International Series in
Operations Research & Management Science

H.A. Eiselt
Vladimir Marianov *Editors*

Foundations of Location Analysis

Springer

Foundations of Location Analysis

International Series in Operations Research & Management Science

Volume 155

H. A. Eiselt • Vladimir Marianov
Editors

Foundations of Location Analysis

Springer

Editors
Prof. Dr. H. A. Eiselt
University of New Brunswick
Faculty of Business Administration
Fredericton, New Brunswick
Canada
haeiselt@unb.ca

Prof. Dr. Vladimir Marianov
Pontificia Universidad Católica de Chile
Depto. Ingenieria Electrica
Av. Vicuna Mackenna 4860
Macul, Santiago
Chile
marianov@ing.puc.cl

ISSN 0884-8289
ISBN 978-1-4419-7571-3 e-ISBN 978-1-4419-7572-0
DOI 10.1007/978-1-4419-7572-0
Springer New York Dordrecht Heidelberg London

Cover design: eStudioCalamar S.L.

Printed on acid-free paper

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

Writing intellectual history is like trying to nail jelly to the wall.

William Hesseltine

Preface

This book is the final result of a number of incidents that occurred to us while discussing location issues with colleagues at conferences. Frequently we attended presentations, in which the authors quoted the well-known references that helped to make the discipline what it is today. Upon further inquiry, though, it turned out that some of these colleagues had never actually read the original papers. We then discussed among ourselves which contributions could be credited with shaping the field. And, lo and behold, we found that we, too, had neglected to read some of the papers that form the foundation of our science. Whether it was laziness or other things that got in the way, it had become clear that something had to be done.

Our first thought was to collect the original contributions (once we could agree on what they were) and reprint them. When discussing this possibility with a publisher, we immediately ran into a roadblock in the form of copyright. While this appeared to have stopped our enthusiastic effort dead in its tracks, we kept on collecting and reading what we considered original contributions.

This went on until we met Camille Price, who suggested that, rather than reprinting the original contributions, we should invite some of the leaders in our field and ask them to describe the original contribution, explain and interpret it, and comment on the impact that it had to the field. This was, of course, an excellent idea, and the response by our colleagues to our pertinent requests was equally enthusiastic. What you hold in your hands is the result of this effort.

In other words, the purpose of this book is to provide easy access to the main contributions to location theory. The book is organized as follows. The introductory chapter provides an overview of some of the many facets of location analysis. This is followed by contributions in the main three fields of inquiry: minisum, minimax, and covering problems. The next chapters are part of an ever-growing list of nonstandard location models: models including competitive components, those that locate undesirable facilities, those with probabilistic features, and those that allow interactions between facilities. The following chapters discuss solution techniques: after a discussion of exact and heuristic techniques, we devote an entire chapter to Weiszfeld's method, and another to Lagrangean techniques. The last chapters of this book deal with the spheres of influence that the facilities generate and that attract

customers to them, and the last chapter delves back into the origins of location science, when geographers discussed central places.

Since the book is written by different individuals, the style and notations differ. Initial attempts to unify the notation were nipped in the bud. It is apparent that some chapters will be accessible to the laymen, others require substantial mathematical knowledge. However, all are written by competent individuals, who have made a major effort to not only popularize the original work, but also to assess its impact on the field and its implications for theory and practice.

With great sadness we have to report the untimely death of one of our contributors, Professor Roberto Galvão. His chapter is certainly one of the highlights of this book. In order to publish his contribution, it was required to have the usual consent form signed by a relative of his. Alas, none was to be found. Since we are certain that it would have been Professor Galvão's will to see his work in print, Professor Marianov now formally appears as coauthor (and, as such, being able to sign the necessary form), while the chapter was and remains entirely that of Roberto.

Last, but certainly not least, it is our great pleasure to thank all individuals who have contributed to this work and helped to make it reality. First and foremost, there are the contributors to this volume, who have devoted their time and talents to the cause of making the original contributions in our field accessible to those interested in the area. Then, of course, there is Camille Price, without whose suggestions and encouragement this book would never have seen the light of day. Thanks also go to Professor Hillier for his patience, and to Mr. Amboy for his timely help with the preparation of a camera-ready copy of the manuscript.

H. A. Eiselt
Vladimir Marianov

Contents

Contributors

Oded Berman Rotman School of Management, University of Toronto,
105 St. George Street, Toronto, ON M5S 3E6, Canada
e-mail: berman@rotman.utoronto.ca

Joy Bhadury Bryan School of Business and Economics,
University of North Carolina – Greensboro, Greensboro, NC 27402-6170, USA
e-mail: joy_bhadury@uncg.edu

Jack Brimberg Department of Mathematics and Computer Science,
Royal Military College of Canada, Kingston, ON K7K 7B4, Canada
e-mail: jack.brimberg@rmc.ca

Mark L. Burkey School of Business and Economics,
North Carolina A & T State University, Greensboro, NC, USA
e-mail: burkeym@ncat.edu

Abdullah Dasci School of Administrative Studies, York University,
4700 Keele Street, Toronto, ON M3J 1P3, Canada
e-mail: dasci@yorku.ca

Zvi Drezner Steven G. Mihaylo College of Business and Economics,
California State University-Fullerton, Fullerton, CA 92834, USA
e-mail: zdrezner@exchange.fullerton.edu

H. A. Eiselt Faculty of Business Administration, University of New Brunswick,
Fredericton, NB E3B 5A3, Canada
e-mail: haeiselt@unb.ca

Kathrin Fischer Institute for Operations Research and Information Systems
(ORIS), Hamburg University of Technology, Schwarzenbergstr. 95,
21073 Hamburg, Germany
e-mail: kathrin.fischer@tu-harburg.de

Roberto D. Galvão COPPE, Federal University of Rio de Janeiro,
Brazil, deceased

John M. Hodgson Department of Earth and Atmospheric Sciences, The University of Alberta, Calgary, AB T6G 2R3, Canada
e-mail: john.hodgson@ualberta.ca

Lawrence Joseph School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ 85287-0104, USA
e-mail: mikekuby@asu.edu

Bahar Y. Kara Department of Industrial Engineering, Bilkent University, Ankara, Turkey
e-mail: bkara@bilkent.edu.tr

Rex K. Kincaid Department of Mathematics, The College of William and Mary, Williamsburg, VA 23187-8795, USA
e-mail: rrkinc@math.wm.edu

Dmitry Krass Rotman School of Management, University of Toronto, 105 St. George Street, Toronto, ON M5S 3E6, Canada
e-mail: krass@rotman.utoronto.ca

Michael Kuby School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ 85287-0104, USA
e-mail: Lawrence.joseph@asu.edu

Timothy J. Lowe Tippie College of Business, University of Iowa, Iowa City, IA 52242-1994, USA
e-mail: timothy-lowe@uiowa.edu

Vladimir Marianov Department of Electrical Engineering, Pontificia Universidad Católica de Chile, Santiago, Chile
e-mail: marianov@ing.puc.cl

Emanuel Melachrinoudis Department of Mechanical and Industrial Engineering, Northeastern University, 360 Huntington Avenue, Boston, MA 02115, USA
e-mail: emelas@coe.neu.edu

Frank Plastria Department of Mathematics, Operational Research, Statistics and Information Systems for Management, MOSI, Vrije Universiteit Brussel, Pleinlaan 2, B1050 Brussels, Belgium
e-mail: frank.plastria@vub.ac.be

Daniel Serra Department of Economics and Business, Pompeu Fabra University, Barcelona, Spain
e-mail: daniel.serra@upf.edu

Lawrence V. Snyder Department of Industrial and Systems Engineering, Lehigh University, 200 West Packer Ave., Mohler Lab, Bethlehem, PA 18015, USA
e-mail: larry.snyder@lehigh.edu

Mehmet R. Taner Department of Industrial Engineering, Bilkent University, Ankara, Turkey
e-mail: mrtaner@bilkent.edu.tr

Barbaros Ç. Tansel Department of Industrial Engineering, Bilkent University, 6800 Bilkent, Ankara, Turkey
e-mail: barbaros@bilkent.edu.tr

Vedat Verter Desautels Faculty of Management, McGill University, Montreal, Quebec, Canada
e-mail: vedat.verter@mcgill.ca

Jiamin Wang College of Management, Long Island University, 720 Northern Blvd., Brookville, NY 11548, USA
e-mail: jiamin.wang@liu.edu

Richard E. Wendell Joseph M. Katz Graduate School of Business, University of Pittsburgh, Pittsburgh, PA 15260, USA
e-mail: wendell@katz.pitt.edu

Hassan Younies School of Management, New York Institute of Technology, Abu Dhabi, United Arab Emirates
e-mail: hassan.younies@gmail.com

Part I
Introduction

Chapter 1
Pioneering Developments in Location Analysis

H. A. Eiselt and Vladimir Marianov

1.1 Location Problems: Its Problem Statement, Its Components, and Applications

A mathematician would probably define a location problem as solving the following question: "given some metric space and a set of known points, determine a number of additional points so as to optimize a function of the distance between new and existing points." A geographer's explanation might be that "given some region in which some market places or communities are known, the task is to determine the sites of a number of centers that serve the market places or communities." Students of business administration will want to determine "the location of plants and market catchment areas in the presence of potential customers," while computer scientists (or, more specifically, analysts in computational geometry) may want to determine "the minimum number of equal geometrical shapes that are required to cover a certain area, and the positions of their centroids."

All these views have in common the basic components of a location problem: a space, in which a distance measure is defined, a set of given points, and candidate locations for a fixed or variable number of new points. We refer to the known points as "customers" or "demands," and the new points to be located as "facilities."

As far as the space is concerned in which customers are and facilities are to be located, we distinguish between a subset of the d-dimensional real space (most prominently, the two-dimensional plane) and networks. Each of these two categories has two subcategories: one, in which the location of the facilities is continuous, and the other, in which it is discrete. These two subcategories will determine the toolkit needed by the researcher to solve problems. In discrete problems, the decision is

H. A. Eiselt (✉)
Faculty of Business Administration, University of New Brunswick,
Fredericton, NB E3B 5A3, 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

H. A. Eiselt, V. Marianov (eds.), *Foundations of Location Analysis,*
International Series in Operations Research & Management Science 155,
DOI 10.1007/978-1-4419-7572-0_1,

whether or not to locate a facility at that spot, thus modeling the decision with a binary variable is obvious. The result is a (mixed-)integer linear programming problem. On the other hand, continuous problems will have continuous variables associated with them, indicating the coordinates of the facilities that are to be located. Since functions of the distance are typically nonlinear, a nonlinear optimization problem will follow in this case. With the tools for the solution of the different types of problems being so different, it is little surprise that most researchers have decided to either work with continuous or with discrete models. Both types of models are represented in this book.

The space commonly corresponds to a geographical region. An obvious choice for representing such a region is the two-dimensional plane. In some cases, a one-dimensional space is a used to simplify the analysis of complicated problems, as in the case that includes competition between firms. Location problems can also be defined over non-geographical spaces. For example, an issue space in which potential voters for a presidential election are located according to their positions in relation to the issues. The problem would be to optimally locate a candidate so to maximize his vote count in the presence of competing candidates, assuming that voters will vote for the candidate whose position is closest to them. Or consider a skill space in which a number of tasks (demands) have known locations, each location representing the combination of skills required to successfully accomplish the task. The location problem would be to locate company employees in such a way that all tasks are performed by sufficiently skilled employees, and no employee is overloaded with work.

In this book we deal with location problems, which are defined as models, in which the facilities and demands are very small as compared to the space they are located in. Examples of facility location problems are finding the location of an assembly plant within a country; or selecting the site of a school in one of several candidate points in a city. In these cases, the facilities can be considered dimensionless points. A generalization would be the location of lines in a plane, where the lines represent a new road. In some cases, location problems can be cast as geometrical problems: an example is finding the smallest circle that contains a known set of points on a plane, which is equivalent to locating a facility (the center of the circle) in such a way that the largest distance (the radius of the circle) between any customer (the given points) and the facility (the center of the circle) is minimized. In contrast to location problems, layout problems feature facilities, whose size is significant in comparison to the space the facility is to be located in. Examples of layout problems include the siting of a drilling machine in a workshop, the location of tables in a fast-food restaurant, or the location of supply rooms in a hospital. Other things being equal, layout problems are more difficult than location problems with similar features. One reason is that layout problems must consider the shape of the facility to be located, which is unnecessary in location models, where facilities are just points. Layout models are not discussed in this book.

Location problems can be formulated to answer to several different questions. Not only can the location of the facilities be unknown, but also the number of facilities and their capacities. Furthermore, when there is more than one facility, the solution of the location problem also requires finding the assignment of demands to

facilities, called the "allocation" problem. Examples of this problem are the assignment of individual customers to a particular warehouse from which grocery is to be delivered by a phone-order company, or the dispatching of specific ambulances to emergency sites. In these cases, the owner or operator of the facilities does the allocation, but there are other cases in which users decide what facility to patronize, such as the decision which theater to patronize or at what fast-food store to have lunch.

For the location problem to make sense, customers and facilities must be related by distance. Again, this distance can be measured either over a geographical region or over any kind of space. Typical goals are to minimize the distance between facilities and their assigned customers (minisum problems) or maximize the amount of demand (number of customers) that is within a previously specified distance from their assigned facilities. When all the demand is to be served, a goal could be to minimize the number of facilities needed for all the customers to be within a distance from their facility. If competing facilities are involved and distance is an issue for customers, the planner's objective will be to locate his facilities closer than competitors' facilities to as many customers as possible.

Proximity to facilities is, however, not always desirable. When locating a new landfill, most people will object if the facility is to be located close to their homes—the usual NIMBY (not in my back yard) argument applies, even though they usually understand that locating the landfill too far away will cost them, as they are, directly or indirectly, charged for the collection and transportation costs of the solid waste. In this case, the problem could be formulated with conflicting objectives: the landfill should be not too far from the area it serves, but most of the population should be as far as possible from it. These objectives have been described as "push" and "pull".

Besides distances between facilities and demands, there are other factors that can be relevant when seeking good locations. Availability of services and skilled technicians at the candidate locations, land cost, existence of competitors and regional taxes are some examples. While such factors are maybe as important as or possibly even more important than proximity, most standard location problems do not consider them. This stresses the importance of considering the output of the standard location models as an input for the decision maker, who can then include any of the features that were ignored by the mathematical model.

Most of the applications of location models involve, decisions on the strategic level. As such, the decisions tend to be long-term, which implies that many of the data used in the decision-making process, will be quite uncertain. While there usually are few changes concerning distances, future demand tends to be highly uncertain. This problem is exacerbated by the high cost of locating and relocating a facility, meaning that once a location is chosen, it can only be changed at great expense. This argument implies that probabilistic and/or robust models are most likely to result in facility locations that are acceptable to decision makers.

Applications of location models are found in many different fields. Some of these applications are fairly straightforward, such as the locations of trucking terminals, blood banks, ambulances, motels and solid waste transfer points. Others are nontraditional and not at all obvious. Good examples are the location of measuring

points for glaucoma detection, the location of new employees in skill space, the location of advertisements in media and many others.

1.2 A Short History of the Early Developments in Location Analysis

In the early seventeenth century, a question was posed about how to solve the following puzzle: "Given three points in a plane, find a fourth point such that the sum of its distances to the three given points is as small as possible." This question is credited to Fermat (1601–1665), who used to tease his contemporary mathematicians with tricky problems. The earliest (geometrical) solution is probably due to Torricelli (1598–1647), Fermat's pupil and the discoverer of the barometer, although due to the many discoveries and rediscoveries of the problem and its solution, there are different opinions about the true origin of the problem and who solved it first. Because of the many scientists involved in the process, the problem has been called the Fermat problem, the Fermat-Torricelli problem, the Steiner problem, the Steiner-Weber problem, the Weber problem, and many variation thereof. The inclusion of Weber's name follows his generalization of the problem by assigning different weights to the known points, so transforming the mathematical puzzle into an industrial problem, in which a plant is to be located (the unknown point) so to minimize transportation costs from suppliers and to customers (the known points) requiring different amounts of products (the weights). Weber's name stuck, even though the formulation of the model in Weber's book is in an appendix written by Pick. Even if the first formal occurrence of a location problem were due to Fermat or one of his contemporaries, location analysis must be much older than that. For centuries before, people may have wondered in which cave to live, where to build houses, villages, churches, and other "facilities." And they solved these problems using some sort of heuristic method.

Location problems frequently require solving an associated allocation or assignment problem: if locations for more than one facility are known, which facility will serve what customers? A step towards the solution of the allocation problem was taken very early in the seventeenth century. Descartes imagined the universe as a set of vortices around each star—the "heavens"—and illustrated his theory with a drawing that made an informal use of what later would become known as Voronoi polygons and Voronoi diagrams. This concept was subsequently used by Dirichlet in 1850, extended to higher dimensions by Voronoi and rediscovered by Thiessen, both in the early twentieth century. While Thiessen's application involves an improved estimate of precipitation averages, Voronoi diagrams are generally useful tools when consumers have to be assigned to their closest facilities.

Between the 1600s and the 1800s, there was no registered activity related to location problems other than a puzzle in the Ladies Diary or Woman's Almanack in 1755 as reported in a book edited by Drezner and Hamacher in 2002. In 1826, the geographer von Thünen developed a theory concerning the allocation of crops on the land that surrounds a town. His point was that the agricultural activity should

be organized around the town according to transportation costs and value of the products. This theory results in crops cultivated in concentric circles about the town. Clearly, a location issue, although it could also be considered a land layout problem, as the areas or rings to be located around the central place are sizeable in comparison to the space considered.

In 1857, Sylvester asked another location-related question: "It is required to find the least circle which shall contain a given system of points in a plane." This question was later answered by Sylvester himself in 1860 and by Chrystal in 1885. Nowadays, we would call this a one-center problem in the plane.

As in many other fields, the pace of discoveries increased dramatically in the twentieth century. In the late 1920s, the mathematical statistician-turned economist Hotelling wrote a seminal paper on competitive location models that spawned a rich diversity of models that are still under active discussion today. Also during that time, Reilly introduced gravity models into the fray as a way customers gravitate to facilities The 1930s saw contributions by Christaller, who founded central place theory, and Weiszfeld, who developed his famed algorithm that solved Weber problems with an arbitrary number of customers.

Later important contributions included those by Lösch and the regional scientists Isard and Alonso. The birth of modern quantitative location theory occurred in the mid-1960s, when Hakimi wrote his path-breaking analysis of a location model on networks, which, in today's parlance, is a *p*-median problem. Following Hakimi's papers, ReVelle, Church, Drezner, Berman, and many others have made important contributions to location science. Some of those have also contributed to this book.

1.3 Some Standard Location Problems

Facility location problems can take a variety of forms, depending on the particular context, objectives and constraints. It is common to classify location problems in minisum, covering and minimax problems. Most of the numerous remaining facility location problems can be seen as combinations or modified versions of these key problems. We present here prototypes of these basic classes, as well as an assortment of other models that we consider of importance, either from a theoretical viewpoint, or because of their practical interest. This list of models is also a guide for understanding the following chapters and a way to introduce some standard notation, since in most of the chapters of this book, the original notation of the reviewed papers is preserved.

We first state a base formulation, which is used throughout this section to present some of the different problems and models for location on networks. This formulation applies only when demands lie at the nodes and facilities are to be located only at nodes of the network. The problem can be formulated as follows.

$$\text{Min } \sigma \left(\sum_{i,j} c_{ij} y_{ij} + \sum_{j} f_j x_j + gz \right) \tag{1.1}$$

$$\text{s.t.} \sum_{j \in N_i} y_{ij} = 1 \quad i = 1, 2, \ldots n \tag{1.2}$$

$$y_{ij} \leq x_j \quad i = 1, 2, \ldots n, \; j = 1, 2, \ldots m. \tag{1.3}$$

$$y_{ij}, x_j \in \{0, 1\}, \quad z \in \mathbb{R}^+ \quad i = 1, 2, \ldots n, \; j = 1, 2, \ldots m, \tag{1.4}$$

where the variables are:

x_j: a location variable that equals 1, if a facility is located at node j, and 0 otherwise,
y_{ij}: an allocation variable that equals 1, if customer i is assigned to a facility at j, and 0 otherwise, and
z: a continuous variable that takes the value of a maximum or a minimum distance, depending on the problem being solved.

The subscripts are

i, j: a subscript that indicates customers and potential facility sites, respectively,
N_i: is the set of nodes. Its definition depends on the problem, and n and m denote the total number of customers and potential facility locations, respectively.

Finally, the parameters are

σ: a parameter that takes the value 1 or (−1), depending on whether the objective is minimized or maximized, and
c_{ij}, f_j, g: parameters that depend on the problem being solved.

The objective (1.1) of the base problem optimizes linear combinations of the location and allocation variables and the distance variable. Depending on the problem, these linear combinations represent maximum, minimum or average distances, investment or transportation and manufacturing costs. The set of constraints (1.2) state that each customer's demand must be satisfied, and that all of it is satisfied from one facility. In other words, that each customer or demand node is assigned to or served by exactly one facility. The set of constraints (1.3) allows demand node i to be assigned or allocated to a facility at j only if there is an open facility in that location i.e., $x_j = 1$. Constraints (1.4) define the nature of all variables as continuous or binary. Note that these constraints define a feasible set of solutions of the problem of locating a (yet) undetermined number of facilities, choosing their sites from a number of known candidate locations, and finding the right assignment of demand nodes or customers to these facilities.

1.3.1 Minisum Problems

Minisum problems owe their name to the fact that a sum of facility-customer distances is minimized. These problems include single-facility, multiple-facility and

weighted versions of the Weber problem; the *simple plant location problem SPLP* and the *p*-median problem *PMP*.

1.3.1.1 Minisum Problems on the Plane: The Weber Problem

The Weber problem consists in finding the coordinates (x, y) of a single facility X that minimizes the sum of its distances to n known customer or demand points with coordinates (a_i, b_i) and weights w_i, $i = 1, 2, \ldots n$, all of them lying on the same plane. This is an unconstrained nonlinear optimization problem, and its general formulation is

$$\text{Min } z(X) = \sum_{i=1}^{n} w_i d(X, P_i),$$

where $d(X, P_i)$ is the distance between the facility and demand i. If Euclidean distances are used, this distance is $d(X, P_i) = \sqrt{(x - a_i)^2 + (y - b_i)^2}$.

An application of this problem is locating a single warehouse (the facility) that supplies different amounts (weights) of products to a number of dealers (the demands), in such a way that the total transportation cost is minimized, assuming that this cost depends on the distance and amount of transported product.

The solution of this problem is easily found using differential calculus. However, the coordinates of the optimal facility location turn out to be a function of the distance $d(X, P_i)$, which is unknown. The practical solution of this problem, found by Weiszfeld, involves the use of an iterative algorithm that takes as an initial solution the point that minimizes the sum of the squares of the distances.

A natural generalization of the Weber problem is its multiple-facility version. As more than one facility is to be located, an allocation problem must be solved together with the location problem, which makes the problem far more difficult. In the simplest case, the assumption is made that each demand is assigned to its closest facility. This problem was described and a heuristic proposed by Cooper in 1963.

1.3.1.2 Minisum Problems on Networks: Plant Location and Median Problems

The Simple Plant Location Problem

The simple plant location problem, sometimes also referred to as the *uncapacitated facility location problem* seeks minimizing production, transportation and site-related investment costs. The model assumes that the costs of opening the facilities depend on their location, and that the investment budget is not a constraint. Then, the number of facilities to be opened is left for the model to decide. Its integer programming formulation is the base model (1.1)–(1.4), with $\sigma = +1$, $c_{ij} = (e_j + \tilde{c}_{ij} d_{ij}) w_i$,

where e_j denotes the production cost per unit, $\tilde{c}_{ij}$ are the transportation costs from node i to j per unit, d_{ij} symbolizes the shortest distance between i and j, the parameters w_i denote the quantity of the product that must be shipped to customer i, the parameters f_j are the fixed costs of opening a facility at node j, and g equals zero. The set N_i in constraints (1.2) contains all possible candidates to location of facilities. The objective to be minimized, representing the total sum of variable and fixed costs, is now:

$$\text{Min} \sum_{i,j} c_{ij} y_{ij} + \sum_{j} f_j x_j .$$

Since the costs of locating facilities are included in the objective, the solution of the model prescribes the number of facilities to be located, depending on the tradeoff between transportation and fixed costs.

There is also a "weak" formulation of this problem that uses an aggregated version of constraints (1.3) for each facility location j, $\sum_j y_{ij} \leq n x_j$.

There is an important body of literature focusing on this problem, because of its versatility and practical interest. Due to its difficulty, many solution methods have been proposed. In Chap. 2 of this book, Verter describes this problem, and reviews in detail the classical dual ascent method by Erlenkotter, as well as the heuristic proposed by Kuehn and Hamburger.

The 1-Median Problem

The network equivalent of the single-facility Weber problem on the plane is the 1-median problem, whose formulation is exactly the same as that of the single-facility Weber problem. However, in the 1-median problem the demands occurs at the nodes of a network. Hakimi proved that there is always an optimal location of the facility at a node of the network. As a consequence, finding the solution reduces to searching the nodes of the network, and the problem can be stated as follows: find a node v^* such that for all nodes v_k, $k = 1, 2, \ldots n$,

$$\sum_{i=1}^{n} w_i d(v^*, v_i) \leq \sum_{i=1}^{n} w_i d(v_k, v_i).$$

This problem has been solved for locations on tree networks in the 1960s by Hua-Lo Keng, who was looking for the optimal location of a threshing floor for wheat fields, and rediscovered by Goldman in 1971.

The p-Median Problem

The p-median problem is a p-facility generalization of the 1-median problem, in which it is assumed that the decision maker knows how many facilities are to be lo-

cated, and that the cost of locating the facilities is the same no matter where they are located. As in the continuous case, the allocation problem must be solved together with the location problem, i.e., not only the location of the p facilities needs to be found, but also, for each demand, the facility to which this demand is assigned. Let X_p be a set of p points $x_1, x_2, \ldots x_p$. Define distance of a node v_i to its closest point in X_p as

$$d(v_i, X_p) = \min\left\{d(v_i, x_1), d(v_i, x_2), ..., d(v_i, x_p)\right\}.$$

Then the set X_p^* is a "p-median" of the network, if for every X_p on the network,

$$\sum_{i=1}^{n} w_i d\left(v_i, X_p^*\right) \leq \sum_{i=1}^{n} w_i d\left(v_i, X_p\right)$$

i.e. X_p^* is the set of p points on the graph such that, if facilities are open at these points, the total weighted distance between the demands and their closest facility would be minimized. For this problem, Hakimi proved that there is always a solution set containing only nodes of the network.

ReVelle and Swain formulated this problem as a linear integer programming problem. Their model uses the same equations (1.1)–(1.4) of the basic formulation above, and an additional constraint that requires exactly p facilities to be located:

$$\sum_j x_j = p.$$

In their formulation of the p-median, the objective minimizes the sum of the weighted facility-demand distances, i.e., in the objective (1.1) of the basic formulation, the parameters become $\sigma = +1$, $c_{ij} = w_i d_{ij}$, where w_i is the weight associated to each demand node (demand volume or number of customers) and d_{ij} is the shortest distance between nodes i and j, measured along routes on the network, $f_j = 0$, and $g = 0$. The objective (1.1) becomes

$$\text{Min} \sum_{i,j} w_i d_{ij} y_{ij}.$$

Also, the set N_i in constraint (1.2) contains all possible candidates to location of facilities. Note that this formulation assumes that each demand is assigned to its closest facility.

In Chap. 3 of this book, Marianov and Serra review the classic contributions of Hakimi, made in 1964 and 1965, including the definitions of the absolute median and the absolute p-medians of a network. Also, they review the paper in which the first integer programming formulation of the p-median problem was proposed by ReVelle and Swain in 1970.

1.3.2 Minimax Problems

The idea behind minimax problems is to minimize the longest distance between a customer and its assigned facility, expecting that a maximum equity or justice will be achieved. As before, these problems have been defined both on a plane and on a network.

1.3.2.1 Continuous Minimax Problems: Single and Multiple-Facility Center Problems on the Plane

The oldest known single-facility minimax problem was formulated by Sylvester in 1857 in a one-sentence reference. The problem consists in finding the smallest circle containing a set of given points. Finding the solution of this problem is equivalent to finding the location of the center of a circle that encloses all given points, in such a way that its radius is minimized. If a facility is located at the center of the circle, and the points represent demands, the maximum distance between the facility and a demand will be minimized. This problem is now known as the 1-center on the plane or the continuous 1-center. Several geometrical solution methods have been proposed to solve this problem, the first one due to the same Sylvester and rediscovered by Chrystal in 1885.

A natural extension of the 1-center problem is the *p*-center on the plane. Its formulation is as follows. Let X_p be a set of p points $x_1, x_2, \ldots x_p$ on the plane. Define distance of a customer c_i to its closest point in X_p as

$$d\left(c_i, X_p\right) = \min\left\{d(c_i, x_1), d(c_i, x_2), ..., d(c_i, x_p)\right\}.$$

Then the set X_p^* is a "p-center", if for every X_p on the plane,

$$\max d\left(c_i, X_p^*\right) = \max d\left(c_i, X_p\right).$$

In other words, X_p^* is the set of p points on the plane such that, if these points were facilities, the maximum distance between a demand and its closest facility would be minimized. Minimax problems on the plane, as well as their solution methods are analyzed in Chap. 4 of this book. In that chapter, Drezner describes the early research done in the nineteenth century, shows how geometrical methods were used to solve the continuous minimax problems, describes the algorithmic contributions by Elzinga and Hearn in the 1970s, and compares both approaches through new unpublished computational experience.

1.3.2.2 The 1-Center and p-Center Problems on Networks

The 1-center on a network has been first formulated and solved by Hakimi in 1964 and the solution method improved by Goldman in 1970 and in 1972. Hakimi for-

mulates the problem as follows: Let v_i be a node, w_i its potential demand, and x any point on the network. Define $d(v_i, x)$ as the minimum distance between v_i and x. Find a point x^* on the network such that for x,

$$\max_i w_i d(x^*, v_i) \leq \max_i w_i d(x, v_i).$$

This point is called the "absolute center" of the network. Hakimi proved that his result for minisum problems does not apply to the 1-center problem. In other words, the solution can be on an arc of the network, which is even true if all demands are equal.

Later, the 1-center was extended to a multiple-facility problem: the p-center: let X_p be a set of p points $x_1, x_2, \ldots x_p$ anywhere on the network. Define the distance between a node v_i and its closest point in X_p as

$$d(v_i, X_p) = \min\{d(v_i, x_1), d(v_i, x_2), ..., d(v_i, x_p)\}.$$

Then the set X_p^* is a "p-center", if for every X_p on the network,

$$\max_i d(v_i, X_p) \leq \max_i d(v_i, X_p).$$

In other words, X_p^* is the set of p points on the network such that, if these points were facilities, the maximum distance between a demand and its closest facility would be minimized.

If facility locations are restricted to the nodes of the network, the p-center problem can be formulated as an integer programming problem using the base formulation (1.1)–(1.4), and setting $\sigma = +1$, $c_{ij} = 0$, $f_j = 0$, and $g = 1$. Variable z is the maximum distance between a demand and its assigned facility. Two extra constraints must be added:

$$d_{ij} y_{ij} \leq z \quad \text{for} \quad i = 1, 2, \ldots n,\ j = 1, 2, \ldots m, \text{ and} \tag{1.5}$$

$$\sum_j x_j = p. \tag{1.6}$$

Constraint (1.5) requires z, the variable to be minimized, to take the value of the maximum distance between any customer and its assigned (closest) facility. Since constraint (1.2) forces assignment of each customer to exactly one facility, constraint (1.5) can take the aggregated form

$$\sum_j d_{ij} y_{ij} \leq z \quad \text{for} \quad i = 1, 2, \ldots n.$$

Constraint (1.6) sets the number of facilities to be located to p. Pioneering developments on center problems on a network, specifically the works by Hakimi in 1964, Goldman in 1972, and Minieka in 1970 are the subject of Chap. 5 of this

book by Tansel. The chapter first defines the single facility *p*-center problem and its multiple-facility version and then describes their properties. Later, the three classic contributions are reviewed: Hakimi's definition of the absolute center, Goldman's algorithm for locating the center and Minieka's method for finding the *p*-centers on a network when locations are restricted to nodes.

