial processing functions across a network.

In general, rather than specific hardware and software implementations, modern GIS are portrayed as services where the underlying technology is less visible or even less important to the users (see Miller 2018). Hence, although the traditional GIS functions such as management, analysis etc. of spatial data have not been neglected, emphasis is now placed on accessibility, ease of configuration, user friendliness and integration of multiple data sources and formats.

19.2.1 GIS Software

The basic functions described above can be performed by a large variety of GIS packages that have 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. This trend has evolved at an accelerating pace over the years with a lot of GIS applications, developing within the framework of *Free and Open Source Software* (FOSS). The advantages of this approach include decreased software costs, protection of privacy, increased security etc. 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.

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 data collection and representation, as well as real time analysis for supporting decision making, without the need for local GIS software installation. Following the advances in cloud computing, several collaborative platforms have been developed for creating, publishing and sharing maps, data and services over the internet. See https://www.giscloud.com/, https://mangomap.com/ and http://www.qgiscloud.com/ for examples of such platforms. Detailed lists and reviews of GIS products can be found in Wikipedia and in specialist magazines and websites such as CapTerra (https://www.capterra.com/gis-software/) and GIS-Geography (https://gisgeography.com/mapping-out-gis-software-landscape/).

According to de Smith et al. (2013), a frequent criticism of GIS software was that it was over-complicated, resource-hungry and required specialist expertise to understand and use. Indeed, in many applications, only a handful of the capabilities provided by GIS was exploited. As a result, many users preferred 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. However, the emergence of a vast number of web-based tools and the availability of spatial data on the internet have enabled small businesses or even individuals to exploit a significant range of GIS capabilities with minimum training and expertise.

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 he adopts a specific product. Relevant criteria may be the availability of plug-ins that perform specific functions, the support of remote sensing tools, the ability to create web maps, the availability of documentation and examples etc.

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 problem* 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-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 (see Chap. 14 for further coverage on competitive location problems). 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 with 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). The interested reader is referred to Chaps. 2 and 3, respectively, for further details and background on p-median and p-center problems.

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). Further information on covering-type problems is available in Chap. 5.

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), Eiselt and Laporte (1995), Barbati and Piccolo (2016), Barbati and Bruno (2017).

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 Interconnections Between Location Science and GIS: Emerging Trends

As GIS began to evolve, the initial trend was to utilize them as a tool for data generation and visualization of results in various location science problems. In fact, GIS were combined with analytical models in a *loosely coupled* manner in the sense that data was obtained from the GIS and was then imported into an already defined model which was then solved either by an exact or by a heuristic approach. The results were then imported back to the GIS for visual representation.

The recent technological developments have rapidly increased the possibilities for capturing, storing and managing spatial data. This data is now captured by various different platforms ranging from GPS, satellites, aircraft and drones to stationary and mobile videos, road counters or other sensors. Moreover, another class of spatial data includes user generated data either intentionally or unknowingly provided by individuals through the use of smartphones or other electronic equipment. Through platforms such as WikiMapia or OpenStreetMap individual users can create, collect and disseminate spatial data. However, this data may be biased, inconsistent, or subjective and must be treated with caution. Nevertheless, one can safely argue that as far as availability of spatial data is concerned, we are in the age of big data where there is a wealth of information available linking certain phenomena to location and thus creating opportunities for associated analytical models. For instance, Al-Marwani (2014) combines GIS with socio-economic factors to analyze causal relationships that can be used to forecast real estate prices.

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 above. Consequently, apart from suitability analysis and data generation, GIS have inspired the formulation of new location science models, taking into account the wide range of spatial information that was readily available. Moreover, the extended capabilities of GIS were utilized to analyze uncertainty and error propagation in a variety of location science models. Finally, GIS were even used to develop new solution methods for challenging location science problems (Bruno and Giannikos 2015). A clear demonstration of the way GIS has affected the solution of location science problems concerns the Maximal Coverage Location Problem whose continuous version is computationally challenging while the number of its applications was relatively small in comparison to the discrete version. However, as noted by Murray (2016), certain GIS functions including *overlay*, *finite dominating set* (FDS) and *skeleton* have facilitated the solution of special cases of the continuous version thanks to the development of both exact and heuristic techniques.

The range and volume of spatial data that is currently available in various sources and formats has created further possibilities for developing and solving analytical models. Hence, it is evident that the two fields are rapidly converging in a number of ways, some of which are analyzed below.

19.4.1 Location Modeling with Spatio-Temporal Big Data

The term “big data” is loosely used to describe large and complicated data sets that far exceed the capacity of modern computing systems. What constitutes big data largely depends on the capabilities of the users and the available tools. These expanding capabilities have made big data a moving target ranging from few dozen terabytes to many exabytes of data (Everts 2016). The real value in big data lies not so much in the data itself but in the analysis that reveals the information hidden

within it. As pointed out by Lee and Kang (2015), geospatial data has always been big data. Geospatial big data is collected by a large variety of diverse sources such as GPS and GPS-enabled devices, satellite remote sensors, aerial surveys, radar, sensor networks, digital cameras, etc. The emergence of geospatial big data and its significance for developing national and global policies have led to the creation of the United Nations Initiative on Global Geospatial Information Management (UNGGIM), which seeks to guide the development of joint decisions and set directions on the production and use of geospatial information within national and global policy frameworks (http://ggim.un.org/).

According to Percivall (2013), big data analytics is an effective way to enhance the power of location. This has been stated very succinctly by Tobler (1970): "Everything is related to everything else, but near things are more related than distant things". Hence, human behavior is highly predictable as it is unlikely to deviate significantly over time and the future location of individuals can be accurately predicted by a careful analysis of their previous movements. This realization has inspired several successful applications in direct marketing, supply chain management (Provost and Fawcett 2013), vehicle routing (Valdes-Dapena 2011) and other domains.

The increasing availability of such big data has necessitated the development of new methods for modeling and structuring data. More specifically, greater emphasis has been placed on parallel and distributed programming for handling geo-spatial big data sets (Lee et al. 2014; Shekhar et al. 2014) while functional programming concepts or languages such as Haskell Domain-Specific Language (Mintchev 2014) and Map-reduce (Mohammed et al. 2014) have been utilized for managing such data.

The rise of big data has also set the pathway towards location optimization modeling approaches exploiting large spatial datasets. For instance, Cai et al. (2014) analyze travel patterns of a taxi fleet in Beijing to determine the best locations for electric vehicle charging stations. They analyze the trajectories of 10.000+ taxis, determine most common routes and hotspots, and use this information in site selection.

Facility location models in the presence of competitors commonly employ gravity models where customers or demand centers are gravitated towards competing facilities (e.g. stores, shops) based on distance and facility attractiveness. An empirical study is conducted by Suhara et al. (2019) that attempt to validate gravity models using transactional big data with millions of credit card purchasing records obtained from a private bank in Turkey. They show that the real-world transaction activity of consumers indeed confirms gravity models under certain regional parameters and various shopping categories including groceries, gas stations, restaurants and clothing stores. In a recently published study, Ting et al. (2018) analyze geo-spatial and socio-economic and demographic features that are most relevant in predicting the sales performance at a candidate retail site location. They analyze data from various sources and identify feature sets that vary over time as well as the retail location set. They use these findings to develop and evaluate similarity measures to predict sales for a new location.

19.4.2 GIS Tools Integration to Data Analytics Libraries

One of the "traditional" ways of fully exploiting GIS capabilities is the integration of GIS software with data analytics libraries or high level programming languages in order to write scripts or programs that automatically combine a series of GIS tools or utilities to accomplish certain tasks. There are several programming languages such as Python, Jscript or Perl that are particularly suited for writing short scripts since they have more basic syntax and are easier to learn than other languages such as C, Visual Basic or Java. In particular, the introduction into Python of many new programming features, have made the language much easier to deploy. As a result, Python has been combined with commercial GIS software, notably products developed by Esri, as well as open source platforms such as QGIS and GRASS. As noted by Altaweel (2017), most GIS users employ Python for developing short scripts rather than exploit its object oriented or imperative programming style features. Despite its simplicity, Python offers access to a wide range of libraries and facilitates the development of various applications such as GIS for mobile devices, integration of mapping features with web programs, and other tools that require server and cloud based services. Furthermore, Python allows easy access to well known libraries such as Google Maps and other popular Google software. Hence, Python has enabled programmers to more easily develop GIS and mapping tools that are integrated with other popular tools and devices (Altaweel 2017). Although Python is slower compared to other languages since it is an interpreted language, its simplicity and flexibility make it very popular when it comes to developing GIS applications, to the extent that special relevant conferences are organized across the world (http://2018.geopython.net/).

Of the major GIS libraries mentioned above that are available to programmers via Python, Esri's proprietary ArcObjects object library allows users to carry out basic as well as advanced GIS tasks, accessible via the company's suite of products and Python interface arcpy. The GIS library is typically executed as a geo-processing script or batch job where a great variety of spatial tools can be used. Particularly interesting and useful is the Network Analyst extension that encompasses location-allocation modeling and related solution tools, including p-median, p-center models and their variants as well as gravity-based competitive facility location models (www.esri.com/en-us/arcgis/products/arcgis-network-analyst).

Another major GIS library that is accessible via Python is the Geopandas (geopandas.org), an open-source library that offers an extensive set of tools for spatial data processing, geocoding and mapping. Users can apply this library in conjunction with many other supporting libraries for GIS basic functionality (see http://www.data-analysis-in-python.org/t_gis.html). Unfortunately, none of these libraries offer location modeling and optimization functionality.

Finally, open source applications QGIS and GRASS allow access to their underlying GIS function library via Python. They offer a command-line console that allows users to execute Python statements and perform location-based queries

and produce visualizations; they also allow execution of Python scripts against a spatial data source using the underlying libraries.

In addition to Python, the R statistical package has also been integrated with GIS software such as QGIS. Although R was traditionally known as a statistical package, it also has strong spatial analytical tools including point pattern analysis and Bayesian geostatistical modeling. It can read and handle a variety of vector and raster data, including shapefiles, NetCDF, and GDAL supported formats. As a result, its use has recently expanded to applications including natural language processing and web scrapping. By employing R, many popular statistical procedures and more advanced analyses can be implemented directly within GIS tools such as QGIS, thus expanding the statistical capabilities of conventional GIS software packages. Although R and QGIS are both not commonly used in industry, increasingly there are more research applications that integrate these tools. Examples include the papers by Pfeifer et al. (2016) on mapping Borneo's tropical rainforests where a beta-logistic regression was used to assess structural changes evident and Fortelius et al. (2016) on the mammalian fossil records. R itself also supports a library of spatial functions, working with both raster and vector data. These allow users to access various data formats as well as projection types, and manipulate them. Users can also create map visualizations, conduct spatial analysis including spatial joins, spatial queries, spatial statistics and network analysis (www.r-project.org).

Finally, the SAS software for statistical analysis also includes a GIS module for creating and managing spatial data. It also enables the user to interact with the data by selecting features and performing actions that are based on their selections. SAS also offers an extension application that integrates its statistical analysis libraries with Esri's GIS product suite. This extension, called SAS Bridge for Esri, provides a connection between the two systems where users can retrieve data from SAS in GIS analysis and mapping, as well as work in the reverse direction and export GIS data as SAS data sets for statistical analyses using SAS libraries (support.sas.com/rnd/datavisualization/BridgeForESRI/V2/).

The recent developments in storage, processing and analysis of Big Data have also offered profound opportunities for dealing with large geo-spatial datasets. A commonly used open-source cluster-computing framework called Apache Spark (spark.apache.org) has been extended in several ways to perform analyses on such datasets, including Magellan (magellan.ghost.io), GeoPySpark (geopyspark.readthedocs.io), GeoMesa (www.geomesa.org) and Databricks (databricks.com). These platforms offer storage, indexing, querying, transforming and analyzing spatial datasets at scale and are considered as the foundation for the next generation of geo-spatial data handling infrastructures.

In general, it has become obvious that geospatial data is crucial in many situations. As a result, programming languages for developing applications as well as conventional statistical packages have been equipped with plugins and tools for managing spatial data in order to fully exploit their potential.

19.4.3 GIS as Interactive DSS

Given the availability of modern GIS and their capabilities to store, manage and analyze spatial data, it is not surprising that their combination with Decision Support Systems (DSS) was an obvious expectation. In fact, some authors regard GIS as a form of Spatial DSS. However, as stated by Silva et al. (2014) and Ferretti and Montibeller (2016), GIS themselves are limited in terms of decision modeling capabilities which implies that their ability to support realistic decision making processes, involving conflicting objectives, is somewhat limited. Hence, the integration of GIS with decision making methodologies and in particular Multi-Criteria Decision Analysis has been recognized as a growing need by academics and practitioners alike, giving rise to a set of systems termed MC-SDSS (Multi-Criteria Spatial Decision Support Systems) that combine methods and tools for managing spatial data with decision making methodologies for modeling users' preferences and priorities. In fact, several applications have appeared in the literature concerning the development of MC-SDSS to support decisions in various domains including territorial development, urban planning, housing policies, bank branch closures and location of undesirable facilities. An interesting trend is the development of Web based SDSS where GIS information implemented in the World Wide Web (e.g. in Google Maps or similar environment) is exploited by Open Source GIS software and combined by MC Decision Analysis methods (see Malczewski (2004), Ferretti and Montibeller (2016) and the references therein for a detailed review). Although it is practically impossible to enumerate all published SDSS projects, we briefly discuss some typical ones to illustrate the variety of disciplines and the different types of decisions involved. For instance, Wenkel et al. (2013) present an interactive SDSS which supports interactive spatial scenario simulations, multi-ensemble and multi-model simulations at regional scale, as well as the complex impact assessment of potential land use adaptation strategies at local scale. Its main objective is to provide information on the complex long-term impacts of climate change and on potential management options for adaptation by answering "what-if" type questions. In another typical application, Silva et al. (2014) discuss the development of a Web MC-SDSS combining the ELECTRE TRI method of multi-criteria analysis within ArcGIS for analyzing the sustainability of dairy farms in Portugal. In another recent work, Yao et al. (2017) describe a web-based DSS integrated with GIS for preventing and controlling locusts efficiently, accurately, and rapidly. The locust prevention and control DSS is developed to assist farmers and local government agencies in Chinese provinces with high incidence of locust by providing spatial decision-making information. The system offers online access to county, city, provincial and national level data queries and is capable of storing, spatial analyzing, and displaying geographically referenced information of locust data. It can also provide the real-time tracking of GPS location, as well as goods scheduling of locust plagues prevention.

Apart from the applications described in academic papers, several interesting SDSS projects were developed for commercial clients or as part of research projects.

A typical example is the ELVIS (Environmental Values Interrogation System), built using QGIS. This system relies on data concerning locally identified features, along with regionally available GIS datasets. This data can be queried using ELVIS to investigate locations for new developments or activities. The user is able to draw in the "footprint" of a project and get a simple report listing the resource values that will be impacted. According to its developers, in this manner, ELVIS addresses the needs and capabilities of land use planners and upper level decision makers, whilst providing them with information that is relative to local communities (https://research.csiro.au/bismarcksea/step-3-decision-support-tools/). Along these lines, a Web-based GIS, called BDSS, has been developed which assembles data concerning oil production in North Dakota, USA, enabling users to visualize geologic and production information. Analytical tools support the evaluation and interpretation of geological properties such as thickness, depth, structure, and organic content. Production data can be utilized to provide development history and identify areas of low or high production and support relevant decisions (https://www.undeerc.org/Bakken/interactivemap.aspx). Similarly, the INDICATE (*Indicator-based Interactive Decision Support and Information Exchange*) is a European project whose objective was to develop an innovative city-wide decision support system to assist the transition towards smart cities. The system supports stakeholders by providing an interactive decision support tool for urban planning and design. The tool assesses the interactions between urban objects and spaces, buildings, the electricity grid, renewable technologies and information and communications technology (ICT) and recommends options for optimizing infrastructure, installing technology, and providing cost-effective utility services (http://indicate-smartcities.eu/).

Given the rapidly growing interest in MC-SDSS, Ferretti and Montibeller (2016) pointed out that the field was still somewhat fragmented and described the process of developing such a system as a procedure consisting of five steps: (1) Designing the decision process, (2) Structuring the MC-SDSS, (3) Eliciting spatial standardization functions, (4) Aggregation of partial performances and (5) Analysis of results and recommendations. For each of these steps they then stated a number of challenges or issues that need to be addressed. In short, these challenges concern who should participate in the whole process, which MCA method should be selected, what sources should be used to define objectives, etc. It is our belief that these challenges must indeed be faced in a systematic way, perhaps incorporating into the MC-SDSS structure more elements from the theoretical framework of Multi-Criteria Decision Making (MCDM). We anticipate research in this direction to continue within the next few years.

19.5 Using GIS in Location Science Applications

Given the volume and range of spatial information available and the methodological and technological advances, it is no wonder that spatial information is critical to a much wider range of applications than ever before (Chandel 2017). As a

result, geospatial and location technologies are now available across a number of industries, including retail, transportation, government and banking. Consequently, the applications where GIS have been combined with location analysis models for the solution of practical problems have increased rapidly over the last few years and are expected to continue increasing at an even higher rate in the future. As mentioned above, very often GIS are used as MC-SDSS in order to help decision makers in selecting feasible locations and then in choosing the most appropriate ones.

It is practically impossible to enumerate all the domains of applications as new possibilities for combining GIS with location models are constantly emerging. Various sources list literally hundreds of successful applications from the private as well as the public sector (see https://grindgis.com/blog/gis-applications-uses or https://gisgeography.com/gis-applications-uses/). In order to demonstrate the range of applications and the possibilities of expanding across diverse domains, we briefly present some key categories of applications and case studies where GIS and location models have been successfully combined to assist decision makers in practical problems. The objective of this section then is not to present a detailed list of all different applications but to highlight their diversity, flexibility and dynamic evolution and also to demonstrate the certain prospects of even more realistic applications in the immediate future. We also underline that the included examples of applications are not exhaustive of the papers developed in each category.

Emergency Services

Since the early development stages of GIS and their combination with location models, an area with a lot of applications has been the location of emergency services. This trend has continued in recent years as the ability to manage spatial data, perform queries and investigate alternatives is crucial for the location of facilities or servers that need to respond to emergencies. As far as location analysis models are concerned, most approaches rely on some type of covering models to identify the preferred locations. More specifically, Church and Li (2015) develop an integrative approach that uses cyber search, GIS, and spatial optimization to estimate the spatial efficiency of fire protection services in Los Angeles (LA) County, USA. The cyber search tool effectively conducts Web crawling to discover the exact locations for fire stations. The information is handled by means of GIS and the Set Covering Problem is solved to obtain the optimal locations of the fire stations. In another application concerning fire stations, Adesina et al. (2017) determined the optimal locations by solving the Maximum Covering Location Problem with a time and a distance range criterion. They apply their methodology to the city of Minna, Nigeria and conclude that under the time standard constraint, certain parts of the city are not appropriately covered. Similarly, Tali et al. (2017) employ a location-allocation model for the urban area in the city of Karnataka, India. A GIS was used to estimate the served as well as the unserved area of each fire station and to suggest a new configuration of stations that offer better coverage without excessive cost.

In a different type of service, Maghfiroh et al. (2017) investigate the current practices in the Emergency Medical Service (EMS) system of Dhaka, Bangladesh. They formulate the problem of locating EMS facilities and in particular ambulances, as a location-allocation problem. The response time and service coverage are optimized using ArcGIS location-allocation tools and modified K-means clustering. Their study highlights the peculiarities of EMS services location within a still developing city along with the inevitable resource constraints. McCormack and Coates (2015) propose a simulation model to optimize ambulance fleet allocation and use GIS to model response times along the arcs of the transportation network while focusing on increasing patient survival rates. They conduct their analysis on a dataset with millions of EMS call records provided by London Ambulance Services and analyze a variety of scenarios with different resource configurations. Dibene et al. (2016) use geo-spatial tools to model the demand for EMS in Tijuana, Mexico, and use integer linear programming to solve an ambulance location problem. They consider over 10.000 EMS calls, but use hierarchical clustering to simplify the demand model for optimization purposes. Esmaelian et al. (2015) also consider the EMS station location problem and propose a spatial DSS which is an integration of GIS with PROMETHEE IV. Finally, Lei et al. (2016) consider a generalized version of the problem known as Vector Assignment Ordered Median Problem (VAOMP), and propose a Tabu Search based heuristic approach to solve it. They discuss many dimensions of implementing the solution as a generalized tool in a GIS framework, replacing existing p-median based solution approaches in GIS software.

A different type of service is discussed in Fraser et al. (2018) who present a method for the siting of official cooling center facilities that offer protection in cases of extreme heat waves. Methods related to the Maximal Covering Location Problem are utilized to addresses various issues. In particular they show how disparate and large datasets, describing neighborhood level heat vulnerability and residential level access to public air-conditioned spaces, may be used to locate cooling centers more effectively. In addition, they evaluate the efficiency of the current network, indicating which facilities should be expanded in each count. The method is applied in Los Angeles County, CA and Maricopa County, AZ, USA.

Another active research and practice area in emergency services is disaster relief and humanitarian logistics. For instance, Xu et al. (2016) propose a multi-criteria location model for locating earthquake evacuation shelters. They use GIS to calculate spatial coverages of each candidate shelter location using a road network as well as the population capacity of the shelter. Similarly, Kılcı et al. (2015) also consider locating earthquake evacuation shelters and propose a mixed integer linear programming based methodology to select their locations. They obtain road network distances from ArcGIS using the Network Analyst extension and use GIS to also visualize solutions they produce.

Energy and Environment

Due to a wider policy recognition about the role of renewable energy to reduce greenhouse gas emission, in the recent years researchers have paid much attention to methodologies for identifying suitable locations for these sources of energy.

With this aim in view, GIS can support decision makers considering its ability to combine and to integrate different kinds of data and information. Of the various renewable energy resources, wind is highly favored as a result of the competitive cost of produced energy and availability of wind resources in almost every region. For this reason, applications on wind farm location and design have been gaining a significant interest. The suitability of locations for siting wind farms is mainly based on the presence of wind speeds that should be adequately mapped on a given study area. To this end, GIS can represent an important tool in combination with various MCDM methods, especially to generate criteria scores, to select potential sites and to generate constraints to extract suitable sites. For instance, Ayodele et al. (2018) propose an interval type-2 fuzzy Analytic Hierarchy Process (AHP) method, Gigovic et al. (2017) and Villacreses et al. (2017) combine different MCDM methodologies while Sanchez-Lozano et al. (2016) adopt a fuzzy TOPSIS approaches. Bina et al. (2018) and Noorollahi et al. (2016) show the capability of GIS to cross a numerous set of information appropriately organized into layers, to provide support for the site selections. Shaheen and Khan (2016) analyze large datasets originating from potential sites for wind turbines in their search for the best site in Pakistan. Due to the number of features and the magnitude of the data collected, they apply a principal component approach for finding the most relevant set of attributes, and use the resulting input variables in a multiple regression model. In a related study, Vasileiou et al. (2017) propose a GIS-based decision analysis approach for conducting site selection of offshore wind and wave energy systems in Greece. They develop a GIS database with the relevant data collected for geographical areas where these systems can be installed, and evaluate these areas using AHP. The approach is very similar even when the problem concerns the selection and the ranking of candidate areas for solar farms deployment (Tahri et al. 2015; Watson and Hudson 2015; Sindhu et al. 2017).

Besides these most popular applications, other proposals are provided in which GIS represent the fundamental support to locate renewable facilities. For instance, Franco et al. (2015) illustrate a multi-criteria decision problem to identify the most suitable facility locations for biogas plants on the basis of the positions of a set of farms considered as sources of biomass. In this case GIS is used for measuring the attributes of the alternatives according to a given set of criteria that are aggregated through a combined AHP-Fuzzy methodology. The same problem is addressed by Hohn et al. (2014) that use GIS to elaborate and provide data needed to solve a p-median problem.

Mohib-Ul-Haque Khan et al. (2018) present a methodology for determining suitable locations for waste conversion facilities considering waste availability as well as environmental and social constraints. A GIS allows to identify the most suitable areas and to screen out unsuitable lands while AHP is used for a multi-criteria evaluation of the relative preferences of different environmental and social factors. A case study is conducted for Alberta, Canada, by performing a province-wide waste availability assessment. The selection of appropriate areas to build an incinerator to serve healthcare facilities located in Kenya is the objective of an MCDM approach combining AHP, VIKOR and PROMETHEE methodologies (Ali

Hariz et al. 2017) in which GIS is exploited to eliminate unsuitable land that do not satisfy given criteria.

Gwak et al. (2017) present a framework for the selection of optimal locations for green roofs to achieve a sustainable urban ecosystem. The proposed framework selects building sites that can maximize the benefits of green roofs, based not only on the socio-economic and environmental benefits to urban residents, but also on the provision of urban foraging sites for honeybees. The final building sites are selected by solving the maximal covering location problem to determine the optimal locations for green roofs as urban honeybee foraging sites.

Finally, Hsieh et al. (2015) consider the problem of determining new air quality monitoring station locations by collecting and analyzing data from a sparse set of stations. They propose an entropy minimization model to determine the best locations for the new stations and evaluate their approach with data from Beijing.

Financial Services

According to de Villiers et al. (2016), Financial Service Providers (FSPs) use a variety of GIS layers to decide where to locate new branches or agents. GIS data allows an FSP to identify locations for investment based on existing infrastructure, mobile coverage and the socio economic conditions in the area. The decision to determine new locations also involves other information such as the distance to other branches or agents and the concentration of similar services in the catchment area. The analysis takes into account customer travel elasticity (which basically expresses how many branches the customer is willing to pass before arriving at their preferred bank branch) and potential physical barriers (rivers, highways, etc.) that prevent customers who appear close to a service point from accessing it. In another application, a South African bank uses the locations of bus or minibus-taxi ranks as a leading indicator for the placement of ATMs. Moreover, network optimization analysis is performed periodically to determine areas that were once profitable but now no longer require as many service points or perhaps different types of service points (e.g. Bank vs ATM). Bozkaya (2017) reports a GIS-based implementation by a private bank in Turkey, where the bank uses its underlying datasets for existing private and corporate customers, their home and work locations, merchants with its POS (point-of-sale) devices and certain other population statistics to forecast demand spatially for its branch and ATM services. They then use the gravity-based models implemented within the GIS application in conjunction with spatial regression models to select the best locations from a grid overlay to locate new ATMs and branches. Figure 19.1 reflects the two visualizations of this analysis, where the left image reflects a grid-based ATM suitability heatmap (red = more revenue-generating locations) and the right image represents a district-based branch suitability map.

In another implementation, Bozkaya (2018) proposes a clustering application for insurance agent segmentation for a major insurance company in Turkey, which takes into account spatial distribution of existing agents, the points of interest (POI) around them and the general economic indicators to assess the business potential of agents. Recommendations are then made as to which out-of-network agents located in what parts of the city, province and country should be pursued as new agents. As

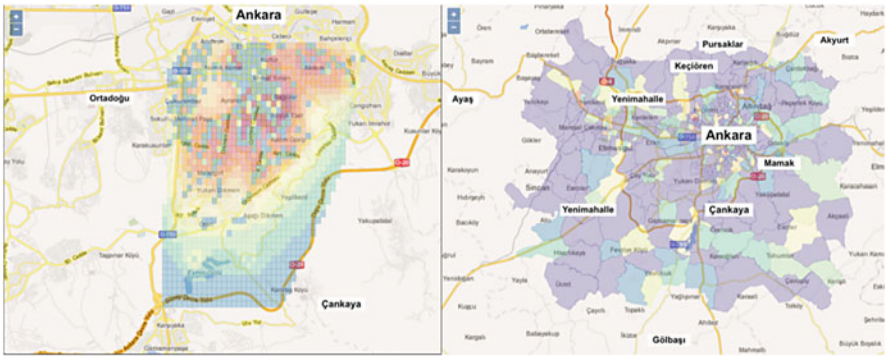


Fig. 19.1 Visualizations for ATM suitability (left) and branch suitability (right) analysis

already mentioned, Suhara et al. (2019) use GIS in a gravity model to analyze data concerning an application for a private bank.

Health Care

Another domain where the combination of GIS with location analysis models results in numerous applications is the location of health care facilities. As an indication, we refer to the study by Dodson et al. (2017) who examine the initial deployment of antiretroviral therapy in a particularly hard-hit region of Mozambique. Their analysis employs GIS with location-allocation modeling to examine alternative definitions of need for rural populations and how they might impact the allocation of this vital health service. The main conclusion is the fact that the definition of need matters when allocating limited healthcare resources and the use of need-based metrics can help ensure a significantly better distribution of services. Luis and Cabral (2016) also study healthcare accessibility in Mozambique and with the help of GIS identify accessibility problems across the country on various modes of transportation, On a similar note, Jankowski and Brown (2014) examine the effects of aggregating population demand for primary health care, ranging from census tract to aggregated census block, on estimates of primary health care accessibility. They employ GIS to manage different spatial representations of aggregated demand and incorporate them into a location-allocation model in order to determine a measure of accessibility represented by the unmet demand for primary health care services. The model is implemented for the U.S. State of Idaho, based on the allocation of Idaho residents' demand for primary health care to the state's existing primary health care facilities. Another application concerning accessibility to health care facilities is the work by Polo et al. (2015) who integrated location-allocation and spatial accessibility models in GIS to develop a comprehensive strategy proposal for assisting in the spatial planning of public health care services and for facilitating the population's accessibility to different public health interventions. The proposed methodology is evaluated using the data from the public sterilization program for the dogs and cats of Bogota, Colombia. Zhu et al. (2016) consider the problem of optimally locating trauma centers in Shenzen, China, on the basis

of a hierarchical location-allocation model. They analyze the spatial distribution of trauma demand and apply an ant-colony approach to determine the proposed solutions in combination with ArcGIS network analyst extension tools. Kim et al. (2018) use QGIS for conducting spatial analysis to determine accessibility of healthcare service locations to citizens in the city of Seoul, Korea. They propose accessibility measures and with the help of GIS identify vulnerable regions of low accessibility. They also conduct regression analysis to explore the relationship between accessibility and demographic indicators such as income. A similar study is performed for the city of Jeddah, Saudi Arabia by Murad (2018), who uses ArcGIS to assess the variations in healthcare accessibility. Using a geodatabase with a transportation network, the locations of service facilities and the population distribution, Murad (2007, 2018) reports the relative imbalance throughout the city between locations of facilities and potential demand for service. Bruno et al. (2018) exploit QGIS capabilities to support decisions related to the reorganization of regional Blood Management systems in Italy. In particular, they solve a version of the covering model to identify the potential catchment areas of blood stations and blood centers, in terms of attracted donors. GIS is also used to generate different scenarios according to various calibration parameters.

Public Services

Different types of applications involve the use of GIS to support decision making process in the field of provision and organization of services in the public sector. For instance, through the Bureau of Geographic Information (MassGIS), the Commonwealth of Massachusetts has created a comprehensive, statewide database of geospatial information that have been used in a large variety of applications concerning the merging of State offices in underused facilities, school re-districting decisions, etc. For an overview of MassGIS and a detailed list of the implemented projects, see https://www.mass.gov/orgs/massgis-bureau-of-geographic-information.

Many problems related to public service organization in various sectors (school, hospitals, justice, waste) are formulated in terms of districting models (Caro et al. 2004; Kalcsics et al. 2005). As this class of problems has a strong spatial component, it is natural to integrate models and algorithms in a GIS. In particular, in this context they can be used to generate and/or analyze instances, to represent produced districts maps, to verify the contiguity property, that is a distinguishing feature in this class of problems. A typical application of districting models is represented by the political districting, i.e. the formation of districts for the election of political representatives. Solutions to this problem are often characterized by partisan distortion with the aim of building districts with peculiar racial and socio-economic characteristics that can offer political advantage in a vis-a-vis competition. This phenomenon is known as "gerrymandering" from the name of Governor Elbridge Gerry which, in earlier nineteenth century, notoriously manipulated the shape of the political district of Boston to secure victory. With reference to this problem, Shapiro and Bliss (2016) use GIS to examine the remapping of Chicago from early 2012 to assess how potentially gerrymandered districts were formed in relation to the distribution of

racially homogeneous groups. In this context GIS is used to measure the extent and intensity of gerrymandering, focusing not only on the irregularity of a district but also on the changes occurred over the time. Bruno et al. (2017a,b) address a problem of the territorial reorganization aimed at redefining boundaries of Italian administrative districts, with the aim of reducing the administrative sites, in a general context of spending review. These studies are performed in order to show how the application of various kinds of models can lead to different scenarios. In particular GIS are fundamental to measure appropriate indicators to compare solutions provided by different models.

Masron et al. (2016) present the conceptual design and application of a web-based tourism decision support system concerning the Langkawi Island, Malaysia. The system offers online interactive maps that provide information on hotels, restaurants, shopping places, public facilities and others tourist attractions in the island. It can also assist local retailers to reach potential customers and at the same time help customers to plan their trip or visit that meets their preferences.

In fact, the ability to efficiently handle spatial data and use it for supporting decisions has been recognized by some educational authorities as an essential skill for pupils in primary and secondary education. For instance, Kolvoord et al. (2017) present a program in Virginia, USA that provides opportunities for high school students to become deeply immersed in geospatial technologies and spatial thinking and problem-solving. This program features mobile technologies and location-based services (LBS) in students' coursework and projects and helps them build their spatial thinking and problem-solving skills.

Another interesting field of application in which the use of GIS is strongly increasing concerns the security sector. A first aspect is related to the crime mapping in which GIS is a fundamental tool to cross data in order to detect crime hotspots, i.e. spatial locations that are good targets for police. Jefferson (2017) describes the evolution of the development of GIS in US police departments and, in particular, focuses on the Chicago police's digital mapping application, CLEARmap. In the same field Bunch et al. (2017) underline the importance of GIS in techniques of searching for missing persons as tool for the spatial connections between the location where the victim was last seen (VLS) and the body recovery site (BR). Khalid et al. (2018) show a method, based on spatial statistics provided by a GIS, for spatio-temporal hotspots mapping of crime events distributed over the road network (network constrained crimes) to discover high-density road segments. The methods are applied to the city of Faisalabad, Pakistan.

Smart Cities Organization

Given that the majority of residents in big cities own smartphones, several municipal authorities exploit the huge amount of location related data generated by these devices in order to offer a variety of services to their citizens. The potential for such applications is virtually endless, given the capabilities of smartphones and wearable devices. It is interesting to note that the Ash Center for Democratic Governance and Innovation at Harvard Kennedy School launched the Data-Smart City Solutions project which seeks to promote the combination of integrated, cross-agency data

with community data to better discover and preemptively address civic problems. The project web site lists several cases where location data has been used by cities to solve a wide spectrum of location problems ranging from crime prevention to public works. For instance, the police department in Huntington Beach, CA, is monitoring real-time social media activity and uses this data to identify locations where trouble might start. If they anticipate a problem, they can deploy officers or contact on-site security at the location thus making more effective use of their limited resources. In Boston, the authorities make use of data generated by an app called Street Bump that residents can install on their phones. The app collects vibration data to identify patterns associated with potholes. The phone geo-tags possible potholes and uploads these sites to an aggregation system, thus providing the authorities with an updated map of the city's roadways. For a more detailed description of such applications, see https://datasmart.ash.harvard.edu/news/article/learning-from-location-806.

There are other public initiatives and commercial applications where the location-based information is collected or contributed by the citizens for the improvement of public services. The smart citizen initiative (https://smartcitizen.me), which is launched in various locations around the globe, helps cities and communities contribute to data collection and environmental monitoring via smart sensor-based devices. The resulting data about pollution, noise and traffic levels are shared with the community for taking necessary actions for improving the environment. Another research group at MIT Media Lab (https://www.media.mit.edu/groups/city-science) aims to formulate strategies that make cities more livable by analyzing extensive datasets on people's mobility, telecommunication and commercial activities at various levels of resolution. The spatio-temporal databases collected by this group are analyzed to conduct simulations for determining more efficient configurations for workplaces, communities and even entire cities, and offering incentives to their respective inhabitants. Apps with citizen-contributed location-based data also play a role in achieving better efficiencies in day-to-day life. One such application Waze (https://www.waze.com) provides up-to-date traffic and journey information via data shared by its users on congestions, road closures, one-time events, or even "police traps", all through a mobile mapping interface. For further applications and services, see http://geoawesomeness.com/lbs-and-smart-city-initiatives-on-a-global-scale/.

Supply Chain Management

Since today's supply chains tend to be global and automated, the availability of accurate date and the ability to process it effectively can have a significant impact on the performance of a firm's supply chain. This has been long recognized by a lot of firms that operate GIS modules in order to design, monitor and improve their supply chains. For instance, S Group is the largest retailer in Finland, with business sectors that include grocery stores, service stations, utility goods, hotels, restaurants, tourism, car dealerships, and agricultural trade stores. S Group's development and support organization, SOK Corporation, adopts Esri ArcGIS and Business Analyst to automate profiling reports for each of its 1600 business locations. These reports are then analyzed using the ArcGIS system. SOK can assess the area of influence

of any business location; forecast annual sales volumes; and improve network planning, including both opening and closing stores. ArcGIS also helps SOK's marketing division understand its customer base and better target its distribution of catalogs. Similarly, GIS technology from Esri helps Werner Enterprises keep track of its fleet of more than 9000 trucks. Using ArcGIS and a tractor tracking device traditionally used by long-haul trucking companies, Werner can now bill mileage to customers more accurately and route its fleet more efficiently. Werner implemented Esri ArcGIS for Server, which integrates geographic location into business data to better manage information. Werner uses the software to keep track of its very large fleet and outfits its trucks with transmitters that provide two-way text and data communications between the vehicles and Werner's headquarters in Omaha, Nebraska, USA. For more GIS applications along these lines, see the relevant documentation at esri.com/retail.

Transportation

Due to the need of using and integrating socio-economic spatial data, transportation planning is one of the most traditional fields of application of GIS. GIS functionalities, indeed, are useful tools for data collection which is a fundamental activity to obtain reliable measures to support planning and management decisions. In particular, composite indicators of accessibility and/or efficiency are usual metrics to evaluate transportation facility or network. In this context Chen et al. (2018) propose the introduction of an index system built for quantifying public transport supply from multiple dimensions (i.e., service coverage, service level, and service accessibility) through GIS. Saghapour et al. (2016) present a conceptual framework for the definition of a Public Transport Accessibility Index (PTAI) in which GIS plays the central role to provide and integrate data of different categories. The proposed index is compared to others available in literature considering the case of Melbourne, Australia. Similarly, Magalhaes (2016) develops a spatial coverage index for assessing national and regional transportation infrastructures which can be applied to different modes and geographical aggregations (countries, states, municipalities, or any other area), whose calculations require various GIS functionalities.

However, as accessibility is essentially a dynamic concept characterized by dramatic variations throughout the day and/or week, this opens up a growing field of research in which new data sources (satellite navigations, websites, social networks) may be dynamically combined and elaborated through GIS (see for instance, Gomez et al. 2018).

The role of GIS within transportation facility location models is essential when detailed information should be used at urban level, especially in the context of innovative and sustainable transportation modes. Terh and Cao (2018) propose a GIS based path planning support framework incorporating multiple criteria to address the location of new cycling paths. They use GIS to build a composite indicator to rank the areas where these additional cycling paths are prioritized and apply the model to the case of Singapore. In the same context MCDM methods supported by GIS are also used to tackle the problem of bike-share station site

selection. In particular Kabak et al. (2018) attempt to evaluate the present status of Iran's Mashhad City bike-share stations ranking potential sites applying AHP in combination with GIS. Similarly, Guerreiro et al. (2018) develop a three-step method to design and compare cycling networks. Wang et al. (2016) make use of biking rental and trajectory data in the city of Taipei to identify, through a spatio-temporal hot spot analysis, bike-lacking and bike rack-lacking locations. They then use this information to assess the locations of existing bike rental stations and determine most suitable locations for additional installation of rental stations.

Other interesting applications concern the proposal of a GIS-based fuzzy MCDA approach to determine the optimal site of electric vehicle charging stations from environmental/geographical, economic and urbanity perspectives (Erbas et al. 2018) and the solution of a relocation optimization problem of electric cars in one-way car-sharing systems by a bi-level tabu search algorithm, where GIS tools are used to estimate station catchment areas (see Ait-Ouahmed et al. 2018).

Finally, Burciu et al. (2015) locate a hub terminal linking producers of cereals to be exported by naval transportation with the Romanian fluvial-maritime ports of Galati and Braila. A GIS environment is used to integrate and analyze the relevant data and a location-allocation model is implemented to determine the optimal location of the hub terminal.

19.6 Conclusions

In the previous edition of this chapter, we had anticipated that advancements in the technology of GIS and the prospects of linking them with location science would be rapid. Looking back over the years, it can be argued that developments have been even more accelerated than our predictions. Indeed, GIS are no longer viewed as a simple input-output tool in location science but are recognized as a domain which is closely related to certain types of location analysis models and which offers opportunities for enhancing and further advancing these models.

The technological advances in GIS caused significant changes in their functionality and user requirements. What used to be the privilege of large corporations or well-trained specialists is rapidly becoming available to a much wider range of users. As a result, GIS are perceived as services and issues such as accessibility and integration of multiple data sources and formats are becoming critical.

The development with the biggest impact in the capabilities of GIS and their connection with location science, has been the explosive growth in the use of smartphones and mobile devices and the resulting huge availability of spatial data. In connection with the emerging methodologies and techniques for handling big data, this development has inspired a multitude of applications involving location science models utilizing large spatial datasets. The transition towards a digitized society implies that a large part of these location-based data is contributed by the users themselves. In turn, this evolution will probably increase the popularity of GIS

even further as more and more people wish to take advantage of their capabilities and benefit from the available services.

The integration of GIS with high level programming languages and environments such as Python or R has enabled developers to customize GIS tools and employ them to perform a variety of tasks in different location science settings. As these programming environments become more accessible to non-experts, we expect this trend to continue in the future and expect that more sophisticated techniques will be incorporated into commercial as well as open source GIS software for solving particular location science problems.