1.3.3 Covering Problems

In minisum and minimax location models, the distance between each and every customer and his closest facility is explicitly considered. While minisum problems minimize the average distance between a customer and its closest facility, minimax problems minimize the longest customer-facility distance. In contrast, covering models do not explicitly include customer—facility distances in the model and these distances only matter if they exceed a preset value $\bar{D}$. A common example appears in emergency medical services: service is considered adequate if an ambulance reaches the site of the emergency in no more than, say 8 minutes. In the food business, a pizza parlor could offer free pizzas when the delivery time exceeds 30 minutes. In synthesis, the concept of coverage implies that a customer can and will be adequately served ("covered") when a facility is located within a preset distance or travel time. Distance is no longer included in the objective function, but appears as a constraint.

1.3.3.1 The Location Set Covering Problem

The *location set covering problem LSCP* was first introduced by Hakimi, and formulated as an integer programming problem by Toregas et al. in 1971. Its goal is to find the minimum number of facilities and their locations, so that all customers are covered, meaning that all customers are no farther than a preset distance $\bar{D}$ of their closest facility. Although the original formulation uses only location variables x_j, the problem can be reformulated using the base model (1.1)–(1.4). To reproduce the *LSCP* with this base model, we set $\sigma = +1$, $c_{ij} = 0 \; \forall \; i, j, f_j = 1 \; \forall \; j$, $g = 0$, and the set N_i in constraint (1.2) contains all candidates to location of facilities that are within distance $\bar{D}$, i.e., $N_i = \{j : d_{ij} \bar{D}\}$. The objective is now

$$\text{Min} \sum_j x_j .$$

The formulation provided by Toregas et al. is simpler. In that formulation, only location variables are used and the allocation problem is not solved, i.e. the model does not prescribe which facility provides the service, which can be provided by any of the facilities within range. In consequence, it does not matter if a customer has more than one facility within covering distance.

Although at first sight there is no apparent relation between minimax and covering problems, Minieka found that the solution of the node constrained *p*-center can be found by solving a sequence of location set covering problems.

1.3.3.2 The Maximal Covering Location Problem (MCLP)

If the minimum number of facilities needed to cover all the demand cannot be achieved, because of a limited budget or any other constraints on the total number of facilities, the *maximal covering location problem* of Church and ReVelle solves the problem of maximizing covered demand when there are a limited number of facilities to be sited. In terms of the base model (1.1)–(1.4) $\sigma = -1$, $c_{ij} = h_i$ if $d_{ij} \leq \bar{D}$ and 0 otherwise, where h_i is the demand volume of customer *i,* the fixed costs $f_j = 0$ $\forall\, j$, the value of $g = 0$, and the set N_i in constraint (1.2) contains all candidates to location of facilities. The objective is now

$$\text{Max} \sum_{i,j} c_{ij} y_{ij}.$$

This objective maximizes the demand-weighted number of customers that have at least one facility located within distance $\bar{D}$.

Church and ReVelle's model was formulated using only location variables and covering variables. Again, the allocation or assignment problem is not solved by this model.

Seminal contributions on covering problems are reviewed in Chap. 6 of this book. In that chapter, Snyder thoroughly reviews the contributions by Hakimi in 1965, in which the location set covering problem is defined, the work by Toregas et al. in 1971 with the first integer programming formulation of the same problem, and the contribution by Church and ReVelle in 1974, which defines the maximum covering location problem and formulates it as an integer programming problem. The author of the chapter then adds some insight into these problems by running new computational experiments on variants of the original problems, including tradeoff between coverage and number of facilities; Lagrangean approach as a solution method; inclusion of budget constraints; and relationship to the *p*-median problem.

1.3.4 Other Relevant Location Problems

Minimax, minisum and covering models are considered the key formulations from which the remaining location problems are descendants. There are some nonstandard location problems and models that have attracted considerable attention from researchers. We briefly review some of these in the remainder of this section.

1.3.4.1 Competitive and Conditional Location Problems

None of the key location problems considers the existence of other facilities, either previously located or to be located in the same region by a competing firm. If this is the case, a whole new range of problems can be posed, in which the relationship between facilities plays an important role: competitive and conditional location problems.

In order to analyze the locational pattern of the competing firms, assumptions need to be made about consumers' behavior. A typical assumption is that customers prefer closest facilities over more distant ones, and facilities at which the goods or service can be obtained at a lower price over the more expensive facilities.

Competitive Location: Equilibrium Problems

Suppose a market over which the demand is distributed, either continuously or at discrete points. Two firms are entering the market, each one with one facility that can move at any time without incurring costs. There are two fundamental questions. The first asks that given that customers patronize the facility at which they can satisfy their demand for the lowest full price (i.e., mill price plus transportation costs), is there an equilibrium? In other words, is there a situation, in which both facilities have decided on a location and a price, so that neither facility has an incentive to unilaterally change its price and/or its location? The second question is then, what the location and price situation in such an equilibrium is, given that it exists.

These questions—whose answer is now known as Nash equilibrium—were first answered for two competing facilities by Hotelling in 1929, who studied the problem in its simplest form: a linear market (i.e., a simple line segment) with uniformly distributed customers. Each of the two firms locates a single branch, and both firms use mill pricing.

In Chap. 7, Eiselt discusses the details and applications of the problem of finding a locational equilibrium on a linear market; he provides a detailed review of Hotelling's seminal paper, assesses its impact and describes subsequent work on the subject, including the article by d'Aspremont et al. Their paper was published no less than fifty years after Hotelling's findings, and it invalidated some of Hotelling's results.

Conditional Location Problems

When facilities can move freely without costs, the search for Nash equilibrium is the natural approach. However, this is not always the case. Most facilities stay where they were first located, as relocation costs are significant. If a firm is planning to enter a market with *immobile* facilities, there are two different aspects to be considered. The first aspect concerns firms that intend to locate in the presence of

already existing competitors. this is a conditional location problems, i.e., the location of facilities, *given* that other facilities are already located in the region. This is the *follower's problem*, who will make his location choice based on whatever profit or revenue-maximizing objective he has.

On the other hand, the already existing firms are the leaders in the location game. The *leader's problem* is to locate under the consideration that after he has decided where to locate his branch(es) under the consideration that a follower will locate his own competing branches later. This means that the leader will have to take a cautious approach and guard against future entrants onto the market.

Solutions of this location game are usually referred to as von Stackelberg solutions, named after the German economist of that name. The main feature of the game is its sequential nature. In Chap. 8, Younies and Eiselt describe sequential location problems and explain the von Stackelberg concept of leader and follower firms. this is followed by a review of the classic application of these concepts in location analysis, made by Prescott and Visscher (1977), where they extend Hotelling results on a line to the sequential case.

In Chap. 9, Dasci discusses the conditional problems in a plane and a network, and offers a thoroughful revision of the seminal works by Drezner in 1982 and Hakimi in 1983. In these works, the follower's problem (resulting in what is now known as an $(r|X_p)$-medianoid) and the leader's problem (dubbed an $(r|p)$-centroid) are defined on the plane and a network, respectively, given that the leader locates p facilities and the follower locates r facilities.

1.3.4.2 Location of Undesirable or Semi-obnoxious Facilities

In many everyday situations, customers want to be as close as possible to the facilities that are to be located. It is the case of grocery stores and shopping centers, schools, primary care centers, and similar facilities. However, the situation changes when considering facilities such as prisons, landfills, or power plants. Facilities of this nature are considered undesirable by most people. In the early days, they were referred to as noxious and obnoxious. Some people nowadays refer to these facilities as semi-obnoxious, because although nobody wants them in the neighborhood, they cannot be too far away, because their operation becomes too expensive. However, while realistic solutions should incorporate desirable and undesirable features in a model, undesirable facility location models push facilities as far away from customers as possible. More realistic models will have to balance some of the usual criteria (minisum, minimax, and covering) with either distance maximization (maximin or maxisum), coverage minimization, or a minimization of the costs of compensating all the affected population.

An example of a maximin formulation of undesirable facility location problems is, again, derived from the base model (1.1)–(1.4), by setting the variable z to be the minimum distance between a demand and its assigned facility, $c_{ij} = 0, f_j = 0$, and $g = 1$, and maximizing the objective, which becomes simply z. The following two constraints must be added:

$$d_{ij} y_{ij} \geq z \quad i = 1, 2, \ldots n, \; j = 1, 2, \ldots m. \tag{1.7}$$

$$\sum_j x_j = p. \tag{1.8}$$

Constraint (1.7) forces z to take the value of the smallest distance between a customer and his closest facility.

Early contributions on undesirable facility location are the subject of Chap. 10 by Melachrinoudis. The classical contribution of Church and Garfinkel 1978 (maximizing the sum of the distances from the closest facility) is reviewed, as well as the string of papers by Shamos published in 1975, Shamos and Hoey's work from 1975, Dasarathy and White's paper from 1980, and Drezner and Wesolowski's contribution from 1980, all of them dealing with the maximin problem, i.e., maximizing the shortest distance between a customer and its closest facility.

1.3.4.3 Probabilistic or Stochastic Location Problems

The pool of location problems would not be complete without the consideration of uncertainty. Future demand, the time of appearance of emergency calls, service times at the facilities, and travel times between demand nodes and facilities tend to be non-deterministic. Disruptions of facilities and connecting routes are also everyday random occurrences.

There is a large body of literature dealing with problems that include one or more of these sources of uncertainty. The design of emergency services is a popular application, as well as the location of immobile facilities that can become congested. For instance, service may be refused because the facility is busy, or lines are formed in which customers have to wait until they can be served. After the 9/11 events and the destruction of New Orleans by hurricane Kathrina, many researchers have turned their attention to the location of facilities so as to mitigate such catastrophic effects, or to the search for robust locations in the sense that disruptions do not interrupt their functioning.

Chapter 11 by Berman, Krass and Wang is devoted to the description of these problems. The original contribution by Frank from 1966 is reviewed in detail, and its impact analyzed. Generalizations of Frank's work are presented, and open problems are discussed.

1.3.4.4 Hub Location Problems

Hub location problems occur mainly in transportation and telecommunications networks. A hub is a transfer point at which either traffic from several origins is added up and forwarded to another hub, or disaggregated into several streams that are forwarded to their destination. Airlines use hub airports because they allow tak-

ing advantage of economies of scale, by using high capacity planes in high traffic routes, as opposed to using only small planes between every origin and destination city. Computer networks are frequently organized as a set of small hub-and-spoke networks, joined together by a backbone network.

Naturally, hub location problems have been formulated having in mind the same principles as the remaining location problems. Thus, there are *p*-hub location problems, median-hub location problems, uncapacitated hub location problems, *p*-hub-center location problems, and so on. The first formulation of a hub location problem by O'Kelly in 1982 was nonlinear. Later, different linearizations have been proposed. The base model (1.1)–(1.4) needs to be modified to represent an uncapacitated hub location problem (analogous to the uncapacitated facility location problem) as follows. Facilities are now hubs, and customers need to be assigned to hubs for both their incoming and outgoing traffic. We assume each customer is assigned to his closest hub. The traffic between two customers, h_{ij}, starts at a node i, first goes to the assigned hub k, continues from there to another hub m, and finally to the destination node j. The traffic into the opposite direction uses the same path. Both hubs could be the same, i.e., traffic could move through only one hub. We use constraints (1.2)–(1.4), but the objective is now:

$$\text{Min} \sum_i \sum_j h_{ij} \left(\sum_k c_{ik} y_{ik} + \alpha \sum_k \sum_m c_{km} y_{ik} y_{jm} + \sum_m c_{jm} y_{jm} \right) + \sum_j f_j x_j \tag{1.9}$$

where h_{ij} is the traffic between customers or demand nodes i and j, c_{ij} is the unit traffic cost on the arc (i, j), i.e. on the leg of the trip that goes between nodes i and j, and α is a factor <1, accounting for economies of scale. Note that the first term in parentheses in the objective function is the cost of sending traffic from the origin to the first hub. The second term in the cost of carrying traffic between hubs, where because of the traffic concentration, it is assumed that the cost is reduced by a factor $(1 - \alpha)$. The third term is the cost of sending the traffic from the second hub to the destination, and the last term is the cost of opening hubs. This objective is nonlinear, but different linear versions have been proposed in the literature.

Chapter 12 of this book, written by Kara and Taner, focuses on hub location problems. O'Kelly's seminal paper is reviewed and its contribution assessed.

1.4 An Outline of Classic Solution Methods

Most of the problems related to facility location are known to **NP**-hard. It is natural then to look for procedures that can find adequate optimal or approximate solutions to these problems. Many solution methods have been proposed in the literature. It is important to describe some of these methods, either because they have a broader applicability or because they led the way to the discovery of a whole line of solution procedures. This book presents some of the more important techniques.

1.4.1 Exact Techniques

The best known exact solution technique for location problems written as integer programming problems is branch-and-bound. The technique was first proposed by Land and Doig in 1960. The first two contributions that apply this technique to the location field are those by Gavett and Plyter in 1966, who solve the quadratic assignment problem, and Efroymson and Ray in 1966, who solved the uncapacitated facility location problem or, as it is frequently called today, the simple plant location problem. These two contributions are reviewed in Chap. 13 by Lowe and Wendell. The authors describe the principles behind the branch-and-bound technique, *viz.*, partitioning the set of solution into smaller subsets, and provide a detailed description of the two seminal papers. They also show other techniques to solve the quadratic assignment problem. The Efroymson and Ray approach to solving the *SPLP* is complemented by other branch-and-bound techniques discovered later. One section of their contribution is devoted to branching strategies.

When the problem has a special structure, as is the case when the space is a tree-shaped network, efficient techniques can be developed to find solutions to the location problems more efficiently. An example of these techniques is the method for finding the 1-median on a tree. In Chap. 14, Kincaid first defines some graph theory concepts, presents a notation for location problems, and reviews three classical contributions: the Harary and Norman article of 1953, where central points on a tree are defined, the paper by Hua Lo-Keng and others in 1962—although known at an earlier date in Chinese—where the first method for location of a 1-median on a tree (and on networks with some cycles) is proposed and proven to be optimal, and the paper by Goldman in 1971, who rediscovered Hua Lo-Keng's findings.

1.4.2 Heuristic Solution Methods

Although the maximum size of the problems that can be solved using exact methods has increased in time almost like Moore's law for semiconductors (that states that the number of transistors in an integrated chip doubles every two years), exact methods are still not capable of finding a solution in many realistic cases, in which the number of variables is too large. Heuristics are methods that find a good solution, although optimality is not guaranteed or even sought. Many of these heuristics are built on principles that have been proposed a long time ago, but are still valid. Some of these, for location both in the plane and on networks, were sketched in a single paper by Cooper in 1963, and extended later by other authors. For facility location problems in the plane, a giant step forward was the heuristic for the Weber problem, which was discovered by Weiszfeld in the 1930s.

The heuristics by Cooper, and their formalization by Maranzana and by Teitz and Bart, are the subject of Chap. 15 of this book, by Brimberg and Hodgson. In that chapter, the authors describe in detail the several methods proposed by Cooper

for the continuous minisum problem, as well as the rediscoveries of these methods and their application to the discrete problem, made by Maranzana and by Teitz and Bart. Then, they assess the impact these early discoveries had, their application to other problems and the current status of the heuristic methods for the continuous and discrete *p*-median problem.

In Chap. 16, Plastria first offers a short history of the Weber problem, a brief listing of different methods that have been proposed to solve it, and a description of its optimality conditions. As a way to introducing the Weiszfeld algorithm, the problem is presented of finding the point at which the weighted squares of the distances to the given points is minimum. Then, the author describes in detail the method found by Weiszfeld, together with its properties and proofs of convergence. After that, he presents a modern view of Weiszfeld's algorithm and the rediscoveries, additions and improvements made by the researchers that followed Weiszfeld. The author then focuses on applications of the algorithm in other areas of knowledge.

Finally, there are heuristics that not only find a solution, but tell the user how far could be this solution from the real optimum. One such heuristic is Lagrangean Relaxation. As Galvão outlines in Chap. 17, this technique was first used in the field of location by Bilde and Krarup in 1967. Later, Diehr, and also Marsten, applied this technique also to location problems. the author then offers a brief review of the technique, surveys the seminal papers by Bilde and Krarup, Diehr and Marsten, and provides an account of the works that followed them.

1.5 Customer Choice and Location Patterns

Most location models that allow customers to choose among a number of facilities assume that customers patronize the closest facility or the facility they can satisfy their demand from most cheaply. However, these simplifying assumptions are not always true in practice. In fact, customers at a demand point can patronize different facilities at different times; or different customers at the same demand point can have different patterns of behavior, when choosing a facility. This phenomenon is the subject of a number of studies, and the problems that include a consideration to it, have been called models with customer choice.

Some studies are oriented to describe in the best possible form the behavior of customers. Some others, focus on the locational patterns and their interaction with markets and customer concentration.

1.5.1 Gravity Models or Spheres of Influence

Joseph and Kuby address gravity modeling in Chap. 18. In other words, they study models that include choice rules other than simple proximity or price. In particular, their focus is on that customers are attracted to facilities according to laws whose

expression resembles that of Newton's laws of physics. The authors describe the contribution by Reilly in 1931, who introduced the so-called *law of retail gravitation* and validated this model with an empirical study. They also review the contribution by Huff in 1963, who proposes a method to determine trade areas for retail, later known as the Huff model. In the chapter, other techniques are also sketched, and applications to location-allocation models are described.

1.5.2 Voronoi Diagrams

As mentioned above, Voronoi diagrams are a tool that allows modelers to allocate customers to facilities based on proximity. In Chap. 19, Burkey et al. first review the original work by Thiessen written in 1911. They then formally define Voronoi diagrams and explore some of their properties and extensions. This is followed by a case study of the Triad Region in North Carolina. Here, a number of weighted and unweighted Voronoi diagrams are constructed and compared using concentration indices and statistical tests.

1.5.3 Central Places

In the nineteenth and the first half of twentieth centuries, some authors attempted to explain why and how economic activities are located in relation to the markets. These contributions by von Thünen in 1826, Christaller in 1933 and Lösch in 1940, began long strings of work on land use, central places theory and regional planning. Truly, these must be considered as major works in the facility location field. In Chap. 20, Fischer synthesizes these works and assesses their contribution to the field.

Acknowledgements This contribution was in part supported by the National Research Council of Canada under grant number #9160, and by Instituto Milenio "Complex Engineering Systems" under grants ICM P-05-004-F and CONICYT FBO16. This support is gratefully acknowledged.

Part II
Minisum Problems

Chapter 2
Uncapacitated and Capacitated Facility Location Problems

Vedat Verter

2.1 Introduction

The *uncapacitated facility location problem* (*UFLP*) involves locating an undetermined number of facilities to minimize the sum of the (annualized) fixed setup costs and the variable costs of serving the market demand from these facilities. *UFLP* is also known as the "simple" facility location problem *SFLP*, where both the alternative facility locations and the customer zones are considered discrete points on a plane or a road network. This assumes that the alternative sites have been predetermined and the demand in each customer zone is concentrated at the point representing that region. *UFLP* focuses on the production and distribution of a single commodity over a single time period (e.g., one year that is representative of the firm's long-run demand and cost structure), during which the demand is assumed to be known with certainty. The distinguishing feature of this basic discrete location problem, however, is the decision maker's ability to determine the size of each facility without any budgetary, technological, or physical restrictions. Krarup and Pruzan (1983) provided a comprehensive survey of the early literature on *UFLP*, including its solution properties. By demonstrating the relationships between *UFLP* and the set packing-covering-partitioning problems, they established its **NP**-completeness.

The seminal paper of Erlenkotter (1978), which is reviewed in Sect. 2.2 of this chapter, presents a dual-based algorithm for solving the *UFLP* that remains as one of the most efficient solution techniques for this problem. Prior to Erlenkotter (1978), the best-known approaches for solving the *UFLP* were the branch-and-bound algorithm developed by Efroymson and Ray (1966) and the implicit enumeration technique of Spielberg (1969). Efroymson and Ray (1966) use a compact formulation of *UFLP* to take advantage of the fact that its linear programming relaxation can be solved by inspection. Nonetheless, this linear programming relaxation does not provide tight lower bounds for *UFLP*; Efroymson and Ray's model is therefore

V. Verter (✉)
Desautels Faculty of Management, McGill University, Montreal, Quebec, Canada
e-mail: vedat.verter@mcgill.ca

H. A. Eiselt, V. Marianov (eds.), *Foundations of Location Analysis,*
International Series in Operations Research & Management Science 155,
DOI 10.1007/978-1-4419-7572-0_2, © Springer Science+Business Media, LLC 2011

known as the "weak formulation." Khumawala (1972) developed efficient branching and separation strategies for the branch-and-bound algorithm. Erlenkotter (1978), however, uses the "tight formulation" of *UFLP* that is known to often produce natural integer solutions. This property of the tight formulation was first highlighted by Schrage (1975) and was used effectively by Cornuejols et al. (1977). Here, it is important to credit the work of Bilde and Krarup (1977), which led to the development of a dual-based algorithm for *UFLP* that is quite similar to Erlenkotter's procedure.

In many cases, it is more realistic to incorporate the capacity limitations on the facilities to be established. This version of *UFLP* is called the *capacitated facility location problem* (*CFLP*). Section 2.3 reviews the contribution by Kuehn and Hamburger (1963). Their paper presents one of the earliest models and a heuristic procedure for the *CFLP*. Branch-and-bound procedures for this problem were developed by Akinc and Khumawala (1977) using linear programming relaxation, and by Nauss (1978) through Lagrangean relaxation. The cross-decomposition algorithm of Van Roy (1986) and the Lagrangean-based approach of Beasley (1988) are among the most effective techniques that were subsequently devised for solving the *CFLP*. The basic idea of Van Roy's algorithm is to obtain a *UFLP* structure by dualizing the capacity constraints. This Lagrangean relaxation provides values for the location and allocation variables given a set of multipliers. The location decisions are then used to fix the integer variables and solve the *CFLP* as a transportation problem, obtaining improved multiplier values. It is necessary, however, to solve an appropriately defined linear program at some of the iterations to update the multipliers.

The *UFLP* and *CFLP* constitute the basic discrete facility location formulations, and there is an abundance of papers based on their extensions by relaxing one or more of the underlying assumptions mentioned above. Section 2.4 presents an overview of the prevailing literature. Aikens (1985) presented a survey of the early work on discrete location models for distribution planning. He reviewed 23 models covering a wide range of problems from the single-commodity *UFLP* to the multi-commodity, capacitated, multi-echelon versions. Although the *UFLP* and *CFLP* formulations have been used for tackling a wide range of problems, the most common context for their use has been the production-distribution network (i.e., supply chain) design problem. In a supply chain that comprises suppliers, plants, distribution centers, warehouses and customers, these basic formulations are relevant for making location decisions involving two consecutive echelons. For example, notwithstanding the focus of a majority of the literature on warehouse location, the *UFLP* and *CFLP* formulations are equally relevant for choosing suppliers to satisfy the needs of a firm's plants (Gutierrez and Kouvelis 1995). The next two sections review two classical papers that form the basis of this chapter.

2.2 Erlenkotter 1978: A Dual-Based Procedure for the UFLP

Let I denote the set of m alternative facility locations, indexed by i, and J denote the set of n customer zones, indexed by j. The *UFLP* has two sets of decision variables:

x_{ij}: the fraction of customer zone j's demand satisfied by the facility at i, and
y_i: binary variables that assume a value of 1, if a facility is to be established at location i, and 0 otherwise.

Note that the demand data pertaining to each customer zone j is implicit in the definition of the facility-customer allocation variables x_{ij}. The cost data is represented by the following notation:

f_i: the (annualized) fixed cost of establishing a facility at location i, and
c_{ij}: the total capacity, production and distribution cost for supplying all of customer zone j's demand by the facility at i.

The variable costs c_{ij} are assumed to be linear functions of the quantities produced and shipped at each facility, thus ignoring any possible economies of scale in the variable costs. Erlenkotter (1978) presents the following formulation of *UFLP*:

$$\text{Max} \sum_i \sum_j c_{ij} x_{ij} + \sum_i f_i y_i \tag{2.1}$$

$$\text{s.t.} \sum_i x_{ij} = 1 \quad \text{for all } j \tag{2.2}$$

$$x_{ij} \leq y_i \quad \text{for all } i, j \tag{2.3}$$

$$x_{ij} \geq 0, \quad y_i \in \{0, 1\} \quad \text{for all } i, j.$$

The objective function (2.1) represents the total fixed and variable costs, whereas constraints (2.2) ensure that the demand at each customer zone is satisfied. Constraints (2.3) guarantee that customer demand can be produced and shipped only from the locations where a facility is established, i.e., if $y_i = 1$, and in such a case, the firm incurs the associated fixed costs. The weak formulation of *UFLP* uses a more compact formulation of these constraints by aggregating the constraints (2.3) into a single constraint for each facility location i:

$$\sum_j x_{ij} \leq n y_i \quad \text{for all } i.$$

In developing the solution approach, Erlenkotter (1978) utilizes a condensed dual formulation to the linear programming relaxation of *UFLP*. To this end, let v_j and w_{ij} represent the dual variables associated with constraints (2.2) and (2.3), respectively. By relaxing y_i as non-negative variables, the dual problem can be formulated as follows:

$$\text{Max} \sum_j v_j \tag{2.4}$$

$$\text{s.t.} \sum_j w_{ij} \leq f_i \quad \text{for all } i \tag{2.5}$$

$$v_j - w_{ij} \leq c_{ij} \quad \text{for all } i,\ j \tag{2.6}$$

$$w_{ij} \geq 0 \quad \text{for all } i,\ j$$

Note that the w_{ij} variables are not part of the dual objective, and hence can be safely fixed at the minimum levels permitted by the values of v_j. Erlenkotter assumes that $w_{ij}=\max\ \{0, v_j - c_{ij}\}$ and develops the condensed dual formulation below that has a single set of decision variables:

$$\text{Max} \sum_j v_j$$

$$\text{s.t.} \sum_j \max\{0, v_j - c_{ij}\} \leq f_i \quad \text{for all } i \tag{2.7}$$

The *dual ascent procedure* that constitutes the core of Erlenkotter's algorithm aims at increasing the values of v_j so as to maximize their sum. The idea is to use a quick and simple heuristic for solving the condensed dual rather than searching for an exact solution. To this end, the heuristic starts by setting the v_j values to the smallest c_{ij} for each customer zone j. At each iteration of the dual ascent procedure, the customer zones are processed one by one and the v_j value at each zone is raised to the next higher c_{ij} value, unless such an increase is constrained by (2.7). When the inequality (2.7) becomes binding during this process, the v_j value is increased to the highest level allowed by the constraint. The heuristic terminates when no further increase is possible for the v_j values.

To illustrate the dual ascent procedure, consider a *UFLP* instance with eight customer zones and five alternative facility sites, which was also used by Erlenkotter. Table 2.1 depicts the variable costs c_{ij} and fixed costs f_i for this problem instance. At the initialization, the v_j values are set at the lowest c_{ij} value at each column in Table 2.1. As a result, s_i, the slack of constraint (2.7), is equal to the fixed cost f_i at each location. The initialization step is denoted as *Iteration 0* in Tables 2.2 and 2.3, which depict the progress of the v_j and s_i values during the course of the algorithm.

The bolded entries in Table 2.2 indicate the v_j values blocked by (2.7) from further increase. Note that in iteration 1, all v_j values are raised to the next higher c_{ij} value (under column j in Table 2.1), except v_8. We would normally raise v_8 from 120

Table 2.1 Cost data for the illustrative example

i/*j*	Variable cost								Fixed cost
	1	2	3	4	5	6	7	8	
1	120	180	100	–	60	–	180	–	100
2	210	–	150	240	55	210	110	165	70
3	180	190	110	195	50	–	–	195	60
4	210	190	150	180	65	120	160	120	110
5	170	150	110	150	70	195	200	–	80

Table 2.2 The values of the dual variables v_j

Iter/v_j	1	2	3	4	5	6	7	8
0	120	150	100	150	50	120	110	120
1	170	180	110	180	55	195	160	**155**
2	180	190	**110**	**180**	60	**195**	**160**	**155**
3	**180**	**190**	**110**	**180**	**65**	**195**	**160**	**155**

to 165, but this would violate (2.7). Therefore, the value of v_8 is raised to 155 reducing the dual slack s_4 to zero, as indicated in Table 2.3 under *Iteration 1*.

At *Iteration 2*, the dual variables for customer zones 3, 4, 6 and 7 are blocked, and the heuristic terminates after *Iteration 3* when no further increase is possible. Table 2.3 indicates that the dual constraints for locations 4 and 5 are binding at the end of the dual ascent procedure.

It is helpful to analyze the complementary slackness conditions for the condensed dual and the linear programming relaxation at this point. The bolded terms in (8) and (9) indicate the optimal values of the primal and dual decision variables.

$$\boldsymbol{y_i}\left[f_i - \sum_j \max\{0, \boldsymbol{v_j} - c_{ij}\}\right] = 0 \quad \text{for all } i \tag{2.8}$$

$$[\boldsymbol{y_i} - \boldsymbol{x_{ij}}]\max\{0, \boldsymbol{v_j} - c_{ij}\} = 0 \quad \text{for all } i, j \tag{2.9}$$

The dual ascent produces a feasible solution v_j with at least one binding constraint (2.7). For each associated location i, the slack of the dual constraint is zero, and using (2.8) it is possible to set $y_i = 1$. Examining (2.9) for these open facilities, we hope that there is only one facility i with $c_{ij} \le v_j$ for all j, because in this case it is possible to set $x_{ij} = y_i = 1$ and obtain a primal integer solution that satisfies both complementary slackness conditions. It is likely, however, that the dual ascent procedure terminates with a solution where, among open facilities, there is more than one facility i with $c_{ij} \le v_j$ for some j. This would violate (2.9), since each customer zone must be served from the lowest-cost open facility. Therefore it is possible to set $x_{ij} = y_j = 1$ for only the smallest value of c_{ij}, and the primal integer solution is not optimal.

In the illustrative example, customer zones 1, 2, 3, and 4 are served from facility 5 and zones 5, 6, 7, and 8 are served from facility 4. A comparison of the v_j values at *Iteration 3* of Table 2.2 with the c_{ij} values in Table 2.1 reveals that there are no

Table 2.3 The values of the slack of (2.7)

s_i/iter	0	1	2	3
1	100	40	20	15
2	70	20	15	10
3	60	55	50	45
4	110	0	0	0
5	80	20	0	0

complementary slackness violations and the solution produced by the dual ascent procedure is optimal. Consider another instance with fixed costs f_i=(200, 200, 200, 400, 300) and the same variable costs. At the termination of the dual ascent procedure, $s_2=s_5=0$ and $v_6=285$ (the other v_j values are irrelevant here). Given that $c_{26}=210$ and $c_{56}=195$ (see Table 2.1), there is more than one c_{ij} with a smaller value than v_j, and hence (2.9) would be violated.

To close the duality gap in such cases, Erlenkotter first uses a *dual adjustment procedure*, and if this does not suffice, he resorts to a simple branch-and-bound. The dual adjustment procedure focuses on a customer zone *j* for which (2.9) is violated. Reducing the value of v_j can create slack for some of the binding dual constraints (2.7), which in turn can be used for increasing the value of other dual variables. As a result, the dual solution can be improved. Even if the dual solution remains the same, the associated primal integer solution would be altered because a different set of dual constraints would be binding after the adjustment. Continuing the above illustrative example, the value of v_6 is reduced to 210 in the adjustment procedure, creating slacks for three of the dual constraints (2.7) that are then used for improving the dual solution. The dual adjustment procedure processes each customer zone *j* associated with a complementary slackness violation and terminates when no further improvement to the dual solution is possible. If the duality gap persists, a standard branch-and-bound is utilized to identify the optimal solution. The solutions generated by the dual ascent and dual adjustment procedures serve as bounds during this final phase of the algorithm.