The class of location science problems where the linkages with GIS are most evident is probably the class of coverage problems. The expanding capabilities of GIS have contributed significantly to the formulation of new models and the solution of problems that are challenging from a computational point of view such as the continuous version of the Maximal Coverage Location Problem (MCLP). It is not surprising that a large part of location science applications involving the use of GIS fall within the general class of coverage problems. Such applications typically include the location of emergency services or health care facilities. A particular class of applications that is becoming more evident refers to the location of disaster relief facilities or the configuration of humanitarian logistics networks. The use of GIS tools for assessing accessibility, estimating coverage or exploring paths has facilitated the development of methods that produce robust solutions which may perform well under different scenarios. As such natural or man-made disasters (e.g. forest or urban fires) cannot be avoided, we expect research in this area to continue as modern societies attempt to minimize the effects of these disasters.

The expansion of smartphones and wearable devices and the availability of spatial data virtually worldwide have also brought a proliferation of applications aiming to offer better services to citizens in the context of smart cities. Various public initiatives and commercial applications have been developed that make use of location-based information and analyze it in order to improve certain aspects of city life such as the reduction in noise and pollution levels, balancing of traffic flows, prevention of crime etc. In this setting, it has become desirable if not necessary for different agencies to be able to seamlessly share and integrate geographically referenced data. We expect applications along these lines to continue growing as citizens expect a better quality of life within modern digitized societies.

In order to exploit the full potential of GIS as a DSS, it is important to successfully combine them with decision making methodologies and in particular MCDA techniques that consider multiple conflicting criteria. Since the technical capabilities of GIS have evolved considerably over the last few years, the development of several such methods is to be expected. Given that programming environments such as Python or R can be used to develop systems that employ GIS tools, we expect this trend to continue at an accelerating pace in the future.

In the concluding section of the previous edition of this chapter, one of the main questions posed was whether or not it would be beneficial for location analysts to invest time and effort in GIS. As far as the theoretical aspects of location science are concerned, some advancements have been recorded mainly in the class of

coverage problems. However, in terms of practical approaches, the vast number of applications that have appeared since the publication of that chapter conclusively confirms that the investment is worthwhile. It remains a challenge to exploit the ever expanding capabilities of GIS to enhance existing models or develop new ones that reflect more realistic aspects of location science.

References

Adesina EA, Odumosu JO, Morenikeji OO, Umoru E, Ayokanmbi AO, Ogunbode EB (2017) Optimization of fire stations services in Minna Metropolis using maximum covering location model (MCLM). J Appl Sci Env Sustain 3(7):172–187

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. Accessed 23 Oct 2018

Ait-Ouahmed A, Josselin D, Zhou F (2018) Relocation optimization of electric cars in one-way car-sharing systems: modeling, exact solving and heuristics algorithms. Int J Geogr Inf Sci 32:367–398

Ali Hariz H, Donmez CC, Sennaroglu B (2017) Siting of a central healthcare waste incinerator using GIS-based multi-criteria decision analysis. J Clean Prod 166:1031–1042

Al-Marwani HA (2014) An approach to modeling and forecasting real estate residential property market, Ph.D. Thesis, Brunel University

Altaweel M (2017) The use of python in GIS, in GISLounge. Accessed 31 July 2018

Ayodele TR, Ogunjuyigbea ASO, Odigiea O, Mundab JL (2018) A multi-criteria GIS based model for wind farm site selection using interval type-2 fuzzy analytic hierarchy process: the case study of Nigeria. Appl Energ 228:1853–1869

Barbati M, Bruno G (2017) Exploring similarities in discrete facility location models with equality measures. Geogr Anal https://doi.org/10.1111/gean.12151

Barbati M, Piccolo C (2016) Equity measure properties for location problems. Optim Lett 10:903–920

Bina SM, Jalilinasrabady S, Fujii H, Farabi-Asl H (2018) A comprehensive approach for wind power plant potential assessment, application to northwestern Iran. Energy 52:77–88

Bozkaya B (2017) Using a GIS-based framework to determine ATM and branch locations. Personal Communication

Bozkaya B (2018) A clustering approach to insurance agent segmentation and identifying new target insurance agents. Personal Communication

Brandeau ML, Chiu SS (1989) An overview of representative problems in location research. Manage Sci 25:645–674

Bruno G, Giannikos I (2015) GIS and location. In: Laporte G, Nickel S, Saldanha da Gama F (eds) Location science. Springer International Publishing, New York, pp 509–536

Bruno G, Diglio A, Melisi A, Piccolo C (2017a) A districting model to support the redesign process of Italian Provinces. In: International conference on optimization and decision science (pp. 245–256). Springer, Cham

Bruno G, Genovese A, Piccolo C (2017b) Territorial amalgamation decisions in local government: models and a case study from Italy. Socio Econ Plan Sci 57:61–72

Bruno G, Diglio A, Piccolo C, Cannavacciuolo L (2018) Territorial reorganization of regional blood management systems: evidences from an Italian case study. Omega. https://doi.org/10.1016/j.omega.2018.09.006

Bunch AW, Kim M, Brunelli R (2017) Under our nose: the use of GIS technology and case notes to focus search efforts. J Forensic Sci 62:92–98

Burciu S, Stefanica C, Rosca E, Dragu V, Rusca F (2015) Location of an intermediate hub for port activities. In: IOP conference series: materials science and engineering, vol 95, No. 1. IOP Publishing, Bristol, p 012064

Cai H, Jia X, Chiu ASF, Hu X, Xu M (2014) Siting public electric vehicle charging stations in Beijing using big-data informed travel patterns of the taxi fleet. Transp Res Part D: Transp Environ 33:39–46

Caro F, Shirabe T, Guignard M, Weintraub A (2004) School redistricting: embedding GIS tools with integer programming. J Oper Res Soc 55(8):836–849

Chandel K (2017) Location intelligence is making geospatial technology go main-stream. https://www.geospatialworld.net/article/location-intelligence-geospatial-collaborations/. Accessed 10 Aug 2018

Chang K-T (2018) Introduction to geographic information systems 9e. McGraw-Hill Education, New York

Chen Y, Bouferguene B, Li HX, Liu H, Shen Y, Al-Hussein M (2018) Spatial gaps in urban public transport supply and demand from the perspective of sustainability. J Clean Prod 195:1237–1248

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

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, Li W (2015) Estimating spatial efficiency using cyber search, GIS, and spatial optimization: a case study of fire service deployment in Los Angeles County. Int J Geogr Inf Sci 30(3):535. https://doi.org/10.1080/13658816.2015.1083572

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

Cooper L (1963) Location-allocation problems. Oper Res 11:311–343

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 4e. Winchelsea Press, Winchelsea

de Villiers L, Motsomi A, Berkowitz B (2016) 7 applications of GIS data by financial service providers, insight2impact (i2i). http://www.i2ifacility.org/, Accessed 10 Aug 2018

Dibene JC, Maldonado Y, Vera C, de Oliveira M, Trujillo L, Schütze O (2016) Optimizing the location of ambulances in Tijuana, Mexico. Comput Biol Med 80:107–113

Dodson ZM, Agadjanian V, Driessen J (2017) How to allocate limited health care resources: lessons from the introduction of antiretroviral therapy in rural Mozam-bique. Appl Geogr 78:45–54

Drezner Z, Wesolosky GO (1980) Single facility l_p distance minimax location. SIAM J Algebra Discr 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

Erbaş M, Kabak M, Ozceylan E, Çetinkaya C (2018) Optimal siting of electric vehicle charging stations: a GIS-based fuzzy multi-criteria decision analysis. Energy 163:1017–1031

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

Esmaelian M, Tavana M, Arteaga FJS, Mohammadi S (2015) A multicriteria spatial decision support system for solving emergency service station location problems. Int J Geogr Inf Sci 29(7):1187–1213

Everts S (2016) Information overload. Distillations 2(2):26–33. Retrieved 30 July 2018

Farahani RZ, SteadieSeifi M, Asgari R (2010) Multiple criteria location problems: a survey. Appl Math Model 34:1689–1709

Ferretti V, Montibeller G (2016) Key challenges and meta-choices in designing and applying multi-criteria spatial decision support systems. Decis Support Syst 84:41–52

Fortelius M, Žliobaitė I, Kaya F, Bibi F, Bobe R, Leakey L, Werdelin L (2016) An econometric analysis of the fossil mammal record of the Turkana Basin. Philos Trans R Soc B 371(1698):20150232

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

Franco C, Bojesen M, Hougaard JL, Nielsen K (2015) A fuzzy approach to a multiple criteria geographical information system for decision support on suitable locations for biogas plants. Appl Energ 140:304–315

Fraser AM, Chester MV, Eisenman D (2018) Strategic locating of refuges for extreme heat events (or heat waves). Urban Clim 25:109–119

Gigovic L, Pamucar D, Bozanic D, Ljubojevic S (2017) Application of the GIS-DANP-MABAC multi-criteria model for selecting the location of wind farms: a case study of Vojvodina, Serbia. Renew Energy 103:501–521

Goldman AJ (1971) Optimal center location in simple networks. Transp Sci 5:212–221

Gomez CD, González CM, Osses M, Aristizábal BH (2018) Spatial and temporal disaggregation of the on-road vehicle emission inventory in a medium-sized Andean city. Comparison of GIS-based top-down methodologies. Atmos Environ 178:142–155

Goodchild MF (2010) Twenty years of progress: GIScience in 2010. J Spat Inf Sci 1:3–20

Guerreiro TCM, Providelo JK, Pitombo CS, Rui Ramos AR, Rodrigues da Silva AN (2018) Data-mining, GIS and multicriteria analysis in a comprehensive method for bicycle network planning and design. Int J Sust Transp 12:179–191

Gwak JH, Lee BK, Lee WK, Sohn SY (2017) Optimal location selection for the installation of urban green roofs considering honeybee habitats along with socio-economic and environmental effects. J Environ Manag 189:125–133

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

Hohn 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 Energ 113:1–10

Hsieh H-P, Lin S-D, Zheng Y (2015) Inferring air quality for station location recommendation based on urban big data. Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp 437–446

Jankowski P, Brown B (2014) Health care accessibility modeling: effects of change in spatial representation of demand for primary health care services. Quaestiones Geographicae 33(3):39–53

Jefferson BJ (2017) Digitize and punish: computerized crime mapping and racialized carceral power in Chicago. Environ Plann D 35(5):775–796

Kabak M, Erbas M, Çetinkaya C, Ozceylan E (2018) A GIS-based MCDM approach for the evaluation of bike-share stations. J Clean Prod 201:49–60

Kalcsics J, Nickel S, Schröder M (2005) Towards a unified territorial design approach—applications, algorithms and GIS integration. Top 13:1–56

Khalid S, Shoaib F, Qian T, Rui Y, Bari AI, Sajjad M, Shakeel M, Wang J (2018) Network constrained spatio-temporal hotspot mapping of crimes in Faisalabad. Appl Spat Anal 11:599–622

Kılcı F, Kara BY, Bozkaya B (2015) Locating temporary shelter areas after an earthquake: a case for Turkey. Eur J Oper Res 243:323–332

Kim Y, Byon Y-J, Yeo H (2018) Enhancing healthcare accessibility measurements using GIS: a case study in Seoul, Korea. PLoS One. https://doi.org/10.1371/journal.pone.0193013

Kolvoord R, Keranen K, Rittenhouse P (2017) Applications of location-based services and mobile technologies in K-12 classrooms. ISPRS Int J Geo Inf 6(7):209

Lee J-G, Kang M (2015) Geospatial big data: challenges and opportunities. Big Data Res 2(2):74–81

Lee K, Ganti RK, Srivatsa M, Liu L (2014) Efficient spatial query processing for big data. In: Proceedings of the 22nd ACM SIGSPATIAL international conference on advances in geographic information systems – SIGSPATIAL'14, pp 469–472

Lei TL, Church RL, Lei Z (2016) A unified approach for location-allocation analysis: integrating GIS, distributed computing and spatial optimization. Int J Geogr Inf Sci 30(3):515–534

Luis AA, Cabral P (2016) Geographic accessibility to primary healthcare centers in Mozambique. Int J Equity Health 15:173

Magalhaes MT (2016) Spatial coverage index for assessing national and regional transportation infrastructures. J Transp Geogr 56:53–61

Maghfiroh MF, Hossain M, Hanaoka S (2017) Minimising emergency response time of ambulances through pre-positioning in Dhaka city, Bangladesh. Int J Log Res Appl 21(1):53–71

Malczewski J (2004) GIS-based land-use suitability analysis: a critical overview. Prog Plann 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

Masron T, Norhasimah I, Azizan M (2016) The conceptual design and application of web-based tourism decision support systems. Theor Empirica 11:21–35

McCormack R, Coates G (2015) A simulation model to enable the optimization of ambulance fleet allocation and base station location for increased patient survival. Eur J Oper Res 247:294–309

Miller G (2018) Teaching modern GIS. news.aag.org/2018/02/teaching-modern-gis/

Minieka E (1970) The m-center problem. SIAM Rev 12:138–141

Mintchev S (2014) User-defined rules made simple with functional programming. In: Abramowicz W, Kokkinaki A (eds), Business Information Systems. Springer International Publishing, New York, pp 229–240

Mohammed EA, Far BH, Naugler C (2014) Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends. BioData Mining 7(1):22. http://dx.doi.org/10.1186/1756-0381-7-22

Mohib-Ul-Haque Khan M, Vaezi M, Kumar A (2018) Optimal siting of solid waste-to-value-added facilities through a GIS-based assessment. Sci Tot Environ 610–611:1065–1075

Murad A (2007) A GIS application for modeling accessibility to health care centers in Jeddah, Saudi Arabia. In: Lai PC, Mak ASH (eds) GIS for health and the environment. Lecture notes in geoinformation and cartography. Springer, Berlin

Murad A (2018) Using GIS for determining variations in health access in Jeddah city, Saudi Arabia. Int J Geo Inf 7:254. http://dx.doi.org/10.3390/ijgi7070254

Murray AT (2010) Advances in location modeling: GIS linkages and contributions. J Geogr Syst 12:335–354

Murray AT (2016) Maximal coverage location problem: impacts, significance and evolution. Intern Reg Sci Rev 39:5–27

Noorollahi Y, Yousefi H, Mohammadi M (2016) Multi-criteria decision support system for wind farm site selection using GIS. Sustain Energy Tech Assess 13:38–50

Percivall G (2013) The power of location. http://www.opengeospatial.org/blog/1817 (April 2013). Open geospatial consortium

Pfeifer M, Kor L, Nilus R, Turner E, Cusack J, Khoo M, Chey VK, Ewers RM (2016) Mapping the structure of Borneo's tropical forests across a degradation gradient. Remote Sens Environ 176:84–97

Polo G, Acosta CM, Ferreira F, Dias RA (2015) Location-allocation and accessibility models for improving the spatial planning of public health services. PLoS One 10(3):e0119190

Provost F, Fawcett T (2013) Data science for business: what you need to know about data mining and data-analytic thinking, 1st ed. O'Reilly Media, Sebastopol

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

Saghapour T, Moridpour S, Thompson RG (2016) Public transport accessibility in metropolitan areas: a new approach incorporating population density. J Transp Geogr 54:273–285

Sanchez-Lozano JM, García-Cascales MS, Lamata MT (2016) GIS-based onshore wind farm site selection using fuzzy multi-criteria decision making methods. Evaluating the case of Southeastern Spain. Appl Energy 171:86–102

Shaheen M, Khan MZ (2016) A method of data mining for selection of site for wind turbines. Renew Sustain Energy Rev 55:1225–1233

Shapiro MA, Bliss D (2016) Rewards and consequences: redistricting on the Chicago City Council. Local Gov Stud 42:139–163

Shekhar S, Evans MR, Gunturi V, Yang K, Cugler DC (2014) Benchmarking spatial big data. In: Rabl T, Poess M, Baru C, Jacobsen H-A (eds) Specifying big data benchmarks. Springer, Berlin, pp 81–93

Silva S, Alcada-Almeida L, Dias LC (2014) Biogas plants site selection integrating multicriteria decision aid methods and GIS techniques: a case study in a Portuguese region. Biomass Bioenerg 71:58–64

Sindhu S, Nehraa V, Luthrab V (2017) Investigation of feasibility study of solar farms deployment using hybrid AHP-TOPSIS analysis: case study of India. Renew Sust Energ Rev 73:496–511

Suhara Y, Bahrami M, Bozkaya B, Pentland A (2019) Validating gravity-based market share models using large-scale transactional data. arXiv preprint arXiv:1902.03488

Tahri M, Hakdaoui M, Maanan M (2015) The evaluation of solar farm locations applying geographic information system and multi-criteria decision-making methods: case study in southern Morocco. Renew Sust Energ Rev 51:1354–1362

Tali JA, Malik MM, Divya S, Nusrath A, Mahalingam B (2017) Location–allocation model applied to urban public services: spatial analysis of fire stations in Mysore urban area Karnataka, India. Int J Adv Res Develop 2(5):795–801

Terh SH, Cao K (2018) GIS-MCDA based cycling paths planning: a case study in Singapore. Appl Geogr 94:107–118

Ting C-Y, Ho CC, Yee HJ, Matsah WR (2018) Geospatial analytics in retail site selection and sales prediction. Big Data 6(1). https://doi.org/10.1089/big.2017.0085

Tobler W (1970) A computer movie simulating urban growth in the Detroit region. Econ Geogr 46(2):234–240

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

Valdes-Dapena P (2011) GPS systems that save gas. http://money.cnn.com/2011/03/03/autos/navigation_gps_fuel_economy/, March 2011, CNN Money

Vasileiou M, Loukogeorgaki E, Vagiona DG (2017) GIS-based multi-criteria decision analysis for site selection of hybrid offshore wind and wave energy systems in Greece. Renew Sust Energ Rev 73:745–757

Verma S, Verma R K, Singh A, Naik N (2012) Web-based GIS and desktop open source GIS software: an emerging innovative approach for water resources management. In: Wyld DC et al (eds) Advances in computer science, Engineering & Application, AISC, vol 167, pp 1061–1074

Villacreses G, Gaona G, Martínez-Gomez J, Jijon DJ (2017) Wind farms suitability location using geographical information system (GIS), based on multi-criteria decision making (MCDM) methods: the case of continental Ecuador. Renew Energy 109:275–286

Wang J, Tsai CH, Lin PC (2016) Applying spatial-temporal analysis and retail location theory to pubic bikes site selection in Taipei. Transp Res A 94:45–61

Watson JJW, Hudson MD (2015) Regional scale wind farm and solar farm suitability assessment using GIS-assisted multi-criteria evaluation. Landscape Urban Plan 128:20–31

Wenkel KO, Berg M, Mirschel W, Wieland R, Nendel C, Köstner B (2013) LandCaRe DSS e an interactive decision support system for climate change impact assessment and the analysis of potential agricultural land use adaptation strategies. J Envir Mang 127:s168–s183

Xu J, Yin X, Chen D, An J, Nie G (2016) Multi-criteria location model of earthquake evacuation shelters to aid in urban planning. Int J Disaster Risk Reduct 20:51–62

Yao X, Zhu D, Yun W, Peng F, Li L (2017) A WebGIS-based decision support system for locust prevention and control in China. Comput Electron Agr 140:148–158

Zhu Y, Du Q, Tian F, Ren F, Liang S, Chen Y (2016) Location optimization using a hierarchical location-allocation model for trauma centers in Shenzhen, China. Int J Geo Inf 5(10):190

Chapter 20
Green Location Problems

Sibel A. Alumur and Tolga Bektaş

Abstract This chapter discusses aspects of sustainability and "green" that are relevant to and arise within location problems. More specifically, it describes ways in which some environmental criteria, in particular emissions, can be quantified and integrated with location models. The chapter also presents design problems in which location decisions arise as one of the key ingredients in improving the environmental performance of distribution systems.

20.1 Sustainability and "Green" in Location Problems

Facility location problems arise in the broader context of logistics network design, where the primary aim is to move freight from points of origin to points of destination. The activities relevant to production, transportation and distribution of the goods over a logistics network inevitably results in undesirable effects on the environment, generally referred to as *externalities*, which include the following:

- Logistics operations *deplete natural resources* that are needed for the underlying infrastructure and the various activities run on the network. The former includes construction of facilities such as hubs, depots or ports, as well as highways and railway tracks, relying on the extraction and the use of various materials needed for the construction, use of land on which they are built, and potential change or damage to the ecosystem around them. The activities themselves require energy for the movement of goods, primarily fuel.
- Most logistics activities induce some form of *noise* that is detrimental to human health and well-being. In addition, the vibration caused by the movement of

S. A. Alumur (✉)
Department of Management Sciences, University of Waterloo, Waterloo, ON, Canada
e-mail: sibel.alumur@uwaterloo.ca

T. Bektaş
University of Liverpool Management School, Liverpool, UK
e-mail: T.Bektas@liverpool.ac.uk

G. Laporte et al. (eds.), *Location Science*,
https://doi.org/10.1007/978-3-030-32177-2_20

goods, e.g., freight trains passing through cities, lorries operating in urban areas, may damage the buildings around them in the long run.

- *Pollution* is perhaps the most prominent of all externalities, not least for their direct and indirect impacts on the environment, but also the role they are deemed to play in climate change. Air pollution is caused by various types of gases, including greenhouse gases such as methane, ozone, carbon dioxide and nitrous oxide, others such as carbon monoxide, nitrogen oxide and sulphur oxide, and particulate matter, emitted by production and transportation activities. They are responsible for environmental impacts such as the greenhouse gas effect, acidification, summer smog and other toxic effects.
- *Accidents* occur predominantly within transportation, but also manufacturing and production environments, and are responsible for injury and death for all forms of life. Their severity could be far more significant in the context of obnoxious facilities and hazardous materials, with potentially disastrous consequences.

There are various ways to mitigate or lessen the externalities above within the broader context of logistics management, including technological solutions and management strategies, which is beyond the remit of this chapter. The interested reader is referred to McKinnon et al. (2015) and Psaraftis (2016) for a broader and a more in-depth treatment of the topic.

To achieve greener ways of operation in the particular context of facility location, it is pertinent at this point to differentiate between the two types of research directions listed below:

- Reducing externalities from a logistics network *by locating* facilities, caused by the amount of inbound and outbound of goods and the way in which they are moved within the network.
- Reducing externalities *from located* facilities that appear within a logistics networks, mainly from energy consumption (lighting, heating, cooling, use of equipment for production and handling of goods), water consumption, and use of land.

The rest of this chapter will focus on the first research direction above, namely those that concern decisions around facility location and the impact thereof on the environmental performance of the logistics networks within which the facilities are to be installed and used. The second research direction will not be covered within this chapter for two reasons. First, some of the decisions to reduce the externalities from facilities requires adopting new technologies, including improving energy efficiency (e.g., use of eco-friendly lighting), switching to energy-efficient mechanical equipment, harnessing green energy sources and incorporating sustainability considerations into building design (Baker and Marchant 2015). Second, even if non-technological solutions were to be adopted, the issues relate to the interior design, as opposed to the location of a facility, and involve decisions ranging from the type and shape of the construction, to optimizing layouts within facilities.

It is important at this point to stress on the fact that this chapter does not concern itself directly with the broader issues around *sustainability* apart from those that are indirectly linked with the *green* agenda. From an environmental perspective, our understanding of the former term following Jaehn (2016) revolves around the ability of the environment around a logistics system that allows the system to maintain its operations ad infinitum, or, in practical terms, for very long periods of time. It is obvious that, in the long run, the scarcity of the resources involved in running a logistics system makes it unsustainable ipso facto, in the strict sense of the word. Improving sustainability of a system may be achieved through novel solutions, including the use of unconventional technologies (such as use of alternative fuels or harnessing and using new sources of energy), and we will indeed touch on these aspects later in the chapter when discussing location problems in the context of alternative fuel vehicles. In contrast, greening of a given system, at least in the way that we interpret and treat in this chapter, will be to *lessen* or *mitigate* the environmental externalities of that system through better planning, and one that does not necessarily involve a fundamental change in the way that the operations are set up and run. One example in the context of facility location may simply be to change the number and location of facilities to improve the environmental performance of a distribution system.

20.2 Environmental Considerations in Location Problems

An explicit consideration of externalities within location problems is possible to the extent that their impacts are quantifiable, and that it is possible to estimate the quantity of the amount through analytical models as a function of decisions made within location problems. If such analytical models exist, then they can generally be integrated within existing models of various facility location problems. In the rest of this section, we will focus on emissions as the main environmental impact, not least given their prominence within environmental externalities, but also the relative easiness with which they can be quantified.

Emissions are, in most cases, proportional to the amount of energy consumed by a given logistics operation. In conventional road transport, for example, the amount of pollutants emitted is dependent on the amount of fuel consumed. This makes it easier to estimate the amount of emissions from a given operation, if the level of activity is known and there exist emission factors. This is the main principle behind what is known as *emission factor models*, for which there exist two types of models:

- The first type is used when the actual amount of energy or fuel α is known (e.g., in kWh or litres), which is then multiplied by the emission factor ϕ (e.g., in grams per kWh), yielding the total emissions $E = \alpha \times \phi$ for a given activity. This actual energy consumed can be calculated using historical data, such as readings

from storage tanks for lorries or electricity bills for facilities, and is therefore calculated retrospectively.

- If the actual amount of energy consumed cannot be calculated, or is not available, then one can resort to the second type of emissions model that uses average conversion factors that are pre-defined depending on the type of activity. The UK Government's Department for Business, Energy & Industrial Strategy defines the emission factors separately for fuel, electricity, heat, steam, passenger transport, freight land transport, sea transport and air transport, and for various types of gases such as carbon dioxide (CO_2), methane (CH_4) and nitrogen oxide (N_2O), and for different types of vehicles and for various load levels. For road transport, the emission factors are defined for one kg of CO_2 per vehicle kilometers traveled, or one kg of CO_2 per tonne.kilometer. For air freight, rail transport or sea transport, the emissions factors are defined for one kg of CO_2, CH_4 or N_2O per tonne.kilometer. The reader is referred to Hill et al. (2017) for further details.[1] For illustrative purposes, Fig. 20.1 shows the resulting amount of CO_2 emissions for two different types of vehicles under different load levels estimated by using the factors given by Hill et al. (2017) traveling from 10 to 100 km.

There exist other types of analytical models to estimate emissions that are more detailed as compared to emission factor models. One such type that is used within road transportation is the *macroscopic* or *average speed models*, a class of models that are primarily regression based, and use average speed v of a vehicle as a primary determinant to estimate emissions. One such model appears in an emissions inventory guidebook by the European Environment Agency (Ntziachristos et al. 2017), where hot emissions $E(v)$ (g/km) are calculated on the basis of average speed v (km/h) using the following generic expression,

$$E(v) = \left(\frac{a_1 v^2 + a_2 v + a_3 + a_4/v}{a_5 v^2 + a_6 v + a_7} \right) \beta, \tag{20.1}$$

where a_1–a_7 are coefficients that differ by fuel, vehicle class and engine technology,[2] and β is a correction factor applied, if necessary, to account for different types of road (i.e., urban, rural and highway). Figures 20.2 and 20.3 show the resulting CO emissions output by the average speed model (20.1) for different types of goods vehicles.

At a micro-level, a more detailed class of models is available that generally named as *microscopic* or *instantaneous emissions models*. These models are derived from mechanical physics of automobile engines, and take a significant number and range of parameters into account, such as vehicle characteristics (mass, drag force, rolling resistance, engine efficiency) as well as external factors (air density,

[1] For the full set of factors, see https://www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2017.

[2] For the full set of parameters, see https://www.eea.europa.eu/publications/emep-eea-guidebook-2016/part-b-sectoral-guidance-chapters/1-energy/1-a-combustion/1-a-3-b-i-1/at_download/file.

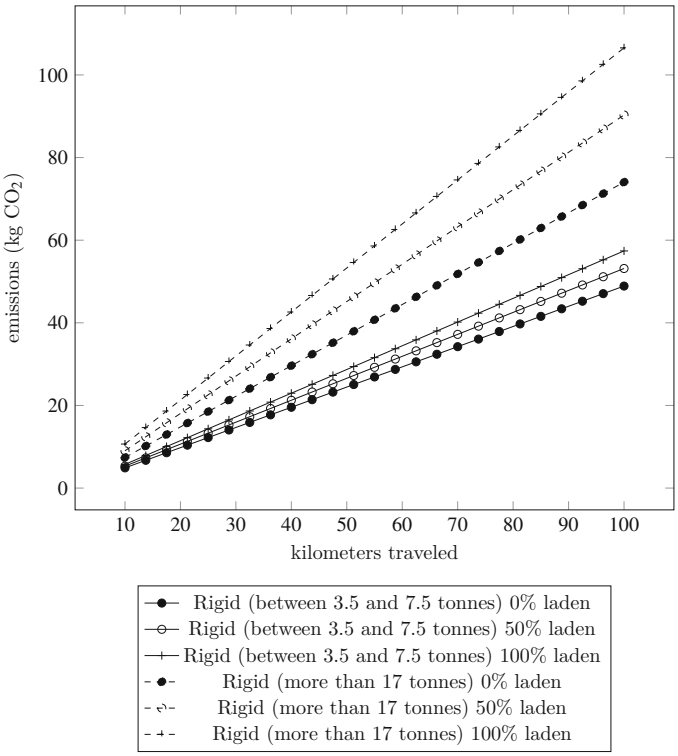


Fig. 20.1 CO_2 emissions for various goods vehicles estimated using the factor model

gravitational constant), in order to estimate emissions, often on a second-by-second basis. Further details on microscopic models can be found in Demir et al. (2014). However, as facility location problems typically involve strategic (and sometimes tactical) decisions made to last over relatively long time-spans, micro-level models may be too detailed a representation of vehicle dynamics to influence such long-term decisions and may not necessarily be the most suitable types of models to use. There may, however, be exceptions to this situation if location problems arise at an operational level of decision making, in which case an integration of facility location and micro-level emission models may be appropriate.

20.2.1 Accounting for Emissions in Facility Location Problems

Location problems almost always involve decisions pertaining to installation of facilities that are typically modeled by a vector y of binary variables, which induces a fixed cost $f(y)$ that generally, but not always, linearly increases with the number and type of facilities. The second set of decisions relate to the assignment of

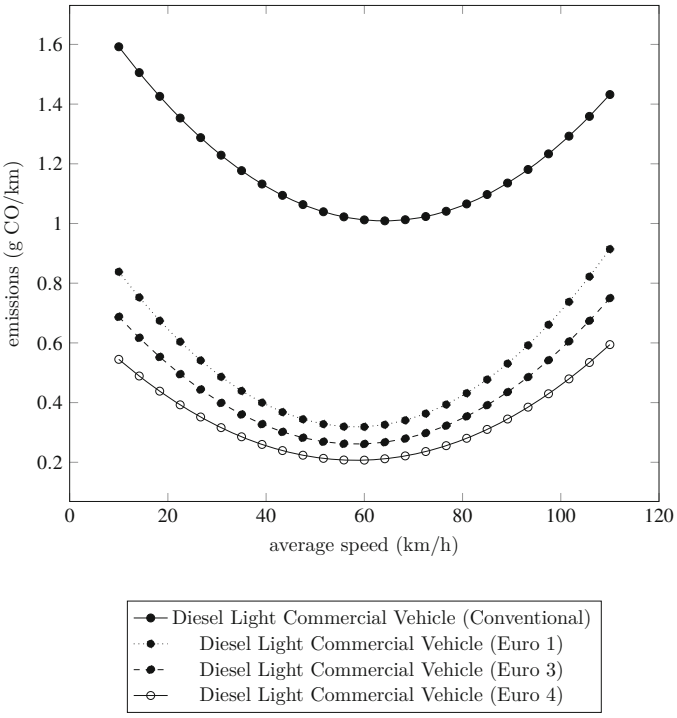


Fig. 20.2 Emissions calculated using the average speed model for vehicles with conventional and older engine technology

customers to installed facilities, represented by the vector x, which induces an operational cost $c(x)$ that may include, amongst others, shipment costs, including drivers, fuel and vehicle acquisition. The assignment decisions x also dictate the volume of products that flow between a facility and a customer, which affects both fuel consumption and emissions, either through choice of vehicle or load, or both. If $g(x)$ denotes the amount of emissions arising from the shipment of products, then a simplified representation of a facility location model that captures the trade-off between operational costs and emissions can be presented as follows:

$$
\begin{aligned}
\psi_1 = \text{minimize} \quad & f(y) + c(x) \\
\psi_2 = \text{minimize} \quad & g(x) \\
\text{subject to} \quad & \\
& Ax = 1 \\
& Bx \leq Dy \\
& y \in \mathbb{B} \\
& x \geq 0,
\end{aligned}
$$

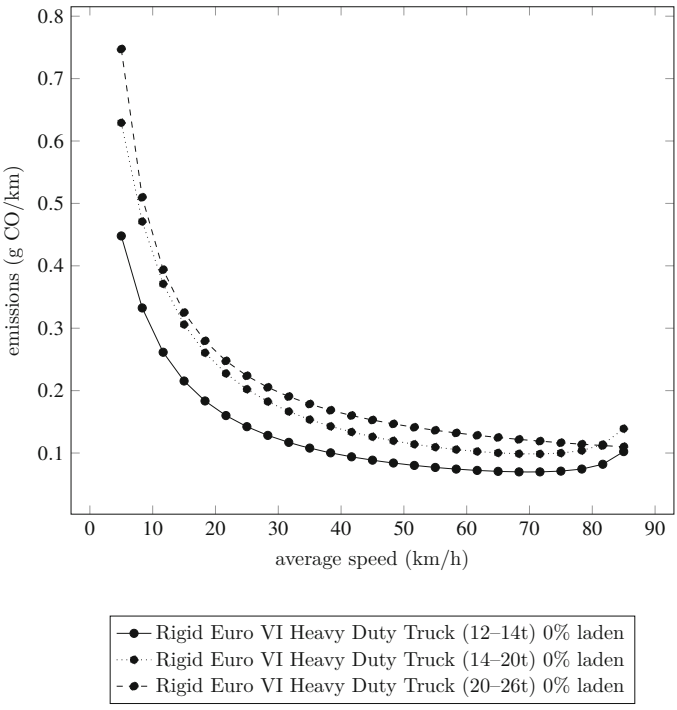


Fig. 20.3 Emissions calculated using the average speed model for vehicles with newer engine technology

where A, B and D are matrices of suitable proportions. The model above is a bi-objective formulation, with the first objective ψ_1 reflecting the operational costs and the second objective ψ_2 denoting the emissions. The first set of constraints are relevant to the assignment of customers to facilities and the second set of constraints ensure that each customer is assigned to a facility that is installed. If there is a suitable weighting γ of emissions (e.g., cost) to make it commensurate with $c(x)$, then the two objectives can be re-written as follows,

$$\begin{aligned} \psi_1' &= \text{minimize} \quad f(y) \\ \psi_2' &= \text{minimize} \quad c(x) + \gamma g(x), \end{aligned}$$

particularly as, in most cases, the amount of emissions from a vehicle is proportional to fuel consumption. One extreme solution to the model above is to reduce the number of facilities to the minimum possible (e.g., only one if there are no capacity restrictions), which will achieve objective ψ_1' but significantly increase the transportation costs $c(x)$ and emissions $g(x)$. In the other extreme, increasing the number of facilities (e.g., one at the site of each customer) will maximize the facility costs $f(y)$ but simply drive ψ_2' equal to 0. The two extremes show that the problem is truly bi-objective, i.e., the two objectives ψ_1' and ψ_2' are contradictory, and the

solution highly depends on the relative importance of the operational costs and the weight of the environmental factors.

The approach adopted by Tricoire and Parragh (2017) is along the lines of the general model introduced above in formulating a bi-objective hub location routing problem, where one objective is related to minimization of operational costs that include that of establishing hubs and space acquisition costs, and the other minimizes CO_2 emissions induced by the empty and loaded running of the vehicles. This study treats the emissions as being fixed per unit of distance and per unit load of weight, and derives insights on the trade-offs between the number of hubs used and the resulting emissions, with an overall conclusion that "investing in facilities does reduce future pollution".

The studies by Koç et al. (2014) and Toro et al. (2017) investigate similar problems in that they integrate facility location decisions with those of fleet size and mix and routing (in the former) and vehicle routing (in the latter). In contrast to the work of Tricoire and Parragh (2017), however, they both use a more detailed representation of fuel consumption and emissions that is minimized as part of the overall problem. In Koç et al. (2014), this entails the use of a comprehensive modal emissions model as part of the (single) objective for estimating fuel consumption from heavy-good vehicles operating in urban areas, part of which requires to optimize speeds of the vehicles. The authors conclude by stating that it is preferable to locate depots outside the city centre and to use heterogeneous fleets over homogeneous fleets. This study shows the impact choice of the location of the facilities can have on the overall choice of a distribution strategy within urban areas. In Toro et al. (2017), vehicle speed is not part of the decision problem whereby the vehicles are assumed to travel at constant speed, and that fuel consumption is estimated on the basis of the distance traversed and the load carried by a vehicle. The latter work treats the problem as being bi-objective, with one objective minimizing operational costs, and the other minimizing fuel and emissions. One interesting finding of Toro et al. (2017) is that increasing the number of vehicles used results in improved fuel economy and hence less emissions, mainly due to the shorter trips that the vehicles will have to perform.

Studies that look at the integration of facility location problems with consideration of externalities are few and far between, some of which are described above, but more research is required in this area, not only for development of methodological approaches to solve such problems but perhaps more importantly, understanding the trade-offs involved in making strategic and operational decisions to improve economical and environmental performance of distribution systems.

20.3 Reverse Logistics Network Design

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, repair-

ing, refurbishing, re-manufacturing, 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.

Reverse logistics activities aim to lessen or mitigate the environmental externalities as such operation of reverse logistics networks lead to reduced use of natural resources as well as pollution prevention through the reduction of waste. Major driving forces behind reverse logistics include not only environmental consciousness but also economic factors and government legislations. 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.

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. 20.4.

The design of a reverse logistics network is a complex problem comprising the determination of the optimal locations of different types of facilities as well as the integration between these facilities. The facilities to be located include but

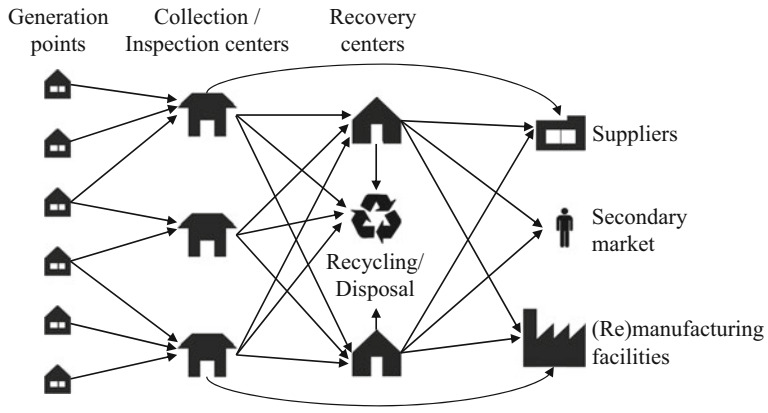


Fig. 20.4 A generic reverse logistics network

not limited to collection, inspection, recovery, (re)manufacturing, recycling, and disposal centers. The decisions to be made include determining the number, size, and capacities of the facilities to be located, the amount of products to be recovered at each facility, and the amount of products and/or components to be sent in-between these facilities.

In the next sub-section, we present a generic reverse logistics network design model and in the following section we discuss its possible extensions.

20.3.1 A Generic Reverse Logistics Network Design Model

Multiple commodities need to be considered in the configuration of a reverse logistics network. These commodities include used, inspected, repaired, or refurbished products, components, or raw materials and are represented by the set P. In order to represent a different state (inspected, repaired, refurbished, etc.) of a certain item, a different product type needs to be defined within this set.

Let R represent the set of available recovery options. This 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 plausible in real-life reverse logistics networks, we include these in the set R.

Other sets of parameters include N_r the set of potential locations for recovery option $r \in R$; E_r the set of existing facilities with recovery option $r \in R$; I_r the set of selectable facilities with recovery option $r \in R$, $I_r = N_r \cup E_r$; J_r the set of non-selectable locations with recovery option $r \in R$ (e.g. secondary market, disposal); and L the set of all locations, $L = \cup_{r \in R} (I_r \cup J_r)$. Some recovery options may be operated by third-party logistics providers. Such 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.

The parameters required for the mathematical model are as follows:

$g_{\ell p}$	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 $q \in P$ using recovery option $r \in R$
$K_{r\ell}$	Capacity of recovery option $r \in R$ at location $\ell \in L$
T_{rp}	Recovery target for product $p \in P$ with recovery option $r \in R$
$H_{r\ell p}$	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)
$S_{r\ell p}$	Cost of recovering one unit of product $p \in P$ with recovery option $r \in R$ at location $\ell \in L$
F_{ri}	Fixed setup cost of establishing recovery option $r \in R$ in location $i \in N_r$
$C_{\ell\ell' p}$	Unit cost of transporting product $p \in P$ from location $\ell \in L$ to location $\ell' \in L$

Transitions between the stages of products and reverse bills-of-materials (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. There are recovery targets, usually set by the legislations, for each type of product and recovery option. Revenues may be obtained through some recovery options, e.g., by selling products or components to recycling facilities, to the secondary market or to external (re)manufacturing facilities. Some recovery options may also incur costs as in the case of product disposal.

The decision variables of the model are:

$$y_{ri} = \begin{cases} 1, \text{ if recovery option } r \in R \text{ is operated at the selectable facility } i \in I_r, \\ 0, \text{ otherwise.} \end{cases}$$

$x_{\ell\ell' p} =$ Quantity of product $p \in P$ shipped from location $\ell \in L$ to location $\ell' \in L$.

$v_{r\ell p} =$ Amount of product $p \in P$ recovered with recovery option $r \in R$ at location $\ell \in L$.