Erlenkotter solved *UFLPs* of up to 100 alternative facility sites and 100 customer zones, including the classical problem instances provided Kuehn and Hamburger (1963). In all but two of the instances, there was no duality gap at the end of the dual ascent and adjustment procedures and hence branch-and-bound was not necessary. Among the largest problem instances, two required branching and 21 nodal solutions were evaluated for the most challenging *UFLP*. Perhaps more importantly, the solution from the dual ascent procedure was within 1% of the optimal objective value in all reported instances. The quality of the lower bounds obtained from the condensed dual formulation, coupled with the ease of constructing primal integer solutions from a dual solution, underlies the efficiency of Erlenkotter's algorithm.

2.3 Kuehn and Hamburger (1963): A Heuristic Program for Locating Warehouses

Kuehn and Hamburger's classical paper presents, perhaps, the earliest heuristic solution approach for discrete facility location and describes in detail a set of twelve problem instances. Focusing on warehouse location, they highlight the potential advantages of these facilities due to (1) economies of scale in transportation costs between factories and warehouses, (2) economies of scope from combining products

from different factories into a single shipment in serving customer demand, and (3) improved delivery times by increased proximity to customer locations. In determining the locations for a set of capacitated warehouses, Kuehn and Hamburger trade off these potential cost savings associated with the new facilities with the costs of establishing and operating them.

They state the following three principles concerning the proposed heuristic:

1. most geographical regions are not promising sites for a regional warehouse, as locations with promise will be at or near concentrations of demand,
2. near optimum warehousing systems can be developed by locating warehouses one at a time, adding at each stage of the analysis that warehouse which produces the greatest cost savings for the entire system; and
3. only a small subset of all possible warehouse locations needs to be evaluated in detail at each stage of the analysis to determine the next warehouse site to be added.

In essence, Kuehn and Hamburger assume that the set of M alternative facility sites is a subset of the set of demand locations. They adopt a myopic approach as the basis of their heuristic, and confine the detailed evaluation at each iteration of the heuristic to a small subset of N location alternatives that they call the "buffer" (where $N<M$). The heuristic comprises a constructive phase ("the main program") and an improvement phase ("the bump and shift routine").

At the beginning of the constructive phase the buffer is initialized with the N sites, where serving the local demand with a local warehouse results in the highest cost savings. Then the N sites in the buffer are assessed one by one in terms of the system-wide cost savings that can be attained by opening a warehouse. The site that brings in the highest cost savings to the distribution network is assigned a warehouse, while the sites that do not offer any cost savings are eliminated from further consideration. The algorithm cycles between re-constructing the buffer from the remaining sites and the detailed evaluation step until all the sites are either eliminated or assigned a warehouse. The resulting solution is evaluated in the improvement phase to determine whether it is possible to attain cost savings by closing any of the open warehouses and/or by shifting each warehouse to another alternative site within its service region.

Kuehn and Hamburger propose 12 problem instances comprising combinations of three sets of factory locations and four levels of warehouse setup costs. The sample problems involve a single commodity and the transportation costs are assumed to be proportional to the railroad distances. The set of customer zones comprise 50 large cities across the United States, and 24 of these are also identified as alternative warehouse locations. The computational experiments were carried out with a buffer of 5 facilities. The Kuehn and Hamburger problem instances are available through the OR-Library at http://people.brunel.ac.uk/~mastjjb/jeb/info.html (developed and maintained by J. Beasley). These problems still constitute benchmark instances for comparing computational efficiencies of different algorithms for *UFLP* and *CFLP*.

2.4 Major Works that Followed

The classical *UFLP* and *CFLP* models have been extended in a number of ways by relaxing one or more of their underlying assumptions mentioned in Sect. 2.1. Here we provide an overview of the major works that extend the classical formulations by increasing the number of products, the number of facility echelons, and the number of time periods included in the model, as well as by more realistic representation of problem parameters through incorporation of possible scale and scope economies and uncertainties.

An immediate generalization of *UFLP* is the *multi-commodity* facility location problem that relaxes the single product assumption. Although Neebe and Khumawala (1981) and Karkazis and Boffey (1981) offered alternative formulations for this problem, both papers assumed that each facility deals with a single product. Klincewicz and Luss (1987) was the first paper that studied a multi-commodity facility location model without any restrictions on the number of products at each facility.

Another important extension involves increasing the number of echelons incorporated in the problem formulation. One of the earliest *multi-echelon* formulations is by Kaufman et al. (1977), which determined the locations of a set of facilities and a set of warehouses simultaneously. Tcha and Lee (1984) presented a model that could represent an arbitrary number of echelons. Both of these papers ignored the cost implications of possible interactions among the facilities at different echelons. Generalizing Erlenkotter's dual-based method, Gao and Robinson (1992) proposed an efficient dual-based branch-and-bound algorithm for the two-level facility location problem. Barros and Labbe (1994) presented a profit maximization version of the same problem and developed a branch-and-bound procedure based on Lagrangean relaxation as well as various heuristics.

Perhaps the most influential paper following the sketchy *CFLP* formulation in (the Appendix of) Kuehn and Hamburger (1963) was the contribution by Geoffrion and Graves (1974). Their model aimed at minimizing the total cost of transportation and warehousing over a distribution network comprising three echelons; factories, distribution centers (*DC*s), and customers. Given the existing plant and customer locations, Geoffrion and Graves (1974) devised a Benders decomposition approach for determining the optimal number and locations of the distribution centers to be established. They assumed a single-sourcing policy that requires serving each customer from a single *DC*. Their model contained both lower and upper bounds on *DC* throughput, which enabled modeling piecewise linear concave operation costs for the distribution centers. The differentiating feature of Geoffrion and Graves (1974) from earlier multi-echelon models was the way they modeled the flow variables. In earlier work, the flows between each pair of consecutive echelons were represented by a different set of decision variables, which required the use of flow conservation constraints at each facility. In contrast, Geoffrion and Graves (1974) used a single set of variables to represent the flows from the factories through the *DC*s to the customer zones. Although this leads to a considerable increase in the number of

decision variables, the resulting model is a tighter formulation of the problem that enables the development of efficient algorithms. Moon (1989) extended the model and solution procedure in order to incorporate possible economies of scale in *DC* throughput costs. To this end, he used general concave cost functions to represent the *DC* throughput costs. Pirkul and Jayaraman (1996) provided another extension that enables facility location decisions at both the *DC* and the plant echelons. However, they imposed limits on the number of *DC*s and plants that could be opened and relaxed the lower bound used by Geoffrion and Graves (1974) on *DC* throughput levels. In a subsequent paper, Jayaraman and Pirkul (2001) also incorporated supplier selection in a multi-commodity problem setting. Both papers used Lagrangean relaxation as a solution framework. Recently, Elhedhli and Goffin (2005) highlighted the efficiency of interior point techniques in solving multi-echelon formulations.

A number of researchers focused on relaxing the single period assumption of the *UFLP* and *CFLP*, and developed models and solutions for the *dynamic* facility location problem. The objective was to determine the spatial distribution of the facilities at each time period so as to minimize the total discounted costs for meeting the customer demand over time. The earliest work on this problem is by Van Roy and Erlenkotter (1982), who extended the dual-based algorithm of Erlenkotter to handle multiple time periods. Lim and Kim (1999) and Canel et al. (2001) proposed alternative methods for solving the problem with capacity restrictions at the facilities. Recently, Melo et al. (2005) presented a dynamic and multi-commodity formulation as an extension of the *CFLP* and investigated the possible use of the model as a framework for strategic supply chain planning.

Another stream of research to extend the classical *UFLP* and *CFLP* formulations focuses on improving the realism of the *cost representations* in these models. These efforts are motivated by the possible economies of scale and scope in the fixed and variable costs, as well as the potential cost implications of the interactions between a plant's location and the other structural decisions including capacity acquisition and technology selection. Soland (1974) is one of the earliest attempts to develop an extension of the *UFLP* that incorporates scale economies by representing the fixed facility costs as a concave function of facility size. Holmberg (1994) and Holmberg and Ling (1997) extended the *CFLP* by formulating the capacity acquisition costs as arbitrary piecewise linear functions. Verter and Dincer (1995) proposed a model where the capacity costs are assumed to be general concave functions of the capacity acquired at each facility. Erlenkotter's dual based algorithm is utilized as a subroutine during the progressive piecewise linear under-estimation technique developed in this paper. Dasci and Verter (2001) and Verter and Dasci (2002) provide extensions to a multi-product setting, where the firm is enabled to select among product-dedicated and flexible technology alternatives. At each alternative facility location, the technology options present different forms of scale and scope economies. More recently, a number of authors studied the integration of inventory control and logistics decisions with facility location. Shen (2005) used concave functions to represent economies of scale in the costs pertaining to the firm's inventories, whereas Snyder et al. (2007) and Sourirajan et al. (2007) presented facility location models that also considered the logistics costs.

An important stream of efforts to extend the classical *UFLP* and *CFLP* models involves the incorporation of *uncertainties* in the problem parameters. This is particularly relevant for global manufacturing firms that diversify their operations and facilities across many countries. Globalization has many potential advantages: access to cheap labor, raw material, and other production factors; presence at regional markets, and access to locally available technological resources and know-how. The resulting production-distribution networks are, however, increasingly exposed to price, exchange rate, and demand uncertainties in the international domain. The earliest efforts to incorporate exchange rate uncertainty in the *UFLP* are by Hodder and Jucker (1985) and Hodder and Dincer (1986). They used scenario-based approaches in modeling a risk-averse decision maker's structural choices. To this end, the expected profit is penalized by a term that corresponds to the constant portion of profit variability. Gutierrez and Kouvelis (1995) also used a scenario-based approach to find robust solutions under all possible scenario realizations. Canel and Khumawala (2001) and Kouvelis et al. (2004) studied the inclusion of subsidies and tariffs in international facility location models. Despite the popularity of the scenario-based approach in modeling the various types of uncertainties in the international domain, the prevailing papers show that the proliferation of the set of possible scenarios as a function of the problem size remains the major challenge from both academic and practical perspectives.

This section is an overview of the major works that followed the two classical papers reviewed in the preceding sections. The reader is referred to the recent reviews by Goetschalckx et al. (2002), Klose and Drexl (2005), Meixell and Gargeya (2005), Snyder (2006), Sahin and Sural (2007), and Shen (2007) for more exhaustive and comprehensive accounts of the state of the art in discrete facility location.

2.5 Potential Future Research Directions

In line with the classical *UFLP* and *CFLP* formulations, an overwhelming majority of the proposed extensions aim at minimizing the total fixed and variable costs relevant to the location problem under consideration. Using the categorization in Fisher (1997), these models are certainly suitable for designing efficient supply chains with functional products. The cost minimization objective, however, does not seem to be appropriate in the context of responsive supply chains that typically deal with innovative products. Note that many of the reported practical applications of discrete facility location models are associated with plant closure decisions, resulting in improved efficiency but mostly ignoring the possible ramifications concerning customer response. According to Ferdows (1997), the access to skills and knowledge and the proximity to markets are at least as important as the access to low-cost production factors in the firms' plant location decisions. Among the list of factors provided in Ferdows (1997), improving customer service, preemption of potential competitors, learning from supply chain partners, and attraction of a skilled workforce are typically not incorporated in the prevailing discrete facility location

models. It is necessary to improve the location modeling paradigms in order to better represent all the factors deemed important by firms in current practice. The need to improve the realism of the objective functions utilized in location models is also highlighted in Avella et al. (1999), summarizing the personal views of 20 young location researchers.

There is a need for increased empirical research in order to develop a better understanding of the factors that impact the facility location decisions of manufacturing and service firms and their decision making processes. Based on the location decisions of foreign-owned manufacturing plants in the United States in the 1990s, three factors seem to be most significant: the presence of a skilled workforce; the existence of a manufacturing base comprising suppliers, competitors and relevant industries, and the quality of transportation infrastructure. Interestingly, some of the past research reported rather conflicting empirical findings. For example, based on a survey of 73 plant managers, Brush et al. (1999) identified proximity to markets as the most significant location determinant, and concluded that subsidies and free trade zones are among the least important factors. Other authors, however, have pointed out that firms have been quite sensitive to subsidies, free trade zones, taxes and labor costs in making their location decisions (Coughlin and Segev 2000; Head et al. 1994). This calls for more empirical research and is perhaps due to the differences between the strategic priorities of the industries represented in the sample populations. If this observation can be confirmed through empirical studies, the development of industry-specific models rather than locating "generic" facilities would arise as a fruitful avenue for future research in location science.

References

Aikens CH (1985) Facility location models for distribution planning. Eur J Oper Res 22:263–279

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

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

Barros AI, Labbé M (1994) A general model for uncapacitated facility and depot location problem. Locat Sci 2:173–191

Beasley JE (1988) An algorithm for solving large capacitated warehouse location problems. Eur J Oper Res 33:314–325

Bilde O, Krarup J (1977) Sharp lower bound and efficient algorithms for the simple plant location problem. Ann Discrete Math 1:79–97

Brush TH, Maritan CA, Karnani A (1999) The plant location decision in multinational manufacturing firms: an empirical analysis of international business and manufacturing strategy perspectives. Prod Oper Manag 8:109–132

Canel C, Khumawala BM (2001) International facilities location: a heuristic procedure for the dynamic uncapacitated problem. Int J Prod Res 17:3975–4000

Canel C, Khumawala BM, Law J, Lo A (2001) An algorithm for the capacitated, multi-commodity multi-period facility location problem. Comput Oper Res 28:411–427

Cornuejols G, Fisher M, Nemhauser GL (1977) Location of bank accounts to optimize float: an analytic study of exact and approximate algorithms. Manag Sci 23:789–810
Coughlin CC, Segev E (2000) Location determinants of new foreign-owned manufacturing plants. J Reg Sci 20:232–251
Dasci A, Verter V (2001) The plant location and technology acquisition problem. IIE Trans 11:963–974
Efroymson MA, Ray TL (1966) A branch and bound algorithm for plant location. Oper Res 14:361–368
Elhedhli S, Goffin J-L (2005) Efficient production-distribution system design. Manag Sci 51:1151–1164
Erlenkotter D (1978) A dual-based procedure for uncapacitated facility location. Oper Res 26:992–1009
Ferdows K (1997) Making the most of foreign factories. Harv Bus Rev 75:73–88
Fisher M (1997) What is the right supply chain for your product? Harv Bus Rev 75:105–116
Gao L, Robinson EP (1992) A dual based optimization procedure for the two-echelon uncapacitated facility location problem. Nav Res Logist 39:191–212
Geoffrion AM, Graves GW (1974) Multicommodity distribution system design by Bender's decomposition. Manag Sci 20:822–844
Goetschalckx M, Vidal CJ, Dogan K (2002) Modeling and design of global logistics systems: a review of integrated strategic and tactical models and design algorithms. Eur J Oper Res 143:1–18
Gutierrez GJ, Kouvelis P (1995) A robustness approach to international sourcing. Ann Oper Res 59:165–193
Head C, Ries J, Swenson D (1994) The attraction of foreign manufacturing investments: investment promotion and agglomeration economies. www.nber.org/papers/w4878.pdf
Hodder JE, Dincer MC (1986) A multifactor model for international plant location and financing under uncertainty. Comput Oper Res 13:601–609
Hodder JE, Jucker JV (1985) A simple plant location model for quantity-setting firms subjected to price uncertainty. Eur J Oper Res 21:39–46
Holmberg K (1994) Solving the staircase cost facility location problem with decomposition and piecewise linearization. Eur J Oper Res 75:41–61
Holmberg K, Ling J (1997) A Lagrangean heuristic for the facility location problem with staircase costs. Eur J Oper Res 75:41–61
Jayaraman V, Pirkul H (2001) Planning and coordination of production and distribution facilities for multiple commodities. Eur J Oper Res 133:394–408
Karkazis J, Boffey TB (1981) The multi-commodity facilities location problem. J Oper Res Soc 32:803–814
Kaufman L, Eede MV, Hansen P (1977) A plant and warehouse location problem. Oper Res Quart 28:547–554
Khumawala BM (1972) An efficient branch and bound algorithm for the warehouse location problem. Manag Sci 18:B718–B731
Klincewicz JG, Luss H (1987) A dual based algorithm for multiproduct uncapacitated facility location. Transp Sci 21:198–206
Klose A, Drexl A (2005) Facility location models for distribution system design. Eur J Oper Res 162:4–29
Kouvelis P, Rosenblatt MJ, Munson CL (2004) A mathematical programming model for global plant location problems: analysis and insights. IIE Trans 36:127–144
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 9:643–666
Lim SK, Kim YD (1999) An integrated approach to dynamic plant location and capacity planning. J Oper Res Soc 50:1205–1216

Meixell MJ, Gargeya VB (2005) Global supply chain design: a literature review and critique. Transp Res E 41:531–550
Melo MT, Nickel S, da Gama FS (2005) Dynamic multi-commodity capacitated facility location: a mathematical modeling framework for strategic supply chain planning. Comput Oper Res 33:181–208
Moon S (1989) Application of generalized benders decomposition to a non-linear distribution system design problem. Nav Res Logist 36:283–295
Nauss RM (1978) An improved algorithm for the capacitated facility location problem. J Oper Res Soc 29:1195–1201
Neebe AW, Khumawala BM (1981) An improved algorithm for the multi-commodity location problem. J Oper Res Soc 32:143–169
Pirkul H, Jayaraman V (1996) Production, transportation, and distribution planning in a multi-commodity tri-echelon system. Transp Sci 30:291–302
Sahin G, Sural H (2007) A review of hierarchical facility location models. Comput Oper Res 35:2310–2331
Schrage L (1975) Implicit representation of variable upper bounds in linear programming. Math Program Study 4:118–132
Shen ZJM (2005) A multi-commodity supply chain design problem. IIE Trans 7:753–762
Shen ZJM (2007) Integrated supply chain design models: a survey and future research directions. J Ind Manag Optim 3:1–27
Snyder LV (2006) Facility location under uncertainty: a review. IIE Trans 8:537–554
Snyder LV, Daskin MS, Teo CP (2007) The stochastic location model with risk pooling. Eur J Oper Res 179:1221–1238
Soland RM (1974) Optimal facility location with concave costs. Oper Res 22:373–382
Sourirajan K, Ozsen L, Uzsoy R (2007) A single-product network design model with lead time and safety stock considerations. IIE Trans 39:411–424
Spielberg K (1969) Algorithms for the simple plant location problem with some side constraints. Oper Res 17:85–111
Tcha D, Lee B (1984) A branch and bound algorithm for the multi-level uncapacitated facility location problem. Eur J Oper Res 18:35–43
Van Roy TJ (1986) A cross decomposition algorithm for capacitated facility location. Oper Res 34:145–163
Van Roy TJ, Erlenkotter D (1982) Dual-based procedure for dynamic facility location. Manag Sci 28:1091–1105
Verter V, Dasci A (2002) The plant location and flexible technology acquisition problem. Eur J Oper Res 136:366–382
Verter V, Dincer MC (1995) Facility location and capacity acquisition: an integrated approach. Nav Res Logist 42:1141–1160

Chapter 3
Median Problems in Networks

Vladimir Marianov and Daniel Serra

3.1 Introduction

Suppose a number of geographically distributed customers are demanding a service or good, and facilities providing it need to be optimally located. Once facilities are deployed, either customers travel to the facilities to satisfy their needs, or vehicles travel from the facilities to customers' locations, carrying the goods to be delivered. The *p*-median problem finds the optimal location of exactly *p* facilities, so that the sum of the distances between customers and their closest facilities, measured along the shortest paths, is minimized. Since the number *n* of customers is known, by dividing the objective by *n*, the minimum average distance between customers and facilities is obtained too.

The *p*-median problem has became one of the most well-known and studied problems in the field of facility location. Its uses include a large number of applications, both geographical (the location of schools and warehouses) and non-geographical (defining the best clusters of objects, tasks, events, see Hansen and Jaumard 1997).

The *p*-median problem is a good model for many practical problems, provided some assumptions are made. The first assumption is that exactly *p* facilities are to be located. The decision on how many facilities are to be located either comes from political considerations, or simply because there is a fixed budget and the cost of locating a facility is the same no matter where it is located. In both cases, *p* is decided exogenously to the model, although it could be made also endogenous, as in Marianov and Taborga (2001).

As we explain below, further assumptions need to be made when more than one facility is to be located, since two questions must be answered: where to locate

V. Marianov (✉)
Department of Electrical Engineering, Pontificia Universidad Católica de Chile, Santiago, Chile
e-mail: marianov@ing.puc.cl

D. Serra
Department of Economics and Business, Pompeu Fabra University, Barcelona, Spain
e-mail: daniel.serra@upf.edu

H. A. Eiselt, V. Marianov (eds.), *Foundations of Location Analysis,*
International Series in Operations Research & Management Science 155,
DOI 10.1007/978-1-4419-7572-0_3, © Springer Science+Business Media, LLC 2011

the *p* facilities (the "location" problem), and what customer or demand node is assigned to which facility (the "allocation" problem). Regarding the allocation problem, there are two clearly distinct cases of application of the *p*-median. In the first case, the planner decides the location of the facilities, the route between facilities and customers, and the facility-demand allocation. No assumptions need to be made in this case. An example of this "central planning" case is the deployment of telephone switching centers: the planner decides not only their location, but also to which facility each customer will be physically connected through a pair of wires, as well as the layout of the wires. As opposed to the switching centers case, there are some situations in which it is not the planner who decides the customer-facility allocation and the route the customers follow to reach a facility. When stores are located, it is the customers who decide which store to patronize and the route they will follow from their homes to the chosen stores. In this case, if the *p*-median is to be used, the assumption is made that customers will patronize their closest facilities, and in its integer-programming formulation, the *p*-median naturally chooses this allocation. However, the fact that multiple facilities are located allows for alternatives, including allocation of a demand to more than one facility, which could be optimal if facilities have a limited capacity, or a better representation of the situation in which customers patronize different facilities on different occasions.

A last assumption often made is that customers will travel along the shortest paths between their origins and the facilities. This assumption can be relaxed as long as the planner knows exactly what route each customer would follow to each potential facility location. In most cases, travel or connection costs are assumed to be linear with distance. In the integer programming formulation of the *p*-median, there is no need for the costs to be linear with distance, and any non-decreasing cost function of the distance can be used.

Some authors have identified the objective of the *p*-median with a "public sector" objective. From the point of view of public decision-making, the *p*-median maximizes accessibility, if this is defined as average proximity of customers to a facility. If a region is represented by a network whose nodes are patient locations, and whose edges are roads, locating *p* hospitals according to the solution of a *p*-median will minimize the total or average travel distance for patients attending those hospitals. Ambulances will minimize their travel time or travel distance if *p* emergency rooms are located as determined by the *p*-median solution.

However, the *p*-median can also be used in the private sector. In this setting, the *p*-median objective represents minimization of transportation costs. A company that needs to locate a fixed number of warehouses and deliver its products from the warehouses to the customers will find the optimal solution to its problem using the *p*-median. If each node of a network represents a customer, and *p* maintenance centers housing each a vehicle have to be located, the *p*-median solution will provide the locations that minimize the total distance traveled by the vehicles when customers have to be served one at a time.

The name for the *p*-median first used by Hakimi, derives from the concept of a *median vertex*, which is the vertex of a network or graph for which the sum of

the lengths of the shortest paths to all other vertices is the smallest. On a network, finding the median vertex solves a problem similar to that posed by Fermat on an Euclidean plane in the 1600s, consisting of finding the location of the point on the plane that minimizes the sum of its distances to three known points. A first generalization to the Fermat problem is the Weber problem, in which weights are added to the three known points to represent the amount of demand aggregated at them. Assuming that transportation costs are proportional to both distance and demand, if a facility is located at the weighted median, it will satisfy the demand of the three points with the minimal transportation cost. A further generalization of the Weber problem includes more than three demand points, and more than one facility. The version with multiple facilities became known as the Multi-Weber problem. In the twentieth century, Cooper (1963, 1964) provided heuristic solutions for it, which are discussed in Chap. 15 of this volume.

On a network, as opposed to the problem on a plane, the demand is located only on vertices or nodes, each of them having a weight representing the total amount of demand that it houses. In the most general version of the *p*-median, the facility can be located on a node or at a point on an edge of the network; this distinction does not exist when the problem lies on the plane. Hakimi proved, however, that there is always an optimal solution considering only nodes or vertices of the network. The problem consists of finding this optimal solution. The *p*-median objective, which minimizes the sum of the distances between the customer or demand nodes and their closest facilities weighted by the amount of demand at the demand nodes, has been called a "minsum" or "minisum" objective, an objective also employed by the *Simple Plant Location Problem*, studied in Chap. 2 of this book.

Although now it seems a natural step, the *p*-median was not always formulated as an integer programming problem. The first formulation is due to ReVelle and Swain (1970) who, not being familiar with the results of Hakimi (1964, 1965), assumed node-only locations for what they called central facilities. Their formulation is now well known and used profusely, in the following form:

$$\text{Min} \sum_{i,\,j} h_i d_{ij} x_{ij} \tag{3.1}$$

$$\text{s.t.} \sum_{j} x_{ij} = 1, \quad i = 1, 2, \ldots n \tag{3.2}$$

$$x_{ij} \leq y_j, \quad i = 1, 2, \ldots n, \quad j = 1, 2, \ldots m \tag{3.3}$$

$$\sum_{j} y_j = p \tag{3.4}$$

$$x_{ij}, y_j \in \{0, 1\}, \quad i = 1, 2, \ldots n, \quad j = 1, 2, \ldots m, \tag{3.5}$$

where the subscripts, parameters, and variables are defined as follows:

i: index of customers,
j: index of potential facility sites,
m: total number of potential facility locations,
n: total number of customers,
p: total number of facilities to be located,
h_i: weight associated to each demand node (demand or number of customers),
d_{ij}: distance between customer i and potential facility at j,
x_{ij}: allocation variable equal to 1 if customer i is assigned to a facility at j, and 0 otherwise, and
y_j: location variable equal to 1 if there is an open facility at j, and 0 otherwise.

The set of constraints (3.2) forces each demand node to be assigned to exactly one facility. The set of constraints (3.3) allows demand node i to be allocated to a facility at j only if there is an open facility in that location. Constraint (3.4) sets the number of facilities to be located, and constraint (3.5) states that all variables are integers (binary).

The set of constraints (3.3) is known as the Balinski constraints, since Balinski (1965) was the first to write them in this form when studying the *Simple Plant Location Problem*. An alternative condensed version of the problem can be formulated by substituting the Balinski constraints with the following set:

$$\sum_{j=1}^{n} x_{ij} \leq m y_j, \quad i = 1, 2, \ldots, m \tag{3.6}$$

This constraint precludes customers being allocated to node j, unless there is an open facility on that node. While this set of constraints substantially reduces the size of the problem, when solving the linearly relaxed problem, by relaxing constraints (3.6), these constraints will tend to produce all fractional x_{ij}. On the other hand, the Balinski set of constraints increases the size of the problem in terms of the number of constraints, but when solving the linearly relaxed p-median problem, most variables x_{ij} tend to be integer in the solution. ReVelle and Swain (1970) observed that when branch-and-bound was required, the extent of branching and bounding needed was very small, always less than 6 nodes of a branch-and-bound tree. Therefore, the expanded form of the constraint makes integer solutions far more likely. In fact, in this formulation, ReVelle and Swain demonstrated that only the location variables y_j need to be declared binary, because once the facilities are located (and variables y_j have a value zero or one), there is always an optimal allocation that considers each customer fully allocated to a single facility. We outline the reasoning in the section dedicated to ReVelle and Swain's (1970) original contribution. Morris (1978) solved 600 randomly generated problems of the Simple Plant Location Problem (also a minsum problem) with the extended form of the constraint and found that only 4% required the use of branch-and-bound to obtain integer solutions. Rosing et al. (1979) proposed several ways to reduce both the number of variables and constraints in order to make the p-median problem more tractable.

Many solution procedures have been proposed for the *p*-median problem. Heuristic methods can be found in Chap. 15, and exact methods in Chaps. 13 and 14 of this volume. In this chapter, we first review and synthesize the early work by ReVelle and Swain (1970). Although chronologically speaking it is not the first contribution to the *p*-median, it is the most natural formulation and it implicitly includes branch-and-bound as a solution method. Next, we review Hakimi (1964, 1965) separately, following each section with an assessment of the impact these works had on the discipline. Finally, we review the major contributions that followed the first papers and finish with conclusions.

3.2 The Original Contributions to the p-Median Problem

This section will review three original contributions that have laid the foundations to what we now call the *p*-median problem, probably one of the best-known models in location science.

3.2.1 ReVelle and Swain (1970): An Integer Formulation of the p-Median Problem

ReVelle and Swain (1970) addressed the problem they call "central facility location," consisting of designating p of n communities in a geographical region as centers, so that the average time or distance travelled by people to go to these centers is minimal. They also suggest that the formulation they use is applicable to the case in which the facilities are supply points from where goods are distributed to the communities.

The average distance travelled by people is

$$\bar{d} = \frac{\sum_{i=1}^{n} h_i d(v_i, V_p^*)}{\sum_{i=1}^{n} h_i},$$

where i is the index of communities, n is the total number of communities and potential facility locations, h_i is the weight (demand) associated to each community, and $d(v_i, V_p^*)$ is the distance between the community (demand node) i and its closest center, belonging to the set of centers V_p^*.

Note that ReVelle and Swain never use the term "median" to denote the points that minimize the distance traveled by customers from communities to central facilities. As long as no confusion can arise, we will follow their lead. Also, references are included to previous works in the incipient area of discrete location analysis, most of them on the *Simple Plant Location Problem* (*SPLP*), also called

Uncapacitated Facility Location Problem (*UFLP*). Of particular interest is the reference to the work of Efroymson and Ray (1966), who used the Land and Doig (1960) method, which was later called branch-and-bound, applied to a new formulation of the *Simple Plant Location Problem*.

Linear programming tools are used for solving the central facility location (*p*-median) problem and, in the unlikely event of a non-integer solution, a branch-and-bound scheme is recommended. Also, using the same linear programming methods, the authors suggest that a heuristic solution can be tested for optimality. Alternatively, heuristic solutions can be used as a good starting point for the optimal solution.

The assumptions are: travel is performed using the shortest path between a community and a center, and allocation cannot be partial, i.e., a community (or demand node) is assigned fully to one and only one center (later proven to be an optimal choice, provided that communities with a center allocate to themselves). An additional assumption is that all centers are located at communities, and there are no candidate locations other than communities.

Once the matrix of shortest distances $d(v_i, v_j)$ between communities (vertices) v_i and v_j is computed for all i and j, and the allocation variables are defined as

$$x_{ij} = \begin{cases} 1, \text{ if community } i \text{ is assigned to center } j \\ 0 \text{ otherwise} \end{cases},$$

the *p*-median problem can be formulated as a linear programming problem:

$$\text{Min} \sum_{i,\,j} h_i d(v_i, v_j) x_{ij}$$

$$\text{s.t.} \sum_{j} x_{ij} = 1, \quad i = 1, 2, \ldots n$$

$$x_{ij} + \sum_{k \neq j} x_{jk} \leq 1; \quad i,\, j = 1, 2, \ldots n,\; i \neq j.$$

$$x_{ij} \in \{0, 1\}; \quad i,\, j = 1, 2, \ldots, n.$$

The last constraint requires that, if community i assigns to community j, the last one must be assigned to itself. Note that if a community j assigns to itself ($x_{jj}=1$), then the community must house a facility or center. This constraint can be replaced by the simpler one

$$x_{ij} \leq x_{jj}; \quad i,\, j = 1, 2, \ldots n.$$

Finally, a constraint is added to enforce the required number of facilities (p):

$$\sum_{j} x_{jj} = p.$$

If all self-assignments (x_{jj}) are either zero or one, then there is an optimal solution that considers each of the communities assigned fully to one facility (variables x_{ij} are either zero or one). The basic argument is that, if a community's demand were divided among two or more facilities in the solution, this solution could not be optimal. More specifically, unless the community is equidistant from these two or more facilities, that proportion of the demand assigned to the farther of the two facilities can be reassigned to the closer, and the objective will decrease. If a community is equidistant from two or more facilities in an optimal solution, then there exists an alternative optimum, and the present solution can be substituted by a solution with full assignment (integer variables x_{ij}). The consequence is that, if branch-and-bound is needed, only the self-assignment or location variables need to be declared integer.