The reverse logistics network design problem can be formulated as a mixed-integer linear program as follows:

$$\text{Maximize} \quad \sum_{r\in R}\sum_{\ell\in L}\sum_{p\in P}(H_{r\ell p} - S_{r\ell p})\, v_{r\ell p} - \sum_{r\in R}\sum_{i\in N_r} F_{ri}\, y_{ri}$$

$$- \sum_{\ell\in L}\sum_{\ell'\in L\setminus\{\ell\}}\sum_{p\in P} C_{\ell\ell' p}\, x_{\ell\ell' p} \tag{20.2}$$

$$\text{subject to} \quad g_{\ell p} + \sum_{r\in R}\sum_{q\in P}\beta_{rqp}\, v_{r\ell q} + \sum_{\ell'\in L\setminus\{\ell\}} x_{\ell'\ell p} =$$

$$\sum_{r\in R} v_{r\ell p} + \sum_{\ell'\in L\setminus\{\ell\}} x_{\ell\ell' p}, \quad \ell \in L,\ p \in P \tag{20.3}$$

$$\sum_{\ell\in L} v_{r\ell p} \geq T_{rp}, \quad r \in R,\ p \in P \tag{20.4}$$

$$\sum_{p\in P} v_{rip} \leq K_{ri}\, y_{ri}, \quad r \in R,\ i \in I_r \tag{20.5}$$

$$\sum_{p\in P} v_{rjp} \leq K_{rj}, \quad r \in R,\ j \in J_r \tag{20.6}$$

$$x_{i\ell p} \leq \sum_{r\in R} K_{ri}\, y_{ri}, \quad i \in \cup_{r\in R} I_r,\ \ell \in L \setminus \{i\},\ p \in P \tag{20.7}$$

$$x_{\ell ip} \leq \sum_{r\in R} K_{ri}\, y_{ri}, \quad \ell \in L \setminus \{i\},\ i \in \cup_{r\in R} I_r,\ p \in P \tag{20.8}$$

$$y_{ri} \in \{0, 1\}, \quad r \in R,\ i \in I_r \tag{20.9}$$

$$x_{\ell\ell' p} \geq 0, \quad \ell, \ell' \in L\ (\ell \neq \ell'),\ p \in P \tag{20.10}$$

$$v_{r\ell p} \geq 0, \quad r \in R,\ \ell \in L,\ p \in P. \tag{20.11}$$

The objective function (20.2) 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 reverse logistics network. The latter comprises the cost of recovery (e.g. disposal), setting up new recovery options at facilities, and transporting products.

Equalities (20.3) 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 (20.4) 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 (20.5) and (20.6) are the capacity constraints for new and existing recovery options, respectively. Constraints (20.7)–(20.8) impose that products can only be shipped from operated facilities. Lastly, conditions (20.9)–(20.11) set the domains of the decision variables.

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 model is NP-hard, being a generalization of the simple plant location problem (see Chap. 3). General purpose optimization software (e.g., CPLEX or Gurobi) can however be used to solve small to medium-sized instances of this model within reasonable times. For large-sized instances there may be a need for customized algorithms and heuristics. The following section discusses some extensions of the above model.

20.3.2 Extensions

The reverse logistics network design model introduced in the previous section can be extended in manifold ways. Analogous with the traditional facility location models, the above formulation can be generalized to include capacity selection and extension decisions, a multi-period/dynamic setting, uncertainty associated with the

problem parameters, multiple objectives, etc. These extensions are already well-discussed within the other chapters of this book. We briefly discuss some of such extensions within the context of reverse logistics network design below to provide some exemplary references.

The design of a reverse logistics network can be embedded 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. A multi-period setting is also appropriate for planning the re-design of a reverse 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. In turn, new facilities can be established through successive sizing. Multi-period models in reverse logistics network design were proposed, for example, by Lee and Dong (2009), Salema et al. (2010), and Alumur et al. (2012).

A distinguishing feature of reverse logistics network design problems is that there are various sources of uncertainty for the supply arising at the upper echelon facilities (e.g., uncertainty in the amount and in the quality of returned products). Studies addressing uncertainty issues in the context of reverse logistics network design include Realff et al. (2004), Listeş and Dekker (2005), Listeş (2007), Salema et al. (2007), El-Sayed et al. (2010), and Fonseca et al. (2010).

Many actors are involved in the design and operation of a reverse logistics network. Even though extended producer responsibilities defined in the legislations of 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. Multiple actors lead to decision problems with multiple objectives. Although 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 can certainly require further attention.

A major extension of reverse logistics network design is to integrate reverse flows with forward flows of the supply chain. The term *closed-loop supply chain* refers to a network comprising both forward and reverse flows. Figure 20.5 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 ideally be designed by integrating forward and reverse logistics activities. The generic model introduced above for the reverse logistics network design problem can certainly be extended 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) presenting models determining the locations of facilities within closed-loop supply chain networks.

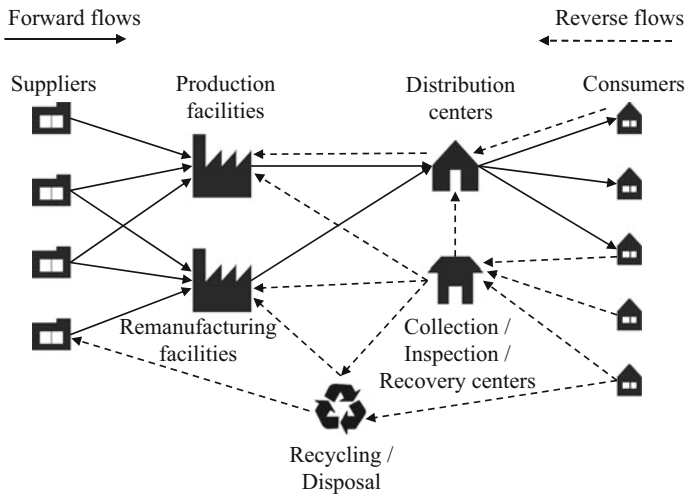


Fig. 20.5 A closed-loop supply chain network

For other extensions and special cases on reverse logistics network design, 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).

20.4 Location Problems Related to Alternative Fuel Vehicles

One recent class of problems within which location arise as part of the planning decisions is relevant to alternative fuel vehicles (AFVs), which either run on fuel as opposed to traditional petroleum-based fuels (petrol or Diesel fuel) or alternative technologies to power an engine that does not involve solely petroleum (Wikipedia 2017). The types of AFVs include those running on biofuel, natural gas, hydrogen (fuel cell electric), hybrid electric, plug-in hybrid electric (PHE), and battery electric (BE), also known as Zero Emission Vehicles (ZEVs) (Hackbarth and Madlener 2013; Guerra at al. 2016). AFVs are readily used in various applications including goods distribution and public transportation as well as personal transport (e.g., Pelletier et al. 2016, Tzeng et al. 2005). This section briefly discusses location problems that arise within the context of AFVs.

Similar to the conventional vehicles that are powered by petroleum fuels, refueling is also important for AFVs, if not more critical, to ensure continuity of operation without disruption. Location analysis therefore plays a significant role for installing refueling or recharging stations, in particular for PHEVs and BEVs where the refueling (charging) times can be significant. Basic location models like the p-median problem (see Chap. 2) can certainly be employed for determining the optimal locations of refueling stations (e.g., Goodchild and Noronha 1987, Nicholas

et al. 2004). The motivation behind using the p-median model is the assumption that the consumers generally prefer to refuel near their homes (Upchurch and Kuby 2010).

In traditional location problems, demand is generated at nodes of the network. In contrast, the demand arising from the need to refuel AFVs does not originate from the fixed nodes but from traffic flows. Hodgson (1990) describes a so-called *flow capturing location model* (*FCLM*) to model the flow-based demand, where the aim is to locate p facilities to maximize the total demand captured (covered), and where a unit of demand to be covered is defined as the fixed path from a given origin to a destination. It is assumed that a facility placed at a node in the network covers all the flow which passes through that node; it suffices therefore to locate one facility on a path. Location problems that assume flow-based demand have later been named as *flow interception problems* (see, e.g., Berman et al. 1995). Kuby and Lim (2005) introduced the *flow refueling location model* (*FRLM*) by extending the FCLM to the context of generic range-limited AFVs. The FRLM defined by Kuby and Lim (2005) optimally locates p refueling stations on a network so as to maximize the total flow volume refueled for predetermined origin-destination paths. The FRLM recognizes the fact that it may be necessary to stop at more than one facility for refueling and, thus, unlike the FCLM, it allows for locating more than one facility on a path. Kuby and Lim (2005) also describe a mixed-integer programming model for the problem which assumes that all feasible facility combinations that can be used to refuel vehicles on each given origin-destination path are exogenously determined. Since generation of all these feasible combinations requires significant computational memory and time, different solution methods are proposed in the literature to overcome the difficulty (see Capar and Kuby 2012; Capar et al. 2013; Kim and Kuby 2012; Lim and Kuby 2010; MirHassani and Ebrazi 2012). In particular, Capar et al. (2013) and MirHassani and Ebrazi (2012) provide alternative formulations for the FRLM both of which drastically increase the computational efficiency.

Variations of the FRLM include those that allow locating refueling stations along arcs as well as on nodes of the network (Kuby and Lim 2007), imposing capacity constraints that limit the number of vehicles refueled at each station (Upchurch et al. 2009), incorporating locomotive refueling scheduling decisions in railroad networks (Nourbakhsh and Ouyand 2010), allowing deviation from the shortest path up to a tolerance that the drivers are willing to accept (Yıldız et al. 2016), generalizing it to account for plug-in hybrid electric vehicles (Arslan and Karaşan 2016), and assuming a probabilistic travel range for the vehicle (Lee and Han 2017).

The FRLM is a maximal covering type model (see Chap. 5 for more information on the covering location problems). An alternative way to approach the AFV refueling station location problem is through a *flow-based set covering model*, as was done by Wang and Lin (2009). The objective of this problem is to minimize the total cost of locating refueling stations where all flow-refueling demand is to be covered by the stations within a specific coverage distance for fixed origin-destination paths. Other studies using a flow-based set covering model for the location of refueling stations include Wang and Wang (2010), Wang and Lin (2013), MirHassani and Ebrazi (2012) and Li and Huang (2014).

The refueling station location problem can also be formulated within location-routing type models, referred as location-routing problem with intermediary facilities (see Chap. 15), more specifically, by extending such models to include additional considerations specific to AFVs (see, e.g., Kand and Recker 2014; Schiffer and Walther 2017).

In addition to determining the locations of refueling stations, one other relevant problem arising within the context of AFVs is to determine the locations of battery swapping or switching stations, where depleted batteries can be exchanged for recharged ones during a journey. Mak et al. (2013) introduced this problem and developed robust optimization models that aid the planning process for deploying battery-swapping infrastructure. Variants of this problem can be found in Yang and Sun (2015) and Hof et al. (2017).

The particular characteristics of AFVs can be incorporated within any of the location models as additional constraints, for example, considering multiple types of charging facilities with varying charging rates (Liu and Wang 2017), partial or full charging options (Keskin and Çatay 2016; Schiffer and Walther 2017), battery life-span or battery degradation concepts (Kong et al. 2017).

Finally, and although not necessarily a location problem in the traditional sense, it is worth mentioning a study by Chen et al. (2016) that introduces a network design problem related to AFVs, which consists of determining an optimal deployment of charging lanes for electric vehicles in transportation networks. This follows a recent development in the 'charging-while-driving' technology, which envisages deploying charging lanes in regional or even urban road networks of the future which electric vehicles can use. In this case, the lanes themselves may be seen as facilities. It is clear that technologies that are fast developing for AFVs will give rise to other such interesting problems in the near future.

20.5 Research Prospects

Environmental issues arising within location problems are broad and complex, but need to be captured and addressed nevertheless. Green location problems necessitates an explicit consideration of micro-level and firm-based environmental performance measures, such as internal consumption of resources including energy, water, land and building materials, as well as the wider macro-level impacts that extend beyond a facility, such as "land use, atmospheric emissions, waste management, traffic and congestion, public transport, visual intrusion and ecology" as highlighted by Baker and Marchant (2015) and captured within the environmental assessment framework proposed by the same authors. In terms of modeling, the difficulties reside in (1) being able to represent the impact of internal and external externalities in quantifiable terms, (2) their integration within existing or new models of facility location, and (3) the ability to bring together the impact of long-term decisions along with those of day-to-day operations on the environment. This

brief chapter has touched upon some of these issues and described ways in which they can be addressed from the point of view of location analysis.

Other relevant problems for which location decisions are integral, which were not discussed in this chapter due to space limitations, offer further research prospects. At this point, we suffice to briefly mention three research directions below, but also recognize that they are inherently related (and possibly overlapping):

- *Waste management*, which includes determining the locations of waste disposal sites (landfills, incinerators, etc). Such problems can be regarded as part of reverse logistics networks, but with their own challenges in relation to location decisions, including the location of treatment sites and landfills as well as allocation decisions. For further information, the reader is referred to the review by Ghiani et al. (2014).
- *Undesirable facility location*, which involves locating semi-obnoxious facilities, such as a garbage dump, a chemical plant or a nuclear reactor, that may have adverse effects on people or the environment. Locating such facilities within close proximity to people or other forms of life is undesirable, for which the aim of such problems is to minimize the nuisance and the adverse effects on existing facilities or population centers (see e.g., Erkut and Neuman 1989, Melachrinoudis 2011).
- *Hazardous materials logistics* which entails determining the location, size, and the technology type of potentially hazardous facilities as well transportation of hazardous materials. These problems typically involve multiple objectives, the most prominent ones being minimization of cost and of risk, and equitable distribution of risk. The interested reader may refer to the book chapter by Erkut and Verter (1995).

References

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 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. CRC Press, Boca Raton, chap 5, pp 67–98

Arslan O, Karaşan OE (2016) A benders decomposition approach for the charging station location problem with plug-in hybrid electric vehicles. Transp Res B Methodol 93:670–695

Baker P, Marchant C (2015) Reducing the environmental impact of warehousing. In: McKinnon et al (eds) Green logistics: improving the environmental sustainability of logistics. Kogan Page, London, pp 194–226

Berman O, Hodgson MJ, Krass D (1995) Flow-interception problems In: Drezner Z, Hamacher HW (eds). Facility location: Applications and Theory. Springer, New York, pp 389–426

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. Springer, New York, chap 6, pp 171–212

Capar I, Kuby M (2012) An efficient formulation of the flow refueling location model for alternative-fuel stations. IIE Trans 44(8):622–636
Capar I, Kuby M, Leon VJ, Tsai Y (2013) An arc cover–path-cover formulation and strategic analysis of alternative-fuel station locations. Eur J Oper Res 227(1):142–151
Chen Z, He F, Yin Y (2016) Optimal deployment of charging lanes for electric vehicles in transportation networks. Transp Res B: Methodol 91:344–365
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. Springer, Berlin, chap 1, pp 3–27
Demir E, Bektaş T, Laporte G (2014) A review of recent research on green road freight transportation. Eur J Oper Res 237(3), 775–793
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
Erkut E, Neuman S (1989) Analytical models for locating undesirable facilities. Eur J Oper Res 40:275–291
Erkut, E, Verter V (1995) Hazardous materials logistics In: Drezner Z, Hamacher HW (eds). Facility location: a survey of applications and methods. Springer, New York, pp 467–506
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. Springer, Berlin, chap 4, 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
Ghiani G, Laganà D, Manni E, Musmanno R, Vigo D (2014) Operations research in solid waste management: a survey of strategic and tactical issues. Comput Oper Res 44:22–32
Goodchild MF, Noronha VT (1987) Location-allocation and impulsive shopping: the case of gasoline retailing. Spatial analysis and location-allocation models, 121–136
Guerra CF, García-Ródenas R, Sánchez-Herrera EA, Rayo DV, Clemente-Jul C (2016) Modeling of the behavior of alternative fuel vehicle buyers. A model for the location of alternative refueling stations. Int J Hydrog Energy 41(42):19,312–19,319
Hackbarth A, Madlener R (2013) Consumer preferences for alternative fuel vehicles: a discrete choice analysis. Transp Res D Transp Environ 25:5–17
Hill N, Bramwell R, Harris B, (2017) 2017 Government GHG conversion factors for company reporting. UK Government, Department for Business, Energy & Industrial Strategy, London. Available at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/650244/2017_methodology_paper_FINAL_MASTER.pdf. Accessed 18 Feb 2018
Hodgson JM (1990) A flow-capturing location-allocation model. Geogr Anal 22(3):270–279
Hof J, Schneider M, Goeke D (2017) Solving the battery swap station location-routing problem with capacitated electric vehicles using an AVNS algorithm for vehicle-routing problems with intermediate stops. Transp Res B Methodol 97:102–112
Jaehn F (2016) Sustainable operations. Eur J Oper Res 253:243–264
Jayaraman V, Luo Y (2007) Creating competitive advantages through new value creation: a reverse logistics perspective. Acad Manage Perspect 21:56–73
Kang JE, Recker W (2014) Strategic hydrogen refueling station locations with scheduling and routing considerations of individual vehicles. Transp Sci 49(4):767–783
Keskin M, Çatay B (2016) Partial recharge strategies for the electric vehicle routing problem with time windows. Transp Res C: Emerg Technol 65:111–127
Kim JG, Kuby M (2012) The deviation-flow refueling location model for optimizing a network of refueling stations. Int J Hydrog Energy 37(6):5406–5420
Koç Ç, Bektaş T, Jabali O, Laporte, G (2014) The fleet size and mix pollution-routing problem. Transp Res B: Methodol 70:239–254
Kong C, Jovanovic R, Bayram IS, Devetsikiotis M (2017) A hierarchical optimization model for a network of electric vehicle charging stations. Energies 10(5):675

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
Kuby M, Lim S (2005) The flow-refueling location problem for alternative-fuel vehicles. Socio Econ Plan Sci 39(2):125–145
Kuby M, Lim S (2007) Location of alternative-fuel stations using the flow-refueling location model and dispersion of candidate sites on arcs. Netw Spat Econ 7(2):129–152
Lee DH, Dong M (2009) Dynamic network design for reverse logistics operations under uncertainty. Transp Res E-Log 45:61–71
Lee C, Han J (2017) Benders-and-price approach for electric vehicle charging station location problem under probabilistic travel range. Transp Res B: Methodol 106:130–152
Li S, Huang Y (2014) Heuristic approaches for the flow-based set covering problem with deviation paths. Transp Res E: Log Transp Rev 72:144–158
Lim S, Kuby M (2010) Heuristic algorithms for siting alternative-fuel stations using the flow-refueling location model. Eur J Oper Res 204(1):51–61
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
Liu H, Wang DZW (2017) Locating multiple types of charging facilities for battery electric vehicles. Transp Res B: Methodol 103:30–55
Mak HY, Rong Y, Shen ZJM (2013) Infrastructure planning for electric vehicles with battery swapping. Manag Sci 59(7):1557–1575
Melachrinoudis E (2011) The location of undesirable facilities In: Eiselt HA, Marianov, V (eds). Foundations of location analysis. Springer, New York, pp 207–239
McKinnon A, Browne M, Whiteing A, Piecyk M (eds) (2015) Green logistics: improving the environmental sustainability of logistics. Kogan Page, London
MirHassani SA, Ebrazi R (2012) A flexible reformulation of the refueling station location problem. Transp Sci 47(4):617–628
Nicholas M, Handy S, Sperling D (2004) Using geographic information systems to evaluate siting and networks of hydrogen stations. Transp Res Rec: J Transp Res Board 1880:126–134
Nourbakhsh SM, Ouyang Y (2010) Optimal fueling strategies for locomotive fleets in railroad networks. Transp Res B: Methodol 44(8–9):1104–1114
Ntziachristos L, Samaras Z (2017) EMEP/EEA air pollutant emission inventory guidebook. European environment agency: part 1.A.3.b.i-iv Road transport 2017, Luxembourg. Available at https://www.eea.europa.eu/publications/emep-eea-guidebook-2016/part-b-sectoral-guidance-chapters/1-energy/1-a-combustion/1-a-3-b-i. Accessed 19 May 2018
Pati RK, Vrat P, Kumar P (2008) A goal programming model for paper recycling system. Omega 36:405–417
Pelletier S, Jabali O, Laporte G (2016) 50th anniversary invited article – goods distribution with electric vehicles: review and research perspectives. Transp Sci 50(1):3–22
Psaraftis H (ed) (2016) Green transportation logistics: the quest for win-win solutions. Springer, Cham
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
Schiffer M, Walther G (2017) The electric location routing problem with time windows and partial recharging. Eur J Oper Res 260(3):995–1013
Tari I, Alumur SA (2014) Collection center location with equity considerations in reverse logistics networks. INFOR: Inf Syst Oper Res 52:157–173

Toro EM, Franco JF, Echeverri MG, Guimarães FG (2017) A multi-objective model for the green capacitated location-routing problem considering environmental impact. Comput Ind Eng 110:114–125

Tricoire F, Parragh SN (2017) Investing in logistics facilities today to reduce routing emissions tomorrow. Transp Res B: Methodol 103:56–67

Tzeng GH, Lin CW, Opricovic S (2005) Multi-criteria analysis of alternative-fuel buses for public transportation. Energy Policy 33(11):1373–1383

Upchurch C, Kuby M (2010) Comparing the p-median and flow-refueling models for locating alternative-fuel stations. J Transp Geogr 18(6):750–758

Upchurch C, Kuby M, Lim S (2009) A model for location of capacitated alternative-fuel stations. Geogr Anal 41(1):85–106

Wang YW, Lin CC (2009) Locating road-vehicle refueling stations. Transp Res E: Log Transp Rev 45(5):821–829

Wang YW, Lin CC (2009) Locating multiple types of recharging stations for battery-powered electric vehicle transport. Transp Res E: Log Transp Rev 58:76–87

Wang YW, Wang CR (2010) Locating passenger vehicle refueling stations. Transp Res E: Log Transp Rev 46(5):791–801

Wikipedia, the free encyclopedia. Available at https://en.wikipedia.org/wiki/Alternative_fuel_vehicle. Accessed 5 April 2018

Yıldız B, Arslan O, Karaşan OE (2016) A branch and price approach for routing and refueling station location model. Eur J Oper Res 248(3):815–826

Yang J, Sun H (2015) Battery swap station location-routing problem with capacitated electric vehicles. Comput Oper Res 55:217–232

Chapter 21
Location Problems in Humanitarian Supply Chains

Bahar Y. Kara and Marie-Ève Rancourt

Abstract In this chapter, we first present a general description of humanitarian supply chains. This includes the main purpose and components (facilities and transportation flow) of humanitarian supply chains within different contexts. This description also aims to classify the types of facilities that need to be located for supporting disaster relief operations as well as development programs. We then describe the location decisions that need to be made and some important metrics to consider. We also present a general model to solve location problems, which is a formulation that serves as a base for humanitarian network design problems involving location decisions. Finally, we discuss some extensions of this basic location problem.

21.1 General Description of Humanitarian Supply Chains

The objective of deploying humanitarian supply chains is to provide assistance in order to maintain life, improve health and support the population affected by a disaster or a crisis. This is true when responding to disasters and when setting up long-term development programs—both require the management of complex supply chains to achieve their aid objectives. The level of response and extent of the programs depends on several factors, including the scale of the disaster, the socio-economic conditions of the affected area, the vulnerability of the population, the state of the critical infrastructures, the situational awareness of the different stakeholders, and the available funding and resources. In this chapter, we use the term *humanitarian supply chains* to describe the logistics networks designed and managed to support disaster relief as well as development programs; see Kara and

B. Y. Kara
Department of Industrial Engineering, Bilkent University, Ankara, Turkey
e-mail: bkara@bilkent.edu.tr

M.-È. Rancourt (✉)
Department of Logistics and Operations Management, HEC Montréal, Montréal, QC, Canada
e-mail: marie-eve.rancourt@hec.ca

G. Laporte et al. (eds.), *Location Science*,
https://doi.org/10.1007/978-3-030-32177-2_21

Savaser (2017), Çelik et al. (2012) and Kovács and Spens (2007) for distinctions between logistics operations to support disaster response and development programs.

The main logistics activities conducted following a disaster are (1) the assessment of needs and damage, (2) the rescue and evacuation of people, (3) the transport of resources (staff and equipment) and supplies to and from several facilities, (4) the provision of assistance services, and (5) the distribution of relief supplies. Nevertheless, successful responses are planned, not improvised. Indeed, mitigation and preparedness are crucial phases in reducing the negative impacts of a forthcoming disaster and in facilitating efficient response and recovery phases. Relief operations and the issues related to each of these phases are different, leading to phase-specific types of supply chain decisions (Çelik et al. 2012).

Affected states, the governments of territories in which the disaster occurred, play a leading role in disaster management and are supported through the efforts of several domestic actors (national disaster management agencies, NGOs, the military, etc.). Yet some disasters require the involvement of the international community. Indeed, international support must be solicited when the national response capacity is insufficient or overwhelmed in the face of a major disaster (International Federation of Red Cross and Red Crescent Societies 2017). The international community is also heavily involved in various development programs implemented in multiple countries to address important challenges, usually related to one or several of the 17 sustainable development goals set by the United Nations (United Nations 2018). Consequently, several actors are involved within humanitarian supply chains (donors, governments, local and international NGOs, United Nations agencies, private companies, media, beneficiaries, etc.), and different organizations assist the population in need according to their main sectors of activity (e.g. health, food security, logistics, telecommunications, protection, shelter, education, water, sanitation, and hygiene).

To provide humanitarian assistance, the effective deployment of supply chains is crucial to move the necessary resources and supplies and to set up dispensing facilities. Figure 21.1 depicts the general structure of a humanitarian supply chain and its main transportation flows to reach people in need located in underserved areas (i.e., areas affected by a disaster, remote areas, or areas where living conditions need to be improved). In this figure, the main physical infrastructures are represented by black boxes, and the main transportation flows are represented by dashed blue arrows and symbols. Alternative transportation flows are represented by dashed grey arrows and symbols.

21.1.1 International and Regional Distribution Centers

International distribution centers (depots) are strategically located to ensure that worldwide coverage is available within only short delays following a disaster (the first response). This is usually accomplished by means of air transportation. In the

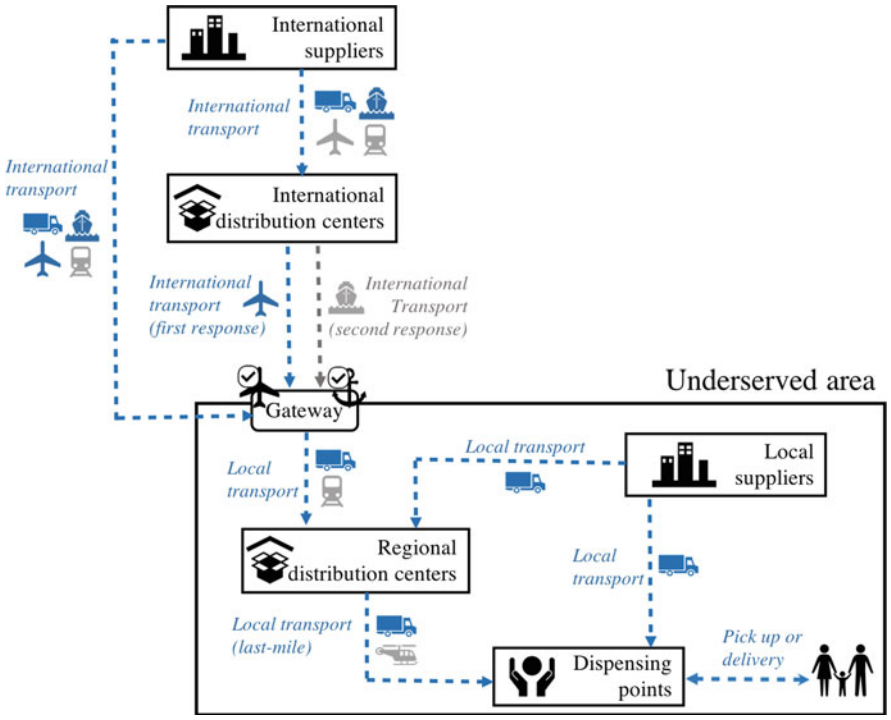


Fig. 21.1 Humanitarian supply chain

case of a major disaster, some of the relief supplies and equipment required for the longer-term response (the second response) can sometimes be shipped by sea to reduce the transportation costs. International distribution centers also serve as storage facilities for stockpiling resources in anticipation of sudden needs, which is known as *prepositioning* in the humanitarian sector. Humanitarian organizations implement prepositioning strategies to eliminate procurement delays and ensure the relief supplies are available when they need them, thus enhancing their preparedness and emergency-response capacity, particularly for sudden-onset disasters (Duran et al. 2011). For example, the United Nations Humanitarian Response Depot (UNHRD) provides various logistics services related to prepositioning to its partners throughout its network of depots located in Brindisi (Italy), Dubai (UAE), Panama City (Panama), Accra (Ghana), Kuala Lumpur (Malaysia) and Las Palmas (Spain) (Dufour et al. 2018). Thus, international distribution centers are usually located in relatively stable areas from a political point of view, which are not prone to disaster and have well-developed logistics infrastructures, often close to an airport or a port.

Local organizations, such as civil societies, NGOs, local government authorities, and community organizations, also preposition relief supplies and equipment in regional distribution centers to strengthen local capacities. For example, the Federal Emergency Management Agency (FEMA), an agency of the United States

Department of Homeland Security that coordinates responses to major disasters occurring in the US, maintains its inventory of relief supplies in nine distribution centers located across the country and its territories (US Department of Homeland Security 2009). Local capacities depends on several factors, including the organization of the local governments, the presence of any international organizations in the country (e.g., UN entities), local and regional economic resources and logistics infrastructures, predisposition to disasters, and the availability of human and technological resources. Regional distribution centers can be owned and managed by a local organization, or warehouse space can be rented from a third-party logistics provider. These centers usually receive supplies from both international and local suppliers.

21.1.2 Dispensing Points

Dispensing points consist mainly of points of distribution or services, where beneficiaries can receive relief supplies or where various services (e.g., healthcare and shelter) can be provided to them during a disaster response or a development program. Points of distribution are usually deployed to distribute life-sustaining commodities, such as water, food, tarps, and other resources, to the public within the first 72 h following a disaster, whereas points of dispensing are designed to quickly distribute medications, vaccines, and medical supplies to a large number of people within a short time frame during a public health emergency. Note that dispensing points can also be set up for long-term development programs. For example, Rancourt et al. (2015) conducted a field study in the region of Garissa (Kenya), in collaboration with the World Food Programme and the Kenya Red Cross, to determine a set of distribution points where food aid has been directly handed out to beneficiaries for several years in order to alleviate severe food insecurity. They present location models, considering the welfare of the different stakeholders, and analyze the results obtained using real data. GIS data were processed to determine the parameters of the last-mile distribution network, see Chap. 18 for an overview of location and GIS.

Following a disaster, portable and temporary medical facilities (field hospitals) can also be deployed by different organizations (e.g., IRFC, Doctors Without Borders, and military bodies) to replace or supplement the destroyed local healthcare capacity and to support relief efforts. In an urban setting, field hospitals can be installed in accessible and visible buildings, such as schools, town halls, and stadiums. Dispensing points are also used for similar purposes in the context of development programs lasting for longer periods of time than disaster responses.

Relief shelters are used to provide protection, safety, security and privacy to people who have left or lost their housing as a result of a disaster. Shelters are used until displaced people can be rehoused in either their restored dwellings or new, permanent houses, which means that shelter locations may be required for several months or even years following a disaster (Abdulrahman et al. 2014). Refugee

camps are temporary settlements built to receive internally or internationally displaced people, and they can host thousands of people for several years (e.g., Suruc in Turkey, Bidi Bidi in Uganda, Dadaab and Kakuma in Kenya). In general, refugees seek asylum to escape hostile conflicts in their home countries, but some camps also house migrants because of difficult environmental or economic conditions. The United Nations High Commissioner for Refugees reported that, at the end of 2017, 71.1 million people worldwide had been forcibly displaced because of conflict and persecution. Among them, 19.9 million people were classified as refugees, whereas 39.1 million people were classified as internally displaced persons (UNHCR 2018). These numbers provide an idea of the scale of the operations required to support refugee crises, which combine relief and development forms of aid (e.g., medical relief and food aid distribution as well as children's education and micro-enterprise programs). There exists a wide range of shelter types, and several factors (environmental, economic, technical, and sociocultural) must be taken into account when planning and designing shelters to ensure they are appropriate for the situation, i.e., they meet the needs and conditions of the beneficiaries (Jahre et al. 2018).

In the remainder of this chapter, we will use the term *dispensing points* to account for the different types of locations where beneficiaries can receive various forms of humanitarian aid. In most cases, the beneficiaries have to walk or use another mode of transportation to reach the dispensing points. The aid is either directly distributed or provided at the household level, depending on the context.

21.1.3 Transportation Flows

International suppliers or distribution centers send the relief items to an entry port or airport of an affected country (known as a gateway) by sea or air depending on the time sensitivity of demand. Often, air transportation is the only possible means of transportation to ensure the relief resources reach the affected area within a short enough time frame, i.e., usually less than 72 h for the first response. When feasible, for the second response or the recovery operations, relief items can be shipped by sea, which is less costly.

Once the relief items have reached the affected area, they are usually shipped by road from the gateway to regional distribution centers (warehouses). Local suppliers can transport items to regional distribution centers or directly to the dispensing points, where beneficiaries can collect their relief supplies. Airlifts can also be used to provide assistance in remote regions in a timely manner. This allows for enhanced transportation response capacities in hostile or inaccessible zones, where infrastructure has been severely damaged or destroyed. Although prohibitively expensive compared to other modes of transportation, airlift is sometimes the only possible means to quickly support humanitarian response efforts in difficult-to-reach areas. For example, helicopters were used to reach remote communities located

in mountainous areas after the 2015 Nepal earthquake, as well as in the conflict-affected areas of South Sudan. Airdrop is a last resort for aid organizations because it is expensive and inefficient compared to road transportation. Most of the time, beneficiaries must go to the dispensing points, either through their own means or in some organized fashion to receive health or shelter services or to collect their relief supplies. In some specific cases, such as community health care programs (Cherkesly et al. 2017; VonAchen et al. 2016; McCoy and Lee 2014) or food airdrops, aid is delivered at the household level or closer to the beneficiaries.

21.2 Humanitarian Facility Location Problems

Designing humanitarian supply chains involve a number of location decision problems. Indeed, determining the locations of the international and regional distribution centers and the locations of the dispensing points can have a major impact on response effectiveness in terms of service quality and logistics costs. Depending on the purpose of these storage and distribution locations, different criteria and metrics must be taken into consideration. For example, international and regional distribution centers have to be located in areas where logistics infrastructure is well developed and connected, whereas dispensing points need to be easily accessible for the beneficiaries.

The extent to which location problems are encountered in humanitarian supply chains differ depending on the decision level (strategic, tactical or operational) and the scope of the problem. For example, the location decision for an international distribution center for a large-scale NGO is a strategic decision based on the global service network of the NGO. Likewise, the location decision for NGOs mostly serving beneficiaries on a local level is also strategic, except that regional distribution centers are located based on the potential locations of the beneficiaries. Dispensing points, however, can be temporarily set up after a disaster and modified depending on the evolution of the situation. These are tactical decisions.

The dynamics of a supply chain and the performance measures depend on the intended scope of the operations. Again, for a local NGO, the last-mile performance—the timing of the distribution or the equity in aid provided to beneficiaries—can be more of a concern, whereas for an internationally active large-scale NGO, the resilience or the agility of its supply chain may be a more important issue. Thus, the specifics of the location decisions at the international level and within an underserved area can differ considerably. There can also be important differences among the various stakeholders, who may not have the same objectives (Gralla et al. 2014; Holguín-Veras et al. 2013).

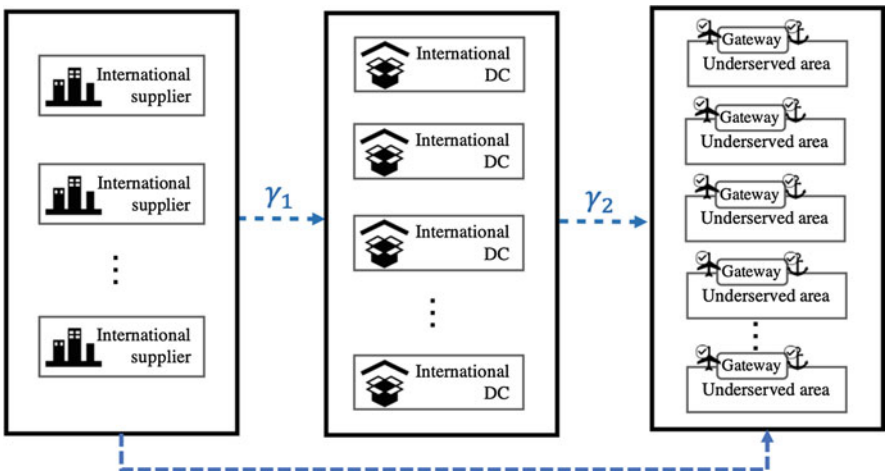


Fig. 21.2 The general structure of the global supply chain

21.2.1 Locations in Global Humanitarian Supply Chains

At the global level, the main location decisions involve the international distribution centers (DCs). As explained earlier, these DCs are mainly used for prepositioning inventory in order to reach an underserved area on time. They are also used to support long-term development programs. For these systems, every potential *underserved area* will be aggregated into a demand node, and the location decision will be based on the network, where the nodes are these aggregated points. Figure 21.2 depicts such a structure and the overall flow. The supplies will either flow through the DCs or will move from suppliers directly to affected areas. At this global level, the main criteria in determining the locations of DCs are the travel times and costs to bring the resources from the suppliers to the affected areas. In Fig. 21.2. γ_1 represents the value of the metric being used (typical metrics are cost, distance, time, or any combination of these) between the suppliers and the DC, and γ_2 represents the value of the metric between the DC and the underserved areas.

21.2.2 Locations in Local Humanitarian Supply Chains

At the local level, determining the best site for a regional distribution center (RDC) is an important location problem. Moreover, one of the crucial decisions is to establish where services or supplies need to be distributed to beneficiaries. These facilities (dispensing points) can be categorized as (1) shelter sites, (2) field hospitals, or (3) points of distribution (PoDs). Depending on the decision maker and

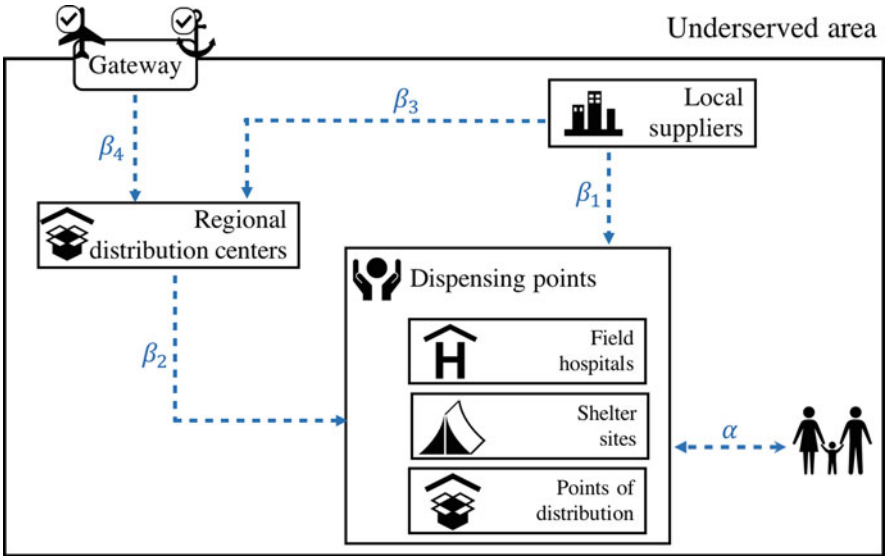


Fig. 21.3 The details of the affected area

the problem at hand, the location decision can be for the RDC or for the dispensing points. Figure 21.3 depicts in more detail the underserved area.

Figure 21.3 focuses on the underserved area, and hence the decisions are mainly made at the local level. The location decision of the RDC is mostly affected by β_2, β_3 and β_4 in the figure. As in the global supply chain shown in Fig. 21.2, β_i may represent several possible metrics depending on the preferences of the decision maker. At the local level, time and distance are among the typical metrics used. Customarily, the sum of all metrics is used to determine the optimal RDC location. However, it is possible to determine other values in these decisions (e.g., summing some of the β_i metrics and imposing upper bounds on the remaining ones). For example, it could be decided that minimizing the total cost or distance from the suppliers and from the gateway, while keeping the travel time or distance to the shelter sites below a certain value, is the most efficient way to make a location decision.

Although the facilities in the underserved area are grouped together in Fig. 21.3 for visual purposes, the location decisions of shelter sites, field hospitals, and PoDs are usually assessed through different performance measures and metrics. For example, in determining the optimal shelter site locations, the main decision maker—the NGO responsible for the shelter sites or the municipalities—may wish to minimize the total cost or distance to local suppliers and the RDC ($\beta_1 + \beta_2$). At the same time, the walking distance α of the beneficiaries cannot exceed a predetermined value (Kinay et al. 2018; Kinay et al. 2019). On the other hand, the distance to beneficiaries is more important in determining the optimal location of PoDs since the beneficiaries will come and collect supplies on a regular basis.

Thus, a typical performance measure in locating PoDs is the sum of α's. For these problems, the distance to local suppliers and the RDC (β_1 and β_2) can also be controlled by forcing predetermined upper bounds.

In sum, depending on the decision maker and the location problem considered, any combination or function of the values of the metrics given as α, β_1, β_2, β_3, and β_4 can appear as a performance measure or as a constraint. From the beneficiaries perspective, α is always important. For some applications, the sum of those distances may appear as an objective (Rancourt et al. 2015). The maximum value among the α's is usually considered to measure equity.

21.2.3 General Overview

As detailed in Sects. 21.2.1 and 21.2.2, the dynamics of the facility location problems encountered in global and local humanitarian supply chains may be different. Table 21.1 summarizes potential measures and the relationships among the metrics given in Figs. 21.2 and 21.3. In Table 21.1, $f(.)$ represents a general function of the different metrics, and studies typically used a weighted sum. For some applications, one echelon (either from the supply side or the demand side) may have more impact on the overall performance, and the decision maker may wish to distinguish that effect. For these applications, weighted sum is the most widely used function. For example, for international DC locations, if both echelons have the same effect, a typical objective function could be min $\gamma_1 + \gamma_2$. However, if the distance from suppliers has more effect on the overall decision, then a typical objective function would be min $a\gamma_1 + b\gamma_2$, where $0 \leq b < a \leq 1$.

Table 21.1 Overview of location decisions

Location decisions	Potential metrics	Potential performance measures	Sources of inbound flow	Destination of outbound flow/potential user
International DCs	Distance, cost	min $f(\gamma_1, \gamma_2)$	International suppliers	Affected areas
Regional DCs	Distance, cost, time	min $f(\beta_2, \beta_3, \beta_4)$, min $f(\beta_3, \beta_4)$ while $f(\beta_2) <$ threshold	Gateway, local suppliers	Locations of facilities
Shelter sites	Distance, cost, time	min $f(\beta_1, \beta_2)$ while $f(\alpha) <$ threshold	Regional DC, local suppliers	People
Field hospitals	Distance, cost, time	min $f(\beta_1, \beta_2)$ while $f(\alpha) <$ threshold	Regional DC, local suppliers	People
PoDs	Distance, cost, time	min $f(\alpha)$ while $f(\beta_1)$ and $f(\beta_2) <$ thresholds	Regional DC, local suppliers	People

As can be seen from Table 21.1, the location decisions of the shelter sites, field hospitals, and PoDs are customarily based on the values of the metrics between the facilities and the RDCs or local suppliers. The function f is usually a weighted sum. The prominent factor is the value of the metrics from and to the RDC, β_2, but depending on the contribution of local suppliers, the value of that metric, β_1, can also be a measure in determining the optimal locations of the shelter sites, field hospitals, or PoDs. The walking distance of the beneficiaries to the facilities, which is a typical metric for α, can also be a factor in determining the locations of these facilities. In particular, for the location of PoDs, the sum of the values of α is a typical performance measure.

A critical look at Table 21.1 immediately indicates the complexity encountered in humanitarian location problems based on the multiple criteria that need to be considered. As can be seen from the table, for some applications some of the criteria are naturally handled within the constraints, while others remain in the objective function. Even though a weighted sum is often used to handle the multiple criteria of the objective functions, recent studies have also used epsilon constraint methods for this purpose (Kinay et al. 2019).

21.3 A Generic Location Model for Humanitarian Supply Chains

As detailed in Sect. 21.2, there exist different types of facility location problems encountered in humanitarian supply chains. We will provide a generic mathematical model with appropriate definitions for sets, parameters, and variables depending on the scope and type of the problem. We will consider the weighted sum of the metrics for the performance measure.

21.3.1 Notation

Table 21.2 provides the set declarations for each type of facility location problem. Generically, K represents the set of suppliers, I the set of demand points, and J the set of alternative locations. Since alternative locations will be determined by the decision maker depending on availability in the local area, there is no need for distinction between the facility types.

Observe from the table that shelter sites act as demand points for field hospital and PoD location problems. Table 21.3 provides the parameter definitions and links these to the metrics given in Figs. 21.2 and 21.3. Table 21.4 provides the variable declarations.

Table 21.2 Set definitions for each type of location problem

Sets	Facility types				
	Distribution centers		Dispensing points		
	International	Regional	Shelter site	Field hospital	PoD
K: set of suppliers	International suppliers	Gateway, local suppliers	Regional DC and local suppliers		
I: set of demand points	Underserved areas	Dispensing points	Homes, gathering points	Shelter sites, homes	
J: set of alternative locations	Alternative locations, globally	Alternative locations in the region			

Table 21.3 Set definitions for each type of location problem

Parameters	Definition	Index sets		Facility type				
				Distribution centers		Dispensing points		
				International	Regional	Shelter site	Field hospital	PoD
c^1_{kj}	Value of the metric between	$k \in K$	$j \in J$	γ_1	β_3, β_4	β_1, β_2		
c^2_{ij}		$i \in I$		γ_2	β_2	α		

Table 21.4 Variable definitions

Variables	Definition	Index sets
y_j	Binary variable being $\begin{cases} 1 & \text{if } j \text{ is selected} \\ 0 & \text{otherwise.} \end{cases}$	$j \in M$
x_{ij}	Binary variable being $\begin{cases} 1 & \text{if } i \text{ is assigned to } j \\ 0 & \text{otherwise.} \end{cases}$	$i \in I,\ j \in J$

21.3.2 Basic Mathematical Model

As in many location problems introduced in Chaps. 2–5, the first decisions to be made are the locations of the facilities and how to allocate demand points to the opened facilities. These two decisions are modelled via the *base model* P_1 presented below. Observe that the variable declaration and the provided model is for a single assignment problem where each demand node receives service from exactly one facility. However, by declaring the allocation variable as $x_{ij} \geq 0$, one can easily allow a demand node to be served by more than one facility, namely the multi-allocation versions:

$$(P_1) \qquad \text{Minimize} \sum_{k \in K,\ j \in J} c^1_{kj} y_j + \sum_{i \in I,\ j \in J} c^2_{ij} x_{ij} \tag{21.1}$$

subject to

$$x_{ij} \leq y_j \qquad \forall\, i \in I,\ j \in J \quad (21.2)$$

$$\sum_{j \in J} x_{ij} = 1 \qquad \forall\, i \in I \quad (21.3)$$

$$x_{ij}, y_j \in \{0; 1\} \qquad \forall\, i \in I,\ j \in J. \quad (21.4)$$

The above model only considers the total sum of the metrics as the performance measure and has no additional requirement. Of course, depending on the application area, there will be many additional constraints to represent the operational dynamics.

An immediate additional restriction is capacity. To consider capacity issues, we also need to define the amount of demand as a parameter for each demand point in addition to the capacity value for node $j \in J$. Let d_i, $i \in I$ denote the amount of demand at a node and Q_j as the capacity of alternative location j. Then, updating constraint (21.2) to the following constraint, (21.5), would suffice to consider the capacity issues:

$$\sum_{i \in I} d_i x_{ij} \leq Q_j y_j \quad \forall\, j \in J. \quad (21.5)$$

This model is just one example of a location problem based on capacity with no apparent specialization in terms of humanitarian supply chains other than the intended meaning of the facility types. This is intentional and managed thanks to the conceptualization of the problem as depicted in Figs. 21.2 and 21.3, and Table 21.1.

On the other hand, one distinguishing feature of humanitarian supply chains is the importance of equity, which usually comes to light at the regional level when we consider the α metric regarding the beneficiaries. The current version of the model P_1 considers the total unweighted sum as the objective function, and thus equity is not recognized separately. One common way of dealing with such equity measures is by forcing the α metric to take values below a predetermined threshold value. Constraint (21.6) is such an example, where T is the threshold:

$$c_{ij}^2 x_{ij} \leq T \quad \forall\, i \in I,\ j \in J. \quad (21.6)$$

Of course including constraint (21.6) may lead to questioning the validity of the second term in the objective function of the model P_1. However, as also observed in Kinay et al. (2019), in order to ensure that the closest possible assignments are attained, this term is still required. For some applications, the performance measure can solely depend on the metric between the suppliers and the facilities, namely the first term in the objective function of P_1 (Kılcı et al. 2015). For such cases, in order to ensure meaningful assignments, one needs to include closest assignment constraints in the model.

Apart from multiple criteria, another distinguishing feature of humanitarian supply chains is the uncertainty that may arise in all aspects of the supply chain,

including the supply, demand, and transportation infrastructure (Kovács and Spens 2007). Most of the studies conducted on humanitarian supply chains focus on demand uncertainty. The uncertainty in supply is usually considered in terms of possible damage to the prepositioned items. Finally, the transportation infrastructure can also be damaged during a disaster, which may lead to uncertainty in travel times, distances, and costs. The studies concerning uncertainty in humanitarian supply chains do not focus solely on location decisions but also consider other paired decisions that will be detailed in the next section.

21.4 Location Problems with Additional Considerations

Both in humanitarian and commercial supply chains, strategic location decisions are often taken considering tactical and, sometimes, operational decisions, such as inventory prepositioning and distribution planning. In this section, we present examples of integrated location problems arising in humanitarian logistics.

21.4.1 Location and Prepositioning

Prepositioning is a distinctive practice used in humanitarian supply chains, where the stockpile levels for the equipment and relief supplies have to be determined based on the anticipated needs and available funding (tactical decisions). In terms of location decisions, organizations have to select where to open the storage facilities, such as international and regional distribution centers, in order to ensure an effective response depending on multiple factors (Richardson et al. 2016). These are usually strategic decisions, but in some cases temporary regional distribution centers can be installed using large relocatable tent-like structures to satisfy a significant surge in demand during a crisis.

The nature of prepositioning requires that the strategic location decisions and the stockpile levels be determined during the pre-disaster phase. Moreover, because future demand, among others, is unknown, solving problems related to prepositioning decisions usually implies stochastic optimization. Balcik and Beamon (2008), Mete and Zabinsky (2010), Rawls and Turnquist (2010) and Salmerón and Apte (2010) are among the pioneering works in stochastic inventory prepositioning and location problems. Two-stage stochastic programming is the most widely used method because it can take into account multiple uncertain parameters, such as demand, supply, and facility and transportation network damages. First-stage decisions take place prior to a disaster (facility locations and stockpile levels), while second-stage decisions are made in the aftermath of a disaster (distribution flow). For an extensive review of literature about two-stage stochastic programming in disaster management, see Grass and Fischer (2016). Some authors, such as Görmez et al. (2011) and Rodriguez-Espindola et al. (2018), have proposed solution

approaches that do not consider uncertainty, and Verma and Gaukler (2015) compare two location models, where the second model extends the first by considering damage intensity as a random variable.

Interesting real-life applications, through studies made in collaboration with different organizations, have been presented in the literature. For example, Duran et al. (2011) have studied the prepositioning network design of CARE International and evaluated the effect of gradually expanding its network based on the average response time, this by means of scenario-based stochastic programming. Acimovic and Goentzel (2016) and Dufour et al. (2018) have studied the prepositioning network of the UNHRD, whereas Jahre et al. (2016) have examined one of the networks of the United Nations High Commissioner for Refugees (UNHCR). Tofighi et al. (2016) provide an extension of the classic prepositioning problem. They consider the location and inventory prepositioning decisions in a two-echelon setting and compare their solutions with the existing relief network for Tehran.

In the following, we present a two-stage stochastic programming model, P_2, for the prepositioning network design problem. Given a set of suppliers and a set of candidate locations for the distribution centers where to preposition different relief items (e.g., blankets, mosquito nets, tarpaulins, family and hygiene kits), the first stage decisions of the model determine the locations of the DCs and the number of relief items shipped from suppliers to the opened DCs. Given a set of demand points and a set of possible demand scenarios, the second stage decisions determine the number of relief items shipped from the opened DCs to the demand points. We next present the notation to formulate the problem.

Sets

R set of relief items; $r \in R$
K^r set of suppliers of relief item r; $k \in K^r$
J set of candidate locations for the DCs; $j \in J$
I set of demand points; $i \in I$
S set of demand scenarios; $s \in S$.

Parameters

l_j fixed location cost for a DC at location $j \in J$

c_{kj}^{1r} cost of acquiring and shipping a relief item of type $r \in R$ from supplier $k \in K^r$ to DC $j \in J$

c_{ij}^{2r} cost of shipping a relief item of type $r \in R$ from DC $j \in J$ to demand point $i \in I$

q_r volume of relief item $r \in R$

Q_j capacity of a DC (measured in volume) at location $j \in J$

d_i^{sr} estimated demand of relief items of type $r \in R$ at demand point $i \in I$ in scenario $s \in S$.

First-Stage Decision Variables

y_j a binary variable equal to 1 if a DC is located in $j \in J$, 0 otherwise

z_{kj}^r amount of relief item of type $r \in R$ delivered to candidate DC location $j \in J$ from supplier $k \in K^r$.

Second-Stage Decision Variables

x_{ij}^{sr} amount of relief item of type $r \in R$ delivered to demand point $i \in I$ from candidate DC location $j \in J$ in scenario $s \in S$.

The two-stage stochastic programming model for the prepositioning network design problem is as follows:

$$(P_2) \quad \text{Minimize} \sum_{j\in J} l_j y_j + \sum_{r\in R}\sum_{k\in K^r}\sum_{j\in J} c_{kj}^{1r} z_{kj}^r + \sum_{s\in S} p_s \sum_{r\in R}\sum_{k\in K^r}\sum_{j\in J} c_{ij}^{2r} x_{ij}^{sr} \tag{21.7}$$

$$\text{subject to} \quad \sum_{r\in R}\sum_{k\in K^r} q_r z_{kj}^r \le Q_j y_j \qquad \forall\, j \in J \tag{21.8}$$

$$\sum_{r\in R}\sum_{j\in J} x_{ij}^{sr} \ge d_i^{sr} \qquad \forall\, i \in I,\ s \in S,\ r \in R \tag{21.9}$$

$$\sum_{i\in I} x_{ij}^{sr} \le \sum_{k\in K^r} z_{kj}^r \qquad \forall\, j \in J,\ r \in R \tag{21.10}$$

$$y_j \in \{0; 1\} \qquad \forall\, j \in J \tag{21.11}$$

$$z_{kj}^r \in \mathbb{Z}^+ \qquad \forall\, r \in R,\ j \in J,\ k \in K^r \tag{21.12}$$

$$x_{ij}^{sr} \in \mathbb{Z}^+ \qquad \forall\, i \in I,\ j \in J,\ s \in S,\ r \in R. \tag{21.13}$$

The first and second terms of the objective function (21.7) represent the sum of the fixed location costs associated with DCs, and the cost associated with acquiring and shipping relief items from suppliers to DC locations, respectively. The last term in (21.7) is the expected shipping costs associated with transportation of relief items after a disaster occurs. Constraints (21.8) ensure that the amount of relief supplies to preposition at each opened DC does not exceed its capacity. Constraints (21.10) limit the amount of relief items of type r that can be shipped from a DC by the amount of available relief items. Constraints (21.9) ensure that the estimated demand is fully met. Finally, constraints (21.11)–(21.13) define the domains of the variables.

21.4.2 Location-Routing Problems

Routing decisions can be integrated into facility location problems to account for transportation activities when designing and planning humanitarian supply chains, especially for locating both international and regional DCs or PoDs. Afshar and Haghani (2012) proposed a location-routing model to control the flow of relief items after a disaster in which several layers of temporary facilities are located based on storage and transportation capacity constraints. Apart from relief distribution, routing decisions can also be considered for evacuation purposes. In this case, emergency facilities where evacuees are transferred must be located, and the flow of people between affected households must be considered for the routing decisions (see An et al. 2013, Bayram and Yaman 2015 and Yi and Özdamar 2007). When location decisions are integrated with routing decisions, on-time deliveries or pick-ups and demand coverage often appear in the objective function because of resource limitations. For example, Abounacer et al. (2014) and Rath and Gutjahr (2014) have conducted studies that aim to minimize the unmet demand while maximizing the demand coverage.

VonAchen et al. (2016) and Cherkesly et al. (2017) addressed a location-routing covering problem that arose in a community healthcare program servicing under-served areas of Liberia. Designing a network in such a context implies determining the locations and density of community healthcare workers (CHWs) and of their supervisors, as well as the routes supervisors need to travel to provide continuous in-service training to CHWs. In this problem, the objective is to minimize the annual program costs (salaries, medical supplies, motorcycle routing, etc.) while ensuring specific coverages. There are two levels of location decisions: the supervisors who train the CHWs, and the CHWs who cover the communities. Thus, as opposed to the classic location problems, the location decisions here involve staff (CHWs and supervisors) not facilities, and the routing decisions also involve staff (supervisors) not vehicles. VonAchen et al. (2016) proposed a two-step heuristic to solve the problem and Cherkesly et al. (2017) solved the problem exactly by integrating all decisions and features into one optimization model. Other applications of healthcare delivery services and relief distribution have been presented where the underlying optimization problem is a coverage problem. For example, Nolz et al. (2010) and Naji-Azimi et al. (2012) adapted the coverage problem for relief distribution in areas affected by a disaster, while Doerner et al. (2007) and Hodgson et al. (1998) adapted it for planning mobile clinics in developing countries.

Uncertainty in the transportation infrastructure is one of the major challenges in humanitarian supply chains. Rawls and Turnquist (2010) conducted one of the first study to consider uncertainty in the network availability due to possible road damages. The authors developed a scenario-based solution approach to determine the location and capacity of the emergency supply storage facilities, as well as the amount of supplies to be prepositioned. Ahmadi et al. (2015) proposed a two-stage stochastic location-routing problem under disrupted networks. Salman and Yucel (2015) also considered random transportation network link failures and proposed a

tabu search heuristics for finding a surviving supply chain under many scenarios. Moreno et al. (2016) and Vahdani et al. (2018) developed a three-echelon location and routing problem under demand uncertainty, whereas Bozorgi-Amiri and Khorsi (2016) considered a prepositioning problem for a two-echelon network structure.

21.5 Conclusion

Humanitarian supply chains are complex systems with features that are specific to each context. Indeed, they are subjected to multiple sources of uncertainty (location and magnitude of the aid needs and available supplies, capacity of the logistics network, impacts of the potential aftershocks, etc.) and some security issues. Multiple stakeholders with different objectives and incentives are involved in humanitarian operations (international organizations, NGOs, government agencies, media, beneficiaries, etc.), which makes humanitarian supply chain management especially challenging. There are also specific considerations in the humanitarian sector that do not apply to their for-profit counterparts. This leads to various location problems at different levels of the supply chain, including international and regional DCs as well as dispensing points. Integrating other considerations with location problems, including routing and coverage decisions as well as sources of uncertainty, offers interesting contributions and opens the door to future research in the field of humanitarian logistics.

References

Abdulrahman B, Garrity S, Moodley K (2014) An overview of the designs of disaster relief shelters. Procedia Econ Finance 18:924–931

Abounacer R, Rekik M, Renauda J (2014) An exact solution approach for multi objective location transportation problem for disaster response. Comput Oper Res 41:83–93

Acimovic J, Goentzel J (2016) Models and metrics to assess humanitarian response capacity. J Oper Manag 45:11–29

Afshar A, Haghani A (2012) Modeling integrated supply chain logistics in real-time large-scale disaster relief operations. Socio Econ Plan Sci 46(4):327–338

Ahmadi M, Seifi A, Tootooni B (2015) A humanitarian logistics model for disaster relief operation considering network failure and standard relief time: a case study on San Francisco district. Transp Res E-Log 75:145–163

An S, Cui N, Li X, Ouyang Y (2013) Location planning for transit-based evacuation under the risk of service disruptions. Transp Res B: Methodol 54:1–16

Balcik B, Beamon BM (2008) Facility location in humanitarian relief. Int J Log Res Appl 11(2):101–121

Bayram V, Yaman H (2015) A stochastic programming approach for shelter location and evacuation planning. Optim Online 2015-09-5088

Bozorgi-Amiri A, Khorsi M (2016) A dynamic multi-objective location-routing model for relief logistic planning under uncertainty on demand, travel time, and cost parameters. Int J Adv Manuf Tech 85:1633–1648

Çelik M, Ergun Ö, Johnson B, Keskinocak P, Lorca Á, Pekgün P, Swann J (2012) Humanitarian logistics. In: Mirchandani P (ed) INFORMS Tutor Oper Res, vol 9. INFORMS, Hanover, pp 18–49

Cherkesly M, Rancourt M-È, Smilovitz K (2017) A set-partitioning formulation for community healthcare network design in underserved areas. Les Cahiers du CIRRELT, CIRRELT-2017–24

Doerner K, Focke AL, Gutjahr WJ (2007) Multicriteria tour planning for mobile healthcare facilities in a developing country. Eur J Oper Res 179(3):1078–1096

Dufour É, Laporte G, Paquette J, Rancourt M-È (2018) Logistics service network design for humanitarian response in East Africa. Omega 74:1–14

Duran S, Gutierrez MA, Keskinocak P (2011) Pre-positioning of emergency items for CARE international. Interfaces 41(3):223–237

Görmez N, Köksalan M, Salman FS (2011) Locating disaster response facilities in Istanbul. J Oper Res Soc 62:1239–1252

Gralla E, Goentzel J, Fine C (2014) Assessing trade offs among multiple objectives for humanitarian aid delivery using expert preferences. Prod Oper Manag 23(6):978–989

Grass E, Fischer K (2016) Two-stage stochastic programming in disaster management: a literature survey. Surv Oper Res Manag Sci 21(2):85–100

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(4):621–638

Holguín-Veras J, Pérez N, Jaller M, van Wassenhove LN, Aros-Vera F (2013) On the appropriate objective function for post-disaster humanitarian logistics models. J Oper Manag 31(5):262–280

International Federation of Red Cross and Red Crescent Societies (2017) Introduction to the Guidelines for the domestic facilitation and regulation of international disaster relief and initial recovery assistance. IFRC, Geneva, Switzerland. https://www.ifrc.org/PageFiles/41203/1205600-IDRL%20Guidelines-EN-LR%20(2).pdf. Cited 19 Sep 2018

Jahre M, Kembro J, Rezvanian T, Ergun O, Hapnes SJ, Berling P (2016) Integrating supply chains for emergencies and ongoing operations in UNHCR. J Oper Manag 45:57–72

Jahre M, Kembro J, Adjahossou A, Altay N (2018) Approaches to the design of refugee camps: an empirical study in Kenya, Ethiopia, Greece, and Turkey. J Humanit Log Supply Chain Manag. https://doi.org/10.1108/JHLSCM-07-2017-0034

Kara BY, Savaşer S (2017) Humanitarian logistics. In: Leading developments from INFORMS communities. INFORMS, Berlin (2017), pp 263–303

Kılcı F, Kara BY, Bozkaya B (2015) Locating temporary shelter areas after an earthquake: a case for Turkey. Eur J Oper Res 243:323–332

Kinay ÖB, Kara BY, Saldanha-da-Gama F, Correria I (2018) Modeling the shelter site location problem using chance constraints: a case study for Istanbul. Eur J Oper Res 270(1):132–145

Kinay ÖB, Saldanha-da-Gama F, Kara BY (2019) On multi-criteria chance-constrained capacitated single-source discrete facility location problems. Omega 83:107–122

Kovács G, Spens KM (2007) Humanitarian logistics in disaster relief operations. Int J Phys Distrib Log Manag 37(2):99–114

McCoy JH, Lee HL (2014) Using fairness models to improve equity in health delivery fleet management. Prod Oper Manag 23(6):965–977

Mete HO, Zabinsky ZB (2010) Stochastic optimization of medical supply location and distribution in disaster management. Int J Prod Econ 126(1):76–84

Moreno A, ALem D, Ferreira D (2016) Heuristic approaches for the multiperiod location-transportation problem with reuse of vehicles in emergency logistics. Comput Oper Res 69:79–96

Naji-Azimi Z, Renaud J, Ruiz A, Salari M (2012) A covering tour approach to the location of satellite distribution centers to supply humanitarian aid. Eur J Oper Res 222(3):596–605

Nolz PC, Doerner KF, Gutjahr WJ, Hartl RF (2010) A bi-objective metaheuristic for disaster relief operation planning. In: Coello CA, Dhaenens C, Jourdan L (eds) Advances in multi-objective nature inspired computing. Studies in computational intelligence, vol 272. Springer, Berlin, pp 167–187

Rancourt M-È, Cordeau J-F, Laporte G, Watkins B (2015) Tactical network planning for food aid distribution in Kenya. Comput Oper Res 56:68–83

Rath S, Gutjahr WJ (2014) A mathematical heuristic for the warehouse location routing problem in disaster relief. Comput Oper Res 42:25–39

Rawls CG, Turnquist MA (2010) Pre-positioning of emergency supplies for disaster response. Transp Res Part B: Methodol 44(4):521–534

Richardson DA, de Leeuw S, Dullaert W (2016) Factors affecting global inventory prepositioning locations in humanitarian operations: a Delphi study. J Bus Log 37(1):59–74. https://doi.org/10.1111/jbl.12112

Rodriguez-Espindola O, Albores P, Brewster C (2018) Disaster preparedness in humanitarian logistics: a collaborative approach for resource management in floods. Eur J Oper Res 264(3):978–993

Salman FS, Yucel E (2015) Emergency facility location under random network damage: insights from the Istanbul case. Comput Oper Res 62:266–281

Salmerón J, Apte A (2010) Stochastic optimization for natural disaster asset prepositioning. Prod Oper Manag 19(5):561–574

Tofighi S, Torabi SA, Mansouri SA (2016) Humanitarian logistics network design under mixed uncertainty. Eur J Oper Res 250(1):239–250

United Nations (2018) About the sustainable development goals. https://www.un.org/sustainabledevelopment/sustainable-development-goals/. Cited 19 Sep 2018

United Nations High Commissioner for Refugees (2018) The World in Number. In: UNHCR Statistics http://popstats.unhcr.org/en/overview#_ga=2.259337609.202445162.1537117566-337189774.1499619310. Cited 24 Sep 2018

US Department of Homeland Security, Office of Inspector General (2009) FEMA's sourcing for disaster response goods & services. In: Report of the US Department of Homeland Security. Office of Inspector General, Washington. https://www.oig.dhs.gov/assets/Mgmt/OIG_09-96_Aug09.pdf/. Cited 19 Sep 2018

Vahdani B, Veysmoradi D, Noori F, Mansour F (2018) Two-stage multi-objective location-routing-inventory model for humanitarian logistics network design under uncertainty. Int J Disast Risk Reduct 27:290–306

Verma A, Gaukler GM (2015) Pre-positioning disaster response facilities at safe locations: an evaluation of deterministic and stochastic modeling approaches. Comput Oper Res 62:197–209

VonAchen P, Smilowitz K, Raghavan M, Feehan R (2016) Optimizing community healthcare coverage in remote Liberia. J Humanit Log Supply Chain Manag 6(3):352–371

Yi W, Özdamar L (2007) A dynamic logistics coordination model for evacuation and support in disaster response activities. Eur J Oper Res 179(3):1177–1193

Chapter 22
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? The 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.

22.1 Introduction

Although Murphy's 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). Other examples of natural disruption include the

M. P. Scaparra (✉)
Kent Business School, University of Kent, Canterbury, United Kingdom
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*,
https://doi.org/10.1007/978-3-030-32177-2_22

hit of Hurricane Harvey on Texas in 2017. Included in that disaster was a chemical plant that was flooded in Crosby, TX, which lost power as well as backup power. The chemicals stored at the plant needed refrigeration, and without power there were significant destructive fires. Harm can also be intentional and simple. For example, in 2015 a cyber-attack shut down three power distribution companies in the Ukraine resulting in loss of electricity to 225,000 customers in winter (Lemos 2018). In another incident, a terrorist was able to drive a vehicle into an Air Products & Chemical plant near Lyon, France, in 2015 that caused an explosion (CEN 2015). Of equal concern is that attackers used phishing emails to gain passwords and compromising information. By doing so, they were able to launch a cyber-attack on a steel mill in Germany in 2014. The attackers had enough familiarity with the system that they caused the plant's control network to fail. In response, plant operators had to perform an emergency shutdown which resulted in significant damage (Lemos 2018). 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 believed 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 transportation 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. Since then 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 call for something simple like providing backup generators for power or providing enough security that it will ward off a would-be attacker. Another possibility is to move 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:

1. *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.
2. *Protection models*. These models optimize the allocation of protective resources among the facilities of already existent systems.
3. *Design models*. These models are used for planning new service and supply systems which are secure and resilient to disruptions.

In this chapter, we 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 remainder of this chapter is organized as follows. Section 22.2 introduces the notation used throughout the chapter. Interdiction, protection and design models are described in Sects. 22.3, 22.4 and 22.5, respectively. In Sect. 22.6, we highlight future trends in modeling location problems under disaster events. Some conclusive remarks are finally provided in Sect. 22.7.

22.2 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
$c_{ij} =$ Unit 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

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

$$s_i = \begin{cases} 1 & \text{if a facility located at } i \text{ is interdicted} \\ 0 & \text{otherwise} \end{cases}$$

$$z_i = \begin{cases} 1 & \text{if a facility located at } i \text{ is protected} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{ij} = \begin{cases} 1 & \text{if the demand of customer } j \text{ is supplied from facility } i \\ 0 & \text{otherwise} \end{cases}$$

$$u_j = \begin{cases} 1 & \text{if customer } j \text{ is covered before disruption} \\ 0 & \text{otherwise} \end{cases}$$

$$v_j = \begin{cases} 1 & \text{if customer } j \text{ is covered after disruption} \\ 0 & \text{otherwise} \end{cases}$$

In the models described in this chapter, single-sourcing is assumed. For some uncapacitated problems, such as the p-median problem, single-sourcing occurs naturally (without imposing binary restrictions on the x_{ij} variables) as customer demands are served by their nearest open facility, unless a customer has the same minimum cost from two or more open facilities (see Chap. 2). The multi-source counterpart of location models under disruption can be easily formulated by relaxing the integrality constraints on the x_{ij} variables.

22.3 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 systems that are designed with an objective that is either based on minimizing total weighted distance of service or maximizing coverage. 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.

The r-Interdiction Median Problem

In addition to the notation introduced in Sect. 22.2, 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:

$$\text{maximize} \quad \sum_{i \in F} \sum_{j \in J} d_j c_{ij} x_{ij} \tag{22.1}$$

$$\text{subject to} \quad \sum_{i \in F} x_{ij} = 1 \quad \forall j \in J \tag{22.2}$$

$$\sum_{i \in F} s_i = r \tag{22.3}$$

$$\sum_{k \in T_{ij}} x_{kj} \leq s_i \quad \forall i \in F, j \in J \tag{22.4}$$

$$x_{ij} \in \{0, 1\} \quad \forall i \in F, j \in J \tag{22.5}$$

$$s_i \in \{0, 1\} \quad \forall i \in F. \tag{22.6}$$

The objective function (22.1) maximizes the demand-weighted total cost after the interdiction of r facilities. Constraints (22.2) ensure that each customer is assigned to a facility after interdiction. Constraints (22.3) stipulate that exactly r facilities are to be interdicted. Constraints (22.4) force each customer j to be assigned to its closest non-interdicted facility. In particular, 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 (22.5) and (22.6) 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 (22.3) 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 (22.4) 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. (2014). 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. A similar problem was considered by Zhang et al. (2016).

The r-Interdiction Covering Problem

The r-Interdiction Covering Problem (r-ICP) can be stated mathematically as follows:

$$\text{minimize} \quad \sum_{j \in J} d_j v_j \tag{22.7}$$

$$\text{subject to} \quad v_j \geq 1 - s_i \quad \forall j \in J, i \in N_j \cap F \tag{22.8}$$

$$\sum_{i \in F} s_i = r \tag{22.9}$$

$$v_j \in \{0, 1\} \quad \forall j \in J \tag{22.10}$$

$$s_i \in \{0, 1\} \quad \forall i \in F. \tag{22.11}$$

The objective function (22.7) minimizes the amount of customer demand which is covered after interdiction. Constraints (22.8) 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 (22.9) force the number of facilities to be eliminated to equal r. The last two sets of constraints (22.10) and (22.11) 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.

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) proposed the Hub Interdiction Median Problem which identifies the most critical hub facilities in hub–and–spoke systems.

22.4 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 systems which are designed using the p-median problem (Scaparra and Church 2008a).

The r-Interdiction Median Problem with Fortification

The bilevel formulation of the r-IMP with Fortification (r-IMPF) is as follows.

$$\text{minimize} \quad H(z) \tag{22.12}$$

$$\text{subject to} \quad \sum_{i \in F} z_i = b \tag{22.13}$$

$$z_i \in \{0, 1\} \quad \forall i \in F, \tag{22.14}$$

where

$$H(z) = \max \quad \sum_{i \in F} \sum_{j \in J} d_j c_{ij} x_{ij} \tag{22.15}$$

$$\text{s.t.} \quad s_i \leq 1 - z_i \tag{22.16}$$

$$(22.2) - (22.6).$$

The leader objective (22.12) is to minimize the highest possible level of demand-weighted service cost, H, following the disruption on r facilities by allocating b protective resources (22.13). The worst-case cost H is computed in the follower problem, which is simply the r-IMP problem defined in Sect. 22.3 with the additional constraints (22.16). 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 the benefits of protection 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 (IE) 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 to optimality and has been successfully applied to problems in different settings as well (e.g., the network protection models in Cappanera and Scaparra 2011).

Note that in the presence of other complicating aspects, such as capacity constraints on the facilities, interdiction problems may require a bilevel formulation. Consequently, the addition of the protection layer results in trilevel models, which are even more challenging to solve. In these cases, the trilevel models are typically solved by using IE for the outer protection level, while other methods, such as decomposition or reformulation, are used for the interdiction bilevel model. Some examples of this are discussed later in this section.

The use of metaheuristics for solving r-IMPF has been recently explored by Cheng et al. (2016), who developed several hybrid approaches where Tabu Search, Simulated Annealing and Genetic Algorithms are used for solving the upper problem, whereas the lower interdiction problem is solved to optimality by a commercial solver. These metaheuristics are more versatile than exact methods based on implicit enumeration, as they do not make any assumption about the follower's problem. As a result, they can be applied to other settings (e.g., problems where a facility may be damaged partially or a facility may be lost only with certain probability).

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 (22.13) 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). A variation of this model can be found in Parajuli et al. (2017) who introduced the notion of gradual capacity backup to hedge against disruption risk in capacitated supply networks. Namely, facilities can be protected at different levels. Protection implies that a facility acquires contingent additional production capacity, and the amount of additional capacity is commensurate to the level of protection investment.

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 trilevel 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 a system which is based upon a p-median problem design. In this model, 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 systems based upon a p-median design, 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.

Recently, considerable attention has been paid to the protection of hub networks (Ghaffarinasab and Atayi 2018; Quadros et al. 2018; Ramamoorthy et al. 2018). These papers built upon the protection model for the multiple allocation hub interdiction median problem introduced by Lei (2013) and proposed different exact solution methodologies for solving it. Ghaffarinasab and Atayi (2018) introduced a two-level implicit enumeration algorithm based on Scaparra and Church (2008a) (one level of IE for protection and one for interdiction); Quadros et al. (2018) proposed a single level integer linear programming formulation for the problem and solved it through a branch-and-cut algorithm; Ramamoorthy et al. (2018) combined IE for the protection model with Benders decomposition for the interdiction model, after improving the lower level using novel closest assignment constraints.

Protection models have also been developed for location problems with hierarchical facilities (Aliakbarian et al. 2015) and for decentralized supply systems (Zhang and Zheng 2018).

22.5 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.

22.5.1 Planning Against Worst-Case Disruptions

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. 22.3 to identify the location of p facilities which maximizes a weighted combination of i) the initial coverage and ii) 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:

$$\text{maximize } \alpha \sum_{j \in J} d_j u_j + (1-\alpha) H(y) \tag{22.17}$$

$$\text{subject to } \sum_{i \in I} y_i = p \tag{22.18}$$

$$\sum_{i \in N_j} y_i \geq u_j \quad \forall j \in J \tag{22.19}$$

$$y_i \in \{0,1\} \quad \forall i \in I \tag{22.20}$$

$$u_j \in \{0,1\} \quad \forall j \in J, \tag{22.21}$$

where

$$H(y) = \min \sum_{j \in J} d_j v_j \tag{22.22}$$

$$\text{subject to } \sum_{i \in I} s_i = r \tag{22.23}$$

$$v_j \geq y_i - s_i \quad \forall j \in J, i \in N_j \tag{22.24}$$

$$s_i \in \{0,1\} \quad \forall i \in I \tag{22.25}$$

$$v_j \in \{0,1\} \quad \forall j \in J. \tag{22.26}$$

The upper-level objective (22.17) is to maximize the weighted sum of covered demand before and after interdiction by locating p facilities (22.18). Initial and post-disruption coverage are weighted in the objective by using a weight α, with $0 \leq \alpha \leq 1$. The demand covered before interdiction is determined by constraints (22.19), whereas the worst-case demand-weighted coverage after interdiction, $H(y)$, is computed in the lower level problem (22.22)–(22.26). This is a simple modification of the r-ICP problem (22.7)–(22.11), where constraints (22.8) are replaced by (22.24). 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. 22.4 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 so-called *supervalid inequalities*.[1]

[1]Supervalid inequalities can be seen as a generalization to bilevel decomposition methods of the standard valid inequalities inside a cutting plane algorithm.

Another example of location-interdiction models can be found in Parvaresh et al. (2014) for p-hub median problems. In this case, the bilevel model is solved heuristically via Simulated Annealing and Tabu Search. Ghaffarinasab and Motallebzadeh (2018) extended this work by introducing the hub interdiction problem under covering and center objectives. A worst-case model for the uncapacitated facility location problem can be found in Hernandez et al. (2014), where a multi-objective optimization approach is used to identify trade-off solutions with respect to the total weighted traveling distance before and after disruptions.

Note that design and protection decisions may be coupled within the same modeling framework. Risk-averse design problems including the option of hardening some of the facilities to be located have received considerable attention. See for example Keçici et al. (2012), Aksen and Aras (2012), Aksen et al. (2013), Shishebori and Jabalameli (2013), Medal et al. (2014), Akbari-Jafarabadi et al. (2017), Zhang et al. (2018) and Jalali et al. (2018). These problems have introduced several novel aspects into the facility protection and robust design literature. For instance, Zhang et al. (2018) considered for the first time the case where the interdictor has no information about the protection resource allocation. Jalali et al. (2018) assumed that facilities fail with some probability which depends on the combined effect of protection and interdiction efforts and used a conditional value-at-risk (CVaR) measure to capture the risk-averse attitude of the system designer.

Design decisions can also be used to identify efficient ways of protecting existing service facilities, as in the problem introduced by Mahmoodjanloo et al. (2016). This problem aims at locating defence facilities at minimum cost, so that each service facility is covered by at least one defence facility. The problem is modeled as a trilevel program, where the bilevel partial interdiction median model introduced by Aksen et al. (2014) is embedded into an outer coverage location model.

Although bilevel location-interdiction models are the most common way of capturing worst-case scenario disruptions, the use of two-stage Robust Optimization (RO) has recently been proposed as an alternative risk-averse approach to hedge against disruptions. RO-based location models use uncertainty sets to capture data uncertainty and seek to determine locations that are robust to any perturbations in the uncertainty sets, including worst-case scenario values. To model situations where some decisions can be made after the uncertainty is revealed, the RO framework can be extended to include second stage recourse decisions. An et al. (2014) proposed the first two-stage RO model to design reliable facility location networks subject to disruptions. Their models, designed for the reliable p-median problem, minimize the weighted sum of the operation costs in normal situations and in the worst disruptive scenario. They also considered two important practical features: facility capacities and demand change due to disruption. The proposed models are solved exactly by Benders decomposition and *column-and-constraint generation* methods. In recent years, two-stage RO approaches have been used to solve other more complex location problems under disruptions. For example, Zarrinpoor et al. (2017) proposed a hierarchical location-allocation model for health service network design which concurrently addresses several key issues, such as service quality, changes in demand patterns, hierarchical structure of networks,

disruption risk and uncertainty associated with demand and service within a queuing theory framework. Cheng et al. (2018) introduced a two-stage RO approach for the reliable logistics network design problem, which includes multiple echelons and facility capacities. To test different levels of conservativeness and study the price of robustness, the authors extended the basic RO scheme and proposed two model variants: the *expanded two-stage RO* model, which uses multiple uncertainty sets, and the *risk-constrained two-stage RO* model, where upper bounds are imposed on the worst-case performance. The application of the models indicates that a considerable decrease in the cost of the worst disruptive situation can be achieved for only a small increase in the normal cost.

22.5.2 Planning Against Random Disruptions

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 facility. Let m be the cardinality of the augmented set $|I|$ and the index of the dummy facility. The dummy facility m never fails and has unit service cost c_{mj} to customer j, which represents the lost-sale cost per unit of demand. As facility m is forced to open, $p + 1$ facilities must be located instead of p as in standard p-median problems. 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 ($0 < l \leq p$) are those made to alternative facilities that can serve a customer if the l closer facilities have failed.

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 RPMP model is as follows.

$$\text{minimize} \quad \sum_{j \in J} d_j \sum_{l=0}^{p} \left[\sum_{i \in I \setminus m} c_{ij} \pi^l (1-\pi) x_{ijl} + c_{mj} \pi^l x_{mjl} \right] \tag{22.27}$$

$$\text{subject to} \quad \sum_{i \in I} x_{ijl} + \sum_{t=0}^{l-1} x_{mjt} = 1 \quad \forall j \in J, l = 0, \ldots, p \tag{22.28}$$

$$\sum_{l=0}^{p} x_{ijl} \leq 1 \quad \forall i \in I, j \in J \tag{22.29}$$

$$x_{ijl} \leq y_i \quad \forall i \in I, j \in J, l = 0, \ldots, p \tag{22.30}$$

$$\sum_{i \in I} y_i = p + 1 \tag{22.31}$$

$$y_m = 1 \tag{22.32}$$

$$y_i \in \{0, 1\} \quad \forall i \in I \tag{22.33}$$

$$x_{ijl} \in \{0, 1\} \quad \forall i \in I, j \in J, l = 0, \ldots, p. \tag{22.34}$$

The objective function (22.27) 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 (22.28) state that each customer j must be assigned to some facility at each level l, unless j has been assigned to the dummy facility at level $t < l$. Constraints (22.29) prevent the assignment of a customer to a given facility at more than one level. Constraints (22.30) prohibit the assignment to facilities which are not open, whereas constraint (22.31) state that exactly p facilities must be opened in addition to the dummy facility, which is forced to be open by constraint (22.32). Constraints (22.33) and (22.34) 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 (22.27)–(22.34) in two aspects: i) some of the facilities may be considered completely reliable and ii) 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 (RFLP), 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. Tran et al. (2017) further extended the concepts of probability chains and introduced a novel network flow structure called a *probability lattice* to solve the reliable single-allocation p-hub median problem.