In order to solve the problem, linear programming is recommended. In the event that there appears a fractional assignment, branch and bound is used on the x_{jj} variables, and the variable to branch on is chosen by a rule that considers to branch first on the variable x_{kk} for which the term $(\min_{j\neq k}\{h_k d(v_k, v_j)\})$ is the largest and variable x_{kk} has not been branched on.

The number of iterations needed to solve the problem using branch and bound may be favorably compared to enumeration, since the number of allocations that need to be evaluated by enumeration are $\binom{n}{p}$, while the number of iterations in the branch and bound scheme can be estimated in this case to be around $2(n^2+1)$, i.e. twice the number of constraints as estimated by Gass (1958). Since problems can grow large, cutting down on constraints is proposed by relaxing the constraints of the type

$$x_{ij} \leq x_{jj}, \quad i, j = 1, 2, \ldots n$$

and solving the following problem:

$$\text{Min} \sum_{i,j} h_i d(v_i, v_j) x_{ij}$$

$$\text{s.t.} \sum_{j} x_{ij} = 1, \quad i = 1, 2, \ldots n$$

$$\sum_{j} x_{jj} = p$$

$$x_{ij} \in \{0, 1\}; \quad i, j = 1, 2, \ldots, n.$$

This problem can be solved to optimality just by inspection:

1. Assign every community to its closest neighbor, without allowing self-assignments.
2. Break the assignments of the p communities with the largest assignment costs and assign them to themselves.

Once this solution is obtained, there will be some communities that receive assignments without being self-assigned. For these cases, add the corresponding constraints $x_{ij} \leq x_{jj}$ and solve, now using linear programming and branch-and-bound.

It was also suggested to start the process from a solution obtained by heuristic methods: set $x_{ij} \leq x_{jj}$ for those communities j that could attract assignments by virtue of being closer to other communities than an existing center and solve. If the solution is optimal, there is no need for more processing.

Regarding location of centers which are not at communities but on the roads between them: if there is a reason to think that there is such a good location, a new node can be added on that road (represented by an edge or an arc in the network). For such nodes, the corresponding variables x_{jj} and x_{ij} are added and the population or demand is set to zero, resulting in an empty node.

In many cases the decision on the number of facilities to locate is political. If a given maximum amount of funds is available, the total cost of a facility is given by

$$L_j = b_j x_{jj} + c_j \sum_i h_i x_{ij},$$

where b_j is the fixed cost and c_j the unitary expansion cost, which is multiplied by the amount of demand assigned to the center. The total cost is the sum of the costs of all the facilities, and this total cost must not exceed the amount of funds, M.

$$\sum_j b_j x_{jj} + \sum_j c_j \sum_i h_i x_{ij} \leq M.$$

If both the fixed cost of establishing a center and the expansion cost of an already located facility are the same, independent of the site of the facility, then

$$b \sum_j x_{jj} + c \sum_i h_i \sum_j x_{ij} \leq M.$$

Or, recalling that $\sum_j x_{ij} = 1$

$$\sum_j x_{jj} \leq \frac{M - c \sum_i h_i}{b}, \text{ i.e.,}$$

$$p \leq \left\lfloor \frac{M - c \sum_i h_i}{b} \right\rfloor.$$

After this analysis, the problem can be solved with different values of p, providing insight into the tradeoff between the travel time (or distance) and the number of centers.

ReVelle and Swain report computational experience with a problem with ten communities and four facilities for which thirty four iterations were needed; they

Table 3.1 Weighted distance matrix—fractional solution

	1	2	3	4	5	6
1	0	0.50	5.00	*M*	*M*	*M*
2	4.50	0	2.00	*M*	*M*	*M*
3	1.00	6.00	0	*M*	*M*	*M*
4	*M*	*M*	*M*	0	1.50	7.00
5	*M*	*M*	*M*	5.50	0	2.50
6	*M*	*M*	*M*	3.00	6.50	0

also report an execution time of 1.51 minutes and only 173 iterations of the linear programming code (as compared to 593,000 possibilities to be enumerated) for a 30-community, 6-node problem.

Finally, although no fractional solutions were encountered while obtaining the computational experience, they may occur for certain cost matrices, such as a matrix whose entries (i, j) are equal to $h_i d(v_i, v_j)$. In fact, the following example shows that fractional solutions could be optimal. Table 3.1 shows the weighted distances matrix from i to j, where $M >> 0$ denotes a sufficiently large constant.

The same cost patterns are shown in Fig. 3.1. In this figure, if node 2 houses a demand assigned to a facility at node 1, the objective increases in 4.5 units; if the demand at node 5 is assigned to a facility at node 6, the objective increases in 2.5 units, while if it is assigned to node 4, the objective grows by 5.5 units. The process continues in this fashion.

Figure 3.2 shows the optimal assignment. At each node there is one half of a facility and one half of the demand at each node is assigned to itself, while the

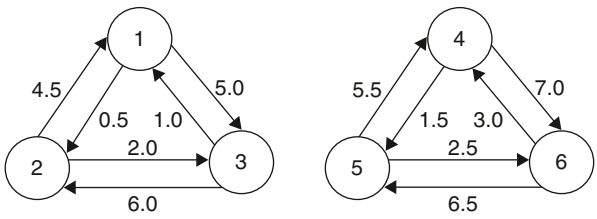


Fig. 3.1 Cost patterns

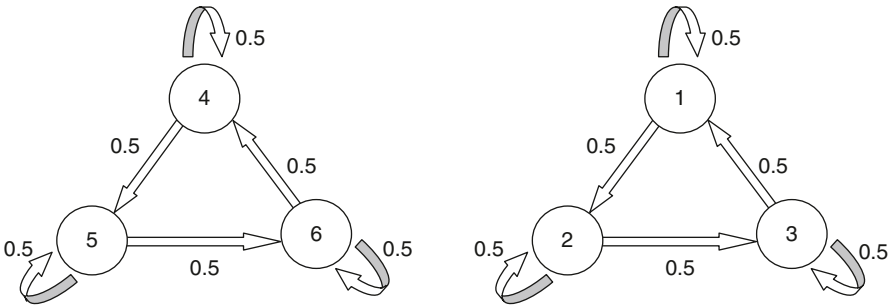


Fig. 3.2 Optimal (fractional) assignments

Table 3.2 Weighted distance matrix—integer solution

	1	2	3	4	5	6
1	0	1.00	2.00	*M*	*M*	*M*
2	1.50	0.00	1.25	*M*	*M*	*M*
3	1.75	1.75	0	*M*	*M*	*M*
4	*M*	*M*	*M*	0	1.00	2.25
5	*M*	*M*	*M*	1.50	0	1.33
6	*M*	*M*	*M*	2.00	2.00	0

remaining half is assigned to the closest node in terms of weighted distance. The optimal solution is then:

$$x_{jj} = 0.5 \forall j : x_{12} = x_{23} = x_{31} = x_{45} = x_{56} = x_{64} = 0.5$$

As the figures show, the appearance of fractional optimal solutions is associated with matrices that have what might be called "counter-cycles" of costs, i.e., cycles of costs running in opposite directions (a fact that is related to non-symmetric matrices). However, counter-cycles are not sufficient, and no general set of conditions was found by the authors for the appearance of fractional solutions. The matrix with counter-cycles of Table 3.2, does not produce fractional solutions:

To close the paper, a paragraph is included on applications: clinics providing therapy for individuals with chronic diseases, warehouses, mail-sorting facilities, central schools and parks, and others.

By the time ReVelle and Swain published their paper, the *p*-median problem and some of its properties had been described by Hakimi (1964, 1965), and two heuristics had been proposed for the *p*-median on a network. The first was that of Maranzana (1964), who proposed a solution method for locating *m* supply points among *n* demand points, and the second was the vertex substitution method by Teitz and Bart (1968). Both are based on heuristic algorithms proposed by Cooper (1964) as starting solutions for the continuous version of the problem. None of these algorithms were optimal.

The main contribution of ReVelle and Swain was proposing a method for solving the problem to optimality, through formulating it as an integer programming problem. The analysis made by ReVelle and Swain showed that their method was much faster than enumeration, the only previously-known exact method, and one that grows tremendously with the size of the instance of the problem. Although they state the problem of not allowing locations on edges of the network (or having to create new nodes along the edges if such locations were desired), thanks to the Hakimi property which we describe in the next section of this chapter, the ReVelle and Swain formulation solves the general problem optimally, using binary location and allocation variables and a branch-and-bound procedure that had then been recently proposed by Land and Doig (1960).

ReVelle and Swain's contribution was not only the integer programming formulation, but also the application to the location of central facilities by relating the *p*-median problem to that solved by Weber (Hakimi does not refer to the problem on the plane). Many applications of the *p*-median were found in the public sector, after ReVelle and Swain (1970) brought awareness to this problem.

A further contribution was likely bringing the problem to the attention of geographers, since all the previous research was published in operations research or mathematical journals.

3.2.2 Hakimi (1964): General Location of a Median on a Network

The concept of the *median vertex* of a graph, as well as some methods for finding the solution for the Multi-Weber problem (including the case with node weights, representing the amounts of demand at the nodes), were known when Hakimi (1964) posed the problem of finding the "absolute median" of a graph. The absolute median was a generalization of the median, in which the facility can be located not only on nodes, but also at any point along an edge of the network. This generalization makes sense only on a network.

Hakimi's (1964) paper also extended the concept of the *center vertex*, which is the vertex whose maximum distance to any other node of the network is minimized. He defines the "absolute center," which is located anywhere on the network. The center problem is addressed in a Chap. 5 on discrete center problems of this book.

The particular application of interest is that of locating a telephone switching center (or switch), S, in a communication network. This communications system is represented as a finite graph or network G. In such a graph, the switching center is directly connected through wires to each vertex v_i. Any message or communication between two vertices must be established through this switch. Each vertex v_i, connected through a branch b_i to the switch S, could need more than a pair of wires to evacuate its traffic. In a telephone network, this number of wires is associated to the number of subscribers at vertex v_i. In other communications networks, it could represent the (discrete) capacity of the branch. The number of wires needed by vertex v_i (its weight) is h_i, and the cost or length of the branch b_i is w_i. Such a network has the shape of a star, having the switch at its center. The problem is to find the optimal location of the switching center in such a way that the total length of the wires is minimal.

The usual concept of the median vertex does not apply to this problem, since the switch S could be located anywhere on the network, including both vertices and branches or edges. Define the distance $d(x, y)$ on the network or graph between points x and y on the network as the length of the shortest path between x and y, where the length of a path is the sum of the weights of the branches on that path; i.e., the sum of the length of the segment of branch connecting the point x to the switch, multiplied by the weight of the branch (weighted length), plus the length of the segment of branch connecting y to the switch $\boldsymbol{S}$, times the weight of that branch. If both points are on the same branch it is simply the length of the segment connecting them times its weight.

Using this notation, the point y_0 on an element of a weighted n-vertex graph G is defined as the *absolute median* of G if the sum of the weighted shortest distances between y_0 and every point y on G, *viz.*,

$$\sum_{i=1}^{n} h_i d(v_i y_0) \leq \sum_{i=1}^{n} h_i d(v_i, y). \tag{3.7}$$

This point is identified with the optimal location of the switch in the communications network.

After defining the absolute median, the following method for finding the location of the median vertex can be used: write the $[n \times n]$-dimensional distance matrix of the graph, adding up all the elements of each column j (distances between the node j and the remaining nodes), and choose as the median the node j corresponding to the column with the least value of the sum of distances.

The main median-related result of the paper is the following theorem, which is the generalization of an unpublished result by Goldstein at Bell Labs. In a private communication to Hakimi, Goldstein proved that an absolute median of a tree is always located at a vertex.

Theorem 1: *An absolute median of a graph is always at a vertex of the graph.*

If x_0 is an arbitrary point on the graph, not on a vertex, there always exist a vertex v_m of G such that

$$\sum_{i=1}^{n} h_i d(v_i, v_m) \leq \sum_{i=1}^{n} h_i d(v_i, x_0) \tag{3.8}$$

i.e. there is an absolute median at a vertex v_m.

Rather than repeating the Hakimi (1964) proof, the rationale is explained, reducing the mathematics as much as possible. The point x_0 is assumed to be located on an edge (v_a, v_b). Assume also that nodes are re-indexed in such a way that the following is true: the point x_0 is now located on the edge (v_p, v_{p+1}), and for all nodes with indices smaller than or equal to p, the shortest path connecting the node and point x_0 goes through node v_p, (connected through the left of x_0) while for all nodes with indices larger than p, the shortest path between the node and point x_0 goes through node v_{p+1} (through the right of x_0). The total weighted distance can then be expressed as the sum of two terms, representing the sum of the weighted distances to the left-side nodes, and to the right-side nodes, respectively:

$$\sum_{i=1}^{n} h_i d(v_i, x_0) = \sum_{i=1}^{p} h_i d(v_i, x_0) + \sum_{i=p+1}^{n} h_i d(v_i, x_0)$$

In turn, each distance can be decomposed in two as follows:

$$\sum_{i=1}^{n} h_i d(v_i, x_0) = \left[\sum_{i=1}^{p} h_i d(v_i, v_p) + \sum_{i=1}^{p} h_i d(v_p, x_0) \right] + \left[\sum_{i=p+1}^{n} h_i d(v_i, v_{p+1}) + \sum_{i=p+1}^{n} h_i d(v_{p+1}, x_0) \right].$$

Since $\sum_{i=p+1}^{n} h_i d(v_{p+1}, x_0) = \sum_{i=p+1}^{n} h_i d(v_{p+1}, v_p) - \sum_{i=p+1}^{n} h_i d(v_p, x_0)$, the full expression is

$$\begin{aligned}\sum_{i=1}^{n} h_i d(v_i, x_0) = &\left[\sum_{i=1}^{p} h_i d(v_i, v_p) + \sum_{i=1}^{p} h_i d(v_p, x_0)\right] \\ &+ \left[\sum_{i=p+1}^{n} h_i d(v_i, v_{p+1}) + \sum_{i=p+1}^{n} h_i d(v_{p+1}, v_p)\right. \\ &\left. - \sum_{i=p+1}^{n} h_i d(v_p, x_0)\right] \\ = &\sum_{i=1}^{p} h_i d(v_i, v_p) + \sum_{i=p+1}^{n} h_i d(v_i, v_{p+1}) + \sum_{i=p+1}^{n} h_i d(v_{p+1}, v_p) \\ &+ \left[\sum_{i=1}^{p} h_i - \sum_{i=p+1}^{n} h_i\right] d(v_p, x_0).\end{aligned}$$

The first three terms are independent of x_0. The term in square brackets is the sum of node weights "on the left," minus the sum of the node weights "on the right" of the point x_0. Without loss of generality, suppose that the sum of node weights on the left is larger than or equal to the sum on the right. Then, the term in square brackets is non-negative, and by reducing the distance $d(v_p, x_0)$ that multiplies the square bracketed term, i.e., moving the point x_0 to the left, the total sum is reduced or, at most, stays the same. The minimum value for this distance is zero, which happens when the median point x_0 is located on the node v_p.

The same argument can be repeated when the sum of the node weights on the right of x_0 is strictly larger than the sum of node weights on the left. In that case, the term in square brackets is strictly negative, and moving the point x_0 to the right strictly reduces the value of the total sum. The best value is obtained when x_0 is located on top of v_{p+1}.

This proves that there is always a median point on a vertex of the graph, either on the left or the right of a point x_0 on an edge. In other words, for any point x_0,

$$\sum_{i=1}^{n} h_i d(v_i, v_m) \leq \sum_{i=1}^{n} h_i d(v_i, x_0),$$

and, although an absolute median can be defined, there is always one at a vertex median.

Note that the previous result does not preclude other absolute medians existing on the network; however, it can be concluded that a median vertex is an optimal location for a switch in a communications network. It also could be a good location for a police station if h_i were the average number of daily automobile accidents in

community i, and the police must visit the scene of each accident to make a report. A mixed approach could be used too, in which a combination between the median and the center points is sought.

3.2.3 Hakimi (1965): Multiple Facilities and Vertex Optimality

The just described single median problem answers the question of the optimal location of a single facility. When more than a facility is to be located (say p facilities), the problem becomes known as the "p-median" problem, a term that was first used by Hakimi (1965) in the sequel to his 1964 paper. As before, Hakimi studies the p-median as a model that solves the problem of locating p switching centers on a communications network. Also, in this paper he studies a related problem, which we now know as the *Location Set Coverage Problem* (Toregas et al. 1971), applied to finding the least number of policemen to be deployed on a highway network in such a way that nobody is farther away from a policeman than a preset distance. Although the paper breaks ground for two of the most well-known location models, we concentrate on the p-median, while the covering problem is analyzed in Chap. 6 of this book.

When two or more facilities need to be located, on the plane or on a network, there is an extra degree of difficulty as compared to the location of a single facility: the optimal allocation or assignment of demands to facilities must be determined. This decision is to answer the question which facility is to serve the demand at any of the demand points in the problem. The p-median assumes that demands are assigned to their closest facilities.

Following the same lines as in the single facility case, the definition of the multiple median of a graph is generalized. If X_p is a set of p points $x_1, x_2, \ldots x_p$, and the distance of a node v_i to X_p is

$$d(v_i, X_p) = \min \{d(v_i, x_1), d(v_i, x_2), \ldots, d(v_i, x_p)\},$$

i.e. the distance between the node v_i and its closest point x_k in X_p, then the set X_p^* is a "p-median" of the graph G, if for every X_p on G,

$$\sum_{i=1}^{n} h_i d(v_i, X_p^*) \leq \sum_{i=1}^{n} h_i d(v_i, X_p).$$

In other words, X_p^* is the set of p points on the graph such that, if these points were facilities of some sort, the total weighted distance between the demands and their closest facility would be minimized. The p-median is then the optimum locations of p switching centers in a communications network.

The main result of the paper is the extension of the validity of the all-node solution to the p-median case.

Theorem 2: *There exists a subset V_p^* of the set of vertices, containing p vertices such that for every set of p points X on G*

$$\sum_{i=1}^{n} h_i d(v_i, V_p^*) \leq \sum_{i=1}^{n} h_i d(v_i, X).$$

For the proof, let us assume that the allocation problem has been solved, so that there are p clusters of demand points, each cluster j consisting of a point x_j in X and a set of demands for which x_j is the closest point in X. Then, if the point x_j is on an edge, by the theorem in Hakimi (1964), there is always a vertex v_j^* such that

$$\sum_{i \in \text{cluster } j} h_i d(v_i, v_j^*) \leq \sum_{i \in \text{cluster } j} h_i d(v_i, x_j).$$

The same inequality can be derived for each cluster. Note that the allocation of demands has not changed; the demands that were in cluster j are still in the same cluster. Adding up all these inequalities results in

$$\sum_{j} \sum_{i \in \text{cluster } j} h_i d(v_i, v_j^*) \leq \sum_{i=1}^{n} h_i d(v_i, X).$$

The left hand side of this inequality consists of the sum of p terms, one for each one of the original clusters. However, as the median point in each cluster moves toward a vertex, a demand node might become reassigned to a different node in V_p^*. This only happens if the re-allocation contributes to decrease still more the total sum, so that

$$\sum_{i=1}^{n} h_i d(v_i, V_j^*) \leq \sum_{j} \sum_{i \in \text{cluster } j} h_i d(v_i, v_j^*)$$

and

$$\sum_{i=1}^{n} h_i d(v_i, V_j^*) \leq \sum_{i=1}^{n} h_i d(v_i, X).$$

The suggested method for finding the p-median of a graph is enumerating all possible locations and allocating, for every location set, the demands to their closest facilities.

The impact of Hakimi's two contributions is hard to overstate. A common opinion among location researchers is that the paper by Hakimi (1964) strongly contributed to trigger the interest in location theory and analysis, and started a long string of related publications that does not seem to be decreasing. This opinion is somehow confirmed by the increasing yearly frequency of papers on location

since 1964 (Tansel et al. 1983a). Even if this were not the case, it can be safely stated that, at the very least, Hakimi (1964, 1965) brought awareness to the *p*-median problem. Other pioneering works are due to Kuehn and Hamburger (1963), who addressed the heuristic solution of the *Simple Plant Location Problem*, and Maranzana (1964).

Hakimi first generalized the problem of finding the median of a graph, as known up to the date of the publication of his papers, by defining the generalized median. This concept of a median located anywhere on the network, at least when there is a single median to be found, proved later (in the same paper of 1964) not to be too useful, since there is always an optimal location of the facility at a node. However, the question is important, since there are indeed some other problems for which the same property does not hold. Typical examples include the center point problem, addressed in the same Hakimi (1964) paper, the general absolute median problem of Minieka (1977), and the gravity *p*-median of Drezner and Drezner (2007). In these cases, location restricted to nodes could lead to sub-optimal solutions, while in the case of the *p*-median, the property proved by Hakimi (sometimes referred to as "Hakimi property") of existence of optimal solutions on vertices allowed looking for the optimal solution of the problem over a finite set (the nodes), instead of having to search over an infinite and continuous set (anywhere on the network).

Many researchers have focused on the Hakimi property and its applicability to different cases. Levy (1972) proved that the Hakimi property holds when the weights are concave functions of the distance, and Mirchandani and Odoni (1979) do the same when the cost of a path is a concave, nondecreasing function of its total distance and both demands and transportation costs are uncertain. Later, Mirchandani (1980) extends these results for stochastic problems with different assumptions. Goldman (1972) extended the validity of the property for multiple hops between an origin and a destination to what is now known as the hub location problem. Church and Meadows (1979) prove that the Hakimi property holds for covering problems (the *Location Set Covering Problem* and the *Maximum Covering Location Problem*) when the set of nodes is augmented with a set of network intersect points (*NIP*s) located along the arcs. The augmented set is called a Finite Dominating Set. Hooker et al. (1991) further developed the applications of finite dominating sets. A problem for which a finite dominating set is found becomes a problem for which a finite set of solutions must be checked, as opposed to a problem with an infinite number of solutions. In other words, finding a finite dominating set is equivalent to proving the Hakimi property for the problem, defined over an augmented set of nodes. Berman and Odoni (1982) proved that the Hakimi property also holds for the single facility location problem, when travel times are stochastic and the facility can be relocated according to the conditions of different scenarios. Finally, Shiode and Drezner (2003) showed that in a competitive facility location problem on a tree, when the leader faces stochastic demands, the Hakimi property holds for the leader's problem.

3.3 Other Major Contributions to the Field and Extensions

Credit for pioneering work on the *p*-median must be given to Hua Lo-Keng et al. (1962) who proposed an algorithm for locating the 1-median on trees (and networks with cycles), and proved that locating median points on vertices is better than locating them somewhere along the edges. Their paper was intended for practitioners who needed to set up threshing floors for dispersed wheat fields, so it did not include much mathematical insight. Their work was apparently not known in the western world until much later, since Hakimi did not reference it in 1964, and both Goldman (1971) and Kariv and Hakimi (1979) rediscovered the same algorithm several years later (see Chap. 14 of this volume).

An also frequently forgotten contribution is that by Gülicher (1965), who found results similar to those of Hakimi, but in a more restricted context. He was (and currently it is) known better among economists, rather than being a reference among location scientists.

There are several major works presenting generalizations of the original *p*-median problem. Goldman (1969) generalized the *p*-median, defining what is currently known as the hub location problem whereby commodities are transported over a path between origin and destination nodes, and the total transportation costs are minimized. The path goes through one or two medians. The problem addressed by Goldman (1969) has been frequently addressed, and it is now known as the *p*-hub median location problem. In turn, this result was generalized by Hakimi and Maheshwari (1972) to multiple commodities and multiple intermediate medians case.

Holmes et al. (1972) introduced two interesting generalizations of the *p*-median. The first generalization considers elasticity of demand, i.e., situations in which customers lose interest in the service or goods if these are located beyond a threshold distance. This generalization is useful in the case of non-essential goods or services, which are probably more common in practice than strictly essential goods. The second generalization considers a constraint on the capacity of the facilities, such as

$$\sum_i b_i x_{ij} \leq C_j y_j \quad \forall j,$$

where b_i denotes the demand at node i and C_j is the potential capacity of a facility located at j.

This constraint makes the problem much more difficult, since it leads to the appearance of many fractional-valued location and allocation variables in the solution if the integer-programming problem is solved in a linearly relaxed version. A further consequence of a limited capacity of facilities is the allocation of customers to facilities that are not the closest. If closest assignment needs to be forced, the Rojeski and ReVelle (1970) constraints can be used:

$$x_{ij} \geq y_j - \sum_{k \in N_{ij}} y_k,$$

where the set $N_{ij} = \{\text{potential facility sites closer to } i \text{ than to } j\}$.

An interesting extension of the *p*-median model was its application to hierarchical systems, i.e., systems composed by more than one category of facilities. Calvo and Marks (1973) appear to be the first to explore this type of setting. Their resulting model has multiple objectives. Further work with hierarchical systems was performed by Narula (1984).

Probabilistic behavior has also been included in the *p*-median models: Frank (1966), in an early response to Hakimi's contributions, discussed the effect of probabilistic demands on the location problem. Drezner (1987) addressed the "unreliable *p*-median" in which a facility has a certain probability of becoming inactive, and offered a heuristic for solving this problem. Berman et al. (1985) formulated the Stochastic Queue Median, which locates a single facility operating as a *M*/*G*/1 queue on any point of a network. Finally, Marianov and Serra (1998, 2001) investigated the effect of adding a probabilistic constraint to models that have a *p*-median type structure, and Marianov (2003) modified the objective of a *p*-median model to maximize a demand that is elastic to distance and congestion at the facilities.

Wesolowsky and Truscott (1975) introduced the multiperiod *p*-median problem, in which the facilities are relocated in response to predicted changes in demand, considering that relocating facilities has a cost. A loosely related problem is solved by Serra and Marianov (1998), who determined the best locations for *p* facilities when the demand changes through the day.

Very relevant is the analysis of the applicability of the *p*-median problem to practice. A first contribution was that of Hillsman and Rhoda (1978), who studied the effects of data aggregation in the *p*-median, considering the fact that customers are concentrated at the demand nodes. In their paper, the authors identified three classes of aggregation errors: source A, B, and C errors. Source A errors arise due to the approximation of the actual values of distance, source B errors are a particular case that occurs when a demand point coincides with a candidate location and the distance between the demand and the potential facility is considered equal to zero, and source C errors correspond to an incorrect assignment of the demands to facilities.

Another contribution to the analysis of the *p*-median was that of Kariv and Hakimi (1979), who proved that the general *p*-median, where *p* is a variable, is **NP**-hard even in the case of a planar network of maximum vertex degree 3, with vertices of weight 1 and edges of length 1. Also, they rediscovered once again Hua-Lo Keng et al. (1962) algorithm (see Chap. 14 of this volume) for locating one median on a network, and proposed an $O(n^2p^2)$ algorithm to find more than one median on trees.

There are several reviews of results related to the *p*-median. The first review focusing on the *p*-center and the *p*-median problems was that of Tansel et al. (1983a, b). After a classification of the location problems on networks, the authors describe different variants of the problem, as well as solution techniques, and end with results specific to tree networks. Generalizations and extensions of the *p*-median are covered in the excellent review by Mirchandani (1990). They include the multicommodity *p*-median of Hakimi and Maheshwari (1972), in which there are different amounts of demand for different products or commodities, as well as different routes for each product; generalizations that consider some type of constraint (facility capacity, arc capacity, distance constraints and implementation constraints); gen-

eralizations considering probabilistic travel distances and demands; oriented and non-oriented networks; nonlinear transportation costs; and hierarchical *p*-medians. Later, Marianov and Serra (2002) reviewed some of the applications of the *p*-median model, focusing on those in the public sector. Two other reviews are oriented specifically to solution methods: Reese (2006) presents an annotated bibliography of solution methods, while Mladenović et al. (2007) survey metaheuristic approaches to the solution of the *p*-median.

Finally, the reviews of location problems by Brandeau and Chiu (1989), Hale and Moberg (2003), and Snyder (2006) include material about the *p*-median, although the goal of the former two is to overview the research on location problems, while Snyder synthesized the work available on facility location under uncertainty.

3.4 Conclusions

Since the early works of Hakimi and ReVelle and Swain, the *p*-median problem has been, and still is, one of the most studied models in the facility location-allocation academic literature, not only to characterize its properties, but also because it requires sophisticated solution methods when the instances grow large. The *p*-median problem is being used to solve a large variety of applied location problems and also as a decision support tool to make decisions on locations. Unfortunately, only a few real world applications can be found published in the academic literature, compared to the large number of existing theoretical papers.

Among the real world situations that have been reported, interested readers can find studies concerning the location of industrial plants, warehouses and public facilities. A list of applications is provided by Christofides (1975). The *p*-median problem has also be used for cluster analysis, where locations of users are replaced by points in an *m*-dimensional space. Hansen and Jaumard (1997) provide a survey of cluster analysis from the point of view of mathematical programming. Cluster analysis may thus offer a powerful tool for data mining applications, see, e.g., Ng and Han (1994). Other applications of the *p*-median problem are related to the formation of cells (Won 2000), to the detection of glaucoma tests (Kolesar 1980), to the optimal sampling of biodiversity (Hortel and Lobo 2005), and to the assortment and trim loss minimization in the glass industry (Arbib and Marinelli 2004), among others. There is no doubt that many more applications will be found for this problem in the future.

Future work on this problem should include new solution methods for larger problems; theoretical research on the validity of the Hakimi property under different conditions; further relaxation of the assumption of closest assignment so to include different user preferences; and *p*-median models for multiple commodities and different routing policies.

Acknowledgments This research has been possible thanks to grants by the Spanish Ministry of Science and Education, BEC2006-12291, the Chilean CONICYT–FONDECYT 1070741, and support from the Instituto Milenio "Complex Engineering Systems," through grants ICM-MIDEPLAN P-05-004-F and CONICYT FBO16.