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), Berman et al. (2013), Li et al. (2013) and Lu et al. (2015). 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), Berman et al. (2013), Albareda-Sambola et al. (2015) and Yun et al. (2015).

An issue that has been largely neglected in the reliability location literature is the capacity of the facilities. Most existing reliability models assume that the facilities are uncapacitated and able to absorb the demand of disrupted facilities. As a consequence of this assumption, even the issue of partial facility failure has been mostly ignored. An exception is the study by Azad et al. (2013) which considers capacitated facilities, partial capacity loss due to disruption and goods sharing between non-disrupted and partially disrupted facilities. This problem was subsequently extended by Jabbarzadeh et al. (2016) who proposed a hybrid stochastic-robust optimization model, where a robust optimization approach was applied to the stochastic reliable capacitated facility location problem so as to capture additional uncertainties (i.e. demand fluctuations, probability of a disruption occurrence, supply capacity variations). An alternative way of dealing with potentially excessive demand at

non-failing, backup facilities has been considered by Madani et al. (2018) within the context of the reliable p-hub maximal covering problem. In this study, a bi-objective model is introduced, where the primary objective is to maximize the expected covered flow, whereas the secondary objective is to minimize congestion by balancing the flows passing through each hub.

Most existing reliability location models use expected costs or performances in the objective function, thus implicitly assuming that the decision maker is risk-neutral. Yu et al. (2017) argued that risk-averse approaches can provide more robust solutions compared to the risk-neutral approach and proposed two variants of RFLP which use risk-averse measures: conditional value-at-risk (CVaR) and absolute-semideviation (ASD). This study shows that different facility locations are selected under risk-averse measures and that the resulting systems are more reliable than the ones obtained with traditional risk-neutral objectives, but less conservative that the ones obtained with worst-case models.

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, Li and Savachkin 2013, Li et al. 2013 and Jabbarzadeh et al. 2016).

22.5.3 Planning Against Specific Disruption Scenarios

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 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:

$$\text{minimize} \quad \sum_{\omega \in \Omega} \pi_\omega \sum_{i \in I} \sum_{j \in J} d_j c_{ij} x_{ij\omega} \tag{22.35}$$

$$\text{subject to} \quad \sum_{j \in J} x_{ij\omega} \leq (1 - a_{i\omega}) y_i \quad \forall i \in I, \omega \in \Omega \tag{22.36}$$

$$\sum_{i \in I} x_{ij\omega} = 1 \quad \forall j \in J, \omega \in \Omega \tag{22.37}$$

$$\sum_{i \in I} y_i = P \tag{22.38}$$

$$y_i \in \{0, 1\} \quad \forall i \in I \tag{22.39}$$

$$x_{ij\omega} \in \{0, 1\} \quad \forall i \in I, j \in J, \omega \in \Omega. \tag{22.40}$$

The objective function (22.35) minimizes the demand-weighted expected cost across all scenarios. Constraints (22.36) 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 (22.37) 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 (22.35)–(22.40) 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 (see also Chap. 8). 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

$$\sum_{i \in I} \sum_{j \in J} d_j c_{ij} x_{ij\omega} \leq (1 + \beta) z^*_\omega \quad \forall \omega \in \Omega, \tag{22.41}$$

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 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. 22.5.2 cannot. As an example, capacity restrictions can be easily modeled by replacing constraints (22.36) with

$$\sum_{j \in J} d_j x_{ij\omega} \leq (1 - a_{i\omega}) q_i y_i \quad \forall i \in I, \omega \in \Omega, \tag{22.42}$$

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. An example of partial disruption in scenario-indexed models can be found in Fattahi et al. (2017) for a supply chain network (SCN) design problem. The SCN is composed of customers, warehouses and factories and involves multiple products and multiple periods. Lead times are based upon which facility/warehouse combination serves a given customer. Because of possible disruptions, some customers may not be served, which incurs a penalty cost. Although the factories are already located and fixed in number, warehouses are to be located over the planning horizon. Warehouses can be protected at selected fortification levels which limits disruption to certain levels of capacity. Single source delivery is assumed and demands at customers depend on the facilities serving them based on their delivery lead times. The objective is to minimize supply chain costs, including lead times in product delivery and warehouse recovery costs, by locating warehouses, selecting protection levels and assigning factory/customer supply chains to each demand.

Another scenario-based model which considers the effects of disruption on facility capacities is the risk-aware capacitated plant location problem (CPLP-RISK) introduced by Heckmann (2016). CPLP-RISK is a two-stage stochastic model, where the first-stage decisions include which facilities to open and whether to equip them with the option of capacity expansion that can be used when a disruption occurs; the recourse or second-stage decisions involve the selection of the capacity expansion's level and duration. A finite set of scenarios is used to model facility capacity reductions and customer demand fluctuations over time. The objective is to minimize the overall system costs (i.e., facility opening costs, capacity expansion costs and service costs) and the service deterioration level due to unmet demand in case of disruption.

Very recently scenario-indexed models have been studied for hub-and-spoke networks by Rostami et al. (2018) and Zhalechian et al. (2018). Particularly noteworthy is the comprehensive model introduced in the latter paper, which integrates several interesting issues such as: operational risks (i.e. fluctuations in input data) and disruption risks; proactive (mitigation) and reactive (recovery)

strategies to increase resilience; and three different measures of network design quality (network density, network complexity and node criticality).

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 combinations of facilities that can fail). To obviate this difficulty, the scenario space can be approximated using sampling techniques such as Sample Average Approximation (SAA) (Kleywegt et al. 2002). An innovative application of this method can be found in Aydin and Murat (2013) for the capacitated reliable facility location problem. In this study, Particle Swarm Optimization is integrated within the SAA methodology to improve the computational efficiency and solution quality of traditional SAA implementations. 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 fifteen hurricanes.

Note that in standard two-stage stochastic optimization, first-stage decisions must ensure that the solution feasibility is maintained for each scenario realization. A new paradigm, called *Recoverable Robust Optimisation*, has recently been proposed by Liebchen et al. (2009), where first-stage decisions can be revisited once the uncertainty is resolved in the second stage. In particular, the solution built in the first stage can be recovered through a limited set of recovery actions. This paradigm has been used by Álvarez-Miranda et al. (2015) for the uncapacitated facility location problem under disruptions. The objective of the recoverable robust location problem is to minimize the sum of the first-stage cost (i.e. the cost of the initial facility location and customer allocation), plus the second-stage recovery cost (i.e. the worst-case cost to recover the solution over all possible scenarios). The second-stage recovery actions include the opening of new facilities and the re-allocation of customers that were allocated in the first-stage to facilities which are unavailable in the realized scenario.

22.6 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. Work has also involved systems that do not rely on single-source service assignment, like the defender-attacker-defender

model of Scaparra and Church (2012). Their model dealt with the protection of a system of capacitated facilities, with an embedded classical transportation problem. 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 (see for example Heckmann 2016) have addressed the possible duration of a disrupting event as well as how best to cope with it and restore the initial operational level (Heckmann 2016). This too, is an area where more research is needed.

Facilities are but one component in a production and distribution system. Flooding in Thailand in 2011 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. Hurricane Harvey, which hit Texas in 2017, affected more than 13,000 business entities in the flood envelope, including oil refineries, plastic molding facilities, and chemical plants (Chang 2017). Petroleum and coal products manufacturing, chemical manufacturing, and oil and gas extraction suffered the greatest impact. These three critical subsectors provide raw materials for other industries, and their disruption had a ripple effect on the raw materials supply chain. The disruption propagated to other industries and countries that rely on these or related exports from the Port of Houston. Although recent studies have attempted to consider multi-echelon distribution systems, the design of robust risk-optimized supply chain networks and the development of improved supply chain risk management strategies still require additional research to fully capture cross-sector and cross-country business interruption risk.

There are three principal ways in which resilient design has been approached: robust, stochastic 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 approach? 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?

22.7 Conclusions

This chapter has reviewed the research that has evolved over the last 15 years 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

Akbari-Jafarabadi M, Tavakkoli-Moghaddam R, Mahmoodjanloo M, Rahimi Y (2017) A tri-level r-interdiction median model for a facility location problem under imminent attack. Comp Ind Eng 114:151–165

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. CEJOR 18:269–291

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

Aksen D, Aras N, Piyade N (2013) A bilevel p-median model for the planning and protection of critical facilities. J Heuristics 19:373–398

Albareda-Sambola M, Hinojosa Y, Puerto J (2015) The reliable p-median problem with at-facility service. Eur J Oper Res 245: 656–66

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

Aliakbarian N, Dehghanian F, Salari M (2015) A bi-level programming model for protection of hierarchical facilities under imminent attacks. Comput Oper Res 64:210–224

Álvarez-Miranda E, Fernández E, Ljubić I (2015) The recoverable robust facility location problem. Transp Res B Meth 79:93–120

An Y, Zeng B, Zhang Y, Zhao L (2014) Reliable p-median facility location problem: two-stage robust models and algorithms. Transp Res B Meth 64:54–72

Aydin N, Murat A (2013) A swarm intelligence based sample average approximation algorithm for the capacitated reliable facility location problem. Int J Prod Econ 145:173–183

Azad N, Saharidis GKD, Davoudpour H, Malekly H, Yektamaram, SA (2013) Strategies for protecting supply chain networks against facility and transportation disruptions: an improved Benders decomposition approach. Ann Oper Res 210: 125–163

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

CEN, Terrorist attack hits U.S.-owned chemical plant in France (2015). Chemical & Engineering News. https://bit.ly/2NC0DwS

Chang B (2017) Potential supply chain disruptions from hurricane Harvey, AIR. https://airww.co/2xuGm7R

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

Cheng CH, Lai TW, Yang DY, Zhu Y (2016) Metaheuristics for protecting critical components in a service system: a computational study. Expert Syst Appl 54:251–264

Cheng C, Qi M, Zhang Y, Rousseau L-M (2018) A two-stage robust approach for the reliable logistics network design problem. Transp Res B Meth 111:185–202

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, 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. Loc Sci 5:227–246

Dempe S (2002) Foundations of bilevel programming. Kluwer Academic Publisher, Dordrecht

Drezner Z (1987) Heuristic solution methods for two location problems with unreliable facilities. J Oper Res Soc 38:509–514

Fattahi M, Govindan K, Keyvanshokooh E (2017) Responsive and resilient supply chain network design under operational and disruption risks with delivery lead-time sensitive customers. Transp Res E-Log 101:176–200

Ghaffarinasab N, Atayi R (2018) An implicit enumeration algorithm for the hub interdiction median problem with fortification. Eur J Oper Res 267:23–39

Ghaffarinasab N, Motallebzadeh A (2018) Hub interdiction problem variants: models and metaheuristic solution algorithms. Eur J Oper Res 267:496–512

Heckmann I (2016) Towards supply chain risk analytics: fundamentals, simulation, optimization. Springer, Wiesbaden

Hernandez I, Ramirez-Marquez JE, Rainwater C, Pohl E, Medal H (2014) Robust facility location: hedging against failures. Reliab Eng Syst Safe 123:73–80

Jabbarzadeh A, Fahimnia B, Sheu J-B, Moghadam HS (2016) Designing a supply chain resilient to major disruptions and supply/demand interruptions. Transp Res B-Meth 94:121–149

Jalali S, Seifbarghy M, Niaki STA (2018) A risk-averse location-protection problem under intentional facility disruptions: a modified hybrid decomposition algorithm. Transp Res E-Log 114:196–219

Keçici S, Aras N, Verter V (2012) Incorporating the threat of terrorist attacks in the design of public service facility networks. Optim Lett 6:1101–1121

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

Lemos R (2018) Concern rises about cyber-attacks physically damaging industries. eWeek (April 26, 2018). https://bit.ly/2KfefNH

Li X, Ouyang Y (2010) A continuum approximation approach to reliable facility location design under correlated probabilistic disruptions. Transp Res B-Meth 44:535–548

Li Q, Savachkin A (2013) A heuristic approach to the design of fortified distribution networks. Transp Res E-Log 50:138–148

Li Q, Zeng B, Savachkin A (2013) Reliable facility location design under disruptions. Comput Oper Res 40:901–909

Li X, Ouyang Y, Peng F (2013) A supporting station model for reliable infrastructure location design under interdependent disruptions. Transp Res E-Log 60:80–93

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

Liebchen C, Lübbecke M, Möhring R, Stiller S (2009) The concept of recoverable robustness, linear programming recovery, and railway applications. In: Ahuja RK, Möhring RH, Zaroliagis CD (eds) Robust and online large-scale optimization. Lecture Notes in Computer Science, vol 5868. Springer, Berlin, pp 1–27

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

Lu M, Ran L, Shen Z-JM (2015) Reliable facility location design under uncertain correlated disruptions. Manuf Serv Oper Manag 17:445–455

Madani SR, Nookabadi AS, Hejazi SR (2018) A bi-objective, reliable single allocation p-hub maximal covering location problem: mathematical formulation and solution approach. J Air Transp Manag 68:118–136

Mahmoodjanloo M, Parsa Parvasi S, Ramezanian R (2016) A tri-level covering fortification model for facility protection against disturbance in r-interdiction median problem. Comput Ind Eng 102:219–232

Medal HR, Pohl EA, Rossetti MD (2014) A multi-objective integrated facility location-hardening model: analyzing the pre- and post-disruption tradeoff. Eur J Oper Res 237:257–270

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
Parajuli A, Kuzgunkaya O, Vidyarthi N (2017) Responsive contingency planning of capacitated supply networks under disruption risks. Transp Res E-Log 102:13–37
Parvaresh F, Husseini SMM, Golpayegany SAH, Karimi B (2014) 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-Meth 45:1190–1211
Quadros H, Costa Roboredo M, Alves Pessoa A (2018) A branch-and-cut algorithm for the multiple allocation r-hub interdiction median problem with fortification. Expert Syst Appl 110:311–322
Ramamoorthy P, Jayaswal S, Sinha A, Vidyarthi N (2018) Multiple allocation hub interdiction and protection problems: Model formulations and solution approaches. Eur J Oper Res 270:230–245
Rawls CG, Turnquist MA (2010) Pre-positioning of emergency supplies for disaster response. Transp Res B-Meth 44:521–534
Rockafellar RT, Uryasev S (2000) Optimization of Conditional Value-at-Risk. J Risk 2:21–41
Rostami B, Kämmerling N, Buchheim C, Clausen U (2018) Reliable single allocation hub location problem under hub breakdowns. Comput Oper Res 96:15-29
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
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 Str 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
Tran TH, O'Hanley J, Scaparra MP (2017) Reliable hub network design. Transp Sci 51:358–375
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
Yu G, Haskell WB, Liu Y (2017) Resilient facility location against the risk of disruptions. Transp Res B-Meth 104:82–105
Yun L, Qin Y, Fan H, Ji C, Li X, Jia L (2015) A reliability model for facility location design under imperfect information. Transp Res B-Meth 81:596–615
Zarrinpoor N, Fallahnezhad MS, Pishvaee MS (2017) Design of a reliable hierarchical location-allocation model under disruptions for health service networks: a two-stage robust approach. Comput Ind Eng 109:130–150
Zhalechian M, Ali Torabi S, Mohammadi M (2018) Hub-and-spoke network design under operational and disruption risks. Transp Res E-Log 109:20–43
Zhang XY, Zheng Z (2018) A fortification model for decentralized supply systems and its solution algorithms. IEEE T Reliab 67:381–400

Zhang X, Zheng Z, Zhang S, Du W (2016) Partial interdiction median models for multi-sourcing supply systems. Int J Adv Manuf Technol 84:165–181

Zhang C, Ramirez-Marquez JE, Li Q (2018) Locating and protecting facilities from intentional attacks using secrecy. Reliab Eng Syst Safe 169:51–62

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

Chapter 23
Location Problems in Healthcare

Evrim Didem Güneş, Teresa Melo, 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.

23.1 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 the perspective of facility location, healthcare applications bring together different models from location theory and moreover, they give rise to new models as we will see in this chapter.

One of the most discussed topics in healthcare is the equal access to healthcare services and a high level of healthcare 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. 23.2 is devoted to healthcare facility location; we review the literature and present some classical models in that area. The reader needs some

E. D. Güneş
Koç University, Istanbul, Turkey
e-mail: egunes@ku.edu.tr

T. Melo (✉)
Saarland University of Applied Sciences, Saarbrücken, Germany
e-mail: teresa.melo@htwsaar.de

S. Nickel
Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
e-mail: stefan.nickel@kit.edu

G. Laporte et al. (eds.), *Location Science*,
https://doi.org/10.1007/978-3-030-32177-2_23

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 call and the delivery of the patient to an appropriate health service provider. We devote Sect. 23.3 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 deal 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. 23.4, basic models are presented and modern trends are discussed, such as the inclusion of multiple floors, multiple objectives or uncertainty. At the end of the chapter, the reader will find some conclusions and a comprehensive list of references.

23.2 Healthcare Facility Location

In this section, we focus on applications of discrete network location problems to healthcare 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, a topic that is addressed in Chap. 26. We do not discuss continuous location models. In the literature, there are only a few papers applying such models; see, e.g., Dökmeci (1977, 1979).

The location of healthcare facilities can be a critical decision for developing countries since they have scarce resources and the majority of their population lives in rural areas. The low population density in these regions makes the provision of healthcare services a challenge. Within this context, location-allocation models can therefore be successfully applied for the design of healthcare 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 (2005) and Ahmadi-Javid et al. (2017).

In the following, we give an overview of healthcare facility location applications by first discussing the relevant objective functions and then presenting important aspects of these problems with examples from the literature.

23.2.1 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 healthcare 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 healthcare consumers.* This cost type can be defined as travel cost, distance, or travel time from a population district to a healthcare facility, weighted by the population size of that district. When this is the only objective, the standard formulation for the p-median model is commonly used for deciding where to locate a set of healthcare facilities. The following function may represent access cost:

$$\sum_{i \in I} \sum_{j \in J} d_j \, c_{ij} \, x_{ij}, \tag{23.1}$$

  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 healthcare facility, or maximize covered demand.* A covering type objective assumes that a population in a district is covered (has access) only if it can be assigned to a facility within a predetermined maximum distance, and aims at maximizing the covered population. Such a type of objective is appropriate to locate emergency medical services or primary care centers for under-served populations. In fact, when the objective is to minimize the total access cost some districts may end up with too high access costs. A covering objective overcomes this drawback.

  Some healthcare 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 healthcare facilities is another coverage related objective that was first defined by Calvo and Marks (1973). There are several socio-economical factors that affect 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:

$$\sum_{i \in I} \sum_{j \in J} d_j \, \sigma_{ij} \, x_{ij}. \tag{23.2}$$

- *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 objective alone can produce undesirable solutions (Smith et al. 2013). The reader is referred 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 were: 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. The model proposed by Mitropoulos et al. (2013) encompasses three objectives: minimize total distance traveled by patients to their designated facilities, minimize underutilization of capacity of open facilities, and maximize mean efficiency of operating healthcare units. The latter criterion uses efficiency scores that are estimated with data envelopment analysis. 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 (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 organ transplant centers considering distance, waiting list, and maximum waiting list (as a proxy for equity) with different weights in the objective function.

23.2.2 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:

$$\text{minimize} \quad \sum_{i \in I} \sum_{j \in J} d_j \, c_{ij} \, x_{ij} \tag{23.3}$$

$$\text{subject to} \quad \sum_{i \in I} x_{ij} = 1 \qquad \forall j \in J \tag{23.4}$$

$$x_{ij} \leq y_i \qquad \forall i \in I, \; j \in J \tag{23.5}$$

$$\sum_{i \in I} y_i = p \tag{23.6}$$

$$y_i \in \{0, 1\} \qquad \forall i \in I \tag{23.7}$$

$$x_{ij} \in \{0, 1\} \qquad \forall i \in I, \; j \in J, \tag{23.8}$$

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 set of constraints:

$$\sum_{j \in J} d_j \, x_{ij} \leq q_i \qquad \forall i \in I, \tag{23.9}$$

where q_i is the exogenous capacity of the facility at node i. In some situations, decisions regarding the capacities of facilities may also be considered. This case can be modeled by incorporating the corresponding decision variables and the cost associated with building capacity in the objective function.

23.2.2.1 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 (2010) 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:

$$\sum_{s \in S : r \in R_s} \kappa_{sr} \, q_{is} \leq w_{ri} \quad \forall i \in I, \, r \in R \tag{23.10}$$

$$\sum_{j \in J} d_{js} \, x_{ijs} \leq q_{is} \quad \forall i \in I, \, s \in S \tag{23.11}$$

$$\sum_{i \in I} x_{ijs} = 1 \qquad \forall j \in J, \, s \in S, \tag{23.12}$$

where x_{ijs} is a binary variable defining the assignment of patients for service s from district j to the facility at location i, and d_{js} is the demand from district j for service s. In this case, the objective function given in (23.3) should also be changed as follows:

$$\text{minimize} \quad \sum_{i \in I} \sum_{j \in J} \sum_{s \in S} d_{js}\, c_{ij}\, x_{ijs} \tag{23.13}$$

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 image interpretation or surgery), having a high service volume is important to maintain high service quality. See Verter and Lapierre (2002), Güneş and Yaman (2010), Mestre et al. (2012, 2015), and Güneş et al. (2014) for examples on how to incorporate such type of constraints.

23.2.2.2 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. In addition, d_{js} is removed from (23.11) and (23.13), and the assignment constraints (23.12) are changed as follows:

$$\sum_{i \in I} x_{ijs} = d_{js} \qquad \forall j \in J,\ s \in S. \tag{23.14}$$

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 (23.3)–(23.8) (see, e.g., Canovas et al. 2007; Güneş et al. 2014):

$$\sum_{k \in I : c_{kj} > c_{ij}} x_{kj} + y_i \leq 1 \qquad \forall i \in I,\ j \in J. \tag{23.15}$$

These constraints ensure that for a given district $j \in J$, 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) and Smith et al. (2009, 2013).

23.2.2.3 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. In particular, the demand for healthcare 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 to capture 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. Kim and Kim (2013) considered a different form of consumer preference that depends on the income level of patients. Low-income patients are allocated to a public healthcare facility within a given predefined distance, whereas high-income patients may use their preferred public or private facility. Some 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 a 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 healthcare service systems, patients have free choice of where to get service from. If this is the case, 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 (23.15). 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. A related problem was addressed by Vidyarthi and Kuzgunkaya (2015), but in addition to determining the optimal location of preventive healthcare facilities also their capacities are chosen from a set of discrete options using a piecewise linear approximation of the objective function for a network of M/G/1 queues. 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. Aboolian et al. (2015) investigated a similar problem but assumed that service capacity at facilities is continuous. An ϵ-optimal solution method was developed for the problem.

23.2.2.4 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 by 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) provide a comprehensive review on hierarchical systems with a discussion of modeling approaches and applications. The interested reader is further referred to Chap. 13.

Recall also from Chap. 13 that 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 (e.g. 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 (e.g. 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 of model with an objective function that quantifies the total distance traveled:

$$\text{minimize} \quad \sum_{i \in I} \sum_{j \in J} c_{ij} \sum_{s \in S} d_{js}\, x_{ijs} \tag{23.16}$$

$$\text{subject to} \quad \sum_{i \in I} x_{ijs} = 1 \qquad \forall j \in J,\ s \in S \tag{23.17}$$

$$x_{iis} \geq x_{ijs} \qquad \forall i \in I,\ j \in J,\ s \in S \tag{23.18}$$

$$\sum_{i \in I} x_{iik} = \sum_{s \in S : s \geq k} p_s \qquad \forall k \in S \tag{23.19}$$

$$x_{ijs} \in \{0, 1\} \qquad \forall i \in I,\ j \in J,\ s \in S, \tag{23.20}$$

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 (23.17) ensure that all districts are assigned to a facility for all services. Constraints (23.18) ensure that assignments are done to open facilities only, and constraints (23.19) 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 (23.20) 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 and 0 otherwise. 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 patients of population j. A mixed-integer programming formulation to minimize total distance traveled is as follows:

$$\text{minimize} \quad \sum_{i \in I} \sum_{j \in J} c_{ij}\, (x_{ij}^{01} + x_{ij}^{02} + x_{ij}^{12}) \tag{23.21}$$

$$\text{subject to} \quad \sum_{i \in I} (x_{ij}^{01} + x_{ij}^{02}) = d_j \qquad \forall j \in J \tag{23.22}$$

$$\sum_{i \in I} x_{ij}^{12} = \theta \sum_{i \in I} x_{ij}^{01} \qquad \forall j \in J \tag{23.23}$$

$$\sum_{j \in J} x_{ij}^{01} \leq q_1 y_{i1} \qquad \forall i \in I \tag{23.24}$$

$$\sum_{j \in J} (x_{ij}^{02} + x_{ij}^{12}) \leq q_2 y_{i2} \qquad \forall i \in I \tag{23.25}$$

$$\sum_{i \in I} y_{is} = p_s \qquad s \in S \tag{23.26}$$

$$y_{i1} + y_{i2} \leq 1 \qquad \forall i \in I \tag{23.27}$$

$$0 \leq x_{ij}^{01} \leq d_j \qquad \forall i \in I,\ j \in J \tag{23.28}$$

$$0 \leq x_{ij}^{02} \leq d_j \qquad \forall i \in I,\ j \in J \tag{23.29}$$

$$0 \leq x_{ij}^{12} \leq \theta q_1 \qquad \forall i \in I,\ j \in J \tag{23.30}$$

$$y_{is} \in \{0, 1\} \qquad \forall i \in I,\ s \in S. \tag{23.31}$$

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, 2015) for regional and central hospitals. Typically, these models can be solved by commercial optimization solvers. Galvão et al. (2002) applied a tri-level hierarchical model for the delivery of perinatal care in the municipality of Rio de Janeiro (Brazil) 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.

23.2.2.5 Modeling Dynamic Aspects of Location Decisions

A majority of healthcare 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. The model proposed by Cardoso et al. (2016) supports the reorganization of an existing network of long-term nursing care centers over a multi-period planning horizon. Three equity objectives are pursued: minimization of total travel time that includes a penalty for the unserved demand, minimization of the maximum level of unsatisfied demand in a geographical area, and minimization of the total level of unsatisfied demand for low-income users. Demand uncertainty in the context of multi-period planning was captured by Cardoso et al. (2015) via a two-stage stochastic model. 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 context of simulation models. Recently, Intrevado et al. (2019) addressed the problem of adjusting the capacity of an existing network to respond to varying demand for long-term care services. To this end, at each time period, capacity can be added to or removed from service regions.

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 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. Notice that this problem is different from ambulance location problems since mobile facilities here serve for primary care needs as opposed to emergency care situations.

23.2.2.6 Further Reading

In the context of non-emergency healthcare services, facility location problems also arise in other settings than those presented in the previous sections. This includes, for example, the location of blood service facilities and organ transplant centers.

Şahin et al. (2007) developed three mathematical models that together help establish a hierarchical network to collect, process, store, and distribute blood products. Specifically, the decisions to be made involve the location of regional blood centers (RBCs), the determination of service areas for intermediate blood centers, the location of supporting facilities (i.e., blood stations), the calculation of the number of mobile units required to collect and deliver blood, and the homogeneous distribution of mobile units to RBCs among the demand areas. Chaiwuttisaka et al. (2016) addressed the problem of expanding a network of blood centers through the location of two types of service facilities: (1) donation rooms and (2) donation rooms with blood distribution capabilities. The objective function of the proposed binary linear programming model (to be minimized) is a weighted sum of three criteria: total travel distance from the new facilities to RBCs, total demand-weighted distance from RBCs and type 2 facilities to demand points, and total expected amount of blood donations. The location of blood service facilities has also been studied in the context of disaster relief, see e.g., Jabbarzadeh et al.

(2014) and Fahimnia et al. (2017). The interested reader is referred to Chap. 21 for a discussion of facility location models for humanitarian aid, including disaster relief.

Travel time plays a critical role in the location of organ transplant centers. This is due to the fact that there is a maximum allowed time (called ischemia time) between the moment an organ becomes available and its transplant into the recipient's body. Bruni et al. (2006) and Beliën et al. (2013) developed p-median based formulations to find the locations of organ transplant centers, ensuring that the ischemia time is not exceeded. The location-queuing model developed by Zahiri et al. (2014) accounts for alternative transportation modes and uncertainty in demand and supply of organs.

23.3 Ambulance Location

A usual goal of ambulance location problems is to find locations for ambulances (or ambulance bases) minimizing the number of ambulances (or ambulance bases) needed, while fulfilling a certain level of demand. Another possibility is to maximize the coverage having a fixed number of ambulances (or ambulance bases) 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), Li et al. (2011), and Reuter-Oppermann et al. (2017).

In general, ambulance planning can be done at three different levels, the strategic, tactical, and the operational level. At the strategic level, decisions concerning the locations of ambulance bases are made. These decisions often have a long-term effect and last for several decades. The number of ambulances per base and potential movable locations are determined at the tactical level. The operational level includes the dispatching of ambulances to emergencies and the relocation of ambulances to different bases. In the next two sections, exemplary models for the planning problems at the three levels are presented. Section 23.3.1 looks at strategic and tactical models, while Sect. 23.3.2 concentrates on operational aspects.

23.3.1 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 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, which is 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 is as follows:

$$\text{minimize} \quad \sum_{j \in J} x_j \tag{23.32}$$

$$\text{subject to} \quad \sum_{i \in J_j} x_i \geq 1 \quad \forall j \in J \tag{23.33}$$

$$x_j \in \{0, 1\} \quad \forall j \in J. \tag{23.34}$$

23.3.1.1 A Double Coverage Model

The model by Toregas et al. (1971) only ensures 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, respectively, 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 $t_2 \geq t_1$. Notice 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. Furthermore, let d_j denote the demand at node $j \in J$, and define the following two coefficients for $i \in I$ and $j \in J$:

$$\gamma_{ij}^1 = \begin{cases} 1 & \text{if } t_{ij} \leq t_1 \\ 0 & \text{otherwise} \end{cases} \qquad (j \text{ is covered by location } i \text{ within time } t_1) \tag{23.35}$$

and

$$\gamma_{ij}^2 = \begin{cases} 1 & \text{if } t_{ij} \leq t_2 \\ 0 & \text{otherwise} \end{cases} \qquad (j \text{ is covered by location } i \text{ within time } t_2) \tag{23.36}$$

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:

$$\text{maximize} \quad \sum_{j \in J} d_j \, x_{j2} \tag{23.37}$$

$$\text{subject to} \quad \sum_{i \in I} \gamma_{ij}^2 \, y_i \geq 1 \qquad \forall j \in J \tag{23.38}$$

$$\sum_{j \in J} d_j \, x_{j1} \geq \alpha \sum_{j \in J} d_j \tag{23.39}$$

$$\sum_{i \in I} \gamma_{ij}^1 \, y_i \geq x_{j1} + x_{j2} \qquad \forall j \in J \tag{23.40}$$

$$x_{j2} \leq x_{j1} \qquad \forall j \in J \tag{23.41}$$

$$\sum_{i \in I} y_i = p \tag{23.42}$$

$$y_i \leq p_i \qquad \forall i \in I \tag{23.43}$$

$$x_{j1}, \, x_{j2} \in \{0, 1\} \qquad \forall j \in J \tag{23.44}$$

$$y_i \in \mathbb{Z}_0^+ \qquad \forall i \in I. \tag{23.45}$$

The objective function (23.37) 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 imposed by constraints (23.38). Constraint (23.39) states that a proportion α of the demand must be covered within time t_1. A location can only be covered twice within time t_1 if it is covered once, as expressed by constraints (23.40) and (23.41). Exactly p ambulances must be located in total (23.42), and only p_i can be located at node i (23.43). Constraints (23.44) and (23.45) define the domains of the decision variables. The model (23.37)–(23.45) has been tackled in Gendreau et al. (1997) by a tabu search heuristic.

23.3.1.2 Considering Ambulance Utilization

In practice, ambulances are not always available when they are needed. Therefore, the strategic and tactical planning levels should take into account the utilization of ambulances as aggregated data from the operational level. Then, 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; 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. The decision variable x_{jk} is equal to 1 if node j is covered at least k times and 0 otherwise. In addition, γ_{ij} is a binary parameter with:

$$\gamma_{ij} = \begin{cases} 1 & \text{if } t_{ij} \leq T \\ 0 & \text{otherwise} \end{cases} \qquad \text{(an ambulance at } i \text{ covers demand at } j\text{)} \tag{23.46}$$

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:

$$\text{maximize} \quad \sum_{k \in K} \sum_{j \in J} d_j\,(1 - P)\,P^{k-1}\,x_{jk} \tag{23.47}$$

$$\text{subject to} \quad \sum_{k \in K} x_{jk} \leq \sum_{i \in I} \gamma_{ij}\,y_i \qquad \forall j \in J \tag{23.48}$$

$$\sum_{i \in I} y_i \leq p \tag{23.49}$$

$$y_i \in \{0, 1, \ldots, p\} \qquad \forall i \in I \tag{23.50}$$

$$x_{jk} \in \{0, 1\} \qquad \forall j \in J,\ k \in K. \tag{23.51}$$

The objective function (23.47) maximizes the expected demand that is covered. Notice that this expression adds the expected coverage over all possible numbers of ambulances. Constraints (23.48) 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 (23.49) impose that in total at most p ambulances are located. Constraints (23.50) and (23.51) are the variable domain constraints. A heuristic for the problem has also been devised by Daskin (1983).

23.3.1.3 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 covering 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), Noyan (2010), and Nickel et al. (2016).

23.3.2 The Operational Level: Ambulance Relocation

At the operational 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) and Bélanger et al. (2019).

Early relocation approaches 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).

23.3.2.1 Ambulance Preparedness

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 is 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 t_{ij} represents the driving time from ambulance location i to zone j. Moreover, let $t_{[i]j}$ denote the travel time of the i-th closest ambulance to zone j and let x be the matrix form of the decision variables x_{ij}, which are equal to 1 if ambulance i is relocated to zone j and 0 otherwise. Clearly, $t_{[i]j}(x)$ 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 $\gamma^{[i]}$ be the contribution factor of the i-th closest ambulance and let the following two properties be fulfilled:

$$t_{[1]j} \leq t_{[2]j} \leq \ldots \leq t_{[i]p_j}, \tag{23.52}$$

$$\gamma^{[1]} > \gamma^{[2]} > \ldots > \gamma^{[p_j]}. \tag{23.53}$$

The contribution factor $\gamma^{[i]}$ is a user-defined value aiming to describe the influence of the i-th closest ambulance. Therefore $\gamma^{[1]} > \gamma^{[2]} > \ldots > \gamma^{[p_j]}$ is a decreasing sequence. In Andersson and Värbrand (2007), $\gamma^{[i]} = \frac{1}{2^{[i-1]}}$ for $i = 1 \ldots p_j$ is used.

The preparedness in zone j is then defined as

$$\frac{1}{d_j} \sum_{i=1}^{p_j} \frac{\gamma^{[i]}}{t_{[i]j}}. \tag{23.54}$$

Preparedness is a way of evaluating the ability to serve potential patients with ambulances now and in the future. Preparedness in a zone increases if an ambulance moves closer, since the travel time (denominator) decreases. If d_j (the demand) increases then the preparedness decreases.

Andersson and Värbrand (2007) proposed a tree search algorithm for tackling the following relocation model in order to minimize the maximum travel time for the ambulances:

$$\text{minimize} \quad z \tag{23.55}$$

$$\text{subject to} \quad z \geq \sum_{j \in J_i} t_{ij}\, x_{ij} \qquad \forall i \in I \tag{23.56}$$

$$\frac{1}{d_j} \sum_{i=1}^{p_j} \frac{\gamma^{[i]}}{t_{[i]j}(x)} \geq \Delta_{\min} \quad \forall j \in J \tag{23.57}$$

$$\sum_{j \in J_i} x_{ij} \leq 1 \quad \forall i \in I \tag{23.58}$$

$$\sum_{i \in I} \sum_{j \in J} x_{ij} \leq p \tag{23.59}$$

$$x_{ij} \in \{0, 1\} \quad \forall i \in I,\ j \in J. \tag{23.60}$$

In this formulation, J_i (see (23.56) and (23.58)) is the set of zones that can be reached by ambulance i within a given time frame. The objective function (23.55) in conjunction with constraints (23.56) (which ensure that z must not be smaller than any of the driving times t_{ij}) defines the maximum travel time (to be minimized). The preparedness in each zone is at least a value $\Delta_{\min}$ as prescribed in constraints (23.57). In particular, the left-hand side of these constraints can be interpreted as the preparedness for zone j that must be greater than or equal to a minimum value $\Delta_{\min}$. Constraints (23.58) ensure that each ambulance can only be relocated to at most one zone in J_i. Constraints (23.59) guarantee that at most p ambulances are relocated in total. Finally, constraints (23.60) are the variable domain constraints.

23.3.2.2 Further Reading

In recent years, more and more attention has been given to ambulance relocation problems. Several different types of approaches and strategies have been proposed and tested for different countries and different Emergency Medical Systems (EMS), for example, the Netherlands, the U.S., and Canada. Papers give more emphasis to the effectiveness of relocation approaches and the trade-off between response time improvement and additional driving times. Van Barnefeld et al. (2016), for example, studied the effect of the number of relocations on the response time performance for a set of scenarios that originated from a Dutch EMS region. Enayati et al. (2018) proposed a relocation approach that maximizes the expected coverage while minimizing the total travel time and also considering workload restrictions for staff in a shift. A comparison of several relocation and fleet management strategies is presented in Bélanger et al. (2016). Finally, Van Buuren et al. (2018) evaluated two dynamic relocation policies that have been implemented by a Dutch EMS provider.

23.4 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 applications for layout planning of industrial buildings.

Layout planning problems in healthcare were first introduced by Elshafei (1977). The author modeled a hospital layout problem as a quadratic assignment problem and developed heuristics to solve it. In the framework of 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 23.4.2 details a mixed-integer programming (MIP) formulation. Thereafter, in Sect. 23.4.3, 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.

23.4.1 The Quadratic Assignment Problem

The well-known QAP (see Burkard et al. 2009; Drezner 2015), 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 to 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}. For $i \in I$ and $j \in J$, the binary decision variable x_{ij} is equal to 1 if OU j is assigned to location i and 0 otherwise. Moreover, we now assume that $I = J = \{1, \ldots, n\}$ in order to obtain a mathematical formulation of the QAP as follows:

$$\text{minimize} \quad \sum_{h \in I} \sum_{i \in I} \sum_{j \in J} \sum_{k \in J} f_{jk}\, d_{hi}\, x_{hj}\, x_{ik} \tag{23.61}$$

$$\text{subject to} \quad \sum_{i \in I} x_{ij} = 1 \qquad \forall j \in J \tag{23.62}$$

$$\sum_{j \in J} x_{ij} = 1 \qquad \forall i \in I \tag{23.63}$$

$$x_{ij} \in \{0, 1\} \qquad \forall i \in I,\ j \in J. \tag{23.64}$$

The objective function (23.61) minimizes the sum of all flows multiplied by distances which results in the sum of all traveled distances. Constraints (23.62) ensure that each OU is assigned to exactly one room, whereas constraints (23.63) guarantee that each room is only occupied by one OU. Constraints (23.64) 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 is assumed to fit 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.

23.4.2 A Mixed-Integer Programming Formulation

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^L and B^W 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 decision variables α_j define the x-coordinates of the centroid of OU j, whereas its y-coordinates are defined by β_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^a_{jk} (z^b_{jk}) are considered which are equal to 1 if OU j is strictly to the right (top) of OU k and 0 otherwise. The auxiliary decision variable α^+_{jk} (α^-_{jk}) defines the horizontal distance between the centroids of OU j and k if OU j is to the right (left) of OU k, otherwise it is 0. Similarly, β^+_{jk} (β^-_{jk}) defines the vertical distance between the centroids of OU j and k if OU j is to the top (bottom) of OU k, otherwise it is 0. These auxiliary decision variables enable the linearization of the model given by Montreuil (1991).

The layout problem can be formulated as follows:

$$\text{minimize} \quad \sum_{j \in J} \sum_{k \in J} f_{jk} \left(\alpha_{jk}^{+} + \alpha_{jk}^{-} + \beta_{jk}^{+} + \beta_{jk}^{-} \right) \tag{23.65}$$

$$\text{subject to} \quad \alpha_j - \alpha_k = \alpha_{jk}^{+} - \alpha_{jk}^{-} \qquad \forall j,k \in J,\ j \neq k \tag{23.66}$$

$$\beta_j - \beta_k = \beta_{jk}^{+} - \beta_{jk}^{-} \qquad \forall j,k \in J,\ j \neq k \tag{23.67}$$

$$L_j^l \leq \left(a_j'' - a_j' \right) \leq L_j^u \qquad \forall j \in J \tag{23.68}$$

$$W_j^l \leq \left(b_j'' - b_j' \right) \leq W_j^u \qquad \forall j \in J \tag{23.69}$$

$$P_j^l \leq 2 \left(a_j'' - a_j' + b_j'' - b_j' \right) \leq P_j^u \qquad \forall j \in J \tag{23.70}$$

$$0 \leq a_j' \leq a_j'' \leq B^L \qquad \forall j \in J \tag{23.71}$$

$$0 \leq b_j' \leq b_j'' \leq B^W \qquad \forall j \in J \tag{23.72}$$

$$\alpha_j = 0.5 \left(a_j' + a_j'' \right) \qquad \forall j \in J \tag{23.73}$$

$$\beta_j = 0.5 \left(b_j' + b_j'' \right) \qquad \forall j \in J \tag{23.74}$$

$$a_k'' \leq a_j' + M \left(1 - z_{jk}^a \right) \qquad \forall j,k \in J,\ j \neq k \tag{23.75}$$

$$b_k'' \leq b_j' + M \left(1 - z_{jk}^b \right) \qquad \forall j,k \in J,\ j \neq k \tag{23.76}$$

$$z_{jk}^a + z_{kj}^a + z_{jk}^b + z_{kj}^b \geq 1 \qquad \forall j,k \in J,\ j < k \tag{23.77}$$

$$\alpha_j,\ \beta_j,\ a_j',\ a_j'',\ b_j',\ b_j'' \geq 0 \qquad \forall j \in J \tag{23.78}$$

$$\alpha_{jk}^{+},\ \alpha_{jk}^{-},\ \beta_{jk}^{+},\ \beta_{jk}^{-} \geq 0 \qquad \forall j,k \in J,\ j \neq k \tag{23.79}$$

$$z_{jk}^a, z_{jk}^b \in \{0,1\} \qquad \forall j,k \in J,\ j \neq k. \tag{23.80}$$

The objective function (23.65) minimizes the sum of the rectilinear distances of all the flows between the centroids of the OUs. Constraints (23.66) and (23.67) are needed in order to linearize the model given by Montreuil (1991) such that we have $\left| \alpha_j - \alpha_k \right| = \alpha_{jk}^{+} + \alpha_{jk}^{-}$ and $\left| \beta_j - \beta_k \right| = \beta_{jk}^{+} + \beta_{jk}^{-}$.