References

Arbib C, Marinelli F (2004) An optimization model for trim loss minimization in an automotive glass plant. Eur J Oper Res 183:1421–1432

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

Berman O, Larson RC, Chiu SS (1985) Optimal server location on a network operating as an M/G/1 Queue. Oper Res 33:746–771

Berman O, Odoni AR (1982) Locating mobile servers on a network with Markovian properties. Networks 12:73–86

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

Calvo A, Marks H (1973) Location of health care facilities: an analytical approach. Socio-Econ Plan Sci 7:407–422

Christofides N (1975) Graph theory—an algorithmic approach. Academic Press, London

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

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

Cooper L (1964) Heuristic methods for location–allocation problems. SIAM Rev 6:37–53

Drezner Z (1987) Heuristic solution methods for two location problems with unreliable facilities. J Oper Res Soc 38:509–514

Drezner T, Drezner Z (2007) The gravity *p*-median model. Eur J Oper Res 179:1239–1251

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

Frank H (1966) Optimum locations on a graph with probabilistic demands. Oper Res 14:409–421

Gass S (1958) Linear programming, 1st edn. McGraw-Hill, New York

Goldman AJ (1969) Optimal locations for centers in a network. Transp Sci 3:352–360

Goldman AJ (1971) Optimal center location in simple networks. Transp Sci 5:212–221

Goldman AJ (1972) Approximate localization theorems for optimal facility placement. Transp Sci 6:407–418

Gülicher H (1965) Einige Eigenschaften optimaler Standorte in Verkehrsnetzen. Schr Ver Sozialpolit 42:111–137

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

Hakimi SL, Maheshwari SN (1972) Optimum locations of centers in networks. Oper Res 20:967–973

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

Hansen P, Jaumard B (1997) Cluster analysis and mathematical programming. Math Program 79:191–215

Hillsman E, Rhoda R (1978) Errors in measuring distance from populations to service centers. Ann Reg Sci Assoc 12:74–88

Holmes J, Williams F, Brown L (1972) Facility location under maximum travel restriction: an example using day care facilities. Geogr Anal 4:258–266

Hooker JN, Garfinkel RS, Chen CK (1991) Finite dominating sets for network location problems. Oper Res 39:100–118

Hortel J, Lobo JM (2005) An ED-based protocol of optimal sampling of biodiversity. Biodivers Conserv 14:2913–2947

Hua Lo-Keng, others (1962) Application of mathematical methods to wheat harvesting. Chin Math 2:77–91

Kariv O, Hakimi SL (1979) An algorithmic approach to network location problems. Part II: The *p*-medians. SIAM J Appl Math 37:539–560

Kolesar P (1980) Testing for vision loss in glaucoma suspects. Manag Sci 26:439–450

Kuehn AA, Hamburger MJ (1963) A heuristic program for locating warehouses. Manag Sci 9:643–666

Land AH, Doig AG (1960) An automatic method for solving discrete programming problems. Econometrica 28:497–520
Levy J (1972) An extended theorem for location on a network. Oper Res Q 18:433–442
Maranzana F (1964) On the location of supply points to minimize transport costs. Oper Res Q 15:261–270
Marianov V (2003) Location of multiple-server congestible facilities for maximizing expected demand, when services are non-essential. Ann Oper Res 123:125–141
Marianov V, Serra D (1998) Probabilistic maximal covering location-allocation models for congested systems. J Reg Sci 13:401–424
Marianov V, Serra D (2001) Hierarchical location-allocation models for congested systems. Eur J Oper Res 135:195–208
Marianov V, Serra D (2002) Location problems in the public sector. In: Drezner Z, Hamacher H (eds) Facility location: applications and theory. Springer, Berlin
Marianov V, Taborga P (2001) Optimal location of public health centres which provide free and paid services. J Oper Res Soc 52:391–400
Minieka E (1977) The centers and medians of a graph. Oper Res 25:641–650
Mirchandani PB (1980) Locational decisions on stochastic networks. Geogr Anal 12:172–183
Mirchandani PB (1990) The *p*-median problem and generalizations. In: Mirchandani PB, Francis RL (eds) Discrete location theory. Wiley, New York
Mirchandani PB, Odoni AR (1979) Locations of medians on stochastic networks. Transp Sci 13:85–97
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
Morris J (1978) On the extent to which certain fixed-charge depot location problems can be solved by LP. J Oper Res Soc 29:71–76
Narula SC (1984) Hierarchical location–allocation problems: a classification scheme. Eur J Oper Res 15:183–189
Ng RT, Han J (1994) Efficient and effective clustering methods for spatial data mining. Proceedings of the 20th international conference on very large data bases. Santiago pp 144–154
Reese J (2006) Solution methods for the *p*-median problem: an annotated bibliography. Networks 48:125–142
ReVelle CS, Swain RW (1970) Central facilities location. Geogr Anal 2:30–42
Rojeski P, ReVelle C (1970) Central facilities location under an investment constraint. Geogr Anal 2:343–360
Rosing K, ReVelle C, Rosing-Vogelaar H (1979) The *p*-median model and its linear programming relaxation: an approach to large problems. J Oper Res Soc 30:815–823
Serra D, Marianov V (1998) The *p*-median problem in a changing network: the case of Barcelona. Locat Sci 6:383–394
Shiode S, Drezner Z (2003) A competitive facility location problem on a tree network with stochastic weights. European J Oper Res 149:47–52
Snyder LV (2006) Facility location under uncertainty: a review. IIE Trans 38:537–554
Tansel BC, Francis RL, Lowe TJ (1983a) Location on networks: a survey. Part I: the *p*-center and *p*-median problems. Manag Sci 29:482–497
Tansel BC, Francis RL, Lowe TJ (1983b) Location on networks: a survey. Part II: exploiting tree network structure. Manag Sci 29:498–511
Teitz M, Bart P (1968) Heuristic methods for estimating the generalized vertex median of a weighted graph. Oper Res 16:955–961
Toregas C, Swain RW, ReVelle CS, Bergman L (1971) The location of emergency service facilities. Oper Res 19:1363–1373
Wesolowsky GO, Truscott WG (1975) The multiperiod location–allocation problem with relocation of facilities. Manag Sci 22:57–65
Won Y (2000) New *p*-median approach to cell formation with alternative process plans. Int J Prod Res 38:229–240

Part III
Minimax Problems

Chapter 4
Continuous Center Problems

Zvi Drezner

4.1 Introduction

The minimax facility location problem (also called the one center problem) seeks to locate a facility so that the maximum distance to a set of demand points is minimized. Using Euclidean distances in the plane, this problem is equivalent to finding the center of the smallest circle enclosing all points, hence the term "center" regarding this problem. When other metrics are used, the 1-center problem is equivalent to covering all points with a shape similar to the unit ball of the metric. For example, when rectilinear distances are used, the problem is to cover all points with the smallest possible diamond.

Locating several ($p>1$) facilities two problems are discussed in the literature. One formulation, the minimax multifacility problem incorporates distances between the facilities into the objective function. A more commonly investigated problem is termed the minimax location-allocation problem or the p-center problem. Each demand point is allocated to the closest facility and the maximum distance to the closest facility need to be minimized. Another popular objective of the location-allocation type is minimizing the (weighted) sum of distances to the closest facility. Such problems are also called p-median problems and are discussed in Chap. 3 of this volume.

In this chapter, we review the one center and p-center problems in the plane. Early papers investigate the unweighted one center problem using Euclidean distances in the plane. More recent research expands the investigation to weighted problems, multiple facilities location, different distance metrics, and different environments; we, however, restrict our review to continuous spaces. There exists a significant body of literature dealing with discrete location (when there exists a finite set of possible sites for the facilities) and, in particular, location in a network environment.

Z. Drezner (✉)
Steven G. Mihaylo College of Business and Economics,
California State University-Fullerton, Fullerton, CA 92834, USA
e-mail: zdrezner@exchange.fullerton.edu

H. A. Eiselt, V. Marianov (eds.), *Foundations of Location Analysis,*
International Series in Operations Research & Management Science 155,
DOI 10.1007/978-1-4419-7572-0_4,

4.2 Early Research in the Nineteenth Century

The one center unweighted location problem applying Euclidean distances was first suggested by the renowned English mathematician James Joseph Sylvester (1814–1897), who, in 1857, asked the following question in a one sentence "manuscript:"

> It is required to find the least circle which shall contain a given system of points in the plane (Sylvester 1857).

Three years later, Sylvester (1860) published the analysis of this and other related problems.

Sylvester starts with a general discussion of linear approximations to a square root of sum of squares of terms like $\sqrt{x^2+y^2+z^2}$ as approximated by a linear function of the structure $\alpha x+by+cz$. On page 212, he then proceeds to solve the problem of finding the smallest circle enclosing a set of points, observing that points which are not vertices of the convex hull will not be part of the solution. He then notes that the optimal solution involves two or three points. He writes

> If a circle is drawn through three points, then two cases arise. If the three points do not lie on the same semicircle, no smaller circle than this one can be drawn that contain the three points. If the points do lie in the same semicircle, it is obvious that a circle described upon the line joining the outer two as a diameter will be smaller than the circle passing through all three and will contain them all.

This distinction is the same as determining whether the triangle based on three points is acute or obtuse. He then proceeds to describe his algorithm in quite a complicated fashion based on a solution method of Professor Peirce with no citation. The editors of this volume searched for a mathematician named Peirce and found the scholar Benjamin Peirce (1809–1880), a professor at Harvard University, whose scholarly work spanned the techniques needed for Sylvester's approach. His book Linear Associative Algebra (1882) contains an analysis of two-dimensional algebras which are the basis of Sylvester's approach. We therefore believe that Sylvester referred to Benjamin Peirce's work. Here, we prefer to describe the almost identical algorithm provided in the Chrystal (1885) paper because it is explained more simply.

Chrystal (1885) unknowingly reinvented Sylvester's method. Towards the end of his paper Chrystal writes

> I learned a day or two before communicating it to the Mathematical Society of Edinburgh that the problem had originally been proposed by Professor Sylvester (*Quarterly Journal of Mathematics,* 1, p. 79) and that a solution had been given by him in an article in the *Philosophical Magazine* more than twenty years ago (1860, Fourth Series, vol. 20, pp. 206–212). I have since consulted this paper, and find that the solution there given is due to Peirce. It is only briefly indicated, but appears to be substantially identical with the one I have given above.

Note that the last page number of the citation of Sylvester (1860) is wrong in this quote. It should be 206–222.

Chrystal observed several properties of the smallest possible circle enclosing a set of points. First, if a circle encloses all the vertices of the convex hull of the

set of points, it encloses all the points in the set. Therefore, all points that are not vertices of the convex hull can be eliminated from consideration, and only vertices of the boundary of the convex hull need to be included in the problem. This may significantly reduce the number of demand points defining the problem. Second, a circle passing through the vertices of an acute triangle, is the smallest possible circle enclosing the three points. However, if there is an obtuse angle to that triangle, the smallest circle is instead centered at the center of the side opposite the obtuse angle (which is the longest side of the triangle). For a right triangle, the two circles described above are actually the same circle. He then concludes that the smallest circle is either based on the most distant pair of points as a diameter, or the largest circle passing through three points which form an acute triangle.

Chrystal's algorithm starts with a "large" circle that encloses all the points and reduces the radius of the circle iteratively until the smallest circle is obtained. With each iteration, a pair of vertices of the convex hull are examined. The algorithm starts with an arbitrary side of the convex hull; at each iteration the demand point subtending the smallest angle to the selected pair of points is identified. If the smallest angle is obtuse or a right angle, the center of the segment is the center of the minimum circle, because all the points are in the circle defined by the pair of points as the ends of its diameter. If the smallest angle is acute, the other two angles of the triangle are evaluated. If both of them are acute (or if one of them is a right angle), the circle through these three points is the minimum circle. If one angle is obtuse, the side opposite that angle is selected for the next iteration. If there is no obtuse angle, the optimal circle has been found. The process must end with either two or three points defining the minimum circle, because the radius declines at each iteration and therefore the same pair of points cannot be examined again. Even though the concept of complexity was not defined yet in 1885, Chrystal was interested in the maximum possible number of iterations. He determined that the bound on the number of iterations is $\frac{1}{2}m(m-1)$ (where m is the number of vertices in the convex hull of the demand points) because there are $m(m-1)/2$ possible segments connecting vertices of the convex hull and a segment cannot be considered more than once.

We illustrate the Chrystal-Sylvester algorithm in Fig. 4.1. The problem is to cover 6 points marked with full black dots. Two points, A and B are selected as two consecutive points on the convex hull. The angles subtended from each of the other four points are drawn, with the smallest angle subtended from point C. The circle passing through A, B, and C is centered at the empty dot and is drawn as a thick line. The circle is centered at (2, 3.667) with a radius of 3.8006. Note that the center of the circle and its radius need not be calculated. Since the angle ABC is obtuse, points A and C are kept, and all the angles subtended on the segment AC are calculated (not shown in the Fig. 4.1). Only the angle subtended from D is acute, thus the smallest, and therefore the points ACD are selected. They form an acute triangle and therefore the circle passing through them (the thinner line centered at the triangle in the figure) is the solution. The solution is at (2.5, 3.333) with a radius of 3.6553. Again, these values need not be calculated throughout the iterations. One can calculate the center and radius of the optimal circle once the two or three points defining

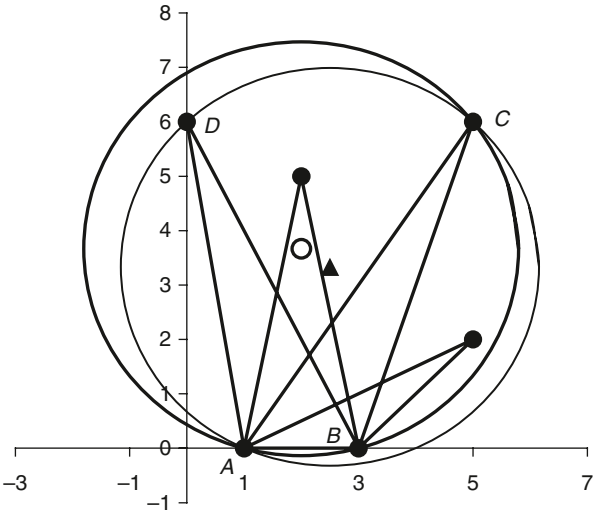


Fig. 4.1 The Chrystal-Sylvester algorithm

it are obtained. Note that the new center is closer to the segment AC, and thus all the points "on the other side of the segment" that were covered before are covered following the shift; the part of the new circle on the other side of AC contains the points covered by the thick circle. The sequence of circles is decreasing and each circle contains all the points.

Chrystal (1885) was also interested in other problems but did not offer solution algorithms for them. He suggested the problem of the minimum possible three dimensional sphere that encloses all points in a three dimensional space. He observed that in three dimensions, the minimum sphere is constructed by two, three, or four points. This result can be derived by Helly's Theorem (Radon 1921). Helly's Theorem states

> Given a finite collection of convex sets in $\mathbb{R}^n$. If the intersection of every $n+1$ of these sets is nonempty, then the whole collection has a nonempty intersection.

For a general discussion see Drezner (1982), who shows that any minimax problem based on convex functions in a k-dimensional space can be determined by solving the minimax problem based on a set of up to $k+1$ functions. The minimum k-dimensional sphere is therefore constructed based on up to $k+1$ points. Chrystal (1885) also suggested the problem of finding an ellipse with the minimum area that encloses a set of points in the plane. We are unaware of a paper addressing this problem.

In order to implement the Sylvester-Chrystal algorithm on a computer, certain derivations are needed. Finding two consecutive vertices of the convex hull to start the algorithm, we can select the first vertex as the one with smallest value of x, and finding the adjacent vertex as the one with the largest slope (positive or negative) of the line connecting the first vertex with all other points. This requires an $O(n)$

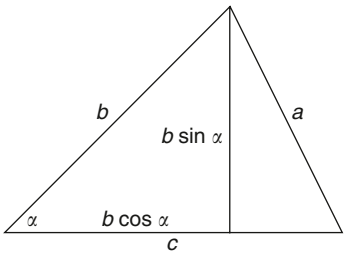


Fig. 4.2 The law of cosines

effort, where n is the number of points. For implementing the iterations, we need to find the point that subtends the smallest angle on a segment connecting two points. This is done by the following observation: let a be the length of the segment, and b and c be the distances between a third point and the end points of the segment. The law of cosines is $a^2=b^2+c^2-2bc\cos\alpha$ (Beyer 1981). For completeness, we provide the simple proof here. Consider Fig. 4.2, where $a^2=(c-b\cos\alpha)^2+b^2\sin^2\alpha$, which leads to the law of cosines.

The angle α opposite side a fulfills

$$\cos\alpha = \frac{b^2+c^2-a^2}{2bc}.$$

Therefore, the smallest angle among all points is the largest value of cosα. Since we are interested only in acute or right angles (because we already have found the solution if none of the angles is acute), when $b^2+c^2-a^2<0$ the point is ignored. To avoid the need to calculate a square root which may increase the run time of the algorithm, the quantity $f=b^2+c^2-a^2$ is calculated and only if $f>0$ we proceed to calculate the maximum among $f^2/(b^2c^2)$. This way, only squares of distances need to be calculated. This approach is very fast and easy to implement. In order to determine whether a triangle is acute or obtuse, a simple rule applies: if $a^2+b^2+c^2>2\max\{a^2, b^2, c^2\}$ the triangle is acute. In case of equality, it is a right triangle; if the inequality is reversed, the triangle is obtuse.

4.3 The Elzinga-Hearn Algorithm (1972)

Elzinga and Hearn (1972a) proposed and solved four different problems. All are based on the following basic formulation. Let n points be located at x_i with associated constants $k_i\geq 0$ for $i=1,\ldots, n$, be given. We need to find a location X that minimizes

$$\max_{1\leq i\leq n}\{d_i(X)+k_i\}$$

where $d_i(X)$ is the distance between point i and the unknown solution point X. Note that the requirement that $k_i \geq 0$ is not necessary, because adding a constant to all k_i does not change the problem and all negative k_i can be converted to nonnegative values.

The unweighted 1-center problem ($k_i = 0$) is called the *Delivery Boy Problem*; when a possibly different constant is added to each distance in the minimax formulation, the problem is called the *Messenger Boy Problem*. The first is called the Delivery Boy Problem since we wish to minimize the maximum distance a delivery boy will need to travel to make deliveries to a set of points. The second is called the Messenger Boy Problem because once the messenger boy reaches the point, he needs to travel an extra distance in order to deliver the message. These names did not catch on and are not repeated in subsequent research papers. Each of these problems is formulated and solved using Euclidean or rectilinear distances.

The algorithm for the Euclidean 1-center (also called single-facility minimax location) is as follows:

Algorithm 1: The Euclidean 1-Center

Step 1: Pick any two points.

Step 2: Construct a circle based on the segment connecting two points as a diameter. If the circle covers all points, then stop, a solution has been found. Otherwise add a point outside the circle to the pair of points to form a set of three points.

Step 3: If the triangle with the three points at its vertices has an angle of at least 90°, drop the point on the obtuse angle and go to Step 2.

Step 4: If the circle passing through the three points covers all points, stop, a solution has been found. If there is a point outside the circle, choose such a point, D, and add it as a fourth point. One of the original three points (A, B, C) must be discarded. The new point D and the farthest point from it, A, remain in the set. To determine the third point, extend the diameter of the current circle through A defining two half planes. Select the point (of the remaining two points B or C) which is not on the same half plane as D.

For the Elzinga-Hearn algorithm, we need a formula for the center of a circle enclosing an acute triangle. Such a formula is given in Drezner and Wesolowsky (1980). We simplify it further by translating the system of coordinates so that the first point is at (0, 0) and then add the original coordinates of the first point to the result. This way the three dimensional determinants (Drezner and Wesolowsky 1980) are reduced to two-dimensional ones. Let (x_2, y_2) and (x_3, y_3) be the coordinates of the second and third point following a translation of the system of coordinates so that the first point is translated to (0, 0). Three quantities are calculated:

$$\Delta = x_2 y_3 - y_2 x_3,$$

$$\Delta_1 = \left(x_2^2 + y_2^2\right) y_3 - \left(x_3^2 + y_3^2\right) y_2, \quad \text{and} \quad \Delta_2 = \left(x_3^2 + y_3^2\right) x_2 - \left(x_2^2 + y_2^2\right) x_3.$$

The center is then at (x_0, y_0) with $x_0=\Delta_1/2\Delta$ and $y_0=\Delta_2/2\Delta$, and the radius of the circle is the distance from the center to the point (0, 0), which equals $\frac{\sqrt{\Delta_1^2+\Delta_2^2}}{2\Delta}$.

We can now make the following observations:

1. The same method described in the last paragraph of Sect. 4.2 can be used to determine whether a triangle is acute or not.
2. Most of the computer time in one iteration is spent on finding a point outside the circle (which is of complexity $O(n)$). All other components are of complexity $O(1)$. Therefore, it may be simpler to check all three points as candidates for removal in Step 4, and select the one which results in the largest radius of the remaining three points rather than the geometric scheme suggested by Elzinga and Hearn (1972a).
3. One can choose the point farthest from the center of the circle rather than any point outside the circle. This guarantees that the newly selected point is a vertex of the convex hull.
4. Convergence of the algorithm is guaranteed because the radii of the circles in consecutive iterations increase.

The algorithm is illustrated using the same six points shown in Fig. 4.1. Suppose that points A and B are selected at the start of the solution algorithm. The center of the smallest circle covering A and B is at the middle of the segment AB and its diameter is the length of AB. The farthest point outside this circle (i.e., the farthest point from the center of the segment) is point C. Since the triangle ABC is obtuse, point B is discarded and the next iteration uses points A and C. The farthest point from the midpoint between A and C is point D, so point D is added to the set. The triangle ACD is acute, so it defines the next iteration with a circle centered at the location marked as a triangle in the figure. All points are enclosed by this circle so the algorithm terminates.

When a different $k_i \geq 0$ is added to the minimax problem (the messenger boy problem, as Elzinga and Hearn (1972a) termed it), the problem can be viewed geometrically as replacing the points with circles centered at the points with radii of $k_i \geq 0$. The problem therefore turns to finding the smallest circle that encloses all circles. The solution is either one circle if there exists a circle that encloses all other circles or a circle that is tangent externally to two or three circles. The proposed algorithm is similar to the algorithm described above for the unweighted 1-center problem. The only difference is that a procedure is designed to find the smallest circle enclosing two or three other circles (rather than points).

Note that the weighted version of this problem, where the distances are multiplied by weights and $k_i \geq 0$ added, is solved in Drezner (1991) by using an Elzinga-Hearn type algorithm. The only issue is finding the solution based on two or three points. While the solution to a two point problem is an explicit formula, an iterative procedure is suggested for the solution of the three points problem.

When distances are rectilinear, the problem is to find the smallest diamond the covers all points. This problem does not require an iterative procedure. As outlined in Elzinga and Hearn (1972a) and Drezner (1987), we must find the maximum and minimum of a_i+b_i and a_i-b_i, where (a_i, b_i) are the coordinates of the points. These

four values define four lines inclined by 45° to the system of coordinates. Two have a positive slope and two have a negative slope. The solution is the center of the diamond defined by these four lines. This can be done in $O(n)$ time. Usually, when not all the sides of the diamond are equal to one another there are multiple solutions to this problem. When constants $k_i \geq 0$ are added to the distances, the points are replaced by diamonds. The problem turns to finding the smallest diamond that encloses all diamonds. The solution is very similar to the unweighted 1-center rectilinear problem and is also solved in $O(n)$ time.

Recall that the Sylvester and Chrystal algorithm starts from a "big" circle that encloses all points and the circle is iteratively shrunk until the optimal circle is obtained. Conversely, Elzinga and Hearn start with a small circle and iteratively increase it by adding points outside the circle until all points are covered. The remainder of this section provides a computational comparison of the two approaches.

The complexities of the Elzinga and Hearn (1972a) or the Sylvester-Chrystal algorithms are not clear. In experiments they seem to behave like $O(n)$ because the number of iterations is quite stable. Drezner and Shelah (1987) constructed a contrived example where the number of iterations of the Elzinga-Hearn algorithm is $O(n)$ and thus the complexity of the algorithm is at least $O(n^2)$. The crude upper bound suggested in Chrystal (1885) leads to a bound of $O(n^3)$ on the complexity of that algorithm.

In their 1972 paper, Elzinga and Hearn claim that the Sylvester-Crystal procedure "was, naturally enough, designed for solving the problem by hand. Our procedure is quite different and is more efficient for implementation on a computer." Actually, it is very easy to code the Sylvester-Chrystal algorithm on a computer as detailed above. Therefore, we tested the relative efficiency of both algorithms.

As indicated, the two or three points defining the smallest circle enclosing all points must be vertices on the boundary of the convex hull. Both algorithms select as the next point in the process a point on the boundary of the convex hull (if in the Elzinga-Hearn algorithm we select the farthest point in Step 2 or 4 of the algorithm) so in both algorithms starting from the first iteration in the Sylvester-Chrystal algorithm, and from the third iteration in the Elzinga-Hearn algorithm, all two or three points defining the circle considered in each iteration are vertices of the convex hull. It may therefore be beneficial to extract the vertices of the convex hull before applying any of the algorithms. This can be done in $O(n \log n)$ time (Graham 1972; Graham and Yao 1983). This may require more effort than the $O(n)$ effort required for each iteration, if the number of iterations is less than $O(\log n)$. In our computational experiments this is not the case. Testing this option is beyond the scope of this chapter.

We evaluated both algorithms for problems with up to 10,000 demand points. First we generated points in a unit square, and then we constructed somewhat harder problems with points uniformly generated in a ring of inner radius of 0.999 and outer radius of 1. These problems have a higher percentage of points that are vertices of the convex hull. Each problem was solved one million times by each algorithm. It turns out that generating the problems themselves required longer computational time than solving them. Therefore, we ran programs that just generated the problems without solving them and reported the extra time needed for the solution process. Programs were coded in Fortran using double precision arithmetic, compiled

Table 4.1 Comparing the Elzinga-Hearn and the Chrystal-Sylvester algorithm

n	Elzinga-Hearn				Chrystal-Sylvester			
	Points in a square							
	Iterations			Time (s)	Iterations			Time (s)
	Min	Max	Aver		Min	Max	Aver	
10	1	7	2.94	1.10	1	7	2.18	0.89
20	1	8	3.46	1.58	1	8	2.58	1.73
50	1	8	3.77	2.42	1	9	3.14	4.48
100	1	8	3.90	3.45	1	9	3.59	9.46
200	1	8	4.00	5.59	1	11	4.04	20.64
500	1	8	4.10	11.83	1	11	4.64	58.17
1,000	1	8	4.16	23.31	1	12	5.10	127.86
2,000	2	8	4.20	45.41	1	14	5.56	278.54
5,000	2	8	4.23	111.67	1	15	6.17	774.78
10,000	2	8	4.26	222.75	1	16	6.63	1,668.69
Points in a ring								
10	1	8	3.62	1.43	1	9	2.81	1.07
20	1	9	4.51	2.16	1	10	3.90	2.44
50	1	10	5.24	3.46	1	13	5.38	7.04
100	1	11	5.68	5.24	1	15	6.42	15.70
200	2	11	6.06	8.64	1	16	7.28	34.15
500	2	11	6.49	19.03	1	18	8.02	91.66
1,000	2	12	6.77	38.26	2	20	8.42	190.56
2,000	2	12	7.03	76.32	2	20	8.82	395.46
5,000	3	13	7.31	193.25	2	21	9.35	1,037.69
10,000	3	12	7.51	393.41	2	21	9.75	2,143.30

by an Intel 9.0 Fortran compiler and run on a 2.8 GHz desk top computer with 256 MB of RAM. The results are summarized in Table 4.1. Note that the run time in seconds is the time required to solve a million problems.

It took about one millionth of a second to solve one problem with $n = 10$ demand points. The Elzinga-Hearn algorithm is more efficient than the Sylvester-Chrystal algorithm, mainly because one iteration of the Sylvester-Chrystal algorithm consumes more computer time than one iteration of the Elzinga-Hearn algorithm. However, the two algorithms are quite comparable, especially for smaller problems.

4.4 More Recent Papers

Proposed algorithms for the unweighted one center problem using Euclidean distances and others have improved complexities over the years. Drezner and Shelah (1987) showed that the complexity of the Elzinga-Hearn algorithm is at least $O(n^2)$. Shamos and Hoey (1975) proposed an $O(n \log n)$ algorithm based on Voronoi diagrams (for a discussion of Voronoi diagrams the reader is referred to Suzuki and Okabe 1995, Okabe et al. 2000, and Chap. 19 of this volume). Megiddo (1983a)

constructed an algorithm that solves the unweighted 1-center problem in $O(n)$ time, which clearly cannot be improved.

Elzinga and Hearn (1972b) constructed an algorithm to solve the unweighted 1-center problem in m-dimensional space. The problem is formulated as a quadratic programming problem and solved by the simplex method for solving quadratic programs.

The rectilinear unweighted 1-center problem can be solved in $O(n)$ time (Elzinga and Hearn 1972a; Drezner 1987). Drezner (1987) also constructed an $O(n)$ algorithm for the 2-center problem and an $O(n \log n)$ algorithm for the 3-center problem.

The weighted one center problem has also received attention in the literature even though it is more difficult to find applications that incorporate weights into the objective function. Dearing (1977) was probably the first one to introduce the weighted 1-center problem in the plane. Following his presentation, Drezner and Wesolowsky (1980) proposed an iterative solution approach similar to the Elzinga and Hearn (1972a) algorithm for the Euclidean, the rectilinear, and the general ℓ_p distances problems. Jacobsen (1981) solved the Euclidean weighted 1-center problem by formulating it as a nonlinear programming problem. Hearn and Vijay (1982) suggested a solution approach based on the Elzinga and Hearn (1972a) algorithm, Megiddo (1983b) developed an $O(n \log^2 n)$ algorithm, and Zemel (1983) developed an $O(n \log n)$ algorithm for its solution. Finally, Dyer (1986) proposed an $O(n)$ time algorithm for the solution of the weighted one center problem using Euclidean distances. He also showed that an $O(n)$ time algorithm exists for the solution of the weighted 1-center problem in a space of any dimensionality. However, the complexity increases exponentially with the dimensionality. For a problem in k dimensions the complexity is $O(3^{(k+2)^2} n)$.

The weighted 1-center rectilinear problem can be solved in $O(n)$ time by employing the technique suggested by Megiddo (1983a). Megiddo showed how to solve linear programs with two or three variables in time linear in the number of the constraints. The weighted 1-center problem in k dimensions can be decomposed into k single dimension problems. Each single dimension problem can be formulated as a linear program with two variables and $2n$ constraints. Let L be the maximum weighted distance from the facility and x the location of the facility on a line. Let x_i be the locations of the demand points on the line and w_i be their associated weights. The linear program is to minimize L subject to the $2n$ constraints $-L \le w_i(x - x_i) \le L$. Therefore, the task of solving the weighted rectilinear 1-center problem in k dimensional space can be performed in $O(kn)$ time.

4.5 Extensions to the 1-Center Problem

There are a few single facility location problems related to the 1-center problem. The k-centrum objective (Andreatta and Mason 1985; Tamir 2001) is to minimize the average of the k largest distances. When $k=1$, the problem reduces to the 1-center problem. These papers investigate the problem in a network environment. Drezner

and Nickel (2009a, b) are the only references we are familiar with that solve this problem in the plane. Another objective is to minimize the k-th largest objective. This is equivalent to finding the smallest circle that covers a certain proportion of the points. These problems can be solved by employing the ordered median objective (Nickel and Puerto 2005). The ordered median objective is defined by a sequence of constants $\lambda_1, \ldots, \lambda_n$. The objective function is a sum of the distances each multiplied by a λ. The smallest distance is multiplied by λ_1, the second smallest by λ_2, and so on, while the largest distance is multiplied by λ_n. Many location problems can be formulated as ordered median problems. The 1-center problem is defined by the vector (0, 0, …, 0, 1). The k-centrum problem is defined by $n-k$ zeroes followed by k ones. The k-th largest objective is defined by all zeroes except for a "1" at the position $n-k$. Drezner and Nickel (2009a, b) suggested a general and efficient approach for optimally solving any single-facility ordered median problem. This general approach can be applied for solving these problems.

4.5.1 Location of Multiple Facilities

There are two main approaches to modeling the location of multiple facilities. The first type of models are called multi-facility location models. Here, the distances between the facilities are included in the objective function, and it is assumed that some interaction exists between facilities. Most papers in this area deal with the minisum objective. Miehle (1958) was the first to define this problem in the minisum context. In the minimax context, the first paper is by Love et al. (1973) using Euclidean distances, and Dearing and Francis (1974) using rectilinear distances. Other early papers on the subject are Drezner and Wesolowsky (1978) for rectilinear, Euclidean, and general ℓ_p distances, and Elzinga et al. (1976) for Euclidean distances.

The second type of models for the location of multiple facilities is the p-center location problem. In this formulation each demand point is serviced by its closest facility and there is no interaction between facilities. The set of demand point is allocated among the facilities and each facility is located at the optimum location for its service set. In early years this problem was termed the location-allocation problem. The minisum version of this problem (p-median) was first suggested by Cooper (1963, 1964). The first reference to the p-center problem is Kariv and Hakimi (1979) in a network environment. The first paper on the p-center problem in the plane is Chen (1983). Other early papers are Drezner (1984a, b) and Vijay (1985). Drezner (1984a) proposes heuristic algorithms that are improvements on the basic ideas suggested by Cooper (1963, 1964). The optimal algorithm proposed in Drezner (1984a) is based on creating a set covering problem and solving it. A solution to a 1-center problem is based on 2 or 3 points. Therefore, all candidate locations are the solutions to the $n(n^2-1)/6$ possible pairs, and triplets of demand points. Many of these solutions can be eliminated from the set if they are inferior to another set. For example, all triplets forming right or obtuse triangles

can be eliminated because the two vertices forming the longest side of the triangle are superior.

Drezner (1984b) develops optimal algorithms for the 2-median and 2-center problems in the plane using Euclidean distances. The algorithms are based on the observation that once the facilities are located, the demand points are separated by the perpendicular bisector to the segment joining the two facilities. Therefore, there are at most $n(n-1)/2$ partitions of the set of demand points and the algorithm evaluates all these partitions in a branch and bound algorithm. Vijay (1985) solves the p-center problem by solving a sequence of set covering problems assuming a given radius. Once the radius is given, a set covering problem is created and solved by a zero-one integer programming code. A binary search on the value of the radius determines the optimal solution to the p-center problem. The computational results reported in Vijay (1985) are excellent for the era of 1985 computers. A recent paper by Chen and Chen (2009) provides the best available techniques to date to solve p-center problems either in discrete space or in the plane.

The p-center problem when demand is generated in an area (defined by a density function depicting the population) is proposed by Suzuki and Drezner (1996). The actual density of demand does not affect the solution as long as the demand is positive at all the points in the area. The problem is equivalent to covering the area with p circles of the minimum possible radius. Suzuki and Drezner (1996) specifically solved the problem to cover a square area, but some of the configurations are surprising in that they resemble hexagonal patterns rather than a grid pattern, see Fig. 4.3 for the configuration for $p=30$. For $p=1$, the best location is clearly at the center of the square. The solution for $p=4$ is to locate the four facilities at the cen-

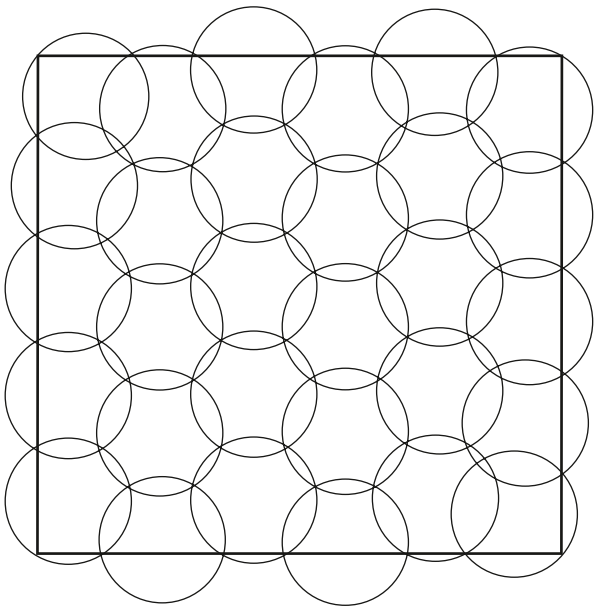


Fig. 4.3 Solution to the 30-center problem

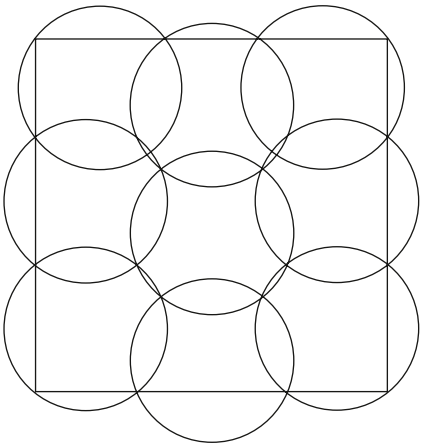


Fig. 4.4 A Solution to the continuous 9-center problem

ters of the four small squares obtained by dividing the square vertically and horizontally in the middle. One might expect that the solution for $p=9$ is to locate the nine facilities in three rows, with three facilities in each row forming a three by three grid, but such a solution is not optimal. The radius required for full coverage by this solution is $\frac{\sqrt{2}}{6} \approx 0.23570$. However, a solution of 0.23064 does exist (see Fig. 4.4). By pushing the middle column down and adjusting the left and right columns, a smaller radius is required for a full coverage of the square. There is a large body of literature about circle packing and circle covering. The p-center problem in an area is called the "circle covering" problem, while the continuous p-dispersion problem is called the "circle packing" problem. The latter problem was solved by Drezner and Erkut (1995) and Maranas et al. (1995). Both papers applied a nonconvex mathematical programming formulation to heuristically solve the problem).

Researchers in the fields of covering and packing are generally not familiar with the location literature and were surprised to find out that their research is done in parallel by location researchers. A recent book on circle packing is Szabó et al. (2007), and http://www.stetson.edu/~efriedma/packing.html is an excellent website that depicts solutions to many continuous problems. The solution for the p-center with $p=9$ is attributed to Nurmela and Östergård (2000). The authors (and the creators of the website) are not aware of the previous results in Suzuki and Drezner (1996).

4.5.2 Conditional Problems

When several facilities exist in the area and p new facilities need to be located, the problem is called the conditional p-center problem (and, for the minisum objective, the conditional p-median problem). Each point is assigned to the closest facility, whether existing or new. The objective of the conditional p-center problem

remains to minimize the maximum distance. The problem is mainly investigated in a network environment (Minieka 1980), and the problem using Euclidean distances was first suggested by Chen (1988). Drezner (1989) showed that any conditional *p*-center problem in any environment (plane, network, globe, or *k*-dimensional space) can be solved by a sequence of $O(\log n)$ *p*-center problems. The following observation can be used to determine whether or not there is a solution whose longest customer-facility distance is $\leq C$ for any given C: first, all customer points for which there is a distance no larger than C to an existing facility, can be eliminated from consideration. The *p*-center based only on points for which no distance $\leq C$ to an existing facility exists, is solved. If the solution to the *p*-center problem is $\leq C$, a solution to the conditional *p*-center $\leq C$ exists, otherwise, it does not. A binary search is conducted as follows: the points are ranked in decreasing order of their distance (can be weighted) to the closest existing facility, $D_1 \geq D_2 \geq \ldots \geq D_n$. A binary search for this vector of distances on the index $1 \leq r \leq n$ using the rule above requires $O(\log n)$ solutions of *p*-center problems.

Berman and Drezner (2008) showed that the conditional *p*-center or *p*-median problem in a network environment can be solved by solving one *p*-center (or *p*-median) problem. They demonstrated that the distance matrix can be modified by incorporating distances to existing facilities leading to an equivalent *p*-center problem.

4.6 Conclusions

Most location problems are interesting mathematical problems and thus were of interest to mathematicians for many years. Therefore, facility location is the oldest topic of Operations Research. The basic Weber problem (finding a point that minimizes the weighted sum of distances from a given set of points) was posed by the famous French mathematician Pierre de Fermat in the early 1600s (Wesolowsky 1993; Drezner et al. 2002). Fermat posed the question of finding a point such that the sum of the distances to three given points is minimum. The great English mathematician Sylvester (1857) posed the question of the smallest circle enclosing a set of points which is the one-center problem and proposed a solution approach in Sylvester (1860). The one-center problem is sometimes called the minimax facility location problem because the solution is the point that minimizes the maximum distance to a set of points. The solution method was actually an algorithm which is very unusual for solving mathematical problems. Chrystal (1885) that solved the problem in a similar way was concerned with the limit on the number of iterations of the algorithm, probably the first instance of complexity analysis.

In more recent years research recognized applications to these basic models in the context of Operations Research. Weber (1909) recognized that the minisum model is applicable to the location of industries such as warehouses. The one-center model was recognized as the suitable model for the location of an emergency facility because we would like to minimize the distance to the farthest facility. These

models were further extended in many ways following their use in solving different applications. The most common extension is the location of multiple facilities which is extensively investigated in recent years.

References

Andreatta G, Mason FM (1985) Properties of the *k*-centrum in a network. Networks 15:21–25
Berman O, Drezner Z (2008) A new formulation for the conditional *p*-median and *p*-center Problems. Oper Res Lett 36:481–483
Beyer HW (1981) Standard mathematical tables. CRC Press, Boca Raton, FL
Chen R (1983) Solution of minisum and minimax location-allocation problems with Euclidean distances. Nav Res Logist Q 30:449–459
Chen R (1988) Conditional minisum and minimax location-allocation problems in Euclidean space. Transp Sci 22:158–160
Chen D, Chen R (2009) New relaxation-based algorithms for the optimal solution of the continuous and discrete *p*-center problems. Comput Oper Res 36:1646–1655
Chrystal G (1885) On the problem to construct the minimum circle enclosing n given points in a plane. Proceedings of the Edinburgh Mathematical Society, third meeting, January 9th, pp 30–33
Cooper L (1963) Location-allocation problems. Oper Res 11:331–343
Cooper L (1964) Heuristic methods for location-allocation problems. SIAM Rev 6:37–53
Dearing PM (1977) Minimax location and Voronoi diagrams. Paper presented at the Joint ORSA/TIMS meeting, Atlanta
Dearing PM, Francis RL (1974) A network flow solution to a multifacility minimax location problem involving rectilinear distances. Transp Sci 8:126–141
Drezner Z (1982) On minimax optimization problems. Math Program 22:227–230
Drezner Z (1984a) The *p*-center problem—heuristic and optimal algorithms. J Oper Res Soc 35:741–748
Drezner Z (1984b) The planar two-center and two-median problems. Transp Sci 18:351–361
Drezner Z (1987) On the rectangular *p*-center problem. Nav Res Logist Q 34:229–234
Drezner Z (1989) Conditional *p*-center problems. Transp Sci 23:51–53
Drezner Z (1991) The weighted minimax location problem with set-up costs and extensions. RAIRO—Oper Res 25:55–64
Drezner Z, Erkut E (1995) Solving the continuous p-dispersion problem using non-linear programming. J Oper Res Soc 46:516–520
Drezner Z, Nickel S (2009a) Solving the ordered one-median problem in the plane. Eur J Oper Res 195:46–61
Drezner Z, Nickel S (2009b) Constructing a DC decomposition for ordered median problems. J Global Optim 45:187–201
Drezner Z, Shelah S (1987) On the complexity of the Elzinga-Hearn algorithm for the one-center problem. Math Oper Res 12:255–261
Drezner Z, Wesolowsky GO (1978) A new method for the multifacility minimax location problem. J Oper Res Soc 29:1095–1101
Drezner Z, Wesolowsky GO (1980) Single facility ℓ_p distance minimax location. SIAM J Algebra Discrete Method 1:315–321
Drezner Z, Klamroth K, Schöbel A, Wesolowsky GO (2002) The Weber problem. In: Drezner Z, Hamacher H (eds) Facility location: applications and theory. Springer, Berlin, pp 1–36
Dyer ME (1986) On a multidimensional search technique and its application to the Euclidean one-centre problem. SIAM J Comput 15:725–738

Elzinga DJ, Hearn DW (1972a) Geometrical solutions for some minimax location problems. Transp Sci 6:379–394
Elzinga J, Hearn DW (1972b) The minimum covering sphere problem. Manag Sci 19:96–104
Elzinga DJ, Hearn DW, Randolph WD (1976) Minimax multifacility location with Euclidean distances. Transp Sci 10:321–336
Graham RL (1972) An efficient algorithm for determining the convex hull of a planar set. Inf Process Lett 1:132–133
Graham RL, Yao FF (1983) Finding the convex hull of a simple polygon. J Algorithm 4:324–331
Hearn DW, Vijay J (1982) Efficient algorithms for the (weighted) minimum circle problem. Oper Res 30:777–795
Jacobsen SK (1981) An algorithm for the minimax Weber problem. Eur J Oper Res 6:144–148
Kariv O, Hakimi SL (1979) An algorithmic approach to network location problems. Part I: the *p*-centers. SIAM J Appl Math 37:513–538
Love RF, Wesolowsky GO, Kraemer SA (1973) A multifacility minimax location method for Euclidean distances. Int J Prod Res 11:37–46
Maranas CD, Floudas CA, Pardalos PM (1995) New results in the packing of equal circles in a square. Discrete Math 142:287–293
Megiddo N (1983a) Linear-time algorithms for linear programming in $\mathbb{R}^3$ and related problems. SIAM J Comput 12:759–776
Miehle W (1958) Link-length minimization in networks. Oper Res 6:232–243
Megiddo N (1983b) The weighted Euclidean 1-center problem. Math Oper Res 8:498–504
Minieka E (1980) Conditional centers and medians of a graph. Networks 10:265–272
Nickel S, Puerto J (2005) Location theory: a unified approach. Springer, Berlin
Nurmela KJ, Östergård PRJ (2000) Covering a square with up to 30 equal circles. Working paper, Helsinki University of Technology, Laboratory for Theoretical Computer Science, Helsinki
Okabe A, Boots B, Sugihara K, Chiu SN (2000) Spatial tessellations: concepts and applications of Voronoi diagrams. Wiley, Chichester
Radon J (1921) Mengen konvexer Körper, die einen gemeinsamen Punkt enthalten. Math Ann 83:113–115
Shamos MI, Hoey D (1975) Closest-point problems. In: Proceedings of the 16th annual IEEE symposium on foundations of computer science. IEEE Computer Society Press, Los Angeles, pp 151–162
Suzuki A, Drezner Z (1996) The *p*-center location problem in an area. Locat Sci 4:69–82
Suzuki A, Okabe A (1995) Using Voronoi diagrams. In Drezner Z (ed) Facility location: a survey of applications and methods. Springer, New York
Sylvester JJ (1857) A question in the geometry of situation. Q J Math 1:79
Sylvester JJ (1860) On Poncelet's approximate valuation of surd forms. Philos Mag 20:203–222 (Fourth Series)
Szabó PG, Markót MC, Csendes T, Specht E, Casado LG, García I (2007) New approaches to circle packing in a square. Springer, New York
Tamir A (2001) The *k*-centrum multi-facility location problem. Discrete Appl Math 109:293–307
Vijay J (1985) An algorithm for the *p*-center problem in the plane. Transp Sci 19:235–245
Weber A (1909) Über den Standort der Industrien. Erster Teil: Reine Theorie der Standorte. Mit einem mathematischen Anhang von G Pick (in German). JCB Mohr, Tübingen
Wesolowsky GO (1993) The Weber problem: history and perspectives. Locat Sci 1:5–23
Zemel E (1983) A linear time randomizing algorithm for local roots and optima of ranked functions. JL Kellog Graduate School of Management, Northwestern University, Evanston

Chapter 5
Discrete Center Problems

Barbaros Ç. Tansel

5.1 Introduction

Our focus in this chapter is on discrete center location problems. This class of problems involves locating one or more facilities on a network to service a set of demand points at known locations in such a way that every demand receives its service from a closest facility, and the maximum distance between a demand and a closest facility is as small as possible. This leads to a minimax type of objective function, which is intrinsically different from the minisum objective that is more widely encountered in location models, for which the primary concern is to minimize the total transportation cost. The term *discrete* in the title refers to a finite set of demand points, while *continuous* versions of center location problems are also possible if the set of demand points to be served constitutes a continuum of points on the network under consideration.

Center location problems most commonly arise in emergency service location, where the concern for saving human life is far more important than any transportation costs that may be incurred in providing that service. Consider, for example, locating a fire station to serve a number of communities interconnected by a road network. If a fire breaks out in any one of these communities, it is crucial for equipment to arrive at the fire as quickly as possible. Similarly, quick delivery of an emergency service is significantly more important in optimally placing, for example, ambulances and police patrol units, than the cost of delivering that service. The common denominator in all of these circumstances is that there is a time delay between the call for service and the actual time of beginning to provide that service that is a direct consequence of the time spent during transportation. All other factors being constant, it makes sense to model such circumstances so that the maximum distance traversed during transportation is as small as possible.

B. Ç. Tansel (✉)
Department of Industrial Engineering, Bilkent University, 6800 Bilkent, Ankara, Turkey
e-mail: barbaros@bilkent.edu.tr

H. A. Eiselt, V. Marianov (eds.), *Foundations of Location Analysis,*
International Series in Operations Research & Management Science 155,
DOI 10.1007/978-1-4419-7572-0_5,

5.1.1 The Single Facility Case: The Absolute Center Problem

To define the center location problem, let us first consider the single facility problem that involves optimally placing an emergency service facility on a road network that interconnects n communities requiring the services of the facility. It is convenient to represent the road network of interest by an undirected connected network $G=(V', E)$ with vertex set $V' = \{v_1, \ldots, v_n, \ldots, v_{n'}\}$ and edge set E consisting of undirected edges of the form $e_{ij}=[v_i, v_j]$ with edge lengths $L_{ij}>0$. Without loss of generality, we assume that the vertex set includes the $n \leq n'$ communities requiring the services of the facility. We further assume, with re-indexing if necessary, that the first n vertices are the vertices that demand service from the facility. Let $V=\{v_1,\ldots, v_n\} \subseteq V'$ be the demand set. Vertices not in V, if any, may represent, for example, intersections of roads. Edges represent road segments connecting pairs of vertices, and their lengths are positive. We take each edge of the network as an infinite set of points (a continuum) connecting the end-vertices of the edge under consideration and refer to each point along an edge as an *interior* point of that edge if the point is not one of the end-vertices. We take the network G as the union of its edges and write $x \in G$ to mean x is any point along any edge of G.

For any pair of vertices v_i and v_j in the network, a path $P=P(v_i, v_j)$ connecting v_i and v_j is a sequence of alternating vertices and edges that begin at v_i and end at v_j. We define the *length* of a path P to be the sum of the lengths of the edges contained in the path. A *shortest path* connecting v_i and v_j, denoted by $SP(v_i, v_j)$, is a path whose length is the smallest among all paths connecting v_i and v_j. Due to the positivity of edge lengths, every shortest path between a pair of vertices is a simple path; meaning no vertex in the path is repeated. In general, there may be many shortest paths between a pair of vertices, each having the same length. We define $d_{ij}=d(v_i, v_j)$ to be the length of a shortest path connecting v_i and v_j, and refer to d_{ij} as the *distance* between v_i and v_j. Vertex-to-vertex distances are computed via well known all-pairs shortest path algorithms, see, e.g., Floyd (1962) or Dantzig (1967). We extend the definition of the shortest path distance to *any* pair of points $x, y \in G$, vertex or not, by defining the length of a path to be the sum of lengths of edges and subedges contained in the path and defining $d(x, y)$ to be the length of a shortest path connecting x and y. The function $d(\bullet, \bullet)$ satisfies the properties of *nonnegativity*, *symmetry*, and *triangle inequality* which are as follows:

$$\forall x, y \in G,$$

- *Nonnegativity:* $d(x, y) \geq 0$; $d(x, y)=0$ iff $x=y$;
- *Symmetry:* $d(x, y)=d(y, x)$;
- *Triangle Inequality:* $d(x, y) \leq d(x, u)+d(u, y) \; \forall \, u \in G$.

The single facility center location problem is referred to as the *Absolute Center Problem,* a term coined by Hakimi (1964) who introduced this problem to the literature. To define the problem, we associate nonnegative constants w_i and a_i with

each vertex v_i, $i=1,\ldots, n$. We refer to each w_i as a *weight* and each a_i as an *addend*. Vertex weights are used as scaling factors to assign relative values of importance to demand vertices based, for example, on population densities. A vertex representing a densely populated business district during work hours may require a more amplified protection against emergency than a vertex representing a rather sparsely populated residential area. Such differences may be reflected into the model by a judicious choice of weights. The addend a_i can be interpreted as preparation time for a fire-fighting squad to get the equipment ready to work at v_i. This preparation time depends in general on the local conditions at a vertex (including access to a fire hydrant, space available for fire engines to position themselves), so that having different addends at different vertices is meaningful. For ambulance services, we may interpret a_i as the time spent transporting the patient from v_i to the closest hospital. If hospital locations are known, this transportation time is a fixed constant that depends only upon the vertex under consideration and a hospital closest to that vertex.

Given $w_i, a_i \geq 0$ $(i=1,\ldots, n)$, define the function f for every $x \in G$ by

$$f(x) = \max\{w_i d(x, v_i) + a_i : i = 1, \ldots, n\} \tag{5.1}$$

and consider the optimization problem

$$r_1 \equiv \min\{f(x) : x \in G\}. \tag{5.2}$$

Any point $x^* \in G$ that solves (5.2) is referred to as an *absolute center* of G, and the minimum objective value r_1 is referred to as the *1-radius* of G. If x is restricted to V in (5.1) and (5.2), the resulting problem is called the *vertex-restricted problem*, and its solution is referred to as a *vertex-restricted center*. If the demand set V in relation (5.1) is replaced by the continuum of all points in G, then the definition of $f(\bullet)$ becomes $f(x)=\max\{d(x, y): y \in G\}$ and any point in G that minimizes this function is referred to as a *continuous center* (see Frank 1967). A different continuous demand version of the center problem is also formulated by Minieka (1977). In his formulation, the objective is to minimize the maximum distance from the facility to a farthest point on each edge. A point in G that minimizes this objective function is referred to as a *general center*. Our focus in this chapter is on the absolute center problem. The continuous and general center problems are briefly discussed in Sect. 5.4.

The absolute center problem is referred to as the *weighted problem* if at least one of the weights is different from one and the *weighted problem with addends* if, additionally, at least one addend is nonzero. The case with $w_i = 1 \; \forall \, i \in I \equiv \{1, \ldots, n\}$ is referred to as the *unweighted problem* or the *unweighted problem with addends,* respectively, depending on if all a_i or not all a_i are zero.

In the unweighted case, the definition of $f(x)$ becomes $f(x)=\min\{d(x, v_i): x \in G\}$ so that $f(x)$ identifies a farthest community and its distance from a facility at x. With $d(x, v_i) \leq f(x) \; \forall \, i \in I$, all communities are covered within a distance of $f(x)$, while there is at least one community whose distance from x is exactly $f(x)$. The optimization in (5.2) seeks to place the facility in such a way that the farthest distance from

it to any community is as small as possible. If x^* achieves this, then $f(x^*)$ supplies the value r_1, which is the smallest possible coverage radius from a facility anywhere on the network. Generally, with weights and addends, each community v_i is covered by a facility at x within a distance of $[f(x)-a_i]/w_i$, while at least one community achieves this bound.

5.1.2 The Multi-facility Case: The Absolute p-Center Problem

Multiple facilities are needed in emergency service location when a single facility is not enough to cover all communities within acceptable distance limits. To model the multi-facility version of the problem, let p be a positive integer representing the number of facilities to be placed on the network. Assume that the p facilities under consideration are identical in their service characteristics and that each is uncapacitated so that communities are indifferent as to which particular facility they receive their services from (provided that the service is given in the quickest possible way). Accordingly, if $x_1,\ldots, x_p$ are the locations of the p facilities, then each community prefers to receive its service from the facility closest to it.

Let $X=\{x_1,\ldots,x_p\}$ and define $D(X, v_i)$ to be the distance of vertex v_i to a nearest element of the point set X. That is,

$$D(X, v_i) = \min\left\{d(x_1, v_i), \ldots, d(x_p, v_i)\right\}. \tag{5.3}$$

Let $S_p(G)$ be the family of point sets X in G such that $|X|=p$. Hence, $X\in S_p(G)$ implies $X=\{x_1,\ldots, x_p\}$ for some choice of p distinct points $x_1,\ldots, x_p$ of G. We extend the definition of $f(x)$ to the multi-facility case as follows: For each $X\in S_p(G)$, define

$$f(X) = \max\{w_i D(X, v_i) + a_i: i = 1, \ldots, n\}. \tag{5.4}$$

The definition in (5.4) reduces to definition (5.2) for the case of $p=1$.

The *Absolute p-Center Problem*, introduced by Hakimi (1965), is the problem of finding a point set $X^*\in S_p(G)$ such that

$$r_p \equiv f\left(X^*\right) = \min\{f(X): X \in S_p(G)\}. \tag{5.5}$$

Any point set $X^*=\{x_1^*,\ldots, x_p^*\}\in S_p(G)$ that solves (5.5) is called an *absolute p-center of G* and each location x_j^* in X^* is referred to as a *center*. The minimum objective value r_p is called the *p-radius* of G. If X is restricted to p-element subsets of V, the resulting problem is referred to as a *vertex restricted p-center problem* and its solution is called a *vertex-restricted p-center*. If each point in the network is a demand point as opposed only to vertices, the resulting problem is called the *continuous p-center problem.* If the maximum distance to a farthest point in each edge is minimized, the resulting problem is *the general p-center problem.* While the

continuous and general center problems are equivalent for $p=1$, different problems result for $p>1$.

Our focus is on the absolute p-center problem. The definition of $f(X)$ in (5.4) implies that $w_i D(X, v_i)+a_i \leq f(X) \forall i$, so that every community v_i is covered by at least one center in X within a distance of $[f(X)-a_i]/w_i$. Note also that there is at least one community which achieves this bound. The optimization in (5.5) seeks to place the p facilities on the network such that the farthest weighted distance of any community from the nearest facility is as small as possible.

Now that we have a clear idea of the type of location models dealt with in this chapter, we focus next on three classical papers that have had significant impact on the literature in this area of research.

5.2 Three Classical Contributions on Discrete Center Location

We give in this section an overview of three early and fundamental papers that had a significant impact on subsequent research in discrete center location. Each of the Sects. 5.2.1, 5.2.2, and 5.2.3 is devoted to one of these papers. The first work that we investigate is the contribution by Hakimi (1964). This is a seminal paper in that it has led to a whole new area of research that we know of today as *network location*. Hakimi poses two problems in his paper, assuming nonnegative weights and zero addends, and calls them the *absolute median* and the *absolute center* problems. Both problems are posed on a network whose edges are viewed as continua of points. The objective in the absolute median problem is to minimize the weighted sum of distances from the facility to all vertices, while the objective in the absolute center problem is to minimize the maximum of such distances. Hakimi provides an insightful analysis for both problems. One consequence of his analysis is the well known vertex optimality theorem for the absolute median problem. Hakimi's analysis for the absolute center problem has led to a methodology that relies on identifying local minima on edges by inspecting piece-wise linear functions. Hakimi's paper is investigated in Sect. 5.2.1.

A second classical contribution is a paper by Goldman (1972). In his work, Goldman gives a localization theorem for the absolute center problem that helps to localize the search for an optimal location to a subset of the network whenever the network has a certain exploitable structure. Repeated application of the theorem results in an algorithm that either finds an optimal location or reduces the problem to a single cyclic component of the network. Goldman's paper is examined in Sect. 5.2.2.

Minieka (1970) focuses on the multi-facility case and gives a well conceived solution strategy for the unweighted absolute p-center problem, which relies on solving a sequence of set covering problems. Minieka's method is directly extendible to the weighted version. Minieka's paper is covered in Sect. 5.2.3.

5.2.1 Hakimi (1964): The Absolute Center Problem

Hakimi's paper is historically the first paper that considers the absolute center problem on a network. The vertex-restricted version of the 1-center problem is posed as early as 1869 by Jordan (1869), and is directly solved by evaluating the objective function at each vertex. The absolute center problem, on the other hand, requires an infinite search over the continua of points on edges and calls for a deeper analysis than simple vertex enumeration.

Hakimi viewed each edge as a continuum of points. This marks a significant departure from the traditionally accepted view of classical graph theory that takes each undirected edge as an unordered pair of vertices. The kind of network Hakimi had in mind is what we refer to today as an embedded network where each edge $[v_i, v_j]$ is the image of a one-to-one continuous mapping T_{ij} of the unit interval [0, 1] into some space S (e.g. the plane) such that $T_{ij}(0)=v_i$, $T_{ij}(1)=v_j$, and each point x in the interior of $[v_i, v_j]$ is the image $T_{ij}(\alpha)$ of a real number α in the open interval (0,1). A formal definition of an embedded network can be found in Dearing, Francis, and Lowe (1976); for details, also see Dearing and Francis (1974). For our purposes, it suffices to view the network of interest as an embedding in the plane with vertices corresponding to distinct points and edges corresponding to continuous curves connecting pairs of vertices. We assume that, whenever two edges intersect, they intersect only at a vertex. A point x in edge $[v_i, v_j]$ induces two subedges $[v_i, x]$ and $[x, v_j]$ with $[v_i, x]\cup[x, v_j]=[v_i, v_j]$ and $[v_i, x]\cap[x, v_j]=\{x\}$.

Hakimi observed that the optimization problem $\min\{f(x)\colon x\in G\}$, where $f(x)\equiv\max\{w_i d(x, v_i)\colon i\in I\}$, can be solved by minimizing $f(\bullet)$ on each edge separately and then choosing the best of the edge-restricted minima. This is an immediate consequence of the fact that the graph G is the union of its edges. It then suffices to develop a solution procedure for the edge restricted problem.

Let $e=[v_p, v_q]$ be an edge of the network. The *edge restricted problem* regarding this edge can then be written as

$$\text{Min}\{f(\text{x})\colon x\in e\}. \tag{5.6}$$

Let $L=L_{pq}$ be the length of the edge e. Observe that as x varies in the edge $\boldsymbol{e}$, the length of the subedge $[v_p, x]$ varies in the interval $[0, L]$. If we denote by x_λ the unique point x in e for which the subedge $[v_p, x]$ has length λ, then we may redefine the edge restricted problem in the equivalent form

$$\text{Min}\{f(x_\lambda)\colon \lambda\in[0, L]\}. \tag{5.7}$$

The form defined by (5.7) is particularly useful for analyzing the structure of $f(\bullet)$ as a function of the real variable λ. We begin the analysis of $f(\bullet)$ by first examining the distance $d(x_\lambda, v_i)$ for a fixed vertex v_i as λ varies in the interval $[0, L]$. Define the function g_i by $g_i(\lambda)=d(x_\lambda, v_i)\,\forall\,\lambda\in[0, L]$. Observe that a shortest path from an interior point x_λ to vertex v_i must include either the subedge $[v_p, x_\lambda]$ or the subedge $[x_\lambda, v_q]$. Accordingly, $SP(x_\lambda, v_i)$ is either $[v_p, x_\lambda]\cup SP(v_p, v_i)$ or $[x_\lambda, v_q]\cup SP(v_q, v_i)$. Figure 5.1 illustrates these two possibilities. It follows that $g_i(\lambda)$ is the minimum of

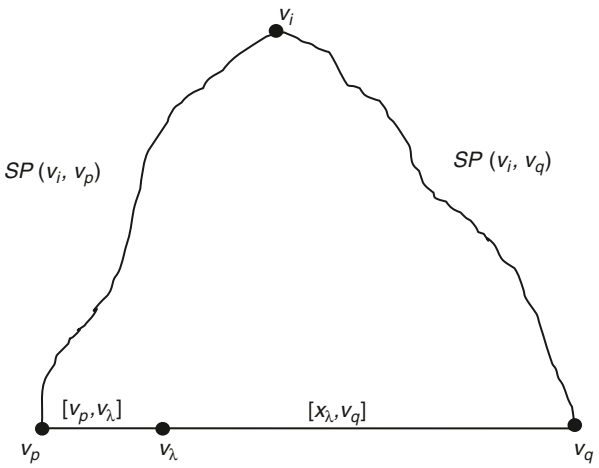


Fig. 5.1 Illustration of shortest path connecting v_i and x_λ

the two path lengths $\lambda+d_{pi}$ and $L-\lambda+d_{qi}$ corresponding, respectively, to the paths $[v_p, x_\lambda]\cup SP(v_p, v_i)$ and $[x_\lambda, v_q]\cup SP(v_q, v_i)$. Accordingly, we have

$$g_i(\lambda) = \min\{\lambda + d_{pi}, L - \lambda + d_{qi}\} \forall \lambda \in [0, L]\,. \tag{5.8}$$

Observe that all quantities in the right side of (5.8) are constants except λ. With this observation, $g_i(\lambda)$ is the pointwise minimum of the two linear functions $\lambda+d_{pi}$ and $L-\lambda+d_{qi}$ in the interval $[0, L]$. The fact that the distance from a fixed vertex v_i to a variable point in a given edge is the pointwise minimum of two linear functions is a key element, observed by Hakimi, that has led to a well-structured theory and solution method.

In general, the pointwise minimum of a finite number of linear functions is a concave piecewise linear function that has, at most, as many pieces as there are linear functions under consideration. Figure 5.2 illustrates a concave piece-wise linear function $h(x)$ that consists of 4 pieces.