Constraints (23.68)–(23.70) 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 are ensured by constraints (23.71) and (23.72). The centroid of each OU is defined by constraints (23.73) and (23.74). The non-overlapping requirements for the OUs are formulated by constraints (23.75)–(23.77). Inequalities (23.75) state that if OU j is to the right of OU k then the x-coordinate of the left side of OU j must be greater than the

x-coordinate of the right side of OU k. Similarly, (23.76) states that if OU j is to the top of OU k then the y-coordinate of the bottom side of OU j must be greater than the y-coordinate of the top side of OU k. Finally, (23.77) ensures that OUs j and k may not overlap since at least one of the z-variables needs to switch on to the value 1. This means that one of the OUs j and k must be either strictly to the right side or above of the other. The domains of the decision variables are given in constraints (23.78)–(23.80). 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.

23.4.3 Further Reading

In this section, some possible extensions to the two models discussed in Sects. 23.4.1 and 23.4.2 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 time 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 Singh and Sharma (2006) and Drira et al. (2007). Textbooks on facility layout planning and design are given by Heragu (2008) and Tompkins et al. (2010).

A general review on dynamic layout problems which takes into account multiple time periods and, thus, changing process flows, is given by Balakrishnan and Cheng (1998). A recent approach for a multi-period ward layout planning problem for hospitals was proposed by Arnolds and Nickel (2013).

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). A graph-theoretical approach for a real-world problem where 25 organizational units have to be located on 6 levels of a hospital building is presented in Arnolds and Nickel (2015).

In the last years, a number of papers have been published with respect to multiple objectives (cf. Chen 1999; Sha and Chen 2001; Tenfelde-Podehl 2002; Chen and Sha 2005; 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, walking distances or times of patients, personnel and materials somehow have to be regarded and balanced.

Two further research directions of hospital layout planning problems are the consideration of multiple connected hospital buildings as well as the possibility to share resources amongst different hospital departments and wards. The former has been tackled by Helber et al. (2016) who developed a hierarchical layout planning approach to find locations for departments and wards in a given system of buildings, while minimizing the consumption of transportation resources. The latter aspect has been approached by Hübner et al. (2018) with respect to bed capacities which can be shared across clinical departments. The aim is to improve bed availability via pooling effects. The authors develop an integer linear programming formulation based on a generalized set partitioning problem to find the cost-minimal combination of departments and wards, while satisfying maximum walking distance thresholds for patients and personnel.

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 (2018) who applied 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. Furthermore, Arnolds and Gartner (2018) connected clinical pathway mining with layout planning. The approach identifies significant pathways that have been observed in the past. Using a generalization threshold, possible future pathways may be inferred from the data. On the other hand, non-significant pathways can be filtered out. The authors present a case study with real-world data which demonstrates the applicability of their approach.

23.5 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 receive 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 third author would like to thank Ines Verena Arnolds and Melanie Reuter-Oppermann for their support in preparing this text.

References

Aboolian R, Berman O, Verter V (2015) Maximal accessibility network design in the public sector. Transport Sci 50:336–347

Ahmadi-Javid A, Seyedi P, Syam SS (2017) A survey of healthcare facility location. Comput Oper Res 79:223–263

Aiello G, Enea M, Galante G (2006) A multi-objective approach to facility layout problem by genetic search algorithm and Electre method. Robot Comput Integr 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, Gartner D (2018) Improving hospital layout planning through clinical pathway mining. Ann Oper Res 263:453–477

Arnolds IV, Nickel S (2013) Multi-period layout planning for hospital wards. Socio-Econ Plan Sci 47:220–237

Arnolds I, Nickel S (2015) Layout planning problems in health care. In: Eiselt H, Marianov V (eds) Applications of location analysis, international series in operations research & management science, vol 232. Springer, Cham, pp 109–152

Arnolds IV, Nickel S (2018) An iterative simulation-optimization approach for hospital layout planning. Paper presented at the 44th annual meeting of the EURO Working Group on Operational Research Applied to Health Services (ORAHS), Oslo, Norway

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 2:48–55

Bélanger V, Kergosien Y, Ruiz A, Soriano P (2016) An empirical comparison of relocation strategies in real-time ambulance fleet management. Comput Ind Eng 94:216–229

Bélanger V, Ruiz A, Soriano, P (2019) Recent optimization models and trends in location, relocation, and dispatching of emergency medical vehicles. Eur J Oper Res 272:1–23

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

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. Manage 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 Manag Sci 9:125–142
Burkard R, Dell'Amico M, Martello S (2009) Assignment problems, chaps 7–8. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, pp 205–286
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 Manag Sci 15:385–412
Cardoso T, Oliveira MD, Barbosa-Póvoa A, Nickel S (2015) An integrated approach for planning a long-term care network with uncertainty, strategic policy and equity considerations. Eur J Oper Res 247:321–334
Cardoso T, Oliveira MD, Barbosa-Póvoa A, Nickel S (2016) Moving towards an equitable long-term care network: a multi-objective and multi-period planning approach. Omega 58:69–85
Chaiwuttisaka P, Smith H, Wu Y, Potts C, Sakuldamrongpanich T, Pathomsiri S (2016) Location of low-cost blood collection and distribution centres in Thailand. Oper Res Health Care 9:7–15
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 37:760–786
Chen CW (1999) A design approach to the multi-objective facility layout problem. Int J Prod Res 37:1175–1196
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 (2005) Heuristic approach for solving the multi-objective facility layout problem. Int J Prod Res 43:4493–4507
Cho CJ (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. Transport Sci 17:48–70
Daskin MS, Dean LK (2005) Location of health care facilities. In: Brandeau ML, Sainfort F, Pierskalla WP (eds) Operations research and health care: a handbook of methods and applications. International series in operations research & management science, vol 70. Springer, Boston, pp 43–76
Dökmeci VF (1977) A quantitative model to plan regional health facility systems. Manag Sci 24:411–419
Dökmeci VF (1979) A multiobjective model for regional planning of health facilities. Environ Plan A 11:517–525
Drezner Z (2015) The quadratic assignment problem. In: Laporte G, Nickel S, Saldanha da Gama F (eds) Location science, 1st edn. Springer, Heidelberg, pp 345–363
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
Elshafei AN (1977) Hospital layout as a quadratic assignment problem. J Oper Res Soc 28:167–179
Enayati S, Mayorga ME, Rajagopalan HK, Saydam C (2018) Real-time ambulance redeployment approach to improve service coverage with fair and restricted workload for EMS providers. Omega 79:67–80
Fahimnia B, Jabbarzadeh A, Ghavamifar A, Bell M (2017) Supply chain design for efficient and effective blood supply in disasters. Int J Prod Econ 183:700–709
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. Transport Res B Methodol 43:798–811
Goetschalckx M, Irohara T (2007a) Efficient formulations for the multi-floor facility layout problem with elevators. Technical Report, H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta. Available from http://www.optimization-online.org/DB_HTML/2007/02/1598.html. Accessed 26 Apr 2019
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 (2010) Health network mergers and hospital re-planning. J Oper Res Soc 61:275–283
Güneş ED, Yaman H, Çekyay 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 healthcare planning and control. In: Hall R (ed) Handbook of health care system scheduling. International series in operations research & management science, vol 168. Springer, Boston, 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
Helber S, Böhme D, Oucherif F, Lagershausen S, Kasper S (2016) A hierarchical facility layout planning approach for large and complex hospitals. Flex Serv Manuf J 28:5–29
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
Hübner A, Kuhn H, Walther M (2018) Combining clinical departments and wards in maximum-care hospitals. OR Spectr 40:679–709
Iannoni AP, Morabito R (2007) A multiple dispatch and partial backup hypercube queuing model to analyze emergency medical systems on highways. Transp Res E Logist 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
Intrevado P, Verter V, Tremblay L (2019) Patient-centric design of long-term care networks. Health Care Manag Sci 22:376–390
Jabbarzadeh A, Fahimnia B, Seuring S (2014) Dynamic supply chain network design for the supply of blood in disasters: A robust model with real world application. Transp Res E Logist 70:225–244
Kim D-G, Kim Y-D (2013) A Lagrangian heuristic algorithm for a public healthcare facility location problem. Ann Oper Res 206:221–240
Koopmans TC, 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 multi-floor 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 Spectr 34:319–348
Mestre AM, Oliveira MD, Barbosa-Póvoa A (2015) Location-allocation approaches for hospital network planning under uncertainty. Eur J Oper Res 240:791–806
Mitropoulos P, Mitropoulos I, Giannikos I, Sissouras A (2006) A biobjective model for the locational planning of hospitals and health centers. Health Care Manag Sci 9:171–179
Mitropoulos P, Mitropoulos I, Giannikos I (2013) Combining DEA with location analysis for the effective consolidation of services in the health sector. Comput Oper Res 40:2241–2250
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, 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
Nickel S, Reuter-Oppermann M, Saldanha-da-Gama F (2016) Ambulance location under stochastic demand: a sampling approach. Oper Res Health Care 8:24–32
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 Manag 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
Reuter-Oppermann M, van den Berg PL, Vile JL (2017) Logistics for emergency medical service systems. Health Syst 6:187–208
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
Şahin G, Süral H, Meral S (2007) Locational analysis for regionalization of Turkish Red Crescent blood services. Comput Oper Res 34:692–704
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 Manag 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. PhD thesis, Universität Kaiserslautern, Germany

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

Van Barneveld TC, Bhulai S, van der Mei RD (2016) The effect of ambulance relocations on the performance of ambulance service providers. Eur J Oper Res 252:257–269

Van Buuren M, Jagtenberg C, van Barneveld T, van der Mei R, Bhulai S (2018) Ambulance dispatch center pilots proactive relocation policies to enhance effectiveness. Interfaces 48:235–246

Verter V, Lapierre SD (2002) Location of preventive health care facilities. Ann Oper Res 110:123–132

Vidyarthi N, Kuzgunkaya O (2015) The impact of directed choice on the design of preventive healthcare facility network under congestion. Health Care Manag Sci 18:459–474

Zahiri B, Tavakkoli-Moghaddam R, Mohammadi M, Jula P (2014) Multi-objective design of an organ transplant network under uncertainty. Transp Res E Logist 72:101–124

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 24
The Design of Rapid Transit Networks

Gilbert Laporte and Juan A. Mesa

Abstract Metros and other rapid transit systems increase the mobility of urban populations while decreasing congestion and pollution. There are now over 210 cities with a metro system in the world. 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 they can be employed to solve some specific subproblems. The scientific literature on rapid transit location planning has grown at a fast rate over the past 25 years. This chapter provides an account of some of the most important results. It first describes the main objectives and indices used in the assessment of rapid transit systems. It then reviews the main models and algorithms used to design such systems. The cases 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.

24.1 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

G. Laporte
GERAD & Canada Research Chair in Distribution Management, HEC Montréal, Montréal, QC, Canada
e-mail: gilbert.laporte@cirrelt.ca

J. A. Mesa (✉)
Departamento de Matemática Aplicada II, Escuela Superior de Ingenería, Universidad de Sevilla, Sevilla, Spain
e-mail: jmesa@us.es

G. Laporte et al. (eds.), *Location Science*,
https://doi.org/10.1007/978-3-030-32177-2_24

car traffic, hence reducing congestion and pollution. The term metro is sometimes used synonymously with rapid transit but the latter has a wider acceptance. In the 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. There are now more than 210 cities with a metro system and this number keeps growing.[1] 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, and they will not be covered in this chapter.

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 its 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. 24.1. 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 jointly in the modeling process. A second specific characteristic of metros in 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

[1] http://mic-ro.com/metro/table.html.

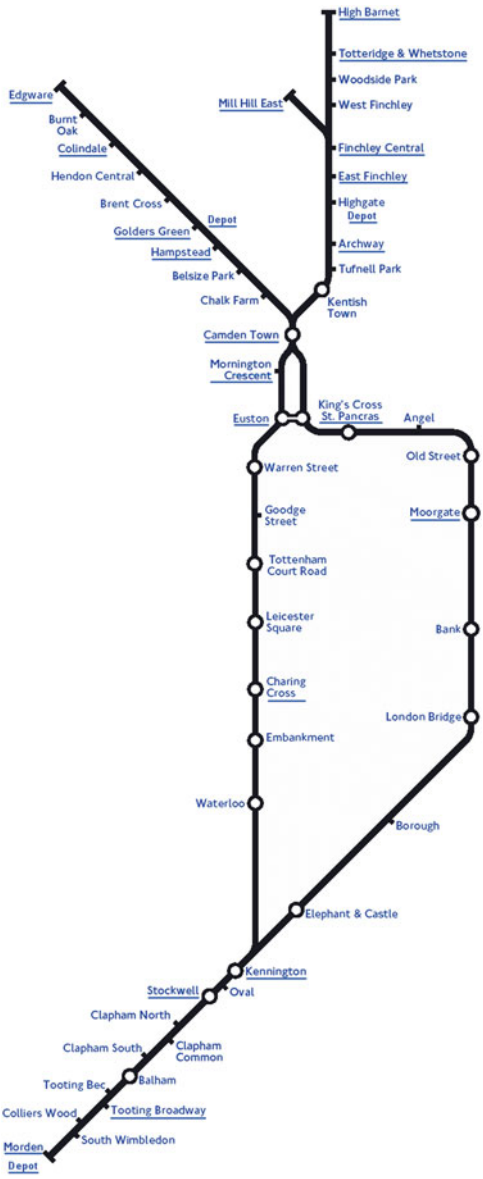


Fig. 24.1 Northern Line, London Underground

(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 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 exist 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 or west 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, as well as algorithmic details. These can be found in the original sources. We will first describe in Sect. 24.2 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. 24.3. In Sect. 24.4 we focus on the location of stations. Conclusions follow in Sect. 24.5.

24.2 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 companies involved in the planning and construction processes, and offering the service. The first group is mainly interested in global advantages such as those mentioned above. 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 one 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. 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 from 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. 24.2) in terms of connectivity and travel directness. Saidi et al. (2016) developed an analytical model to determine the optimal number of radial lines in a ring-radial configuration, which can be viewed as a generalization of a cartwheel in which the radial lines do not necessarily intersect at a unique point.

Gattuso and Miriello (2005) provide a comparative analysis of 13 existing metro networks with respect to 10 indicators. Lee et al. (2008) analyzed the Seoul metro network with respect to characteristic path length, radius, diameter, clustering coefficient, network efficiency, weight of edges, strength of nodes, and maximum flows spanning tree. Other indicators such as regularity, service availability, punctuality

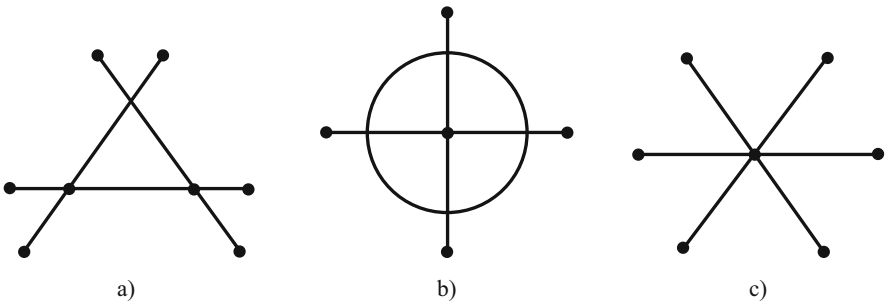


Fig. 24.2 Three basic metro designs. (**a**) Triangle. (**b**) Cartwheel. (**c**) Star

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 new indices based on the assessment of transportation networks from the angle of complex network theory and robustness. In accordance with the glossary of 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. As noted by Yang et al. (2017), metro systems that offer a large diversity of routes to passengers, such as the Beijing metro, are also more robust in the presence of disruptions. In a study of the Shanghai metro system, Sun and Guan (2016) found that the metro lines carrying large volumes of passengers have more impact on the system vulnerability, and lines with a circular topological form have a high impact on passenger flow redistribution in the event of disruptions.

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.

Several researchers have analyzed rapid transit networks in terms of reliability and robustness in the presence of random failures or deliberate attacks. Thus Zhang et al. (2011) analyzed the effect of attacks on nodes or edges of a network: largest degree node based attacks, highest betweenness node based attacks, and random attacks. In this context, the betweenness of a node or of an edge is the number of shortest paths (defined in number of edges) passing through the edge. They found that the Shanghai subway network is robust against random attacks but vulnerable to highest betweenness node-based attacks. Similar conclusions were later found by Sun et al. (2015) and by Xing et al. (2017) who also studied the Shanghai metro network. Jin et al. (2015) considered the problem of allocating protective resources, such as screening detectors at the entrance of some stations, under the threat of deliberate attacks. They cast this problem in a game theoretical framework and illustrated their methodology on the Singapore network. Yang et al. (2015) assessed the robustness of the Beijing subway system. A related research stream is the study of the resilience of a system, i.e., the speed at which it recovers from a failure. D'Lima and Medda (2015) conducted such a study for the London Underground. For an overview of papers on the vulnerability and resilience of transport system, see the paper of Mattson and Jenelius (2015) which contains a section devoted to rail and public transport networks.

Over the past 20 years there has been an increased research interest in the structural properties of the networks representing complex systems, which is interesting for understanding their functioning. 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 $G = (V, E)$ 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 clustering coefficient C and the characteristic path length L are defined as

$$C = \frac{3 \times \text{ number of triangles}}{\text{number of connected triples}} \quad \text{and} \quad L = \frac{1}{|V|(|V|-1)} \sum_{i \neq j} d_{ij},$$

where triangles are triples of vertices in which each node is connected to each of the other two nodes, and connected triples are sets of three vertices linked to one or two of the other two. 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 the 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 another 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 et 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. Xing et al. (2016) studied the connectivity, robustness and reliability of the Shanghai RTS from the viewpoint of complex network theory. Zhang et al. (2016) found that the Minsk and Shanghai metro networks possess the small-world and scale-free properties. They also showed that the hub network is a hierarchical one with a root (the station with the most transfers) which plays an important role controlling some characteristics of the hub network.

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 line 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 network lines. The associated line graphs 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 ease 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. Barrena et al. (2015) explore the transfer system of a collective transportation line network taking into account the passenger level by using hypergraphs and their corresponding line graphs. Finally, Criado et al. (2016) define different line graphs for a multiplex network. This concept is useful to study relationships between the edges of a metro network in which each layer of a multiplex network corresponds to a line. It was applied to the computation of local and global efficiency for the light metro of Calgary.

24.3 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. These problems belong to the class of extensive facility location problems on networks (Puerto et al. 2018).

24.3.1 Location of a Single Alignment

The problem of locating an alignment for a rapid transit system lies within the area of location of one-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 fixed) 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 (a 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. 24.4. 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 point, 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) provided 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 this 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 designed a greedy construction procedure to generate an initial solution which was then provided by 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) considered 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 developed a heuristic consisting of two phases: the construction of the path and the iterative improvement of it. This heuristic was 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 used a continuous variable representing the distribution of the demand of the pair between each mode, according to a logit function which depends on the difference between travel times (or costs) of both modes.

Other objectives have also been employed. For example, 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 was solved with the help of

a Voronoi diagram constructed around the protected sites. More recently, Ortega et al. (2018) considered the problem of locating a single alignment in a sprawled city in order to maximize the functional diversity of the districts covered by the alignment and, indirectly, to reduce the need to travel by car in order to satisfy one's current needs. The authors maximized an objective function defined as an entropy measure. They solved the problem by means of a greedy heuristic akin to the construction phase of the Dufourd et al. (1996) heuristic.

24.3.2 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 was presented in the paper of 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 was 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. Cadarso and Marín (2017) later considered transfer effects in rapid transit network design.

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 considered 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. This work was later extended to account for the number of transfers (Escudero and Muñoz 2014, 2016).

A recent methodological contribution to modeling and solving the transit network design problem can be found in Gutiérrez-Jarpa et al. (2013, 2018). 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 defining preassigned configurations. These segments are later assembled into lines. The authors applied two criteria: minimizing construction cost, and maximizing origin-destination traffic capture, and computed Pareto-optimal solutions. Gutiérrez-Jarpa et al. (2017) solved a related problem incorporating three objectives: infrastructure cost, travel time saving yielded by the use of the metro system, and patronage. They performed a study of the trade-offs between these objectives. Laporte and Pascoal

(2015) described a metaheuristic for the solution of a metro network design problem under two objectives: population coverage and construction cost. As in the previous studies, they worked with a predefined configuration defined as a star, a triangle or a cartwheel. They first constructed non-dominated paths corresponding to the segments of the configuration and then assembled them optimally by solving an integer linear programming problem.

Marín and Jaramillo (2008) studied a multi-period capacity expansion problem. In their 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 the rapid transit network design problems are based on Benders decomposition (Marín and Jaramillo 2009), genetic algorithms (Wang and Lin 2010) and simulated annealing (Fan and Machemehl 2006; Kemanshani et al. 2010). Line configuration with assignment of passengers is studied in Guan et al. (2006).

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 robustness (Cadarso and Marín 2012, 2016; García-Archilla et al. 2013), and of risk aversion (Cadarso and Marín 2016).

Finally, a number of papers now combine two levels of planning in rapid transit network design. The first one, which is strategic, is the location of lines and stations, while the second, which is tactical, is the determination of train capacities and frequencies. These two problems are interrelated because the profitability of a system is a function of the construction and operating costs, and of the revenue related to ridership, which partly depends on travel time and therefore on capacity and frequency. An and Lo (2016) integrate these two planning levels in the framework of stochastic programming. Here the alignments and frequencies are determined in a first phase, and flexible services are offered in a second phase to handle demand overflow. López-Ramos (2014) provides a survey of this line of research. Recent contributions are those of Canca et al. (2016, 2017) and López-Ramos et al. (2017).

24.4 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

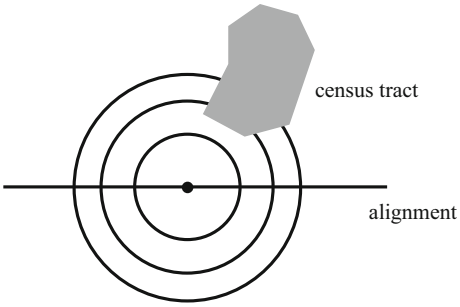


Fig. 24.3 Concentric catchment areas around a station intersecting with a census tract

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) used census tracts coupled with information on population density to estimate the actual walking distances. Different level of attraction were applied in order to obtain a better estimation of the population covered (see Fig. 24.3). For each given location of the stations in a corridor, line coverage was 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.

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) used origin-destination pairs for computing the total population affected by an improvement of a transportation network. In Laporte et al. (2005), trip coverage was 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 was to minimize the number of new stations assuming that each of these covers a demand located 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. The

authors 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. (2014) 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 recent paper by Carrizosa et al. (2016), 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 the passengers subject to the coverage of all demand points. Finally, López-de-los-Mozos et al. (2017) recently solved the problem of locating one or two transfer points in a network in such a way that, under various objective functions, the traffic captured by the network is maximized.

24.5 Conclusions

The design of rapid transit systems is a complex process that 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 25 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 2015-06189, and by project MTM2015-67706-P (MINECO/FEDER, UE).

References

An K, Lo H (2016) Two-phase stochastic program for transit network design under demand uncertainty. Transp Res B Methodol 84:157–181

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 Spec Top 215:93–108

Barrena E, De-Los-Santos A, Laporte G, Mesa JA (2015) Transferability of collective transportation line network from a topological and passenger demand perspective. Netw Heterog Media 10:1–16

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 Á (2012) Recoverable robustness in rapid transit network design. Procedia Soc Behav Sci 54:1288–1297

Cadarso L, Marín Á (2016) Rapid transit network design considering risk. Electron Notes Discrete Math 52:29–36

Cadarso L, Marín Á (2017) Improved rapid transit network model considering transfer effects. Ann Oper Res 258:547–567

Canca D, De-Los-Santos A, Laporte G, Mesa JA (2016) A general rapid network design, line planning and fleet investment integrated model. Ann Oper Res 246:127–144

Canca D, De-Los-Santos A, Laporte G, Mesa JA (2017) An adaptive neighborhood search metaheuristic for the integrated railway rapid transit network design and line planning problem. Comput Oper Res 78:1–14

Carrizosa E, Harbering J, Schöbel A (2016) Minimizing the passenger' travelling time in the stop location problem. J Oper Res Soc 67:1325–1337

Criado R, Hernández-Bermejo B, Romance M (2007) Efficiency, vulnerability and cost: an overview with applications to subway networks worldwide. Int J Bifurcat Chaos 17:2289–2301

Criado R, Flores J, García del Amo A, Romance M, Barrena E, Mesa JA (2016) Line graphs for a multiplex network. Chaos 26:065309

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

D'Lima M, Medda F (2015) A new measure of resilience: an application to the London Underground. Transport Res A Pol 81:35–46

Dufourd H, Gendreau M, Laporte G (1996) Locating a transit line using tabu search. Location 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

Escudero LF, Muñoz S (2014) A survey-based approach for selecting the stations and links for a rapid transit network. Int J Comput Int Sys 7:565–581

Escudero LF, Muñoz S (2016) A survey-based approach for designing the lines of a rapid transit network. Discrete Appl Math 210:14–34

Fan W, Machemehl RB (2006) Using a simulated annealing algorithm to solve the transit route network design problem. J Transp Eng 132:122–132
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
Gutiérrez-Jarpa G, Laporte G, Marianov V, Moccia L (2017) Multi-objective rapid transit network design with modal competition: the case on Concepción, Chile. Comput Oper Res 78:27–43
Gutiérrez-Jarpa G, Laporte G, Marianov V (2018) Corridor-based metro network design with travel flow capture. Comput Oper Res 89:58–67
Hamacher HW, Liebers A, Schöbel A, Wagner D, Wagner F (2001) Locating new stops in a railway network. Electron Notes Theor Comput Sci 50:13–23
IEEE—Institute of Electrical and Electronics Engineers (1990) IEEE standard computer dictionary: a compilation of IEEE standard computer glossaries
Jin J, Lu L, Sun L, Yin J (2015) Optimal allocation of protective resources in urban rail transit networks against intentional attacks. Transp Res E 84:73–87
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 (2014) A maximum trip covering location problem with an alternative mode transportation on tree networks and segments. Top 22:227–253
Laporte G, Pascoal MMB (2015) Path based algorithms for metro network design. Comput Oper Res 62:78–94
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 Á, 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). LNCS 4359, 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 Emerg Technol 44:447–459
Laporte, G, Marín Á, Mesa JA, Perea F (2011) Designing robust rapid transit networks with alternative routes. J Adv Transp 45:54–65
Latora V, Marchiori M (2001) Efficient behavior of small-world networks. Phys Rev Lett 87:198701,1–4
Latora V, Marchiori M (2002) Is the Boston subway a small-world network? Physica A 314:109–113
Lee K, Jung W-S, Park J, Choi M (2008) Statistical analysis of the metropolitan Seoul subway system: network structure and passenger flows. Physica A 387:6231–6234

López-de-los-Mozos MC, Mesa JA, Schöbel A (2017) A general approach for the location of transfer points on a network with a trip covering criterion and mixed distances. Eur J Oper Res 260:108–121
López-Ramos F (2014) Integrating network design and frequency setting in public transportation networks: a survey. SORT Stat Oper Res Trans 38:181–214
López-Ramos F, Codina E, Marín Á, Guarnaschelli A (2017) Integrated approach to network design and frequency setting problem in railway rapid transit systems. Comput Oper Res 80:128–146
Marín Á (2007) An extension to rapid transit design problem. Top 15:231–241
Marín Á, García-Ródenas R (2009) Location of infrastructure in urban railway networks. Comput Oper Res 36:1461–1477
Marín Á, Jaramillo P (2008) Urban rapid transit network capacity expansion. Eur J Oper Res 191:45–60
Marín Á, Jaramillo P (2009) Urban rapid transit network design: accelerated Benders decomposition. Ann Oper Res 169:35–53
Mattson L-G, Jenelius E (2015) Vulnerability and resilience of transport systems – a discussion of recent research. Transp Res A Pol 81:16–34
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 28 Apr 2019
Musso A, Vuchic VR (1988) Characteristics of metro network and methodology for their evaluation. Transport Res Rec 1162:22–33
Ortega FA, Piedra-de-la-Cuadra RA, Ventura S (2018) Applying an entropic analysis to locate rapid transit lines in sprawled cities. Int J Sustain Dev Plan 13:626–637
Puerto J, Ricca F, Scozzari A (2018) Extensive facility location problems on networks: an updated review. Top 26:187–226
Roth C, Kang SM, Batty M, Barthelemy M (2012) A long-time limit for world subway networks. J Roy Soc Interface 9:2540–2550
Saidi S, Wirashinghe S, Kattan L (2016) Long-term planning for ring-radial urban rail transit networks. Transp Res B Methodol 86:128–146
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 December 2002
Sociedad del Metro de Sevilla SA (2001) Proyecto general básico de la red de metro de Sevilla y programacón de fases (in Spanish). UTE Iberinsa and Ghesa
Sun DJ, Guan S (2016) Measuring vulnerability of urban metro network from line operation research. Transp Res A Pol 94:348–359
Sun DJ, Zhao Y, Lu QC (2015) Vulnerability analysis of urban rail transit networks: a case study of Shanghai, China. Sustainability 7:6919–6936
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

Xing Y, Lu J, Chen S (2016) Weighted complex network analysis of Shanghai rail. Discrete Dyn Nat Soc 2016:8pp
Xing Y, Lu J, Chen S (2017) Vulnerability analysis of urban rail transit based on complex network theory: a case study of Shanghai metro. Public Transp 9:501–525
Yang Y, Liu M, Zhou M, Li F, Sun C (2015) Robustness assessment of urban rail transit based on complex network theory: a case study of the Beijing subway. Safety Sci 79:149–162
Yang Y, Chen A, Ning B, Tang T (2017) Measuring route diversity for urban rail transit networks: a case study of the Beijing metro network. IEEE Trans Intell Transp 18:259–268
Zhang J, Xu X, Hong L, Wang S, Fei Q (2011) Characteristics on hub networks of urban rail transit networks. Physica A 390:4562–4570
Zhang J, Zhao M, Liu H, Xu X (2013) Networked characteristics of the urban rail transit networks. Physica A 392:1538–1546
Zhang J, Wang S, Zhang Z, Zou K (2016) Characteristics on hub networks of urban rail transit networks. Physica A 447:502–507

Chapter 25
Districting Problems

Jörg Kalcsics and Roger Z. Ríos-Mercado

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 very diverse 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.

25.1 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

J. Kalcsics (✉)
School of Mathematics, The University of Edinburgh, Edinburgh, UK
e-mail: joerg.kalcsics@ed.ac.uk

R. Z. Ríos-Mercado
Universidad Autónoma de Nuevo León (UANL), Department of Mechanical and Electrical Engineering, San Nicolas de los Garza, NL, Mexico
e-mail: roger.rios@uanl.edu.mx

G. Laporte et al. (eds.), *Location Science*,
https://doi.org/10.1007/978-3-030-32177-2_25

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 such as the p-median or the p-center problem (see Chaps. 2 and 3), only after deciding 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 focus 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 or basic areas, into larger geographic clusters, called districts or territories, 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 25.1 shows an example of a districting plan for streets and for zip code areas.

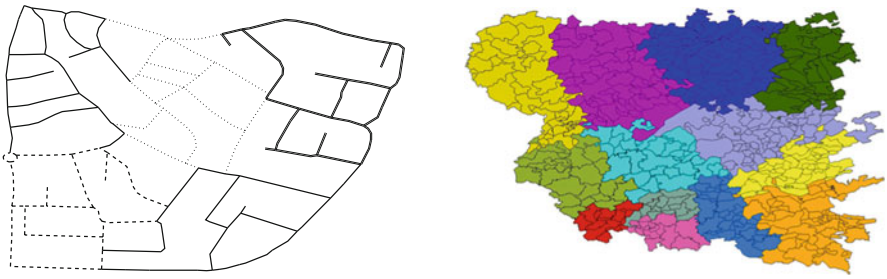


Fig. 25.1 An example of a districting plan for streets and for zip-code areas

Districting problems are motivated by very diverse 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 a few authors consider the districting problem independently from a practical background. Therefore, the aim of this chapter is to 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. Section 25.2 reviews the broad range of districting applications and identifies and motivates the different planning restrictions. In Sect. 25.3, basic notations are introduced. This is followed by Sect. 25.4 that 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. 25.5 presents an overview of the different solution techniques for solving districting problems.

25.2 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. While sales is also a type of service, due to its dominant role in the literature, sales territory design will be discussed separately from service districting. 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.

25.2.1 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 since 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 such as 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.

25.2.2 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 traveling salesman problem (TSP) tour. However, the workload of districts is usually balanced over 3–6 months 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), and 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 a few authors consider more than one balancing criterion: Deckro (1977), Zoltners and Sinha (1983), and Ríos-Mercado and Fernández (2009). Contiguous districts are desired to obtain clearly defined geographic areas of responsibility for each sales person, which should prevent them from competing for customers with a high sales potential. Unfortunately, customers are typically represented by their addresses, i.e., points on the map, and it is not clear how to assess contiguity in this case. Compactness describes the desire for districts that are geographically closely packed. Apart from the visual appeal of compact districts, the criterion often serves, together with contiguity, as a proxy for reducing the unproductive travel time of the sales force. 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 (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 (especially 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).

One important, but only recently addressed aspect of sales territory design is that customers often require service with different frequencies. Some customers have to be visited weekly, while others require service only once per month. As a result, planners not only have to design the districts, but also schedule visits to customers within the planning horizon. For example, if the planning horizon is divided into weeks and days, then we also have to decide which customers should be visited in which week and on which day of that week. The criteria for scheduling customer visits are very similar to the ones for designing the sales territories. The total workload incurred by all customers served in each time period should be the same across all periods and the set of all customers visited in the same time period should be as compact as possible to reduce travel times during each period. While contiguity is still desirable, differing visiting frequencies will make it very difficult, or even impossible, to obtain contiguous sets of customers for each period. For more details, see Bender et al. (2016, 2018).

25.2.3 Service Districting

The problem of designing service districts appears in various contexts. One area of applications focuses on social facilities such as 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 health-care personnel such as nurses or physiotherapists. The goal is to determine contiguous districts that have a good accessibility 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; García-Ayala et al. 2016). 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 (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 the 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; Camacho-Collados et al. 2015).

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 accessibility, 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).

Fernández et al. (2010) study a very unique districting problem arising in the context of collection of waste electric and electronic equipment (WEEE) in Europe. The problem was motivated by a recycling directive adopted in the European Union which states, among other things, that each company selling electrical or electronic equipment in a European country has the responsibility to collect and recycle an amount of returned items proportional to the firm's market share. A districting plan assigns basic units to companies; however, in contrast to classical districting problems, the territories should be as geographically dispersed as possible to avoid regional monopolies. The problem also involves particular balancing constraints and allows splitting basic units to balance territories with respect to different product types. They termed this the maximum dispersion territory design problem. In a related work, Fernández et al. (2013) introduce the maximum dispersion problem which is essentially a partition problem seeking to maximize a dispersion function. In this new problem, no split basic units are allowed, so it can be seen as a special case of the maximum dispersion territory design problem.

25.2.4 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. In the first phase, pickup and delivery districts are created based on uncertain demands. Once the districts are given, the uncertainty is revealed and the routing is done in the second phase on a day-to-day basis (Haugland et al. 2007). 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; Zhong et al. 2007; Jarrah and Bard 2012; Lei et al. 2012, 2015).

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).

25.3 Notations

This section introduces notations for the main components of districting problems.

25.3.1 Basic Units

A districting problem comprises a set $J = \{1, \ldots, n\}$ of *basic units*, sometimes also 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 such as 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 $\{v_i, v_j\} \in E$ if and only if 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 $v_j \in V$.

25.3.2 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 $\mathscr{D} = \{D_1, \ldots, D_p\}$.

25.3.3 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.

25.4 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 $Q = 1$.

25.4.1 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:

$$D_1 \cup \cdots \cup D_p = J \text{ and } D_l \cap D_k = \emptyset, \quad l \neq k,\ 1 \leq l,\ k \leq p.$$

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.

25.4.2 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 $\mu = w(J)/p$:

$$bal(D_k) = \left| \frac{w(D_k) - \mu}{\mu} \right|, \quad 1 \leq k \leq p$$

(cf. Forman and Yue 2003; Ríos-Mercado and Fernández 2009; 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 district sizes also include a solution dependent performance measure—in addition to the activity measure—, then this affects μ and the balance of one and the same district may be different for different districting plans. For example, in sales and service territory design, districts often have to be balanced with respect to workload; workload, in turn, usually consists of service times plus travel times. While the total sum of the former is solution independent, the latter depend on the actual district layout. 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):

$$bal^{\text{sum}}(\mathscr{D}) = \sum_{k=1}^{p} bal(D_k) \quad \text{and} \quad bal^{\text{cv}}(\mathscr{D}) = \lambda\, bal^{\text{sum}}(\mathscr{D}) + (1-\lambda)\, bal^{\max}(\mathscr{D}),$$

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)

$$bal^{\text{rng}}(\mathscr{D}) = \max_{k=1,\ldots,p} w(D_k) - \min_{k=1,\ldots,p} w(D_k).$$

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; 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. While several different activity measures have been discussed in the literature, only a few authors consider more than one criterion simultaneously (Deckro 1977; Zoltners and Sinha 1983). In a recent series of papers, two activity measures have been considered simultaneously: the number of customers and the total demand per district (Salazar-Aguilar et al. 2011b, 2012, 2013b; Ríos-Mercado and Escalante 2016).

Concerning solution dependent performance measures, the most common addition is to include travel times in the district size. Due to the scale of realistic data sets, calculating the exact travel times within each district is usually too costly during optimization. Instead, most authors rely on estimates. A common way to approximate the total travel time (or distance) within a district is to use the Beardwood-Halton-Hammersley formula (Lei et al. 2012, 2015). This formula, however, has the downside that it is non-linear and therefore does not easily admit linear programming formulations. As an alternative, some authors propose to add to the service time of each basic unit a fixed estimate of the travel time to the "next basic unit in the district". This estimate can, for example, be the average (expected) travel time to the k closest basic units, where k is a parameter that has to be tuned for each (set of) data instance(s) (Bard and Jarrah 2009; Jarrah and Bard 2012).

25.4.3 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.

25.4.3.1 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 25.1 An example for these three proximity graphs for a point set with 26 basic units is depicted in Fig. 25.2. 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.

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.

25.4.3.2 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.

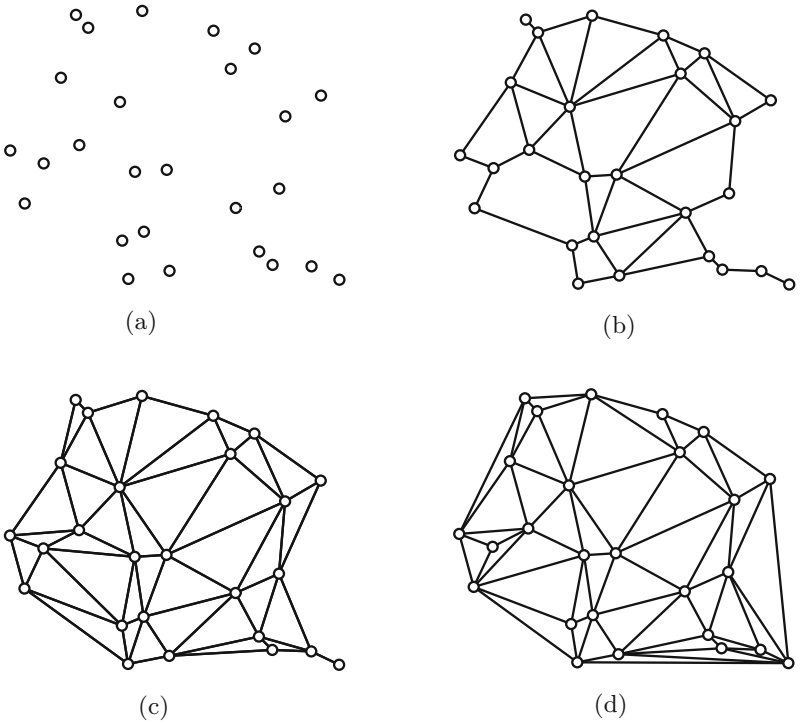


Fig. 25.2 Three different approximate contiguity graphs. (**a**) Point set of basic units. (**b**) Gabriel graph. (**c**) Voronoi-based graph. (**d**) Non-crossing edges graph

25.4.3.3 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 due to Drexl and Haase (1999): 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 basic unit c_k. If all elements of S are assigned to district k (with center in unit c_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 district k:

$$\sum_{j \in \bigcup_{i \in S} N(i) \setminus S} x_{c_k,j} - \sum_{j \in S} x_{c_k,j} \geq 1 - |S| \qquad \forall\, S \subseteq J \setminus \{N(c_k) \cup \{c_k\}\},$$

where $x_{c_k,j}$ is 1 if $j \in J$ is assigned to the district with center c_k and 0 otherwise. A clear drawback of this formulation is that it requires an exponential number of constraints. Nevertheless, this gives naturally rise to cut generation approaches, see Salazar-Aguilar et al. (2011a) and Ríos-Mercado and López-Pérez (2013). A second possibility that only needs a linear number of constraints is based on network flow

constraints (Shirabe 2009). Each basic unit has one unit of supply, and the district centers act as sinks. District k is contiguous if and only if there exists a flow from each of its basic units to c_k that only passes through basic units in D_k:

$$\begin{aligned} \sum_{i \in N(j)} f_{ji} - \sum_{i \in N(j)} f_{ij} &= x_{c_k,j} && \forall\, j \in J \setminus \{c_k\} \\ \sum_{i \in N(j)} f_{ij} &\leq (n-2)\, x_{c_k,j} && \forall\, j \in J \setminus \{c_k\} \\ \sum_{i \in N(c_k)} f_{i,c_k} &\leq n-1, \end{aligned}$$

where f_{ij} is the flow from basic unit i to j and $f_{c_k,j} = 0, \forall\, j \in N(c_k)$.