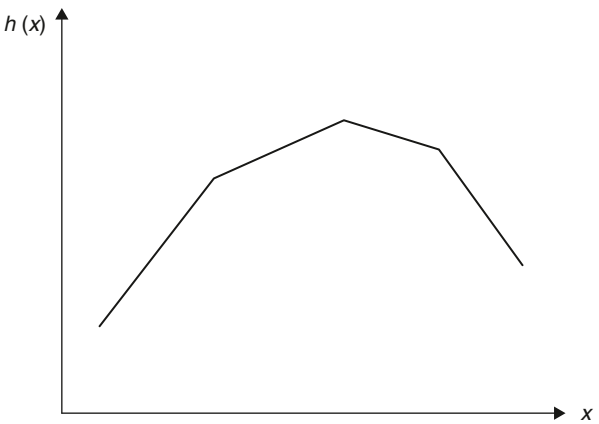


Fig. 5.2 A concave piecewise linear function $h(x)$

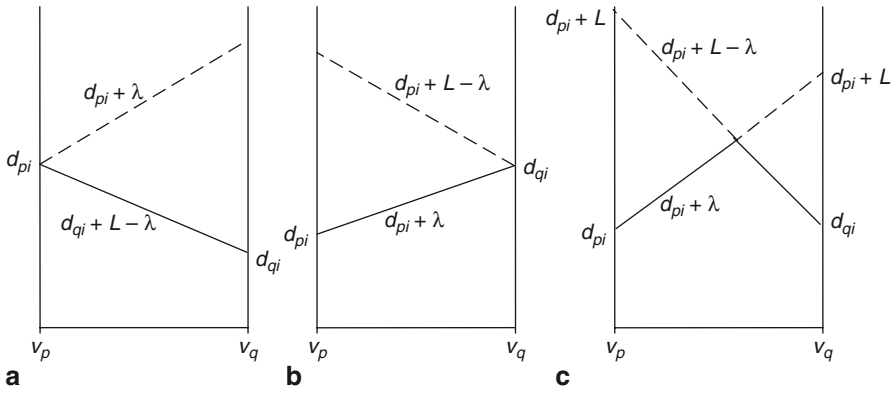


Fig. 5.3 Three possible forms of the function $g_i(\lambda) = \min\{d_{pi}+\lambda, d_{qi}+L-\lambda\}$. **a** Decreasing. **b** Increasing. **c** Two-piece

In our case, g_i is the minimum of two linear functions, so it is either linear or two-piece linear. If g_i is linear, then it is an *increasing* linear function with a slope of +1 whenever $\lambda + d_{pi} \le L - \lambda + d_{qi}\ \forall \lambda \in [0, L]$, while it is a *decreasing* linear function with slope of −1 if the reverse inequality holds. If g_i is a two-piece linear function, then the two linear functions of interest attain the same value at some interior point λ' of the interval $[0, L]$, so that the linear piece in the subinterval $[0, \lambda']$ is increasing while the linear piece in the subinterval $[\lambda', L]$ is decreasing. Figure 5.3 illustrates the three possible forms of g_i. Note that the two linear functions $\lambda + d_{pi}$ and $L - \lambda + d_{qi}$ always intersect at an end-vertex if they do not intersect at an interior point. For example, they intersect at v_p in Fig. 5.3a and at v_q in Fig. 5.3b. In the case of Fig. 5.3a, the linear function $d_{qi} + L - \lambda$ is the smaller of the two linear functions over the entire edge so that $d_{qi} + L \le d_{pi}$ at $\lambda = 0$. However, d_{pi} is the shortest path length between v_i and v_p while $d_{qi} + L$ is the length of a path connecting v_i and v_p via v_q. This implies that $d_{pi} \le d_{qi} + L$. The two inequalities result in the equality $d_{qi} + L = d_{pi}$.

Consider now the weighted distance $w_i d(x_\lambda, v_i) = w_i g_i(\lambda)$ as λ varies in $[0, L]$. Since w_i is positive and g_i is the minimum of two linear functions with slopes ±1, $w_i g_i(\bullet)$ is a concave piecewise linear function with at most two pieces and with slopes of $\pm w_i$. The only difference from the previous case is that the slopes are now $\pm w_i$ rather than ±1.

Let us now focus on the analysis of the function $f(\bullet)$ on edge e. By definition, $f(x)$ is the maximum over $i \in I$ of the weighted distances $w_i d(x, v_i)$. Using again the variable point $x_\lambda \in e$ as λ varies in $[0, L]$, we have

$$f(x_\lambda) = \max\{w_i g_i(\lambda) : i \in I\}. \tag{5.9}$$

Since each $w_i g_i(\bullet)$ is a concave piecewise linear function with at most two pieces, $f(\bullet)$ is the pointwise maximum of n such functions. Accordingly, the restriction of

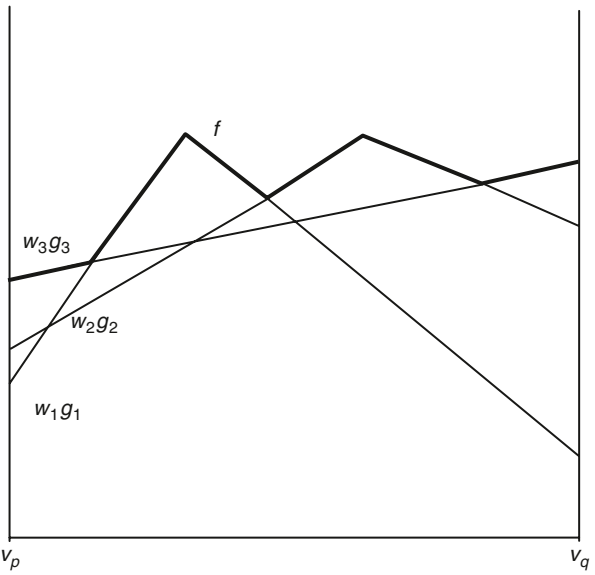


Fig. 5.4 $f(\bullet)$ as the maximum of three concave two-piece linear functions w_ig_i

$f(\bullet)$ to an edge results in a piecewise linear function. Figure 5.4 illustrates this for the case of $n=3$. In general, the maximum of concave functions is not concave, so the only exploitable property of $f(\bullet)$ is piecewise linearity. It is quite clear that the minimum of a piecewise linear function on a closed interval either occurs at a break point or at an end point of the interval. Hakimi's method searches for break points where the slopes to the left and to the right of the point are oppositely signed. The functions $w_ig_i(\bullet)$ are plotted for $i \in I$ on each edge and $f(\bullet)$ is constructed by taking their pointwise maximum. The minimum of $f(\bullet)$ on a given edge is found by inspecting the qualifying break-points of the resulting graph.

Hakimi demonstrates his method on a network with six vertices and eight edges. We reproduce his network from Hakimi (1964) in Fig. 5.5. The edge lengths are shown next to the edges. The vertex-to-vertex distances are shown in Table 5.1.

Let the edges be numbered $e_1,\ldots,e_8$ as shown in Fig. 5.5. The plots of the functions $g_i(\bullet)$ and $f(\bullet)$ on each edge e_j are shown in Fig. 5.6, assuming that all vertex weights are equal to one. The plots of $f(\bullet)$ are in bold. The edge-restricted optimum on edge $e_1=[v_6, v_5]$ shown in Fig. 5.6a is at point x_1, at a distance of 1.5 from v_6 with $f(x_1)=5.5$. For edge $e_2=[v_5, v_3]$ shown in Fig. 5.6b, there are two local optima, one at v_5 and the other at v_3, with $f(v_5)=f(v_3)=6$. For edge $e_3=[v_1, v_6]$ shown in Fig. 5.6c, there is an edge restricted optimum at point x_3, which is at a distance of 2.5 units from v_1 with $f(x_3)=5.5$.

For edge $e_4=[v_1, v_4]$ shown in Fig. 5.6d, there are two edge-restricted optima, one at v_1 and the other at point x_4, at a distance of 2 units from v_1 with $f(v_1)=f(x_4)=6$. For edge $e_5=[v_1, v_2]$ shown in Fig. 5.6e, there are two edge restricted optima, one at v_1 and the other at x_5, at a distance of 2 units from v_1 with $f(v_1)=f(x_5)=6$. For edge $e_6=[v_2, v_4]$ shown in Fig. 5.6f, the two edge-restricted optima are at end-vertices v_2

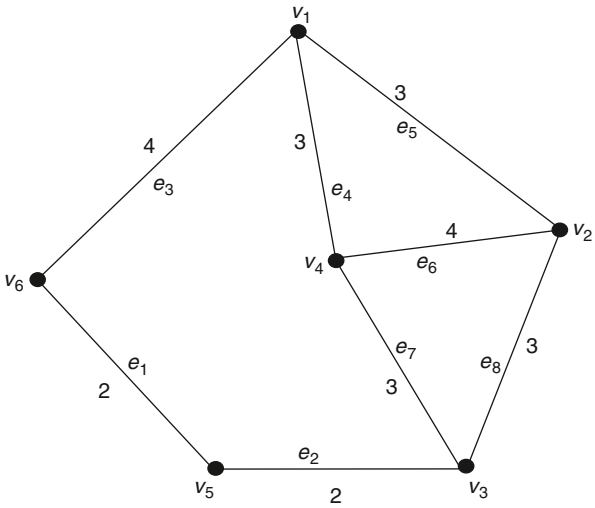


Fig. 5.5 An illustrative network. (Taken from Hakimi 1964)

Table 5.1 Vertex-to-vertex distances for the example network

	v_1	v_2	v_3	v_4	v_5	v_6
v_1	0	3	6	3	6	4
v_2	3	0	3	4	5	7
v_3	6	3	0	3	2	4
v_4	3	4	3	0	5	7
v_5	6	5	2	5	0	2
v_6	4	7	4	7	2	0

and v_4 with $f(v_2)=f(v_4)=7$. For edge $e_7=[v_3, v_4]$ shown in Fig. 5.6g, the edge-restricted optimum is at point x_7, at a distance of 1 unit from v_3 with $f(x_7)=5$. Finally, for edge $e_8=[v_3, v_2]$ shown in Fig. 5.6h, the edge-restricted optimum is at point x_8, at a distance of 1 unit from v_3 with $f(x_8)=5$. Accordingly, there are two absolute centers for the network of Fig. 5.5, one at x_7 and the other at x_8 with $f(x_7)=f(x_8)=5$.

5.2.2 *Goldman (1972): A Localization Theorem for the Absolute Center*

In this section, we continue with a localization theorem for the absolute center problem studied by Goldman (1972). Goldman's localization theorem for the absolute center problem is motivated by a similar localization theorem introduced earlier by Goldman (1971) for the absolute median problem. Goldman's earlier result for the median problem led to a very efficient tree-trimming algorithm for computing optimal medians of tree networks. His result for the absolute center problem is similarly

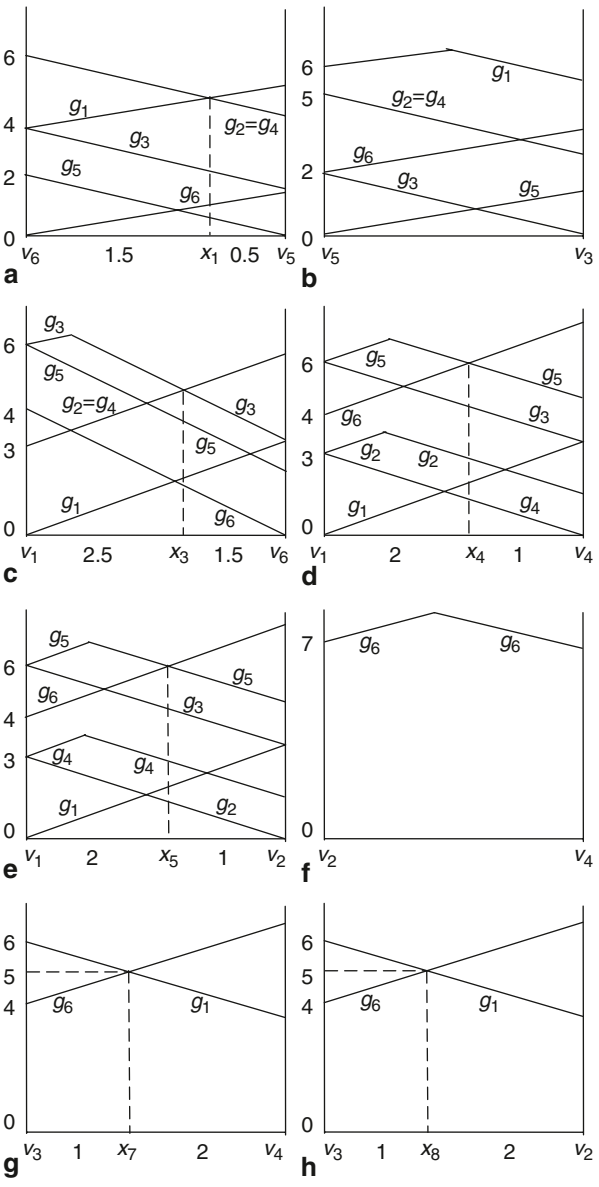


Fig. 5.6 Determining local centers of edges of the network shown in Fig. 5.5

structured and either finds an optimum solution or reduces the problem to a cyclic component of the network.

To begin the analysis, consider the *unweighted* absolute center problem *with addends* on a network $G=(V', E)$. We assume again the first n vertices in V' are the demand vertices and constitute the demand set V. For any point $x \in G$, the objective

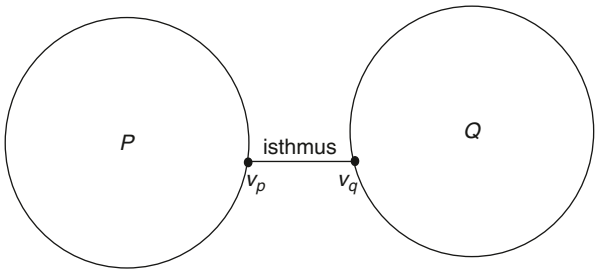


Fig. 5.7 An isthmus $[v_p, v_q]$ with subnetworks P and Q

function is defined by $f(x) \equiv \max\{a_i + d(v_i, x): i \in I\}$, and the objective is to find a point $x^* \in G$ for which $f(x^*) \leq f(x) \; \forall \, x \in G$.

Goldman's localization theorem works best in networks that have edges that are not contained in any simple cycles. Goldman refers to any such edge as an "isthmus." An *isthmus* of G is an edge $[v_p, v_q]$ whereby deleting the interior of this edge results in two disconnected components P and Q. Here, we assume that v_p is in P and v_q is in Q. Figure 5.7 illustrates the definition. An isthmus cannot be contained in any simple cycle of G, otherwise there is a path from a vertex in P to a vertex in Q that does not pass through the edge $[v_p, v_q]$. This, of course, implies that deleting the interior of the edge $[v_p, v_q]$ does not result in two disconnected subsets of G.

Consider an isthmus $e = [v_p, v_q]$ and the associated components P and Q of G where $P \cup e \cup Q = G$, $P \cap e = \{v_p\}$, $Q \cap e = \{v_q\}$, and $P \cap Q = \varnothing$. Let v_i and v_j be a pair of vertices with $v_i \in P$ and $v_j \in Q$. All paths connecting v_i and v_j pass through e so that $d_{ij} = d_{ip} + L + d_{qj}$, where $L \equiv L_{pq}$ is the length of e. Consider a variable point x_λ that moves from v_p to v_q along the edge e as λ varies in the interval $[0, L]$. With λ being the length of the subedge $[v_p, x_\lambda]$ and $L - \lambda$ being the length of the subedge $[x_\lambda, v_q]$, we have $g_i(\lambda) \equiv d(v_i, x_\lambda) = d_{ip} + \lambda$. Hence, $g_i(\bullet)$ is a linear increasing function that begins with value d_{ip} at v_p and ends with value $d_{ip} + L$ at v_q. Similarly, for $v_j \in Q$, we have $g_j(\lambda) = d(v_j, x_\lambda) = d_{jq} + L - \lambda$ so that $g_j(\bullet)$ is a linear decreasing function that begins with the value $d_{jq} + L$ at v_p and ends with the value d_{jq} at v_q.

Consider now the edge restricted problem $\min\{f(x_\lambda): x_\lambda \in e\}$. We may partition the demand vertices into the disjoint vertex subsets $V \cap P$ and $V \cap Q$ so that the definition of $f(x_\lambda)$ becomes

$$f(x_\lambda) = \max\left\{ f_p(x_\lambda), f_q(x_\lambda) \right\} \tag{5.10}$$

where

$$f_p(x_\lambda) \equiv \max\{a_i + g_i(\lambda): v_i \in V \cap P\} \tag{5.11}$$

and

$$f_q(x_\lambda) \equiv \max\{a_j + g_j(\lambda): v_j \in V \cap Q\}. \tag{5.12}$$

Since each g_i is a linear increasing function with identical slopes for vertices $v_i \in V \cap P$, the functions $a_i + g_i(\lambda)$ are also linear increasing with identical slopes and

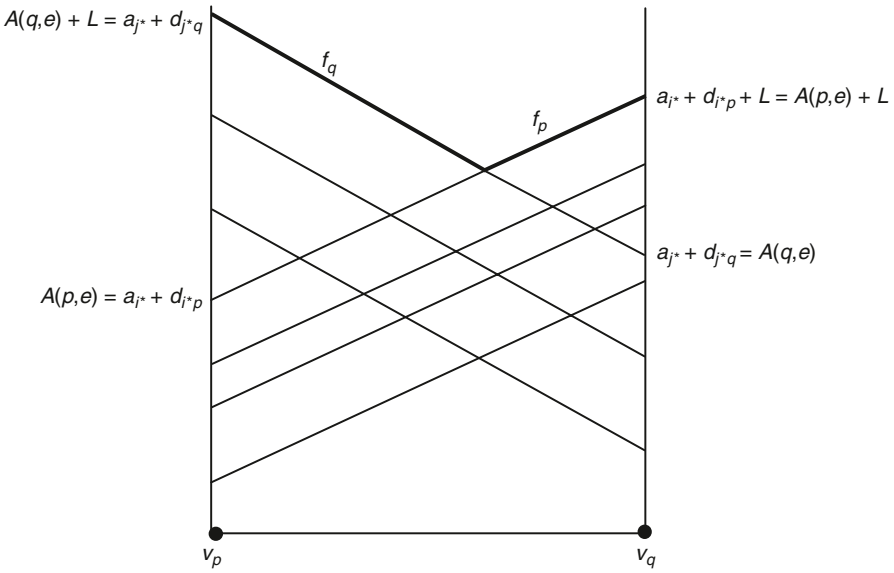


Fig. 5.8 The functions f_p and f_q

with intercepts of a_i+d_{ip} and $a_i+d_{ip}+L$ at v_p and v_q, respectively. Because the slopes are identical, the largest intercept defines $f_p(\bullet)$ on the entire edge. That is, there is a vertex $v_{i^*} \in V \cap P$ such that $a_{i^*}+d_{i^*p}=\max\{a_i+d_{ip}: v_i \in V \cap P\}$ and $f_p(x_\lambda)=a_{i^*}+d_{i^*p}+\lambda$ for $\lambda \in [0, L]$. Similarly, there is a vertex $v_{j^*} \in V \cap Q$ such that $a_{j^*}+d_{j^*q}=\max\{a_j+d_{jq}: v_j \in V \cap Q\}$ and $f_q(x_\lambda)=a_{j^*}+d_{j^*q}+L-\lambda$. Figure 5.8 illustrates the functions $f_p(x_\lambda)$ and $f_q(x_\lambda)$ as the maximum of increasing and decreasing linear functions, respectively, with identical slopes.

Let $A(p, e)$ and $A(q, e)$ be the highest intercepts at v_p and v_q, respectively. That is,

$$A(p,e) = \max\{a_i + d_{ip}: v_i \in V \cap P\} \tag{5.13}$$

and

$$A(q,e) = \max\{a_j + d_{jq}: v_j \in V \cap Q\}. \tag{5.14}$$

We then have $A(p, e)=a_{i^*}+d_{i^*p}$ and $A(q, e)=a_{j^*}+d_{j^*q}$ where the indices i^* and j^* are as defined before. Additionally, we have

$$f_p(x_\lambda) = A(p,e) + \lambda \;\forall\; \lambda \in [0, L], \tag{5.15}$$

$$f_q(x_\lambda) = A(q,e) + L - \lambda \;\forall\; \lambda \in [0, L], \tag{5.16}$$

and

$$f(x_\lambda) = \max\{A(p,e) + \lambda, A(q,e) + L - \lambda\} \forall\; \lambda \in [0, L]. \tag{5.17}$$

Goldman's localization theorem can then be stated as follows.

Theorem 1 (Localization Theorem): *Exactly one of three cases applies:*

(a) $A(q, e)-A(p, e)\geq L$: *Then the problem can be reduced to* Q, *with* a_q *replaced by* $max\{a_q, A(p, e)+L\}$.
(b) $A(p, e)-A(q, e)\geq L$: *Then the problem can be reduced to* P, *with* a_p *replaced by* $max\{a_p, A(q, e)+L\}$.
(c) $|A(p, e)-A(q, e)|<L$. *Then the optimal location is in the interior of edge* e.

In case (a), the lowest value $A(q, e)$ of the linear decreasing function is at least as large as the highest value $L+A(p, e)$ of the linear increasing function so that the value of $f(x_\lambda)$ is defined by the linear decreasing function $A(q, e)+L-\lambda$ on the entire edge. This is sufficient to conclude that any point in $P\cup e-\{v_q\}$ cannot be an optimal location. A more formal justification for this is as follows. Suppose $x\in P\cup e-\{v_q\}$. Then, we have:

$$\begin{aligned} f(x) &= \max\{a_j + d(v_j, x)\colon j \in I\} \\ &\geq \max\{a_j + d(v_j, x)\colon v_j \in V \cap Q\} \\ &= \max\{a_j + d_{jq} + L + d(v_p, x)\colon v_j \in V \cap Q\} \\ &= d(v_p, x) + L + A(q, e) \\ &> A(q, e) \\ &= f(v_q). \end{aligned}$$

This proves that $f(x)>f(v_q)$ for all x in $P\cup e-\{v_q\}$, so this set cannot contain an optimum. We confine the search for an optimum to the subset Q by deleting all points in P and all points in e except v_q. Replacing the addend a_q by the larger of a_q or $A(p, e)+L$ is needed because for any candidate point $x\in Q$, if $f(x)$ is defined by a vertex in P, then $f(x)=d(x, v_q)+L+A(p, e)$ where the quantity $L+A(p, e)$ is the new value of a_q. Note that if $a_q>A(p, e)+L$, then no demand vertex in P can supply the value of $f(x)$ for $x\in Q$ (since $f(x)\geq w_q d(x, v_q)+a_q>w_q d(x, v_q)+L+A(p, e)\geq \max\{w_i d(x, v_i)+a_i\colon v_i\in V\cap P\}$).

Case (b) is similar to case (a) with function $A(p, e)+\lambda$ being at least as large as the function $A(q, e)+L-\lambda$ on the entire edge so that $f(x_\lambda)$ is defined now by $A(p, e)+\lambda$ for $\lambda\in[0, L]$. Using similar arguments as in case (a), it is apparent in case (b) that $f(x)>f(v_p)=A(p, e)\,\forall x\in Q\cup e-\{v_p\}$ implying that no point in $Q\cup e-\{v_p\}$ qualifies as an optimal location. Replacing a_p by $\max\{a_p, A(q, e)+L\}$ is needed to account for the largest a_j+d_{jq} value that can be supplied by demand vertices v_j in $Q\cup e-\{v_p\}$ which is the deleted portion of the network.

In case (c), the linear functions $A(p, e)+\lambda$ and $A(q, e)+L-\lambda$ intersect at an interior point x_{λ^*} of the edge with λ^* defined by $\lambda^*=0.5[A(q, e)+L-A(p, e)]$. Evaluating f at x_{λ^*}, we obtain $f(x_{\lambda^*})=0.5[A(p, e)+A(q, e)+L]=f_p(\lambda^*)=f_q(\lambda^*)$ and letting i^* and j^* be the indices of the two critical vertices in $V\cap P$ and $V\cap Q$, respectively, such that $A(p, e)=a_{i^*}+d_{i^*p}$ and $A(q, e)=a_{j^*}+d_{j^*q}$, we obtain $f(x_{\lambda^*})=0.5[a_{i^*}+d_{i^*p}+a_{j^*}+d_{j^*q}]$. Whenever case (c) occurs, x_{λ^*} is the unique optimal location.

The localization theorem offers a direct computational advantage for tree networks because every edge in a tree network is an isthmus. Let T be a tree network. Any vertex v_t of the tree that is adjacent to exactly one vertex v_s is referred to as a *tip*. It is well known that every tree has at least two tip vertices. The following algorithm uses the localization theorem repeatedly, "trimming" the tree successively by deleting each time a selected tip and the interior of the edge that connects it to the unique adjacent vertex, unless the localization theorem concludes that the optimal location occurs at the selected tip or in the interior of the connecting edge (cases (a) or (c) in the theorem). The process is described in the procedure below.

Algorithm 1: Tree Trimming Procedure

Step 1: If T consists of a single vertex, stop; that vertex is an optimal solution.

Step 2: Select a tip v_p and let v_q be the vertex adjacent to v_p. Let $e=[v_p, v_q]$ and L be the length of e. Take $A(p, e)=a_p$ and calculate $A(q, e)=\max\{w_j d_{qj}+a_j: j\in I, j\neq p\}$. If $A(p, e)\geq A(q, e)+L$, then tip v_p is optimal; stop. If $|A(q, e)-a_p|<L$, then the optimal solution is the interior point x_{λ^*} of e with the length of subedge $[v_p, x_{\lambda^*}]$ given by $\lambda^*=0.5[A(q, e)+L-a_p]$; stop.

Step 3: Delete tip v_p and the interior of edge e from T. Delete p from I. Replace a_q with $\max\{a_q, a_p+L\}$ and return to Step 1.

If the network G under consideration is not a tree, then the localization theorem can be repeatedly used for each isthmus of G, one at a time. Termination occurs when either an optimal location is found or the problem is reduced to a single cyclic component. In the latter case, Hakimi's method is used to solve the reduced problem on the last cyclic component that has persisted. The only computational gain in this case is the reduction of the problem from the initial network with many cycles to a single cyclic component. The number of edge restricted problems that need to be solved is smaller than would have resulted from a direct application of the method on the original network.

An extension of the localization theorem to the weighted case with addends is possible, but its algorithmic utility is limited, because the computational advantages gained in the unweighted case from the updating of the addends do not occur in the weighted case. To outline the weighted version, consider an isthmus $e=[v_p, v_q]$ with associated components P and Q as defined before. In the weighted case, for $x_\lambda \in e$, we have $f(x_\lambda)=\max\{f_p(x_\lambda), f_q(x_\lambda)\}$ where $f_p(x_\lambda)=\max\{w_i(d_{ip}+\lambda)+a_i: v_i\in V\cap P\}$ and $f_q(x_\lambda)=\max\{w_j(d_{jq}+L-\lambda+a_j: v_j\in V\cap Q\}$. It follows that $f_p(x_\lambda)$ is the maximum of *increasing* linear functions with slopes w_i corresponding to demand vertices v_i in P and $f_q(x_\lambda)$ is the maximum of *decreasing* linear functions with slopes $-w_j$ corresponding to demand vertices v_j in Q. It follows that $f_p(x_\lambda)$ is a convex piecewise linear increasing function and $f_q(x_\lambda)$ is a convex piecewise linear decreasing function. Define $A(p, e)=\max\{a_i+w_i d_{ip}: v_i\in V\cap P\}$ and $A'(p, e)=\max\{a_i+w_i(d_{ip}+L)$:

$v_i \in V \cap P\}$. Because it is monotone increasing, $f_p(\bullet)$ has its lowest value at v_p and its highest value at v_q with $f_p(v_p) = A(p, e)$ and $f_p(v_q) = A'(p, e)$. Similarly, define $A(q, e) = \max\{a_j + w_j d_{jq}: v_j \in V \cap Q\}$ and $A'(q, e) = \max\{a_j + w_j(d_{jq} + L): v_j \in V \cap Q\}$. Because it is monotone decreasing, f_q has its highest value at v_p with $f(v_p) = A'(q, e)$ and its lowest value at v_q with $f(v_q) = A(q, e)$. The analogous version of the localization theorem for the weighted case is as follows:

Theorem 2 (Localization Theorem for Weighted Case): *Exactly one of the three cases apply:*

(a) $A(q, e) \geq A'(p, e)$: *Then, the optimum lies in* Q.
(b) $A(p, e) \geq A'(q, e)$: *Then, the optimum lies in* P.
(c) $A(q, e) < A'(p, e)$ *and* $A(p, e) < A'(q, e)$: *Then the optimum is located in the interior of* e.

The assertion in part (a) is a direct consequence of the fact that $f_q(x_\lambda) \geq f_p(x_\lambda)$ on the entire edge because the lowest value of the decreasing function $f_q(\bullet)$ is at least as high as the highest value of the increasing function $f_p(\bullet)$. Part (b) is similar, with $f_p(x_\lambda)$ being at least as large as $f_q(x_\lambda)$ on the entire edge. In part (c), the functions $f_p(\bullet)$ and $f_q(\bullet)$ intersect at an interior point of the edge, and the point of intersection is the minimizer of f. The power of the theorem is partly lost now due to the fact that, even though the optimum can be localized to subsets Q or P, respectively, in parts (a) or (b), the computational advantages available in the unweighted case are no longer available in the weighted case, as the computations of the parameters $A(\bullet, \bullet)$ and $A'(\bullet, \bullet)$ now require the data of the entire network.

5.2.3 Minieka (1970): Solving p-Center Problems via a Sequence of Set Covering Problems

We now focus on the absolute p-center problem where $1 < p < n$. The case $p \geq n$ is trivially solved by placing a center at each of the n demand vertices. Minieka (1970) has solved this problem in a clever way by solving a sequence of set covering problems.

With $S_p(G) \equiv$ set of all subsets of G consisting of p points, $X \in S_p(G)$, $D(X, v_i) \equiv \min\{d(x_j, v_i): x_j \in X\}$, and $f(X) \equiv \max\{w_i D(X, v_i): i \in I\}$, to solve the absolute p-center problem, we look for a point set $X^* \in S_p(G)$ such that $f(X^*) \leq f(X) \forall X \in S_p(G)$. Because each facility can be located anywhere on the network, this calls for an infinite search.

Minieka (1970) considers the unweighted version of the problem, but his approach can be directly extended to the weighted version; see, e.g., Kariv and Hakimi (1979). Minieka reduces the infinite search in $S_p(G)$ to a finite search, by observing that the absolute 1-center of the network occurs at one of a finite number of break points of $f(\bullet)$. Consider an edge $e = [v_p, v_q]$. If x_{λ^*} is an edge-restricted minimum of $f(\bullet)$ in the interior of e, then x_{λ^*} is a break point of $f(\bullet)$ defined by the intersection of two piecewise linear functions associated with a pair of vertices. With this motiva-

tion, we define U to be the set of all points u in G that qualify for en edge-restricted minimum. That is, U is the set of points $u \in G$ such that u is the unique point in its edge for which $d(v_i, u) = d(u, v_j)$ for a pair of vertices $v_i, v_j \in V$ with $i \neq j$. Because the piecewise linear functions have slopes of ± 1, the uniqueness requirement in the definition implies that the slopes of the two intersecting linear pieces are oppositely signed. There exists an absolute 1-center in the set $P \equiv V \cup U$. Clearly, there can be at most $n(n-1)/2$ intersection points in an edge, implying that U has at most $|E|n(n-1)/2$ elements in it. Hence, P is a finite dominating set (i.e., a finite set that supplies an optimum solution) for the unweighted absolute 1-center problem.