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:

$$x_{c_k,j} \leq \sum_{i \in S_j} x_{c_k,i} \qquad \forall j \in J \setminus \{c_k\},$$

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 straightforward 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 25.1 Only a 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 25.2 (Fig. 25.3).

Remark 25.2 It is much easier to ensure strict contiguity if heuristics are used to solve the districting problem. Given a district D_k and the corresponding subgraph of G, it is possible to check in $O(|D_k|)$ time whether D_k is connected or not. If the heuristic is based on local search, then adding a basic unit to a connected district will preserve connectivity. Likewise, removing a basic unit from a connected district will preserve connectivity if the removed basic unit does not coincide with a cut-vertex of the subgraph (Ricca et al. 2013). To reduce the computational effort in the latter case, King et al. (2012) have introduced the concept of geo-graphs for two-dimensional basic units that utilizes information from the planar dual graph of G.

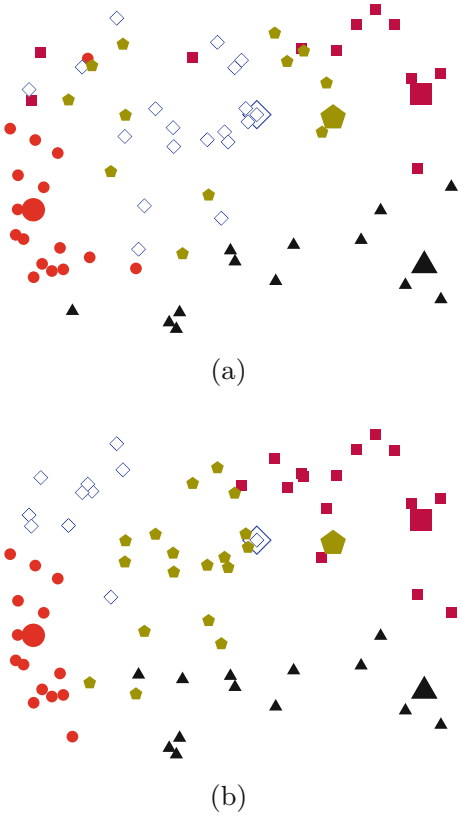


Fig. 25.3 Districting plans for two center-based compactness measures without contiguity. (**a**) Districts for $cmp_{\mathrm{ud}}(\cdot)$. (**b**) Districts for $cmp_{\mathrm{wd}^2}(\cdot)$

25.4.4 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.

25.4.4.1 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:

25.4.4.2 Distance-Based Measures

Distance-based measures are used predominantly in applications where people have to travel within the districts, e.g., salesmen 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 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)

$$cmp_{\mathrm{ud}}(D_k) = \sum_{j \in D_k} d_{c_k,j} \qquad cmp_{\mathrm{ud}^2}(D_k) = \sum_{j \in D_k} d^2_{c_k,j}$$

$$cmp_{\mathrm{wd}}(D_k) = \sum_{j \in D_k} w_j \, d_{c_k,j} \qquad cmp_{\mathrm{wd}^2}(D_k) = \sum_{j \in D_k} w_j \, d^2_{c_k,j}$$

(Bard and Jarrah 2009; Bergey et al. 2003; Hess and Samuels 1971; Zoltners and Sinha 2005). The second and fourth 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 25.2 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 25.3 shows the resulting districting plans that minimize the sum of the two center-based compactness measures $cmp_{\mathrm{ud}}(\cdot)$ and $cmp_{\mathrm{wd}^2}(\cdot)$ 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 $cmp_{\mathrm{ud}}(\cdot)$. 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, $cmp_{\mathrm{wd}^2}(\cdot)$, 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 $cmp_{\mathrm{ud}}(\cdot)$. The results for $cmp_{\mathrm{wd}}(\cdot)$ and $cmp_{\mathrm{ud}^2}(\cdot)$ in terms of 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 $cmp_{\mathrm{ud}}(\cdot)$ and $cmp_{\mathrm{wd}^2}(\cdot)$ are shown in Fig. 25.4. The separation between the districts for $cmp_{\mathrm{ud}}(\cdot)$ 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 $cmp_{\mathrm{wd}^2}(\cdot)$ 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.

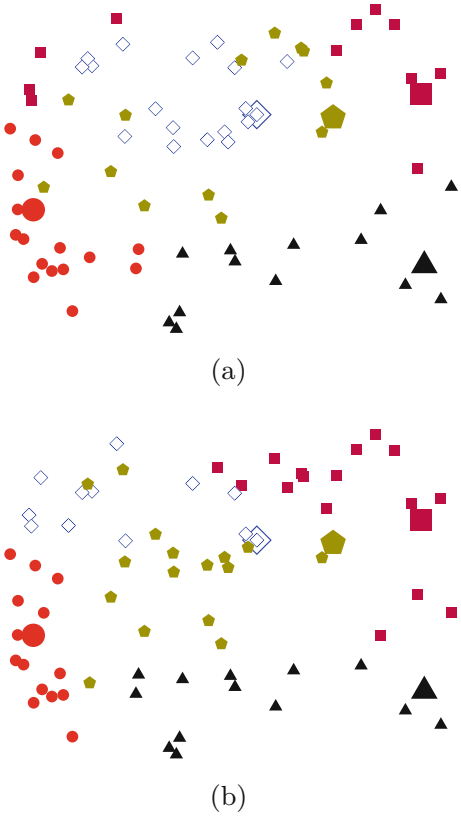


Fig. 25.4 Districting plans for two center-based compactness measures with contiguity. (**a**) Districts for $cmp_{\text{ud}}(\cdot)$. (**b**) Districts for $cmp_{\text{wd}^2}(\cdot)$

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 25.3. 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

$$cmp_{\text{mpw}}(D_k) = \max_{i,j \in D_k,\, i \neq j} d_{ij} \qquad cmp_{\text{spw}}(D_k) = \sum_{i,j \in D_k,\, i \neq j} w_i\, w_j\, d_{ij}$$

(see Ríos-Mercado and Salazar-Acosta (2011), Ríos-Mercado and Escalante (2016) for the former and Blais et al. (2003) for the latter).

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; de Assis et al. 2014; Butsch et al. 2014).

25.4.4.3 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).

25.4.5 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.

25.4.6 Other Criteria

There are a few other criteria for districting problems that are included from time to time in districting models. For example, for redistricting problems the changes in allocation from the old to the new districting plan should be kept small (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 (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). For instance, some applications in healthcare, in particular on the redistricting of liver allocation, attempt to minimize the disparity in liver availability among districts (Gentry et al. 2015). In other areas such as the location of Emergency Medical Service (EMS) the focus is to save lives and to minimize the effects of emergency health incidents. In that context, districting, or designing pre-determined response areas, allows an EMS system to reduce the response time of paramedic support to the incident. An important criterion for these applications is the patient survival probability. Thus, developing both dispatching and districting policies under uncertainty to improve the performance of EMS systems becomes very a very important issue (Mayorga et al. 2013).

25.5 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 metaheuristics such as GRASP, Tabu Search, Scatter Search, 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).

25.5.1 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

$$\text{minimize} \quad \sum_{i,j \in J} w_j \, d_{ij}^2 x_{ij} \tag{25.1}$$

$$\text{subject to} \quad \sum_{i \in J} x_{ij} = 1 \qquad \forall \, j \in J \tag{25.2}$$

$$\sum_{j \in J} w_j \, x_{ij} \geq (1 - \alpha) \, \mu \, y_i \qquad \forall \, i \in J \tag{25.3}$$

$$\sum_{j \in J} w_j \, x_{ij} \leq (1 + \alpha) \, \mu \, y_i \qquad \forall \, i \in J \tag{25.4}$$

$$\sum_{i \in J} y_i = p \tag{25.5}$$

$$y_i, \, x_{ij} \in \{0, 1\} \qquad \forall \, i, j \in J, \tag{25.6}$$

where $x_{ij} = 1$ if basic unit j is assigned to district center i, 0 otherwise, and $y_i = 1$ if basic unit i is selected as district center, 0 otherwise. The objective function (25.1) maximizes the compactness of the districts using the center-based measure $cmp_{\text{wd}^2}(\cdot)$. Constraints (25.2), together with the integrality constraints on the x_{ij}-variables, model the unique and exclusive assignment criterion. Constraints (25.3) and (25.4) restrict the balance of the districts. Finally, Constraints (25.5) 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 Lagrangian 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 mixed-integer linear programming (MILP) 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).

Example 25.3 Consider again the example depicted in Fig. 25.3, 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 $cmp_{\mathrm{wd}^2}(\cdot)$, the new centers that minimize $cmp_{\mathrm{wd}^2}(\cdot)$ over each district are shown on the left-hand side in Fig. 25.5. 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 a 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.

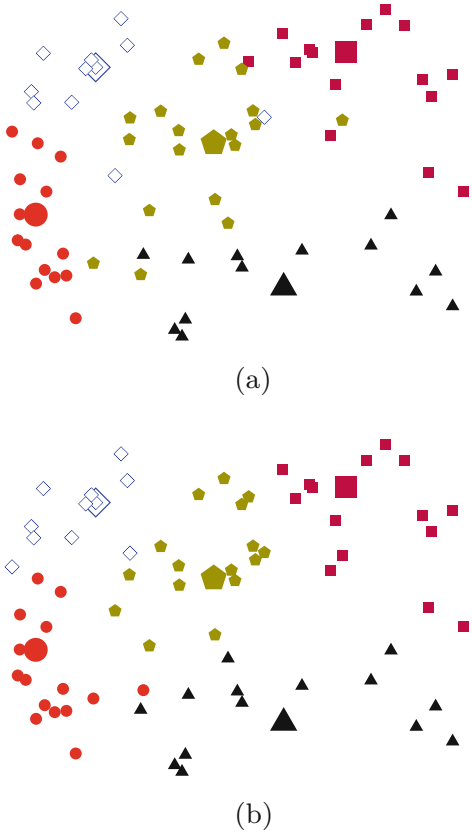


Fig. 25.5 Illustration of one iteration of the location-allocation procedure. (**a**) Location phase: new districts centers. (**b**) Allocation phase: new districts

25.5.2 Exact Methods

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.

More recently, Salazar-Aguilar et al. (2011a) introduced an exact method for handling districting problems subject to the connectivity constraints proposed by Drexl and Haase (1999). The authors present an exact solution framework based on a branch-and-bound algorithm combined with a cut generation strategy. First, the (exponentially many) connectivity constraints are relaxed and then the integer relaxation is solved by branch-and-bound. Afterwards, an easy separation problem

is solved to find unconnected districts. The corresponding violated constraints are then added to the formulation and the iterative process starts again. When no more violated cuts are found, the algorithm stops with an optimal solution. Extensive empirical evidence is presented for several classes of districting models that include multiple balancing constraints and various compactness measures. Two MILP models are assessed: one based on a p-center compactness measure and the other based on a p-median function. The latter turns out to have a stronger linear programming relaxation and results in fewer violated connectivity constraints. The authors also propose two integer quadratic programming formulations for the center and median based compactness measure that result in a smaller number of variables than the linear formulations. These formulations are also solved within the same exact optimization framework. The empirical results show that the quadratic models allow solving larger instances than their linear counterparts. The former also require fewer iterations of the exact method to converge.

Ríos-Mercado and Bard (2019) present an exact optimization scheme for the maximum dispersion territory design problem introduced in Fernández et al. (2010). The exact algorithm takes full advantage of a tighter dual bound and a new reformulation embedded into a biased binary search scheme. Extensive testing indicates that the proposed exact algorithm is able to find optimal solutions to instances with up to 800 basic units and 12 companies and to instances with up to 1400 basic units and 8 companies. Previous to this research, the largest instances optimally solved with off-the-shelf branch-and-bound solvers had between 40 to 100 basic units and 4 companies. This work also extends the results for the maximum dispersion problem introduced by Fernández et al. (2013).

In the context of multi-objective districting, Salazar-Aguilar et al. (2011b) address a commercial districting problem. The authors propose a bi-objective programming model where compactness and balancing with respect to the number of customers are used as performance criteria. Constraints such as connectivity and balancing with respect to product demand are also considered in the model. They propose an improved epsilon-constraint method for generating the optimal Pareto front. Empirical evidence over a variety of instances shows that the improved method is well suited for finding optimal Pareto fronts with no more computational effort than the traditional method. Instances of up to 150 units and 6 territories are solved in relatively short amount of time. For this problem, the improved method finds practically the same fronts than those found by the traditional epsilon-constraint method. This is, to the best of our knowledge, the only exact method for multi-objective districting developed up to date.

25.5.3 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 25.4 Consider again the example in Fig. 25.3 and assume that the district centers are flexible. Figure 25.6 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 $d_{c_k,j}$ 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

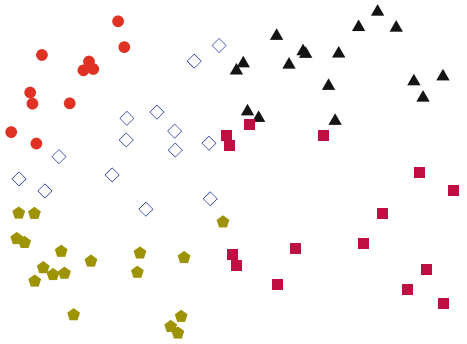


Fig. 25.6 Districting plan with the successive dichotomies algorithm

$w_k\, d_{c_k,j} < w_l\, d_{c_l,j}$, 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 since 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 Lagrangian 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).

25.5.4 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 metaheuristic.

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).

25.5.5 Metaheuristics

Given the NP-hardness of most of the districting problems, it is not surprising that moderate to large scale instances are intractable by exact optimization algorithms. The development of structured heuristic or metaheuristics has been a very important area of research over the past few years. Many interesting ideas and schemes have been developed with great success. A major advantage of these methods is their flexibility to include almost any practical criterion and measure for the design of districts and handle complex constraints. In this subsection we review some of the most relevant works on metaheuristics applied to districting problems in general.

25.5.5.1 Greedy Randomized Adaptive Search Procedure (GRASP)

In recent years, GRASP has been one of the most popular approaches to solve districting problems. An important reason for this is its flexibility to successfully handle connectivity constraints when constructing solutions from scratch.

The first GRASP implementation applied to a districting problem is due to Ríos-Mercado and Fernández (2009). In that work, the authors address a commercial districting problem with connectivity and multiple balancing constraints. They develop a reactive GRASP, where territories are built one at a time during the construction phase and reinsertion and swapping neighborhoods are explored during the improvement phase. The method is enhanced by a reactivity feature that automatically self-tunes the GRASP quality threshold parameter for accepting solutions from the restricted candidate list. The algorithm is tested on data sets coming from a commercial firm that range from 500 to 2000 basic units. It was observed that the algorithm was very robust under many different scenarios, providing solutions of significantly better quality than those from existing practice. In a follow-up work, Ríos-Mercado (2016) provides further experiments by applying the reactive GRASP for solving large scale instances ranging from 1000 to 2000 to basic units under different settings. An interesting finding is that the metaheuristic is able to obtain feasible designs with less than 3% balance deviation.

Fernández et al. (2010) present a GRASP approach for the maximum dispersion territory design problem with three different construction heuristics and several different neighbourhood topologies in its local search phase. Extensive computational testing shows the effectiveness of the proposed algorithm.

Ríos-Mercado and Salazar-Acosta (2011) address a commercial districting problem arising in the bottled beverage distribution industry where a set of city blocks has to be grouped into territories. As planning requirements, the grouping seeks to balance both the number of customers and the product demand across territories, maintain connectivity of territories, and limit the total cost of routing. This work addresses both district design and routing decisions simultaneously by considering a budget constraint on the total routing cost. A GRASP that incorporates advanced features such as adaptive memory and strategic oscillation is presented. Empirical

evidence over a wide set of randomly generated test instances based on real-world data shows a very positive impact of these advanced components, significantly improving the solution quality.

Salazar-Aguilar et al. (2013b) study a commercial districting problem. Each territory must be compact, connected, and balanced according to two activity measures (number of costumers and product demand). Two GRASP heuristics (BGRASP and TGRASP) are proposed for this problem. For each of them two variants are studied: (1) keeping connectivity as a hard constraint during construction and post-processing phases and, (2) ignoring connectivity during the construction phase and adding this as a minimizing objective function during the post-processing phase. The main difference between BGRASP and TGRASP is the way they consider the planning criteria during the construction phase. In BGRASP, the construction attempts to find high quality solutions based on the optimization of two criteria: compactness and balance of the number of customers (product demand balance is treated as a constraint). The construction phase in TGRASP considers three objectives to be optimized: compactness and balance with respect to both activity measures. The proposed procedures are evaluated on a variety of problem instances, with 500 and 1000 basic units. An analysis of these procedures is carried out using different performance measures such as the number of non-dominated points, the k-distance, the size of the space cover (SSC), the coverage of two sets measure, and time. It is observed that SSC, coverage of two sets measure, and time exhibit significant variation depending on the GRASP procedure used. In contrast to that the number of points and k-distance measures did not show any significant variation over all evaluated procedures. BGRASP-I provides good frontiers in short time and BGRASP-II has the best coverage of the efficient points given by the others procedures.

A multi-objective capacitated redistricting problem (MCRP) arising from power meter reading is addressed by de Assis et al. (2014). Two objective functions are considered (compactness and homogeneity of districts) within a bi-objective optimization framework. The redistricting relies on the existence of an original set of districts. The goal of the problem is to partition power utility customers into new districts. The expansion of cities with new developments, population migration, and uneven changes of power demand in the suburbs are examples of forces that pressure the re-definition of districts. Each district refers to the working zone of a group of meter readers that perform readings of power consumption from the customers of that same district. The readings are performed in situ and feed the monthly invoice sent to each customer. The proposed solution method is based on a GRASP and multi-criteria scalarization technique to approximate the Pareto front. The approximate Pareto front is obtained iteratively by solving mono-objective problems in which the objective function is a weighted sum expression of the two criteria under consideration. The GRASP construction phase generates districts, one at a time, by using a greedy function that penalizes both a dispersion measure and district imbalance in weighted manner. If the resulting plan has more than p territories, a repair phase consisting of merging the smallest territories is carried out to ensure p territories are designed. As an improvement phase they use the

reinsertion neighborhood. Computational tests are performed with a diverse set of 24 randomly generated instances with different sizes, demands and densities. A real-life network extracted from the city of São Paulo, Brazil, is also included in the tests. The results demonstrate the effectiveness of GRASP in producing high quality districts with respect to compactness and homogeneity. The results indicate the impact of conformity on the resulting trade-off curve, clearly showing a compromise between attaining compact solutions and maintaining allocations of customers to their current district. The authors conclude that the conformity is thus a relevant criterion and should be included in the optimization and decision making process regarding redistricting problems.

The existing literature reveals that practically all the works on commercial districting use center or median based compactness measures. While these measures yield mixed-integer programming models with some nice properties, they have the disadvantage of being very costly to be evaluated when used within heuristic frameworks. This is due to the center updating operations frequently needed throughout the heuristic search. Ríos-Mercado and Escalante (2016) propose a more robust dispersion measure based on the diameter of the formed territories, allowing for a more efficient heuristic search. For solving this particular territory design problem, they propose a GRASP that incorporates a novel construction procedure where territories are formed simultaneously in two main stages using different criteria. This also differs from previous literature where GRASP was used to build only one territory at a time. The procedure is further enhanced with two variants of forward-backward path relinking, namely static and dynamic. Path relinking is a sophisticated and very successful search mechanism. This idea is novel in any districting or territory design application to the best of our knowledge. The proposed algorithm and its components are extensively evaluated over a wide set of data instances. Experimental results reveal that the construction mechanism produces feasible solutions of acceptable quality, which are improved by an effective local search procedure. In addition, empirical evidence indicate that the two path relinking strategies have a significant impact on solution quality when incorporated within GRASP. The ideas and components of the developed method can be further extended to other districting problems under balancing and connectivity constraints.

25.5.5.2 Tabu Search (TS)

Blais et al. (2003) study a districting problem arising in a local community health clinic in Montreal, Canada, in which five districting criteria must be respected: indivisibility of basic units, respect for borough boundaries, connectivity, visiting personnel mobility, and workload balance. The last two criteria are combined into a single objective function. The authors present a tabu search heuristic considering two different neighborhood topologies. For the case study at hand, the design obtained by the heuristic was able to improve the then current solution in terms of workload balance and personnel mobility.

Bozkaya et al. (2003) propose a tabu search for a districting problem that considers the optimization of four different criteria in a single weighted objective function: population equality, territory compactness, socio-economic homogeneity, and similarity to the existing districting plan. Moreover, connectivity is treated as a hard constraint. The local search is based on a reinsertion and swap neighborhood. Concerning the tabu list, when a given basic unit is used in a move, it remains tabu for the next θ iterations, where θ is chosen randomly. Moreover, an adaptive memory procedure is employed. This procedure is based on the idea that components of high quality solutions can be used to construct other high quality solutions. The method therefore stores in a constantly updated pool a set of districts belonging to some of the best-known solutions. Then, disjoint districts can be extracted from the pool to serve as a basis for a new solution. Each district of the pool, or adaptive memory, is given a larger probability of being selected if it belongs to a better solution. In their empirical work, it was found that the proposed method is robust and powerful since it can easily incorporate a large number of criteria and produces feasible and high quality solutions. When tested on a real-world case study from Edmonton, Canada, the test results indicate that the algorithm can produce maps that dominate the existing districting map of Edmonton with respect to compactness and integrity of communities. It can also reduce the amount of deviation around the average district population from the current 25% to much lower levels (such as 1%), improving on the equality of representation.

Haugland et al. (2007) develop tabu search and multi-start metaheuristics for the problem of designing districts for vehicle routing problems with stochastic demands. In particular, demands are assumed to be uncertain at the time when the districts are made, and these are revealed only after the districting decisions are determined. They use the same neighbourhoods for the local search phase and the same tabu list implementation as in Bozkaya et al. (2003). The authors compare the two heuristics, finding out that tabu search outperforms multi-start.

Ríos-Mercado et al. (2017) present a tabu search metaheuristic as a follow-up to the work on the maximum dispersion territory design problem, first addressed by Fernández et al. (2010). In this paper, the authors significantly improve the previous GRASP approach by incorporating a strategic oscillation component within the tabu search.

25.5.5.3 Simulated Annealing (SA)

D'Amico et al. (2002) address the problem of re-drawing police command boundaries. They model this problem as a constrained graph-partitioning problem involving the partitioning of a police jurisdiction into command districts subject to constraints of contiguity, compactness, convexity and size. Since the districting affects urban emergency services, they also include quality-of-service constraints, which limit the response time (queue time plus travel time) to calls for service. Given the size of the problem, they propose a simulated annealing heuristic to search for good partitions of the police jurisdiction. At each iteration of the algorithm,

they employ a variant of the well-known public domain software tool Patrol Car Allocation Model (PCAM) to optimally assign patrol cars to districts and assess the "goodness" of a particular district design with respect to some prescribed performance measures. For the neighbour topology, they consider moves that reassign a basic unit from a given district to an adjacent district. A computational case study using data from the Buffalo, NY, Police Department (BPD) is carried out revealing the merits of this approach. Among their main findings it was observed that under optimal car allocations, they are able to find an improved district design that lowers the disparity among officer workloads from 30% to only 14%. Also, the proportion of small workloads under 36% is greatly reduced. Hence, officer workloads are better balanced (primarily between 36% and 42%) across all districts and work shifts. At the same time, the response time feasibility constraints ensured no increase in the maximum response time of 29 min under current BPD operations.

25.5.5.4 Genetic Algorithm (GA)

Bação et al. (2005) solve a political districting problem using a genetic algorithm and apply it to a case study in Portugal. Their results indicate that the GA obtains better results when compared to the current practice.

Tavares-Pereira et al. (2007) study a multi-objective districting problem arising for Paris public transportation. The goal is to partition a territory into "homogeneous" zones without inclusions, where each zone is composed of a set of elementary territorial units. They propose a genetic algorithm to approximate the Pareto front based on an evolutionary algorithm with local search. The algorithm presents a new solution representation and new crossover/mutation operators. The algorithm can deal with multiple criteria, allows to solve large-size instances in a reasonable time, and generates high quality solutions. The algorithm is applied to the Paris region public transportation.

Steiner et al. (2015) address a health-care districting problem arising in Parana State, Brazil. The motivation for the problem is to develop a better system for patients by aggregating various health services offered in the municipalities of Parana into micro regions. The problem is formulated as a multi-objective graph partitioning problem, where the municipalities are represented by nodes, and roads connecting them are represented by edges. Their three-objective optimization problem considers maximizing the population homogeneity in the micro regions, maximizing the variety of medical procedures offered in the micro regions, and minimizing the inter-micro region distances to be traveled by patients. They develop a multi-objective genetic algorithm, which yields a significant improvement over the existing health-care system map of Parana State.

Forman and Yue (2003) present a genetic algorithm for a political districting problem, where the encoding of solutions and the genetic operators are based on the ones for Traveling Salesman Problems. This encoding forces near equality of district population and uses the fitness function to promote district contiguity and compactness. A post-processing step further refines district population equality.

Results are provided for three states (North Carolina, South Carolina, and Iowa) using the 2000 census data.

25.5.5.5 Hybrid and Miscellaneous Approaches

Bergey et al. (2003) address an electrical power districting problem arising in the Republic of Ghana. Due to a variety of political, economic, and technological factors, many national electricity industries around the globe are transforming from non-competitive monopolies with centralized systems to decentralized operations with competitive business units. A key challenge faced by energy restructuring specialists at the World Bank is trying to simultaneously optimize the various criteria one can use to judge the fairness and commercial viability of a particular power districting plan. The authors propose a simulated annealing genetic algorithm for this problem. In their empirical work, they observe that the proposed method outperformed a well-known parallel simulated annealing heuristic.

Wei and Chai (2004) present a hybrid approach combining tabu search and scatter search for solving a multi-objective spatial zoning model. The problem considers a scalar function with three objectives: population unbalance, territory compactness, and socioeconomic homogeneity. The model also includes resource capacity constraints, but no connectivity constraints. Later, Bong and Wang (2006) tackle another multi-objective zoning model that optimizes four criteria: population equality, territory compactness, socio-economic homogeneity, and similarity of a solution with the existing plan. The model also includes resource capacity constraints. The authors propose a hybrid algorithm with elements from tabu search, scatter search, and path relinking. A comparative study between the results of multi-objective decision-making and single objective decision-making is conducted for the proposed multi-objective method with a selected single objective method called WAMCF. The empirical results show concrete evidence on two aspects that the proposed method can produce better results for the problem with lower values in the objectives achieved for the minimization problem. It was also observed that a more consistent result for the individual solution was delivered compared to the single objective approach because there is a big difference between the generated maximum and minimum best values.

Ricca and Simeone (2008) present a comparison of several local search meta-heuristics for political districting considering territory connectivity, minimizing measures of population inequality, noncompactness, and nonconformity to administrative boundaries. Experiments on a set of medium to large real-life instances is carried out using descent search, tabu search, simulated annealing, and old bachelor acceptance algorithms. Except for descent, all local search methods show a very good performance. In particular, old bachelor acceptance produces the best results in the majority of the cases, especially when the objective function is focussing on compactness.

Salazar-Aguilar et al. (2012) propose a multi-objective scatter search heuristic for a bi-objective territory design problem. They consider a problem where compact-

ness and balance with respect to product demand are sought. The problem includes also balancing territories with respect to workload and territory connectivity. The proposed scatter search-based framework contains a diversification step based on a greedy randomized adaptive search procedure, an improvement step based on a relinked local search strategy, and a combination step based on a solution of an assignment problem. The proposed metaheuristic is evaluated over a variety of instances taken from literature. This includes a comparison with two of the most successful multi-objective heuristics from literature such as the scatter tabu search procedure for multi-objective optimization by Molina et al. (2007), and the non-dominated sorting genetic algorithm by Deb et al. (2002). Experimental work reveals that the proposed procedure consistently outperforms both existing heuristics from literature on all instances tested.

25.5.6 Lower Bounding Schemes

To the best of our knowledge, the only work on lower bounds for districting problems is due to Elizondo-Amaya et al. (2014). In their work, the authors study a commercial districting problem that minimizes territory dispersion based on a p-center type of function subject to multiple balance constraints. Lower bounds are obtained using a binary search over a range of coverage distances. For each coverage distance a Lagrangian relaxation of a maximal covering model is effectively used. Their computational results indicate that the bounding scheme provides tighter lower bounds than those obtained by the linear programming relaxation.

25.6 Conclusions

In this chapter, we have given 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 was given and the various solution approaches for them that have been used were highlighted.

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) metaheuristic for a districting model with yet another measure for compactness or additional constraint, research should first and 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.

Acknowledgement 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 JR, Clayton ER, Moore LJ (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

Bender M, Meyer A, Kalcsics J, Nickel S (2016) The multi-period service territory design problem – an introduction, a model and a heuristic approach. Transp Res E Logist 96:135–57

Bender M, Kalcsics J, Nickel S, Pouls M (2018) A branch-and-price algorithm for the scheduling of customer visits in the context of multi-period service territory design. Eur J Oper Res 269:382–396

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 PK, Ragsdale CT, Hoskote M (2003) A simulated annealing genetic algorithm for the electrical power districting problem. Ann Oper Res 121:33–55

Blais M, Lapierre SD, 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

Bong CW, Wang YC (2006) A multi-objective hybrid metaheuristic for zone definition procedure. Int J Ser Oper Inf 1:146–164

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

Camacho-Collados M, Liberatore J, Angulo JM (2015) A multi-criteria police districting problem for the efficient and effective design of patrol sector. Eur J Oper Res 246:674–684

Carlsson JG, Delage E (2013) Robust partitioning for stochastic multivehicle routing. Oper Res 61:727–744

D'Amico SJ, Wang SJ, Batta R, Rump CM (2002) A simulated annealing approach to police district design. Comput Oper Res 29:667–684

de Assis LS, Franca PM, Usberti FL (2014) A redistricting problem applied to meter reading in power distribution networks. Comput Oper Res 41:65–75

Deb K, Pratap A, Agarwal S, Meyerivan T (2002) A fast elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans Evol Comput 6(2):182–197

Deckro RF (1977) Multiple objective districting: a general heuristic approach using multiple criteria. Oper Res Q 28:953–961

Dixon RG (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 Recogn 41:3224–3236

Elizondo-Amaya MG, Ríos-Mercado RZ, Díaz JA (2014) A dual bounding scheme for a territory design problem. Comput Oper Res 44:193–205
Ferland JA, Guénette G (1990) Decision support system for a school districting problem. Oper Res 38:15–21
Fernández E, Kalcsics J, Nickel S, Ríos-Mercado RZ (2010) A novel maximum dispersion territory design model arising in the implementation of the WEEE-directive. J Oper Res Soc 61:503–514
Fernández E, Kalcsics J, Nickel S (2013) The maximum dispersion problem. Omega 41:721–730
Fleischmann B, Paraschis JN (1988) Solving a large scale districting problem: a case report. Comput Oper Res 15:521–533
Forman SL, Yue Y (2003) Congressional districting using a TSP-based genetic algorithm. In: Cantú-Paz E, Foster JA, Deb K, David L, Rajkumar R (eds) Genetic and evolutionary computation – GECCO 2003. Lecture Notes in Computer Science, vol 2723. Springer, Berlin, Germany, pp 2072–2083
Galvão LC, Novaes AGN, Souza de Cursi JE, Souza JC (2006) A multiplicatively-weighted Voronoi diagram approach to logistics districting. Comput Oper Res 33:93–114
García-Ayala MG, González-Velarde JL, Ríos-Mercado RZ, Fernández E (2016) A novel model for arc territory design: promoting Eulerian districts. Int Trans Oper Res 23:433–458
Garfinkel RS, Nemhauser GL (1970) Optimal political districting by implicit enumeration techniques. Manag Sci 16:495–508
Gentry S, Chow E, Massie A, Segev D (2015) Gerrymandering for justice: redistricting U.S. liver allocation. Interfaces 45:462–480
George JA, Lamar BW, Wallace CA (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: Bagozzi RP (ed) Sales management: new developments from behavioral and decision model research. Marketing Science Institute, Cambridge, pp 325–343
Gross JL, 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 SC, Laporte G (2007) Designing delivery districts for the vehicle routing problem with stochastic demands. Eur J Oper Res 180:997–1010
Hess SW, Samuels SA (1971) Experiences with a sales districting model: criteria and implementation. Manag Sci 18:41–54
Hess SW, Weaver JB, Siegfeldt HJ, Whelan JN, Zitlau PA (1965) Nonpartisan political redistricting by computer. Oper Res 13:998–1008
Hojati M (1996) Optimal political districting. Comput Oper Res 23:1147–1161
Horn DL, Hampton CR, Vandenberg AJ (1993) Practical application of district compactness. Polit Geogr 12:103–120
Howick RS, Pidd M (1990) Sales force deployment models. Eur J Oper Res 48:295–310
Jarrah AI, Bard JF (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
King DM, Jacobson SH, Sewell EC, Tam Cho WK (2012) Geo-graphs: an efficient model for enforcing contiguity and hole constraints in planar graph partitioning. Oper Res 60:1213–1228
Lei H, Laporte G, Guo B (2012) Districting for routing with stochastic customers. Eur J Transp Logist 1:67–85
Lei H, Laporte G, Liu Y, Zhang T (2015) Dynamic design of sales territories. Comput Oper Res 56:84–92
Lewyn ME (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 58:104–111

Lodish LM (1975) Sales territory alignment to maximize profit. J Mark Res 12:30–36
López-Pérez JF, Ríos-Mercado RZ (2013) Embotelladoras ARCA uses operations research to improve territory design plans. Interfaces 43:209–220
Mayorga ME, Bandara D, McLay LA (2013) Districting and dispatching policies for emergency medical service systems to improve patient survival. IIE Trans Healthc Syst Eng 3:39–56
Mehrotra A, Johnson EL, Nemhauser GL (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
Molina J, Martí R, Caballero R (2007) SSPMO: a scatter tabu search procedure for non-linear multiobjective optimization. INFORMS J Comput 19:91–100
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 SS (1965) Simplified bipartisan computer redistricting. Stanford Law Rev 17:863–899
Niemi RG, 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 FR (1990) Black votes count: political empowerment in Mississippi after 1965. 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 RZ (2016) Assessing a metaheuristic for large scale commercial districting. Cybern Syst 47:321–338
Ríos-Mercado RZ, Bard JF (2019) An exact algorithm for designing optimal districts in the collection of waste electric and electronic equipment through an improved reformulation. Eur J Oper Res 276:259–271
Ríos-Mercado RZ, Escalante HJ (2016) GRASP with path relinking for commercial districting. Expert Syst Appl 44:102–113
Ríos-Mercado RZ, Fernández E (2009) A reactive GRASP for a commercial territory design problem with multiple balancing requirements. Comput Oper Res 36:755–776
Ríos-Mercado RZ, López-Pérez JF (2013) Commercial territory design planning with realignment and disjoint assignment requirements. Omega 41:525–535
Ríos-Mercado RZ, Salazar-Acosta JC (2011) A GRASP with strategic oscillation for a commercial territory design problem with a routing budget constraint. In: Batyrshin I, Sidorov G (eds) Advances in soft computing: proceedings of the 10th Mexican international conference on artificial intelligence (MICAI 2011), Part II. Lecture Notes in Computer Science, vol 7095, Springer, Heidelberg, pp 307–318
Ríos-Mercado RZ, González-Velarde JL, Maldonado-Flores JR (2017) Tabu search with strategic oscillation for improving collection assignment plans of waste electric and electronic equipment. Technical report PISIS-2017-01, Graduate Program in Systems Engineering, UANL, San Nicolas de los Garza, Mexico, May 2017
Salazar-Aguilar MA, Ríos-Mercado RZ, Cabrera-Ríos M (2011a) New models for commercial territory design. Netw Spat Econ 11:487–507
Salazar-Aguilar MA, Ríos-Mercado RZ, González-Velarde JL (2011b) A bi-objective programming model for designing compact and balanced territories in commercial districting. Transp Res C Emerg Technol 19:885-895

Salazar-Aguilar MA, Ríos-Mercado RZ, González-Velarde JL, Molina J (2012) Multiobjective scatter search for a commercial territory design problem. Ann Oper Res 199:343–360
Salazar-Aguilar MA, González-Velarde JL, Ríos-Mercado RZ (2013a) A divide-and-conquer approach for a commercial territory design problem. Computación y Sistemas 16:309–320
Salazar-Aguilar MA, Ríos-Mercado RZ, González-Velarde JL (2013b) GRASP strategies for a bi-objective commercial territory design problem. J Heuristics 19:179–200
Schoepfle OB, Church RL (1991) A new network representation of a "classic" school districting problem. Socio Econ Plan Sci 25:189–197
Schröder M (2001) Gebiete optimal aufteilen. PhD thesis, Universität Karlsruhe, Karlsruhe, Germany (in German)
Shirabe T (2009) Districting modeling with exact contiguity constraints. Environ Plan B 36:1053–1066
Simchi-Levi D, Kaminsky P, Simchi-Levi E (2003) Designing and managing the supply chain: concepts, strategies, and case studies, 2nd edn. McGraw-Hill, New York
Steiner MTA, Datta D, Neto PJS, Scarpin CT, Figueira JR (2015) Multi-objective optimization in partitioning the healthcare system of Parana State in Brazil. Omega 52:53–64
Tavares-Pereira F, Figueira JR, 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
Wei BC, Chai WY (2004) A multiobjective hybrid metaheuristic approach for GIS-based spatial zoning model. J Math Model Algorithms 3:245–261
Williams JC Jr (1995) Political redistricting: a review. Pap Reg Sci 74:13–40
Young HP (1988) Measuring the compactness of legislative districts. Legis Stud Q 13:105–115
Zhong H, Hall RW, 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 AA, Sinha P (2005) Sales territory design: thirty years of modeling and implementation. Mark Sci 24:313–331

Chapter 26
Facility Location in the Public Sector

Knut Haase, Lukas Knörr, Ralf Krohn, Sven Müller, and Michael Wagner

Abstract In this chapter we focus on facility location problems that arise in the public sector. In particular, we consider selected problems in transportation, health care, and education—important sectors of public service. The adequate consideration of demand in these models is of core interest in this chapter. Besides a discussion of selected model formulations we provide a quantitative and qualitative overview of recent publications in the field.

26.1 Introduction

In this chapter, we discuss recent work related to public sector facility location planning. Of course, a location of a public service does not necessarily strictly belong to the public sector. For example, healthcare facilities may also be owned by a private firm while being regulated by a public health agency. The main difference between the planning of public and private facility locations are the objectives that are considered by decision makers. The optimization criteria in private applications are mainly profit and market capture maximization, whereas in public applications social cost minimization, access, efficiency, and equity are the primary goals. Since the measurement of these objectives is relatively difficult, they are frequently simplified by minimizing the locational and operational costs needed for full coverage, or the search for maximal coverage under a given amount of available resources (Marianov and Serra 2002).

K. Haase · R. Krohn
Universität Hamburg, Faculty of Business Administration (HBS – Hamburg Business School), Institute of Transport Economics, Hamburg, Germany

L. Knörr · S. Müller (✉) · M. Wagner
Karlsruhe University of Applied Sciences, Institute for Transport & Infrastructure, Karlsruhe, Germany
e-mail: smueller@europa-uni.de

G. Laporte et al. (eds.), *Location Science*,
https://doi.org/10.1007/978-3-030-32177-2_26

Most of the public facility location models proposed in the selected papers rely on covering problems, *p*-median problems or a combination of both. They are benchmarks in the development of location models. The public sector applications of covering models are based on the concept of acceptable proximity. If a service is provided by a facility located within a maximum distance or travel time, the service is considered adequate – the client demand is covered. Two major types of formulations can be distinguished in such covering models. Set covering models seek to minimize the number of facilities needed for full coverage of the population. In contrast, maximum covering models are limited by the number of facilities or services and maximize the covered population share. Furthermore, a distinction can be made between fixed servers (e.g., schools, hospitals) and systems with mobile servers (e.g., ambulances, see Nickel et al. (2016)). Additionally, a server can be classified as capacitated or uncapacitated. An example of a capacitated service is a primary school that has a limit on the number of students who can enroll in a particular year (Marianov and Serra 2002; Müller et al. 2009). In the following, we discuss selected areas of application of public facility location planning approaches: Bike sharing systems, simultaneous bus scheduling and depot location planning, electric vehicle charging station planning, healthcare facility location planning, and school location planning. The presented models are classified as discrete location-allocation models as well as location choice models.

26.2 Bike Sharing

Since the political interest in the promotion of cyclists continues to increase, it is important to create enough hubs and parking areas for bicycles. People who do not have a bicycle on their own should have an appropriate possibility to use bicycles within cities. Thus, bike sharing models are becoming more and more popular. Bike sharing is often linked to transport hubs, but there are also stand-alone models for rental stations and also station-less approaches.

With the proposal of Sayarshad et al. (2012), an optimization formula to design a bike-sharing system for small communities is presented. This formula can also be used to extend the public transport with incoming and outgoing bike traffic. They try to find a minimum required bike fleet size that also minimizes the unmet demand, non-utilized bikes, and the need to transport bikes between the stations. The mathematical model maximizes the considered company's total benefit where the objective function consists of six terms: (1) the revenue from rented bikes traveling between network nodes, (2) the cost of moving empty bikes within the network, (3) the cost of processing and maintenance of bikes, (4) the bike holding cost at a station, (5) the bikes' capital cost per period, and (6) the penalty cost of unmet demand.

By combining the models for private cars and public bicycles, in Romero et al. (2012) the goal is to achieve an efficient and sustainable transport system that is also economically and socially efficient at the same time. The choice between motor

vehicle and bicycle and route selection is simulated by a user behavior model. Thus, in turn, a combined vehicle-bicycle transport network was created on which a modal split model can be matched. The goal is then to optimize the location of bike stations. The study in Lin and Yang (2011) deals with strategic planning of bike sharing taking into account both the interests of users and investors. Considering those interests, the model attempts to determine the number and location of bike sharing stations, the network structure as well as the travel paths between the stations. Lin et al. (2013) continue expanding this approach by considering the number and location of bicycle stations in the system, the creation of bicycle lanes, and selection of paths between the stations and the inventory levels of the bike sharing facilities. Decisions are made under consideration of total costs and service. An approach to maximize the coverage of a bike sharing facility by also using the available budget as a constraint is proposed by Frade and Ribeiro (2015). They combine the strategic decision for a bike sharing facility and the dimension of the stations with operational decisions. The result is an optimal location as well as the capacity of each station and the number of bikes needed while staying within the budget.

26.3 Location Decisions in Public Transport

Traffic planners face the trade-off between improving accessibility with additional bus stops while simultaneously increasing efficiency so that traffic reach destinations in a reasonable time. Delmelle et al. (2012) address this specific problem with an optimization framework that builds upon facility location coverage models. In contrast to the p-median and maximal covering location problem, the demand can partially be assigned to more than one facility. Furthermore, facility attraction is explicitly integrated. The modeling approach considers the impact of walking distance from a residential location to a stop as well as the transit facility attractiveness (the number of destinations served, for example). Cipriani et al. (2012) deal with the bus network design problem in a multimodal transit context. The approach determines the (near) optimal network configuration regarding bus routes and service frequencies. It aims to minimize the total costs involved in the transport system. A similar method is used by Ciaffi et al. (2012) to solve the feeder-bus network design problem. Their results show that the design procedure could lead to a reduction of the total travel time, an increase in the number of transfers, in a more efficient way.

The locations of bus stops affect travel times and therefore also the expected demand. By using a random utility model (RUM) we can measure the expected impact of travel time and other factors on demand. Klier and Haase (2015) integrate a RUM in line planning that results in a difficult optimization problem. If we assume that distance to the departure stop is the only relevant factor influencing the choice behavior over a given set of potential stop locations, RUM approaches as defined in Haase and Müller (2013, 2014), Müller and Haase (2014) or Ljubić and Moreno (2018) might be appropriate.

Another topic in public transport is the location of bus depots. The depot locations determine the vehicle costs. Therefore, we combine vehicle scheduling and bus-depot location in one integrated approach.

Defining the sets

$\mathcal{N}$ set of nodes representing line trips and potential bus depot nodes,
$\mathcal{M}$ set of potential bus depot nodes,
$\mathcal{I}$ set of nodes representing line trips,
$\mathcal{A}$ set of arcs representing feasible idle trips (compatible with the timetable),

the parameters

c_{ij} costs of idle trip $(i, j) \in \mathcal{A}$,
f_m fixed costs per day of depot m,
k_m maximum number of vehicles in depot m,

and the binary variables

X_{mij} = 1 if a vehicle from depot m serves idle trip $(i, j) \in \mathcal{A}$ (0, otherwise),
Y_m = 1 if depot m is to be established (0, otherwise)

then we formulate the depot location and vehicle scheduling model as follows:

$$\text{Minimize} \quad F = \sum_{m \in \mathcal{M}} \sum_{(i,j) \in \mathcal{A}} c_{ij} X_{mij} + \sum_{m \in \mathcal{M}} f_m Y_m \tag{26.1}$$

subject to

$$\sum_{m \in \mathcal{M}} \sum_{(i,j) \in \mathcal{A}} X_{mij} = 1 \qquad \forall\, i \in \mathcal{I} \tag{26.2}$$

$$\sum_{(i,j) \in \mathcal{A}} X_{mij} - \sum_{(j,i) \in \mathcal{A}} X_{mji} = 0 \qquad \forall\, m \in \mathcal{M};\ j \in \mathcal{N} \tag{26.3}$$

$$\sum_{(m,j) \in \mathcal{A}} X_{mmj} \le k_m Y_m \qquad \forall\, m \in \mathcal{M} \tag{26.4}$$

$$X_{mij} \in \{0, 1\} \qquad \forall\, m \in \mathcal{M};\ (i, j) \in \mathcal{A} \tag{26.5}$$

$$Y_m \in \{0, 1\} \qquad \forall\, m \in \mathcal{M} \tag{26.6}$$

The objective function (26.1) minimizes the total costs per day. Equation (26.2) ensure that each line trip is operated exactly once. Equation (26.3) are flow conservation constraints and Eq. (26.4) ensure that trips can only start from a depot if it is established and the depot capacity is considered.

26.4 Electric Vehicle Charging Station Location

In the application of electric vehicle (EV) charging station location, we find plenty of recent work attributed to the technical developments in the EV industry and to the rising importance of eco-friendly transport modes in times of climate change (Müller and He 2018). With the increasing demand for vehicles with alternative fuel usage, the demand for their charging or refueling stations is also increasing. The papers in Table 26.1 discuss this topic in several approaches by maximizing the coverage, maximizing the traffic flow, minimizing the costs or by combining of these objectives.

Frade et al. (2011) present a charging location problem for parked cars. For this study area, a slow-charging model is suitable, because parked cars are parked for several hours. The proposed model is based on a maximum coverage location model (MCLP) to optimize the demand coverage by simultaneously keeping an acceptable level of service. They optimize the number of stations and the scale of each station. As input parameters, an estimated refueling demand for the day and nighttime is needed. The approach of Giménez-Gaydou et al. (2016) also covers urban areas. Their models consist of a location-allocation model with detailed analysis of charging needs, charging coverage, and adoption potential. Zheng et al. (2017) investigate a network-design-like problem with a bi-level structure. While the upper level aims for optimal locations with minimized general costs calculated from travel time and energy consumption, the lower level aims at minimized individual costs with traffic equilibrium. By adding the lower level to the upper level, those two levels are then combined to a single level model. The hybrid model from Mozafar et al. (2017) handles the optimal allocation and sizing of either renewable energy sources or electric vehicle charging stations. A multi-objective problem is created to obtain several objective variables such as reducing power losses, voltage fluctuations, charging and demand-supply costs, and battery costs. The location and the dimension of the charging stations are handled as decision variables.

In contrast to public charging stations for private vehicles, Yang et al. (2017) introduce a location model for electric powered taxis. With the goal to minimizing the infrastructure costs, an integer linear program (ILP) is formulated. Their key findings include positioning of the charging stations matching the dwell pattern of the taxis, with the combination of charging and waiting spots, fewer chargers are needed and this compromise can be qualified by the cost of charging spots versus parking spots. Another taxi-based approach is proposed by Tu et al. (2016). In contrast to the approach of Yang et al. (2017), their model's goal is to maximize the charging station service within the taxi network. To achieve this, a spatial-temporal demand coverage location model is proposed and the results are analyzed with respect to spatial coverage, temporal demand availability, and waiting and loading behavior. A bus charging model is proposed by Xiang and Zhang (2017). In contrast to the taxi models, for buses with electric drive it is common to replace the battery instead of charging it. A particle swarm optimization algorithm (PSO) is used to

Table 26.1 Charging and refueling station location papers

Reference	Application	Objective	Modeling approach	Solution method	Demand
Frade et al. (2011)	EV charging stations location	Maximize coverage	MCLP, MIP		Distinguished daytime/nighttime, variable
Giménez-Gaydou et al. (2016)	Charging station location	Maximize coverage	Uncapacitated gradual maximal covering model	XPRESS	Willingness and socio-economic factors, variable
Zheng et al. (2017)	Charging network design	Minimize total cost	Bi-level MILP	CPLEX	Choice decisions, variable
Mozafar et al. (2017)	Allocation and sizing of charging stations	Minimize power loss, voltage fluctuation, power supply and total cost	Multi-objective optimization problem	GA-PSO, MATLAB	Variable
Yang et al. (2017)	Taxi charging stations	Minimize infrastructure costs	ILP	MATLAB, YALMIP, Gurobi solver	Fixed
Tu et al. (2016)	Taxi charging stations	Maximize charging network service	STDCLM	GA	Spatial-temporal dynamic demand, variable
Xiang and Zhang (2017)	Bus charging station location	Minimize total costs	IP	PSO	Fixed
Ghamami et al. (2016a)	Charging facilities located on existing parking lots	Minimize total system cost	Fixed CFM	AMPL, Knitro solver	Considering drivers preference, uncertain market penetration rate, variable
Ghamami et al. (2016b)	Charging facilities location along highways	Minimize total system cost	MINLP	SA, B&B, Knitro solver	Considering flow-dependent charging delay, variable

Jeong (2017)	Charging network planning	Minimize total construction costs	SCRLP, VRCP	CPLEX	
Riemann et al. (2015)	Flow refueling location model	Maximize total captured flow of traffic	MCLP, MINLP	B&B, CPLEX	Travelers route choice behavior, variable
Arslan and Karaşan (2016)	Flow refueling location model	Maximize flow volume; minimize total cost	FRLM, CSLP-PHEV	Arc-cover formulation and Benders decomposition	
Miralinaghi et al. (2017a)	Alternative fuel network system	Minimize construction and operational costs	CFLP	B&B, Lagrangian relaxation	Variation in hydrogen refueling demand, variable
Miralinaghi et al. (2017b)	Refueling demand uncertainty	Minimize total costs	RCPM	GA, CPLEX	Variable
Hosseini and MirHassani (2015)	Refueling station location	Maximize total flow	MIP	CPLEX	Variable
Guo et al. (2016)	Charging infrastructure planning	Maximize profit	MOPEC, CDA	PYOMO, Gurobi solver	Multinomial logic model describes choice, variable

Note: *CFM* cost facility model, *SCRLP* set covering version of the refueling-station location problem, *VRCP* vertex restricted covering problem, *STDCLM* spatial-temporal demand coverage location model, *FRLM* flow refueling location modeling, *CSLP-PHEV* charging station location problem with plug-in hybrid electric vehicles

calculate the optimal location for the replacement facilities with a minimum of total costs (transport costs, construction costs, and operating costs).

Another approach is presented by Ghamami et al. (2016a) with the goal to minimize the total system cost. Their idea is to use existing parking lots to install charging facilities. The model becomes complex by introducing costs for uncovered demand and also considering the drivers' preferences for familiar parking lots. Ghamami et al. (2016b) aim to configure charging stations to support long distance intercity travel by using a general corridor model minimizing the total system costs including infrastructure, battery, and user costs. Using a mixed-integer program with non-linear constraints, it is possible to use realistic patterns of origin-destination demands and also considering flow-dependent charging delay caused by traffic jam. With this model, a strategic design of charging stations along highways is possible.

Different to the parking-and-charging models, a wireless-charging model is investigated by Riemann et al. (2015). Based on a mixed-integer non-linear program (MINLP), a method is formulated to find a number of charging facility locations out of a set of candidates and to maximize the total captured flow. Similar to this, a flow refueling location problem for both electric and plug-in hybrid vehicles is introduced by Arslan and Karaşan (2016). With the goal of maximizing the vehicle miles that can be traveled and minimizing the total cost, the presented exact solution is an arc-cover formulation and makes use of a Benders decomposition approach.

To propose a model for locating refueling stations in a transport network, Miralinaghi et al. (2017a) assume that a central planner such as a hydrogen manufacturer or a government agency is planning the locations for refueling stations with alternative fuel type, especially hydrogen. Considering a multi-period travel demand, both the non-linear refueling station operational cost and the deviation of travelers from their shortest routes to refuel are taken into account. The proposed capacitated facility location problem (CFLP) is solved with a combination of Branch-and-Bound and Lagrangian relaxation. Another approach, presented by Miralinaghi et al. (2017b), considers the refueling demand uncertainty with the effect of the deviation of travelers to refuel. A cutting plane algorithm is used to solve the robust centralized planning model (RCPM). The uncertainty model from Hosseini and MirHassani (2015) provides a two-stage stochastic refueling station model for permanent stations in the first stage and portable stations in the second stage. Portable refueling stations are an innovative feature that can be used to close temporary gaps in supply. A business-driven model for charging infrastructure planning is introduced by Guo et al. (2016) by using a multi-agent optimization problem with equilibrium constraint (MOPEC). The goal is to maximize providers' profit. An approach in which a charging network can be planned without existing facilities comes from Jeong (2017). They also provide a dynamic-programming-based algorithm for the case where facilities already exist. The goal is to minimize the total construction costs of the charging network by minimizing the cost of charging stations. In particular, the model underlies the following assumptions: (1) multiple origin-destination round trips along the shortest paths, (2) a single type of alternative fuel vehicle with a constant driving range, (3) uncapacitated stations, (4) possible refueling station locations that are only nodes in the traffic

network (i.e., vertex restricted refueling stations), (5) a linear relationship between fuel consumption and driving distance, and (6) fully fueled vehicles at the point of origin.

Defining the sets

$\mathcal{N}$	nodes of the network,
$\mathcal{E}$	existing refueling stations $\mathcal{E} \subset \mathcal{N}$,
$\mathcal{K}$	considered alternative fuel vehicles,
$\mathcal{P}_k$	sequence of arcs (i, j) along path of vehicle k, and

the parameters

d_{ij}	(Euclidean) distance from node i to node j,
c_i	cost of refueling station at node i,
S	maximum vehicle range, and

the variables

X_i	= 1 if a refueling station is set up at node i (0, otherwise),
Y_{ik}	= 1 if vehicle k is recharged at node i (0, otherwise),
Z_{ik}	remaining driving range of vehicle k at node i, and
W_{ik}	additional driving range of vehicle k if refueled at node i,

the refueling station location problem is to

$$\text{minimize } F = \sum_{i \in \mathcal{N} \setminus \mathcal{E}} c_i X_i \tag{26.7}$$

subject to

$$Y_{ik} \leq X_i \qquad \forall\, i \in \mathcal{N} \setminus \mathcal{E}, k \in \mathcal{K} \tag{26.8}$$

$$W_{ik} = SY_{ik} \qquad \forall\, k \in \mathcal{K}, i \in \mathcal{N} \tag{26.9}$$

$$W_{ik} \leq S - Z_{ik} \qquad \forall\, k \in \mathcal{K}, i \in \mathcal{N} \tag{26.10}$$

$$Z_{jk} = Z_{ik} + W_{ik} - d_{ij} \qquad \forall\, k \in \mathcal{K}, (i, j) \in \mathcal{P}_k \tag{26.11}$$

$$X_i \geq 0 \qquad \forall\, i \in \mathcal{N} \setminus \mathcal{E} \tag{26.12}$$

$$Y_{ik} \in \{0, 1\} \qquad \forall\, i \in \mathcal{N}, k \in \mathcal{K} \tag{26.13}$$

$$Z_{ik}, W_{ik} \geq 0 \qquad \forall\, i \in \mathcal{N}, k \in \mathcal{K}. \tag{26.14}$$

The objective (26.7) is the minimization of the total set up cost of refueling stations. If there is no refueling station at node i, refueling cannot occur at i by constraint (26.8). The refueling amount at node i is S by constraint (26.9), and this amount must not exceed $S - Z_{ik}$ by constraint (26.10) if the vehicle refuels at node i. Constraint (26.11) defines the remaining distance using the remaining fuel at each node i. Considering arc (i, j) the remaining fuel at node j is the sum of the remaining fuel at node i and the fueled amount at node i minus distance between

node i and node j. The authors show that the problem is $\mathcal{NP}$-complete and propose several procedures for its solution. The approach is used to analyze the diffusion of alternative fuel recharging stations in a given market.

26.5 Spatial Planning for Health Care Facilities

One of the key factors to achieve a high standard in healthcare is a systematic and efficient system planning (Shariff et al. 2012). See Chap. 23 for a more detailed discussion. Therefore, it is important to develop methods to facilitate the planners' decision making process in the locating of new healthcare facilities (Zhang et al. 2016) (Table 26.2).

To find a more systematic and efficient way of locating healthcare facilities, Shariff et al. (2012) use a MCLP with capacitated facilities. Zhang et al. (2016) investigate the location problem of healthcare facilities to maximize the equity of accessibility and the total accessibility and to minimize the population outside the coverage range, and minimize the cost of new buildings.

Two location-allocation models to handle the uncertainty in the strategic hospital network planning are proposed by Mestre et al. (2015). The models aim to inform about the (re-) organization of hospital networking systems by improving geographical access (minimize expected travel time) while minimizing costs.

The problem of determining locations for long-term care facilities is investigated in Djenić et al. (2017), where the objective is to minimize the maximum number of patients that are assigned to a single installed facility.

Kim and Kim (2013) focus on public healthcare facilities that can be used by low-income patients. They examine the problem of determining locations of public healthcare facilities within a given budget and allocating the patients to the facilities. The objective is to maximize the number of served patients while considering the patients' preferences of the for the public and private facilities. Basu et al. (2018) focus on socio-economically weaker patients. They aim to quantify the gap in affordable healthcare facilities access. The optimization model shows where new public facilities are required, and the positive impact of the proposed model with increasing coverage is detected.

Besides operations research applications in healthcare operation management, the design of blood supply networks also is important. Hospitals and clinics as demand centers are dependent on blood products and an efficient procurement system is needed. Arvan et al. (2015) intend to locate blood bank components in a network and to determine the allocations among these network components (donation sites, testing and processing labs, blood banks, and demand points). The main objectives are to identify the locations of donation points and central blood banks as well as to decide about the product quantity that is shipped among the facilities. To model the problem a bi-objective approach is proposed not only to minimize the cost but also to minimize the time period in which blood products remain in the network.

Table 26.2 Healthcare facility location papers

Reference	Application	Objective	Modeling approach[a]	Solution method	Demand
Shariff et al. (2012)	Healthcare facility planning	Maximize coverage	CMCLP as a variation of MCLP	GA, CPLEX	Fixed
Zhang et al. (2016)	Healthcare facility location-allocation	Maximize accessibility; minimize inequity of accessibility; minimize the uncovered population, minimize building cost	Multi-objective optimization	GA	Fixed
Mestre et al. (2015)	Location-allocation for hospital planning under uncertainty	Minimize expected travel time, minimize expected cost and capital costs	Model 1: Location as a first-stage decision, Model 2: Location and allocation as first-stage decisions	ε-Constraint method of multi-objective programming	Variable
Djenić et al. (2017)	Long-term care facility location	Minimize the maximum number of patients that are assigned to a single installed facility	LTCFLP LTCFLP-I	Metaheuristic method based on a Variable Neighborhood Search (VNS)	Fixed
Kim and Kim (2013)	Public healthcare facility location	Maximize the number of served patients	IP	Heuristic algorithm based on LR and subgradient optimization methods	Considers patients preference for the public and private facilities, variable
Basu et al. (2018)	Healthcare facility allocation	Maximize the healthcare coverage by minimum number of new public healthcare facilities			Variable
Zhang et al. (2012)	Preventive healthcare facility location	Maximize preventive healthcare program participation	Probabilistic-choice model, optimal-choice model based on MNL, MIP	CPLEX	Variable

(continued)

Table 26.2 (continued)

Reference	Application	Objective	Modeling approach[a]	Solution method	Demand
Haase and Müller (2015)	Preventive healthcare facility location	Maximize preventive healthcare program participation	MNL, MILP, derive lower bound	CPLEX	Variable
Arvan et al. (2015)	Human blood supply chain network	Minimize total cost, minimize times that blood products remain in the network	Bi-objective model, MILP	ε-Constraint method, CPLEX	Deterministic fixed

[a] LTCFLP: long-term care facility location problem

In contrast to immediate medical support, there are also studies regarding preventive healthcare. In this case, clients choose whether to participate in preventive care programs or not. To maximize the total participation in these programs, Zhang et al. (2012) investigate the impact of clients' choice behavior on the preventive care facility network design and the resulting level of participation. They present two alternative models: the probabilistic-choice model and the so-called optimal-choice model. Solving large instances (with CPLEX) can take days. Enhancing the model of Haase (2009), Haase and Müller (2015) show that an alternative formulation of the presented problem can be useful to solve problems considerably faster with commercial solvers. An approach to derive a lower bound to the problem is also presented to accelerate computation time. In the following, we present an extension of this model, which includes variables in patients' utility functions (Krohn et al. 2018).

The locations of client nodes (demand points), the number of eligible patients per node, candidate locations for preventive healthcare facilities and a set of feasible facility modes are given. Different modes represent waiting time for an appointment and quality of care. The problem is to determine the locations and modes of established facilities in a way that maximizes the target population's expected participation in the preventive healthcare program. We integrate quality and waiting time into a deterministic mixed-integer linear problem via discretization of the clients' utility function and consider each combination of a facility's location and its mode as a separate choice alternative, e.g., a single facility with two possible modes results in two alternatives within the client's choice set. The two virtual facilities cannot be established simultaneously, because only exactly one mode is assigned to the facility. Hence, in the solution for this example, only one alternative (the facility located in a specific mode) remains in addition to the no-choice alternative, which is always present. Our approach makes use of the MNL's IIA property (Haase 2009; Aros-Vera et al. 2013; Haase and Müller 2015): The basic idea is to provide in advance calculated choice probabilities as input parameters and to take advantage of their constant ratios.

Defining the sets

$\mathscr{I}$ set of demand nodes,
$\mathscr{J}$ set of candidate facility location nodes $\mathscr{J} \subseteq \mathscr{I}$,
$\mathscr{M}$ set of modes in which a facility can be established (quality of care and waiting time for an appointment), (might also contain capacity levels), and

the parameters

g_i number of clients in node i that are eligible to require health service,
p_{ijm} MNL choice probability of clients in i to access service at a facility located at j being in mode m given that (j, m) is the only facility established, i.e. the choice set consists of the two alternatives $\{(j, m); \text{no}\}$, which results in $p_{ijm} = \frac{e^{v_{ijm}}}{e^{v_{i,\text{no}}} + e^{v_{ijm}}}$ where v_{ijm} is the deterministic utility of clients in i going to a facility located at j being in mode m and $v_{i,\text{no}}$ is the deterministic utility

for demand node i of not attending any facility ("no-choice" or "opt-out" alternative)

$\underline{l}_m$ lower threshold for mode m measured in number of clients

$\bar{l}_m$ upper threshold for mode m measured in number of clients,

p total number of available facilities, and

the variables

X_{ijm} choice probability of clients in i to access service at a facility located at j being in mode m,

Z_i cumulative choice probability of clients in i to refuse to access any facility ("no-choice"),

Y_{jm} = 1 if location j is specified to offer healthcare service in mode m (0, otherwise),

we formulate the healthcare facility location problems as follows:

$$\text{Maximize} \quad F = \sum_{i \in \mathscr{I}} \sum_{j \in \mathscr{J}} \sum_{m \in \mathscr{M}} g_i X_{ijm} \tag{26.15}$$

subject to

$$Z_i + \sum_{j \in \mathscr{J}} \sum_{m \in \mathscr{M}} X_{ijm} \leq 1 \qquad \forall\, i \in \mathscr{I} \tag{26.16}$$

$$X_{ijm} \leq p_{ijm} Y_{jm} \qquad \forall\, i \in \mathscr{I};\, j \in \mathscr{J};\, m \in \mathscr{M} \tag{26.17}$$

$$X_{ijm} \leq \frac{p_{ijm}}{1 - p_{ijm}} Z_i \qquad \forall\, i \in \mathscr{I};\, j \in \mathscr{J};\, m \in \mathscr{M} \tag{26.18}$$

$$\sum_{i \in \mathscr{I}} g_i X_{ijm} \geq \underline{l}_m Y_{jm} \qquad \forall\, j \in \mathscr{J};\, m \in \mathscr{M} \tag{26.19}$$

$$\sum_{i \in \mathscr{I}} g_i X_{ijm} \leq \bar{l}_m Y_{jm} \qquad \forall\, j \in \mathscr{J};\, m \in \mathscr{M} \tag{26.20}$$

$$\sum_{m \in \mathscr{M}} Y_{jm} \leq 1 \qquad \forall\, j \in \mathscr{J} \tag{26.21}$$

$$\sum_{j \in \mathscr{J}} \sum_{m \in \mathscr{M}} Y_{jm} = p \tag{26.22}$$

$$X_{ijm} \geq 0 \qquad \forall\, i \in \mathscr{I};\, j \in \mathscr{J};\, m \in \mathscr{M} \tag{26.23}$$

$$Z_i > 0 \qquad \forall\, i \in \mathscr{I} \tag{26.24}$$

$$Y_{jm} \in \{0; 1\} \qquad \forall\, j \in \mathscr{J};\, m \in \mathscr{M} \tag{26.25}$$

The objective function (26.15) maximizes the expected participation (measured as the number of patients that are expected to access preventive healthcare service). Equations (26.16)–(26.18) in combination with the objective function (26.15) are a linear reformulation of the MNL choice probabilities. Equation (26.16) ensure that a demand node i's final choice probabilities to go to service facilities as well as non-attendance sum up to at most 1. The case where the sum is less than one can be interpreted as rejecting patients at certain facilities. This formulation guarantees feasible solutions if mismatches between mode thresholds and mode demand exist. As an alternative, we might consider a finer mode structure with much more mode levels instead of only a few coarse ones to avoid infeasibility. This is also a possibility to approximate continuous waiting times.

The linking constraints (26.17) allow choice probabilities for a facility to be greater than 0 only if the facility is established. Using p_{ijm} yields a tighter upper bound by the LP-relaxation than just using $X_{ijm} \leq Y_{jm}$ and tighter bounds for X_{ijm} (Haase and Müller 2015), because p_{ijm} is distinctly smaller than 1. Equation (26.18) ensure that the pre-calculated constant substitution ratios between the choice probabilities for any two alternatives are obeyed. They are derived from $\frac{X_{ijm}}{Z_i} = \frac{p_{ijm}}{1-p_{ijm}}$. However, $X_{ijm} \neq p_{ijm}$ and $Z_i \neq (1 - p_{ijm})$ (unless j is the only established facility).

The correct mode in which a facility is established is selected by (26.19) (lower mode interval threshold) and (26.20) (upper threshold). If a certain facility j is established in mode m, $\sum_{i \in \mathcal{I}} g_i X_{ijm}$ has to be between the lower and the upper mode thresholds.

Equation (26.21) ensure that a facility can either only be established in exactly one mode or not at all. Equation (26.22) provides that p facilities are established. We might use a budget constraint instead, with a parameter denoting fixed establishing costs per facility and mode on the left-hand side and replacing the number of desired facilities p with a budget.

26.6 School Location

School networks are expanded or consolidated to meet expected student demand. Müller et al. (2009), Müller (2008) and Delmelle et al. (2014) introduce multi-period capacitated models for school network planning. Müller et al. (2009) consider free school choice and substitution effects between school locations (Müller et al. 2012) whereby school choice probabilities are determined by a mixed multinomial logit model considering scenarios of opened schools. While minimizing total costs, one scenario is selected for each period. Assuming that the students attend the nearest school, the approach of Delmelle et al. (2014) minimizes student travel costs and has the flexibility to modify the maximum capacity of each school, to integrate the

minimum facility age closure, and to reflect the uncertainty of demand projections. Considering free school choice, capacity constraints, a budget, and simulated utility values, Haase and Müller (2013) maximize all students' expected utility. To reduce inefficiencies in school facility location such as travel times, Castillo-López and López-Ospina (2015) present a model of location and modification of school capacity, with the objective to maximize utility (minimize operating costs, minimize travel times, maximize average amount of enrolled students per school, minimize number of schools with multi-grade classes). The process of school choice is modeled by including time and income constraints, and the decisions made by other students (segregation).

Now we discuss a school location model that can be used by private school organizations that want to enter a market or to expand their network. Without loss of generality, we assume that there is one private school provider that competes with public schools. Given already existing own and competing public schools, our objective is to find the optimal location for the establishment of new additional private schools to maximize our market share (number of first-year students that apply for our private schools). We propose to utilize the simulation-based approach introduced in Haase and Müller (2013). We generate a spatial representative (location, numbers) sample of first-year students. We simulate their utility values for all schools by applying a random utility model (e.g., multinomial logit model or mixed-logit model). A student chooses a private school if we establish at least one private school with a utility value larger than the utility values for the public schools.

Defining the sets

$\mathscr{I}$ set of simulated first-year students (spatial representative sample),
$\mathscr{J}$ set of candidate private schools, and
$\mathscr{J}_i$ set of candidate private schools of first-year student i, i.e., for student i, the simulated utility value of school $j \in J_i$ is larger than the largest utility value of all public schools,

the parameters

n number of expected first-year students, and
r number of private schools to be established, and

the variables

X_i = 1 if simulated first-year student i chooses a private school (0, otherwise), and
Y_j = 1 if a private school is to be established at location j (0, otherwise),

we define the following mathematical model:

$$\text{Maximize} \quad F = \frac{n}{|\mathscr{I}|} \sum_{i \in \mathscr{I}} X_i \tag{26.26}$$

subject to

$$X_i \leq \sum_{j \in \mathscr{J}_i} Y_j \qquad \forall\, i \in \mathscr{I} \tag{26.27}$$

$$\sum_{j \in \mathscr{J}} Y_j \leq r \tag{26.28}$$

$$Y_j \in \{0, 1\} \qquad \forall\, j \in \mathscr{J} \tag{26.29}$$

$$X_i \in [0, 1] \qquad \forall\, i \in \mathscr{I} \mid \mathscr{J}_i \neq \emptyset \tag{26.30}$$

The objective function (26.26) maximizes the expected number of all students applying for a private school. Equation (26.27) satisfies that student i selects a private school if at least one of her preferred candidate private schools is available. Equation (26.28) limits the number of private schools to be established. Haase et al. (2018) show that instances with large sample sizes can be solved by this (equivalent) approach within reasonable time.

26.7 Summary

We briefly discussed recent developments in the literature on the public sector facility location planning (2010–2018). They show that a remarkable part of the applications aim at satisfying the needs of the population, minimizing social costs, or ensuring equity. The current focus of the literature particularly lies on topics such as emergency/disaster management and healthcare facility location as well as on transport-related topics like the location of electric vehicle charging stations or bike sharing systems design (Table 26.3). The evolution of approaches enables practitioners to include more and more relevant planning decision factors to build more realistic models. Especially the consideration of stochastic demand modeled with state of the art methods based on behavioral theory is a promising extension of existing facility location proposals.

Table 26.3 Summary of references in public facility location planning 2010–2018

Application area	Number of papers per year									
	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
General	4	1	2	1				2		10
Hub location			1	2			1			4
Bike sharing		1	2	1		1				5
Bus network design			3							3
Charging and refueling stations		1			2	2	4	7		16
Waste management					1	1	1	2		5
Emergency shelter location	1	1			1	1	1	3		8
Disaster management			2			1				3
Emergency medical services	1		3	2			1	2	1	10
Healthcare facility location			2	1		3	2		1	9
School location				1	1	1				3
Other applications		1						2		3
Total[a]	6	5	15	8	5	10	10	18	2	79

[a] Not all references listed here are discussed in the text

References

Aros-Vera F, Marianov V, Mitchell JE (2013) *p*-Hub approach for the optimal park-and-ride facility location problem. Eur J Oper Res 226(2):277–285

Arslan O, Karaşan OE (2016) A benders decomposition approach for the charging station location problem with plug-in hybrid electric vehicles. Transp Res B: Methodol 93:670–695

Arvan M, Tavakkoli-Moghaddam R, Abdollahi M (2015) Designing a bi-objective and multi-product supply chain network for the supply of blood. Uncertain Supply Chain Manag 3(1):57–68

Basu R, Jana A, Bardhan R (2018) A health care facility allocation model for expanding cities in developing nations: strategizing urban health policy implementation. Appl Spat Anal Policy 11(1):21–36

Castillo-López I, López-Ospina HA (2015) School location and capacity modification considering the existence of externalities in students school choice. Comput Ind Eng 80:284–294

Ciaffi F, Cipriani E, Petrelli M (2012) Feeder bus network design problem: a new metaheuristic procedure and real size applications. Proc Soc Behav Sci 54:798–807. Proceedings of EWGT2012 - 15th Meeting of the EURO Working Group on Transportation, September 2012, Paris

Cipriani E, Gori S, Petrelli M (2012) Transit network design: a procedure and an application to a large urban area. Transp Res C: Emerging Technol 20(1):3–14. Special issue on Optimization in Public Transport+ISTT2011

Delmelle EM, Li S, Murray AT (2012) Identifying bus stop redundancy: a gis-based spatial optimization approach. Comput Environ Urban Syst 36(5):445–455

Delmelle EM, Thill J-C, Peeters D, Thomas I (2014) A multi-period capacitated school location problem with modular equipment and closest assignment considerations. J Geogr Syst 16(3):263–286
Djenić A, Marić M, Stanimirović Z, Stanojević P (2017) A variable neighbourhood search method for solving the long-term care facility location problem. IMA J Manag Math 28(2):321–338
Frade I, Ribeiro A (2015) Bike-sharing stations: a maximal covering location approach. Transp Res A: Policy and Pract 82:216–227
Frade I, Ribeiro A, Gonçalves G, Antunes AP (2011) Optimal location of charging stations for electric vehicles in a neighborhood in Lisbon, Portugal. Transp Res Rec 2252(1):91–98
Ghamami M, Nie YM, Zockaie A (2016a) Planning charging infrastructure for plug-in electric vehicles in city centers. Int J Sustain Transp 10(4):343–353
Ghamami M, Zockaie A, Nie YM (2016b) A general corridor model for designing plug-in electric vehicle charging infrastructure to support intercity travel. Transp Res C: Emerg Technol 68:389–402
Giménez-Gaydou DA, Ribeiro ASN, Gutiérrez J, Antunes AP (2016) Optimal location of battery electric vehicle charging stations in urban areas: a new approach. Int J Sustain Transp 10(5):393–405
Guo Z, Deride J, Fan Y (2016) Infrastructure planning for fast charging stations in a competitive market. Transp Res C: Emerg Technol 68:215–227
Haase K (2009) Discrete location planning. Technical Report WP-09-07, Institute of Transport and Logistics Studies, University of Sydney
Haase K, Müller S (2013) Management of school locations allowing for free school choice. Omega 41(5):847–855
Haase K, Müller S (2014) A comparison of linear reformulations for multinomial logit choice probabilities in facility location models. Eur J Oper Res 232(3):689–691
Haase K, Müller S (2015) Insights into clients' choice in preventive health care facility location planning. OR Spectr 37(1):273–291
Haase K, Müller S, Krohn R, Hensher D (2018) School location planning with flexible substitution patterns. Technical report, Working Paper
Hosseini M, MirHassani SA (2015) Refueling-station location problem under uncertainty. Transp Res E: Log Transp Rev 84:101–116
Jeong I-J (2017) An optimal approach for a set covering version of the refueling-station location problem and its application to a diffusion model. Int J Sustain Transp 11(2):86–97
Kim D.-G., Kim Y.-D. (2013) A Lagrangian heuristic algorithm for a public healthcare facility location problem Ann Oper Res 206(1):221–240
Klier MJ, Haase K (2015) Urban public transit network optimization with flexible demand. Or Spectr 37(1):195–215
Krohn R, Müller S, Haase K (2018) Preventive healthcare facility location planning with quality-conscious clients. Working paper
Lin J-R, Yang T-H (2011) Strategic design of public bicycle sharing systems with service level constraints. Transp Res E: Log Transp Rev 47(2):284–294
Lin J-R, Yang T-H, Chang Y-C (2013) A hub location inventory model for bicycle sharing system design: formulation and solution Comput Ind Eng 65(1):77–86
Ljubić I, Moreno E (2018) Outer approximation and submodular cuts for maximum capture facility location problems with random utilities. Eur J Oper Res 266(1):46–56
Marianov V, Serra D (2002) Location problems in the public sector. In Drezner Z, Hamacher HW (eds) Facility location: applications and theory, pp 119–150
Mestre AM, Oliveira MD, Barbosa-Póvoa AP (2015) Location-allocation approaches for hospital network planning under uncertainty. Eur J Oper Res 240(3):791–806
Miralinaghi M, Keskin BB, Lou Y, Roshandeh AM (2017a) Capacitated refueling station location problem with traffic deviations over multiple time periods. Netw Spat Econ 17(1):129–151
Miralinaghi M, Lou Y, Keskin BB, Zarrinmehr A, Shabanpour R (2017b) Refueling station location problem with traffic deviation considering route choice and demand uncertainty. Int J Hydrog Energy 42(5):3335–3351

Mozafar MR, Moradi MH, Amini MH (2017) A simultaneous approach for optimal allocation of renewable energy sources and electric vehicle charging stations in smart grids based on improved GA-PSO algorithm. Sustain Cities Soc 32:627–637

Müller S (2008) Dynamic school network planning in urban areas, Urban and regional planning, vol 5. LIT Verlag, Berlin

Müller S, Haase K (2014) Customer segmentation in retail facility location planning. Bus Res 7(2):235–261

Müller S, He S (2018) Impediments of electric vehicle adoption – evidence from a discrete choice experiment in Beijing, China. Working paper

Müller S, Haase K, Kless S (2009) A multiperiod school location planning approach with free school choice. Environ Plan A: Econ Space 41(12):2929–2945

Müller S, Haase K, Seidel F (2012) Exposing unobserved spatial similarity: evidence from German school choice data. Geogr Anal 44:65–86

Nickel S, Reuter-Oppermann M, Saldanha-da-Gama F (2016) Ambulance location under stochastic demand: a sampling approach. Oper Res Health Care 8:24–32

Riemann R, Wang DZW, Busch F (2015) Optimal location of wireless charging facilities for electric vehicles: flow-capturing location model with stochastic user equilibrium. Transp Res C: Emerg Technol 58:1–12

Romero JP, Ibeas A, Moura JL, Benavente J, Alonso B (2012) A simulation-optimization approach to design efficient systems of bike-sharing. Proc Soc Behav Sci 54:646–655

Sayarshad H, Tavassoli S, Zhao F (2012) A multi-periodic optimization formulation for bike planning and bike utilization. Appl Math Model 36(10):4944–4951

Shariff SSR, Moin NH, Omar M (2012) Location allocation modeling for healthcare facility planning in Malaysia. Comput Ind Eng 62(4):1000–1010

Tu W, Li Q, Fang Z, Shaw S-l, Zhou B, Chang X (2016) Optimizing the locations of electric taxi charging stations: a spatial-temporal demand coverage approach. Transp Res C: Emerg Technol 65:172–189

Xiang Y, Zhang Y (2017) Optimal location of charging station of electric bus in battery replacement mode. In Zeng X, Xie X, Sun J, Ma L, Chen Y (eds) International Symposium for Intelligent Transportation and Smart City (ITASC) 2017 Proceedings, pp 113–125. Springer, Singapore

Yang J, Dong J, Hu L (2017) A data-driven optimization-based approach for siting and sizing of electric taxi charging stations. Transp Res C: Emerg Technol 77:462–477

Zhang Y, Berman O, Verter V (2012) The impact of client choice on preventive healthcare facility network design. OR Spectr 34(2):349–370

Zhang W, Cao K, Liu S, Huang B (2016) A multi-objective optimization approach for health-care facility location-allocation problems in highly developed cities such as Hong Kong. Comput Environ Urban Syst 59:220–230

Zheng H, He X, Li Y, Peeta S (2017) Traffic equilibrium and charging facility locations for electric vehicles. Netw Spatial Econ 17(2):435–457

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, Canada Research Chair in Distribution Management. He has been Editor of Transportation Science, Computers & Operations Research and INFOR. He has authored or coauthored 19 books, as well as more than 550 scientific articles on 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 four occasions. He has been a member of the Royal Society of Canada since 1998, Fellow of INFORMS since 2005, and foreign member of the National Academy of Engineering (United States) since 2019. 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. In 2018, he became a member of the Order of Canada.

G. Laporte et al. (eds.), *Location Science*,
https://doi.org/10.1007/978-3-030-32177-2

Stefan Nickel is a full professor at the Karlsruhe Institute of Technology - KIT (Germany) and one of the directors of the Institute of Operations Research. He obtained his Ph.D. in mathematics at the Technical University of Kaiserslautern (Germany) in 1995. From 2014 to 2016, he was the dean of the Department of Economics and Management at the KIT. He was also member of the scientific advisory board as well as of the management board of the Fraunhofer Institute for Applied Mathematics (ITWM) in Kaiserslautern from 2004 to 2016. Since 2011, he additionally holds the positions of one of the directors of the Karlsruhe Service Research Institute (KSRI) and of the FZI Research Center for Information Technology. From 2006 to 2015, he was editor-in-chief of Computers & Operations Research. Since 2016, he has been editor-in-chief of Operations Research for Health Care. He has coordinated the Health Care working group within the German OR society (GOR) and has been the president of GOR from 2013 to 2014. Moreover, he was coordinator of the EURO working group on locational analysis. Since 2019, Stefan Nickel serves as VP IFORS in the EURO executive committee and is member of the AC of IFORS. He has authored or co-authored five books as well as more than 130 scientific articles in his research areas including location analysis, supply chain management, health care logistics, and online optimization. He has been awarded the EURO prize for the best EJOR review paper (2012) and the Elsevier prize for the EJOR top cited article 2007–2011. In addition, he conducted several industry projects with well-known companies such as Bosch, BASF, Lufthansa, Miele, or SAP.

Francisco Saldanha da Gama is professor of Operations Research at the Department of Statistics and Operations Research at the Faculty of Science, University of Lisbon, where he received his Ph.D. in 2002. He has extensively published papers in scientific international journals mostly in the areas of location analysis, 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 for the EJOR top cited article 2007–2011 (2012), both with 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. Currently, he is editor-in-chief of Computers & Operations Research. His research interests include stochastic mixed integer optimization, location theory, and project scheduling.