Minieka (1970) observed that P is also a finite dominating set for the unweighted absolute p-center problem. To justify this, suppose we have an absolute p-center $X^* = \{x_1^*, \ldots, x_p^*\}$. If not all points of X^* are in P, we may construct an absolute p-center X' from X^* that fulfills this requirement. To do so, partition the demand set V into subset $V_1, \ldots, V_p$ such that all vertices in subset V_i have the i-th element x_{i^*} of X^* as their closest center (ties are broken arbitrarily). Let x_i' be an optimal solution in P to the absolute 1-center problem defined with respect to the demand set V_i. This implies that $\max\{d(x_i', v_r): v_r \in V_i\} \leq \max\{d(x_i^*, v_r): v_r \in V_i\}$. Define $X' = \{x_1', \ldots, x_p'\}$. Since $D(X', v_r) \leq d(x_i', v_r)\ \forall\, v_r \in V$ and $\forall\, i \in \{1, \ldots, p\}$, we have

$$\begin{aligned} f(X') &= \max\{D(X', v_r): v_r \in V\} \\ &\leq \max\{\max\{d(x_1', v_r): v_r \in V_1\}, \ldots, \\ &\quad \max\{d(x_p', v_r): v_r \in V_p\}\} \\ &\leq \max\{\max\{d(x_1^*, v_r): v_r \in V_1\}, \ldots, \\ &\quad \max\{d(x_p^*, v_r): v_r \in V_p\}\} \\ &= \max\{D(X^*, v_r): v_r \in V\} \\ &= f(X^*) \end{aligned}$$

which proves that X' is an absolute p-center solution with $X' \subset P$.

With P supplying an optimal solution to the absolute p-center problem, we may now transform it to a sequence of set covering problems. Given a zero-one matrix **A** and a cost vector **c,** the binary program

$$\text{Min } \mathbf{cy} \tag{5.18}$$

$$\text{s.t. } \mathbf{Ay} \geq \mathbf{1} \tag{5.19}$$

$$\mathbf{y} \in \{0, 1\}^n \tag{5.20}$$

is known to be the *set covering problem*. This problem arises when a given set needs to be covered by the union of a collection of its subsets at minimum cost. Let S be a given set with h elements and let $S_1, \ldots, S_k$ be a collection of nonempty subsets of S. Suppose given costs c_i, $i \in K \equiv \{1, \ldots, k\}$, where c_i is the cost of using subset S_i. If we choose a subset K' of K, the corresponding subcollection $\{S_i: i \in K'\}$ is said to *cover* S if $\cup\{S_i: i \in K'\} = S$. The object is to choose a subset K^* of K, such that the corresponding subcollection $\bigcup_{i \in K^*} S_i$ covers S, and its cost, $\sum_{i \in K^*} c_i$, is as small as

possible among all subcollections that cover S. To convert the problem to the binary program defined by (5.18)–(5.20), define the h by k matrix $\mathbf{A}$ with elements $a_{ij}=1$, if the i-th element of S is an element of the subset S_j, and $a_{ij}=0$ if not. Let y_j be a binary variable with $y_j=1$ if subset S_j is selected and $y_j=0$ if not. To cover all elements of S, we impose the constraint

$$\sum_{j=1}^{k} a_{ij} y_j \geq 1 \quad \forall\, i = 1, ..., h \tag{5.21}$$

which requires at least one y_j for which $a_{ij}=1$ is set equal to 1. This ensures that at least one subset S_j, which contains the i-th element of S is selected by the i-th constraint. The summation on the left side of (5.21) is the dot product of the i-th row of $\mathbf{A}$ with the column vector $\mathbf{y}$ and, accordingly, (5.19) is nothing but a more compact form of the h constraints in (5.21).

In the above formulation, the h rows of $\mathbf{A}$ correspond to the h elements of S. These are the elements that need to be covered. The columns of $\mathbf{A}$ correspond to the k given subsets of S. To make the connection of the set covering problem to the p-center problem, we take S to be V. That is, the elements that need to be covered are the demand vertices $v_1,\ldots, v_n$. The subsets S_j of S are determined on the basis of the finite dominating set P that we identified. Let $p_1,\ldots, p_k$ be an enumeration of the elements of P and let $r>0$ be a selected radius of coverage. Define $S_j, j=1,\ldots, |P|$, to be the set of vertices $v_i \in V$ for which $d(p_j, v_i)\leq r$. Accordingly, the matrix $\mathbf{A}$ in our case has n rows and $k\equiv|P|$ columns and the subsets S_j are defined by the set of demand vertices that are accessible by a facility at p_j within a distance of at most r units. We define the costs $c_j=1\ \forall j\in\{1,\ldots, k\}$ and define $a_{ij}=1$ if $d(v_i, p_j)\leq r$ and $a_{ij}=0$ if $d(v_i, p_j)>r$.

The resulting set covering problem with $\mathbf{A}=[a_{ij}]$, $\mathbf{c}=(1,\ldots, 1)$, and $\mathbf{y}=(y_1,\ldots, y_k)^T$ selects the fewest possible points from P such that every demand vertex has at least one selected point within a distance of at most r units. If the resulting number of points from the set covering solution for a given value of r is at most p while it is strictly greater than p relative to a new radius $r'<r$, then r is, in fact, the p-radius r_p and any optimal solution to the set covering problem relative to this r identifies an absolute p-center solution (by appending as many arbitrarily selected points from P as needed if the set covering solution outputs less than p points). One major question that remains unanswered is how to pick the correct value for r (i.e., the value of r that results in a set covering solution of at most p while any reduction in r results in a set covering solution of more than p points). Minieka has given a well conceived method for accomplishing this. His method relies on modifying $\mathbf{A}$ appropriately and is described in the next paragraph.

Consider a set $X=\{x_1,\ldots, x_p\}$ of p points from P. Put $r=f(X)$ and construct the matrix $\mathbf{A}$ with respect to this choice of r. The resulting set covering problem has a feasible solution $\mathbf{y}=(y_1,\ldots, y_k)$ with $y_j=1$ if $p_j\in X$ and $y_j=0$ if $p_j \notin X$. The objective value defined by $\sum_j y_j$ is equal to p. Suppose now we modify the matrix $\mathbf{A}$ by redefining a_{ij} to be equal to 1 if $d(v_i, p_j)<r$ and $a_{ij}=0$, otherwise. The new version of

A is identical to the old version except that all entries a_{ij} that were one before due to $d(v_i, p_j)$ being equal to r are now replaced by zeroes while $a_{ij}=1$ are retained for all index pairs ij for which $d(v_i, p_j)<r$. Let $\mathbf{A}'$ be the modified version of **A**. Clearly, the new matrix $\mathbf{A}'$ is defined relative to a new radius $r'<r$, but the value of r' is not specified. Even though Minieka does not discuss this issue, some reflection on it reveals that r' is any real number such that $\alpha \le r' < r$ where α is the largest entry in the list of distances $\{d(p_j, v_i): p_j \in P, v_i \in V\}$ that is smaller than r. Solve the set covering problem with matrix $\mathbf{A}'$. Let $\mathbf{y}'$ be an optimal solution and p' be the optimal objective value. If $p'>p$, then clearly X is an absolute p-center since more than p points from P are required to cover each demand vertex within a distance of less than r. This is equivalent to saying that there does not exist a point set X' in P such that $|X'| \le p$ and $f(X') < r = f(X)$.

The same conclusion is also valid if there is no feasible solution to the set covering problem with matrix $\mathbf{A}'$. In the remaining case, there is an optimal solution $\mathbf{y}'$ to the set covering problem of matrix $\mathbf{A}'$ with optimal objective value of $p' \le p$. In this case, X is not optimal because $\mathbf{y}'$ induces a solution $X' \subset P$ with $|X'| = p' \le p$ and $f(X') < r = f(X)$. When this happens, we repeat the process once again with X', $\mathbf{A}'$, and $r' \equiv f(X')$, replacing the roles of X, **A**, and $r = f(X)$, respectively. That is, we modify $\mathbf{A}'$ to obtain a new matrix $\mathbf{A}''$, such that the elements a_{ij} are set equal to 1 if $d(v_i, p_j) < r'$, and 0 otherwise. The set covering problem is re-solved with the new matrix $\mathbf{A}''$ to obtain an optimal solution $\mathbf{y}''$, if it exists, with optimal objective value p''. The optimality of X' is concluded if the set covering problem admits no feasible solution or if it has an optimal solution $\mathbf{y}''$ with optimal value $p'' > p$. In the remaining case, $\mathbf{y}''$ induces a new solution $X'' \subset P$ with $|X''| = p'' \le p$, and the procedure must be repeated. The process must eventually terminate with an optimal p-center solution when either an infeasible set covering problem is encountered or a feasible set covering problem, whose optimal objective value is strictly greater than p, is encountered. The number of repetitions that can occur until termination is at most $n|P|$, since the set of ones in each modified version of **A** is a proper subset of the immediately preceding version of **A**.

5.3 The Impact of the Classical Contributions

Among the three classical papers discussed in the previous section, Hakimi's (1964) contribution is viewed by many, including this author, as a seminal work that has led to the birth and growth of the research area known today as network location.

Hakimi was the first researcher to pose and analyze the absolute center and median problems in the context of a transportation/communication network, where each edge is a continuum of points. Travel occurs in a network along paths composed of sequences of edges, which is intrinsically different from travel paths available in analogous planar location problems. This feature leads to distances on a network defined by shortest path lengths. Hakimi's first fundamental contribution is his concise analysis of the shortest path distance from a fixed point in the network

to a variable point in an edge. The fact that this distance is the minimum of two linear functions results in a concave one or two-piece linear function in a network context, while normed distances in analogous planar location problems are convex. Convexity is a desirable property that leads to strong theory and efficient algorithms in many optimization problems, but it fails in the context of network location unless the network is a tree, as pointed out by Dearing et al. (1976). The theory and algorithms in network location, with certain exceptions of tree location problems, had to be developed with new viewpoints not readily available in analogous planar problems and Hakimi's concave two-piece linear characterization of the edge-restricted distance has provided a foundation for subsequent work.

An immediate consequence of concave piecewise linearity is that multiplication by a positive weight preserves this property. The sum of convex functions is also convex which leads to the well known vertex-optimality theorem for median problems by Hakimi (1964). For the absolute center problem, however, the objective function is defined by the maximum of concave piecewise linear functions and this does not preserve concavity as in the case of the median problem. Even though concavity is lost, piecewise linearity is still retained. This leads to a large, but finite, number of candidate points for local optima on any edge, defined by intersections of pairs of linear pieces with oppositely signed slopes (i.e., directional derivatives). The restriction of local optima to finitely many breakpoints is a fundamental result, initially conceived and used by Hakimi (1964), and exploited later by Minieka (1970) for solving the multi-facility unweighted problem through the solution of finitely many set covering problems. Extensions are given later by Kariv and Hakimi (1979) for the weighted case and by Hooker et al. (1991) for convex nonlinear cost functions.

All subsequent work on 1-centers and *p*-centers have used this result in one way or another. Most of the focus for solving the 1-center problem has been on developing more efficient computational methods that eliminate unnecessary breakpoints or edges during the search for local optima; some pertinent results can be found in Kariv and Hakimi (1979), Handler (1974), Odoni (1974), Halpern (1979), and Sforza (1990). Algorithms for solving *p*-center problems are in one of two categories: set covering based or enumeration base. The set covering approach of Minieka (1970) has initiated a series of contributions on the same or related themes by other researchers including Christofides and Viola (1971), Garfinkel et al. (1977), Toregas et al. (1971), and Elloumi et al. (2004). Enumeration based methods enumerate in different ways *p*-element subsets of the set *P*; see, e.g., Kariv and Hakimi (1979), Moreno (1986), Tamir (1988), and Hooker (1989).

Goldman's paper, discussed in Sect. 5.2.2, focuses on exploiting the structure of the network under consideration. The particular topological element Goldman (1972) has focused on is the type of edge whose removal from the network, except its end-points, results in two disconnected components. Such an edge is referred to as an isthmus by Goldman. An isthmus has a very special feature: Every path originating in one of the resulting components and terminating in the other component must pass through the isthmus. This has an important consequence for the unweighted case. The longest of the shortest paths connecting a pair of vertices, one in

each component, passes through the isthmus under consideration, and its mid-point is either in the isthmus, in which case it is optimum, or in one of the components, in which case the search can be reduced to that component.

The most visible impact of Goldman's paper is that it has drawn attention to special structures in solving location problems on networks, primarily trees. Every edge in a tree is an isthmus. Goldman's algorithm for unweighted trees requires a quadratic number of arithmetic operations in the number of vertices. Handler (1973) and Halfin (1974) developed more efficient linear time algorithms for the unweighted case. The weighted case for tree networks is analyzed and efficiently solved by Dearing and Francis (1974), Hakimi et al. (1978), Hedetniemi et al. (1981), Kariv and Hakimi (1979), and Megiddo (1983). Dearing (1977) and Francis (1977) have extended the problem to incorporate nonlinear monotonic functions of distances and have described efficient solutions methods for tree networks. Goldman's paper has also directed attention to more general structures than trees, but not much can be done unless the cyclic portions of a network (blocks) induce a tree structure when each such component is represented by a single node; see, e.g., the work by Chen et al. (1988), and Kincaid and Lowe (1990). Special structure in multi-facility minimax problems have also led to many elegant results and efficient algorithms for tree networks. Some of the contributions are those by Handler (1978), Hakimi et al. (1978), Kariv and Hakimi (1979), Tansel et al. (1982), Megiddo and Tamir (1983), Frederickson and Johnson (1983), Megiddo et al. (1983), Jaeger and Kariv (1985), and Shaw (1999). As Dearing et al. (1976) point out, convexity of distance is an important property for tree networks and has a significant part in developing theory and efficient algorithms for the single facility case. Convexity does not explain, however, why absolute p-center location problems are so efficiently solvable on tree networks, since the p-center objective function is not convex even on a tree network.

5.4 Subsequent Work in Discrete Center Location

In this section, we survey the subsequent work in discrete center location. We first focus on the single facility case on general networks, followed by problems on tree networks, and finally we consider other specially structured networks. Then the multi-facility problem is covered, again, first on general networks, and then on trees.

5.4.1 The Absolute 1-Center on General Networks

Hakimi's (1964) method requires solving an edge-restricted problem on each edge by inspecting break points that are oppositely signed in either direction. There are at most $n(n-1)/2$ breakpoints per edge which requires evaluating the objective

function at $O(n^2|E|)$ points. This makes Hakimi's method an $O(n^3m)$ algorithm where $m \equiv |E|$. Later, Hakimi et al. (1978) presented an $O(mn^2 \log n)$ version of the same algorithm. This bound is improved to $O(mn \log n)$ for the unweighted case. Kariv and Hakimi (1979) solved the weighted case in $O(mn \log n)$ and the unweighted case in $O(mn)$ time. This is the best known bound for the absolute 1-center problem. The $O(mn \log n)$ bound for the unweighted case is also achieved by Sforza (1990), whose algorithm for the weighted case is $O(kmn \log n)$, where k is a factor that depends on the precision level and weight distribution. This bound does not improve the bound of Kariv and Hakimi (1979), but Sforza's algorithm is more effective in CPU time.

Edge elimination techniques rely on devising lower bounds for each edge and eliminating those edges whose lower bounds are larger than the best objective value attained during the search for optimum. Handler (1974), Odoni (1974), Christofides (1975), and Halpern (1979) made use of edge elimination techniques that have resulted in improved CPU times, where Halpern's bound is stronger than the others. Sforza's (1990) edge elimination technique has been found to be quite successful in practice due to its ability to eliminate 80% of edges in many problems.

All the algorithms mentioned above are improved versions of Hakimi's original technique. Minieka (1981)'s $O(n^3)$ algorithm, on the other hand, only makes use of the distance matrix without using the vertex-to-point cost functions.

An important theoretical contribution is due to Hooker (1986) who analyzed the nonlinear version of the 1-center problem for the problem with convex cost functions and proposed a general purpose algorithm. His analysis is based on decomposing the network into tree-like segments and solving a convex programming problem on each segment. The objective function defined by maximum of convex functions of distances is convex on any tree-like segment, and a local minimum can be found by solving a convex programming problem. Hooker (1986) proved that there are $O(n)$ tree-like segments on an edge.

Shier and Dearing (1983) made another important theoretical contribution in their study of a family of nonlinear single facility location problems on a network that includes, as special cases, the absolute 1-center and absolute 1-median problems. They characterize locally optimal solutions by means of directional derivatives. This characterization is equivalent, in the case of the absolute 1-center problem, for the point under consideration to be a breakpoint of $f(\bullet)$ such that f increases in every "moveable" direction at that point. If the point under consideration is an interior point of some edge, then there are only two directions of movement out of that point. Hence, an interior point is a local optimum if and only if it is a break point of f defined by the intersection of two weighted distance functions associated with a pair of distinct vertices, such that the increase in one of the functions is accompanied by a decrease in the other one if one moves slightly away from the point in either direction.

Continuous demand versions of the absolute 1-center problem are also considered. There are two versions. Minieka (1970) defines the *general absolute center* of a network G as a point whose maximum distance to a farthest point on each edge is minimized. In contrast, Frank (1967) defines a *continuous center* of a network as a

point whose maximum distance to any point on the network is minimized. The two definitions are equivalent for the case of 1-centers. Minieka (1977) showed that Hakimi's algorithm for the absolute 1-center can be used to find the general absolute 1-center if one replaces the distance function $d(x, y)$ with a new distance function $d'(x, e)$ which denotes the distance between x and a farthest point in edge e. Frank (1967) defined the continuous 1-center problem and showed that it can be solved via Hakimi's algorithm.

5.4.2 *The Absolute 1-Center on Trees and Other Special Structured Networks*

Beginning with Goldman's localization theorem, considerable attention has been given to tree networks. Other special structures have also received some attention.

An important property that has led to efficient algorithms for trees has to do, at least in good part, with the convexity of distance on tree networks. Dearing et al. (1976) generalized in a theoretical framework the earlier convexity observations of Goldman and Witzgall (1970) and Handler (1973), as well as nonconvexity observations of Goldman (1971) and Hakimi (1964). Dearing et al. (1976) prove that the function $d(x, y)$ as a function of x alone, or as a function of x and y, is convex if and only if the network is a tree network. This implies that the objective function in the absolute p-center problem is convex on a tree and nonconvex on a cyclic network. Convexity implies that any local minimum on a tree network is also a global minimum.

Goldman's (1972) localization theorem, when applied to a tree, finds an optimum in $O(n^2)$ time. Handler (1973) proves for the unweighted case that the absolute center of a tree is the midpoint of a longest path in the tree and gave an $O(n)$ algorithm. Halfin (1974) modifies Goldman's algorithm and turns it into an $O(n)$ algorithm for trees with unit weights and any addends. Lin (1975) shows that the unweighted problem on a network with addends is equivalent to the unweighted problem on a new network with no addends, where the new network has the same structure as the old one except that for every vertex v_i for which the addend $a_i > 0$, a new vertex v_i' and a new edge $[v_i, v_i']$ is added with length a_i. Hence, addends do not increase the time bounds of proposed algorithms.

Dearing and Francis (1974) analyze the weighted problem on trees and prove that the maximum of the $n(n+1)/2$ numbers $\alpha_{ij} \equiv (d_{ij} + a_i/w_i + a_j/w_j)/(1/w_i + 1/w_j)$, $1 \le i \le j \le n$ is a lower bound for the optimum value of the objective function for any network, and is an attainable lower bound for tree networks. The absolute 1-center of a tree occurs at the point x on the path $P(v_s, v_t\}$, identified by $\alpha_{st} = \max\{\alpha_{ij}: 1 \le i \le j \le n\}$, such that $w_s d(v_s, x) + a_s = w_t d(v_t, x) + a_t$. The computation of α_{st} takes $O(n^2)$ time. Hakimi et al. (1978) propose an $O(n(r+1))$ algorithm for this problem where $r \le n$. Kariv and Hakimi (1979) describe an algorithm that reduces the tree to subtrees until a single edge remains. The local center on the last edge solves the weighted absolute 1-center problem while one of its end-vertices solves the ver-

tex restricted 1-center problem. The time bound is $O(n \log n)$ for weighted trees. Megiddo (1983) solves the weighted absolute 1-center problem on tree networks in $O(n)$ time, which is the best known time bound for this problem. The nonlinear version of the absolute 1-center problem on tree networks, where each weighted distance is replaced by a monotone increasing loss function is considered by Dearing (1977) and Francis (1977). They prove that the maximum of the $n(n+1)/2$ numbers $\beta_{ij} \equiv (f_i^{-1} + f_j^{-1})^{-1}[d_{ij}]$, $1 \le i \le j \le n$, is a lower bound for the minimum objective value of any network, and that this bound is attainable for tree networks. A maximizing pair v_s and v_t identify a path $P(v_s, v_t)$, such that the optimum point is the point x on the path where $f_s[d(v_s, x)] = f_t[d(x, v_t)]$. If $s = t$, this implies v_s is the optimum point.

For special structured networks that are more general than trees, Chen et al. (1988) propose an algorithm, similar in spirit to Goldman's, for linear and nonlinear cost functions. They construct the block diagram of the network in which each block (a maximally connected subgraph that cannot be made disconnected by removing a vertex with its adjacent edges) is represented by a vertex. A block diagram is always a tree. The algorithm either finds an optimum or reduces the problem to a single block. The time bound of the algorithm is $O(n \min\{b, \alpha \log b\})$ for the linear case, where α is the maximum number of cut points in any block and b is the number of blocks. If the algorithm ends with a block, the algorithms of Kariv and Hakimi (1979) or Sforza (1990) may be used to locate the absolute 1-center in the block for the case of linear costs. For nonlinear monotonically increasing convex cost functions, Hooker's (1986 algorithm based on tree-like segments may be used. The time bound of Chen et al.'s (1988) reduction algorithm is $O(n \log n)$ for cactus networks. A more efficient $O(n)$ algorithm is devised for cactus networks that are homemorphic to a 3-cactus by Kincaid and Lowe (1990) that transforms these special networks to trees in which point-to-point distances are preserved.

5.4.3 Absolute p-Centers of General Networks

Kariv and Hakimi (1979) prove that the absolute p-center problem on a network is **NP**-Hard even if the network is planar, unweighted, with unit edge lengths and a maximum vertex degree of 3. Solution approaches are based on the existence of a finite dominating set, initially motivated by Hakimi's (1964) method, formalized by Minieka (1970) for the unweighted case, and extended directly to the weighted case by Kariv and Hakimi (1979). This result is further generalized by Hooker et al. (1991) and a unifying approach is given for identifying finite dominating sets in a rather general setting.

All solution approaches proposed for the absolute p-center problem on general networks are based on the existence of a finite dominating set $P = V \cup U$, where the definition of U is revised for the weighted case to include every interior point x such

that $w_i d(v_i, x) = w_j d(x, v_j)$ for a pair of vertices v_i and v_j, $i \neq j$, and moving in either direction increases one of the functions while decreasing the other one. Solution approaches are either based on solving a sequence of set covering problems, suggested first by Minieka (1970), or enumerating p element subsets of the set P.

Minieka's (1970) algorithm solves the unweighted version by solving a sequence of set covering problems with successively decreasing values of the covering radius r. Garfinkel et al. (1977) also solve a sequence of set covering problems, but they first reduced the search space by finding a heuristic solution X and eliminating from P all those points whose relative radius is greater than $r \equiv f(X)$. This reduces the number of variables in the set covering problem. Christofides and Viola (1971) solve the weighted problem by first generating the set of all feasible regions in the network reachable by at least one vertex within a distance of r, where r is a fixed parameter, and solving a set covering problem that selects the smallest number of points from these regions. This approach is essentially the same as that of Minieka (1970), except that Christofides and Viola do not make use of the finite dominating set P. Toregas et al. (1971) solve the vertex restricted p-center problem by solving the linear programming relaxation of the associated set covering problem, and adding a cut whenever termination occurred with a fractional solution. Elloumi et al. (2004) devise a new integer programming formulation of the problem based on the set covering idea. The linear programming relaxation of their formulation generates better lower bounds for the problem than those of previous models.

Kariv and Hakimi (1979) propose an $O(m^p n^{2p-1} \log n)$ enumeration algorithm for weighted networks. Their algorithm uses the fact that each center in an optimal solution is the 1-center of a subnetwork. They choose $p-1$ arbitrary centers and solve for the p-th one using a 1-center approach. Moreno (1986) provides an algorithm of time bound $O(m^p n^{p+1} \log n)$. Tamir (1988) combines the algorithms of Kariv and Hakimi (1979) and Moreno (1986) to obtain an algorithm with improved time bounds of $O(m^p n^p \log^2 n)$ for the weighted case and $O(m^p n^p \log n)$ for the unweighted case. Further improvements are made by using dynamic data structures resulting in a time bound of $O(m^p n^{p-1} \log^3 n)$. For the case of nonlinear convex cost functions, Hooker (1989) proposes an enumeration algorithm based on tree-like segments, which is practical for small values of p. The algorithm locates p centers for each combination of p tree-like segments by solving linear and convex problems on each segment. The algorithm becomes intractable when p exceeds 4.

5.4.4 Absolute p-Centers of Tree Networks

The minimum distance $D(X, v_i)$ from a vertex to a collection of p points is not convex even though the distance $d(x, v_i)$ is convex for each $x \in X$. Despite the loss of convexity, the absolute p-center problem in tree networks is solved in polynomial time by various algorithms.

Handler (1978) solves the absolute 2-center and the continuous absolute 2-center problems in $O(n)$ time by solving three 1-center problems. His algorithm does not seem to be extendible to larger values of p. Hakimi et al. (1978) describe an $O(n^{p-1})$ algorithm for unweighted tree networks. Kariv and Hakimi (1979) propose an $O(n^2 \log n)$ algorithm for absolute and vertex restricted weighted and unweighted problems on trees. For the unweighted case, they also develop an $O(n \log^{p-2} n)$ algorithm for the absolute p-center problem and $O(n \log^{p-2} n)$ algorithm for the vertex restricted case. Megiddo and Tamir (1983) propose an $O(n \log^2 n \log\log n)$ algorithm for the weighted absolute p-center problem on trees. An $O(n \log n)$ algorithm is presented by Frederickson and Johnson (1983) for the unweighted case. Megiddo et al. (1981) solve the vertex restricted problem in $O(n \log^2 n)$ time. Jaeger and Kariv (1985) devise an $O(pn \log n)$ algorithm for vertex restricted and absolute p-centers on weighted trees for relatively small values of p. If $p < \log n$ for the vertex restricted p-center and $p < \log n \log\log n$ for the absolute p-center, then this time bound is better than the previous ones. Shaw (1999) presents a unified column generation approach for a general class of facility location problems on trees that includes the absolute p-center problem as a special case. The complexity of his algorithm for the weighted p-center problem on trees is $O(n^2 \log n)$. A nonlinear version of the problem, in which each weighted distance is replaced by a monotonic increasing function of the distance, is considered by Tansel et al. (1982) and solved in $O(n^4 \log n)$ time, which is improved to $O(n^3 \log n)$ by the modification given in Chap. 8 of Mirchandani and Francis (1990). For various duality results, see also Shier (1977).

5.5 Future Directions

Tree network absolute p-center problems are well solved in polynomial time both for linear and nonlinear monotonic cost functions. Bozkaya and Tansel (1998) show that there exists a spanning tree of every connected network, such that solving the absolute p-center problem on the tree solves the p-center problem on the network. Trying to find such a tree is a worthwhile undertaking, since solving the problem on it also solves the problem on the original network.

Nonlinear versions with monotonic increasing functions of distances are more realistic versions of p-center problems that may find applications in a wide variety of contexts. Multi-facility versions of such models on general networks have not been considered in the literature and demand attention.

There is an acute need for more realistic models of emergency or covering type of location problems that address major issues of terrorism, pollution, disaster fighting, and fast depletion of natural resources (such as water). Present models seem to be quite short of capturing important aspects of such problems.

Most often, we assume that data for our problems are available in a nice and clean form whereas, large scale realistic applications often require massive amounts of data that are difficult to obtain and process. Methods need to be developed for constructing and maintaining accurate data bases for large-scale applications.

References

Bozkaya B, Tansel B (1998) A spanning tree approach to the absolute *p*-center problem. Locat Sci 6:83–107
Chen ML, Francis RL, Lowe TJ (1988) The 1-center problem: exploiting block structure. Transp Sci 22:259–269
Christofides N (1975) Graph theory: an algorithmic approach. Academic, New York
Christofides N, Viola P (1971) The optimum location of multi-centers on a graph. Oper Res Quart 22:145–154
Dantzig GB (1967) All shortest routes in a graph. Theory of Graphs, International Symposium, Rome, 1966, Gordon and Breach, New York, pp 91–92
Dearing PM (1977) Minimax location problems with nonlinear costs. J Res Natl Bur Stand 82:65–72
Dearing PM, Francis RL (1974) A minimax location problem on a network. Transp Sci 8:333–343
Dearing PM, Francis RL, Lowe RL (1976) Convex location problems on tree networks. Oper Res 24:628–642
Elloumi S, Labbé M, Pochet Y (2004) A new formulation and resolution method for the *p*-center problem. INFORMS J Comput 16:84–94
Floyd RW (1962) Algorithm 97, shortest path. Commun ACM 5:345
Francis RL (1977) A note on nonlinear location problem in tree networks. J Res Natl Bur Stand 82:73–80
Frank H (1967) A note on graph theoretic game of Hakimi's. Oper Res 15:567–570
Frederickson GN, Johnson DB (1983) Finding *k*-th paths and *p*-centers by generating and searching good data structures. J Algorithms 4:61–80
Garfinkel RS, Neebe AW, Rao MR (1977) The *m*-center problem: minimax facility location. Manag Sci 23:1133–1142
Goldman AJ (1971) Optimal center location in simple networks. Transp Sci 5:212–221
Goldman AJ (1972) Minimax location of a facility in a network. Transp Sci 6:407–418
Goldman AJ, Witzgall CJ (1970) A localization theorem for optimal facility placement. Transp Sci 4:406–409
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
Hakimi SL, Schmeichel EF, Pierce JG (1978) On *p*-centers in networks. Transp Sci 12:1–15
Halfin S (1974) On finding the absolute and vertex centers of a tree with distances. Transp Sci 8:75–77
Halpern J (1979) A simple edge elimination criterion in a search for the center of a graph. Manag Sci 25:105–113
Handler GY (1973) Minimax location of a facility in an undirected tree graph. Transp Sci 7:287–293
Handler GY (1974) Minimax network location: theory and algorithms. Ph.D. Thesis, Technical Report No. 107, Operations Research Center, M.I.T., Cambridge, MA
Handler GY (1978) Finding two-centers of a tree: the continuous case. Transp Sci 12:93–106
Hedetniemi SM, Cockayne EJ, Hedetniemi ST (1981) Linear algorithms for finding the Jordan center and path center of a tree. Transp Sci 15:98–114
Hooker J (1986) Solving nonlinear single-facility network location problems. Oper Res 34:732–743
Hooker J (1989) Solving nonlinear multiple-facility network location problems. Networks 19:117–133
Hooker JN, Garfinkel RS, Chen CK (1991) Finite dominating sets for network location problems. Oper Res 39:100–118

Jaeger M, Kariv O (1985) Algorithms for finding *p*-centers on a weighted tree (for relatively small *p*). Networks 15:381–389
Jordan C (1869) Sur les assemblages de lignes. J reine angew Math 70:185–190
Kariv O, Hakimi SL (1979) An algorithmic approach to network location problems: the *p*-centers. SIAM J Appl Math 37:513–538
Kincaid RK, Lowe TJ (1990) Locating an absolute center on graphs that are almost trees. Eur J Oper Res 44:357–372
Lin CC (1975) On vertex addends in minimax location problems. Transp Sci 9:165–168
Megiddo N (1983) Linear-time algorithms for linear programming in $\mathbb{R}^2$ and related problems. SIAM J Comput 12:759–776
Megiddo N, Tamir A (1983) New results on the complexity of *p*-center problems. SIAM J Comput 12:751–758
Megiddo N, Tamir A, Zemel E, Chandrasekaran R (1981) An $O(n \log^2 n)$ algorithm for the *k*-th longest path in a tree with applications to location problems. SIAM J Comput 12:328–337
Minieka E (1970) The *m*-center problem. SIAM Rev 12:138–139
Minieka E (1977) The centers and medians of a graph. Oper Res 25:641–650
Minieka E (1981) A polynomial time algorithm for finding the absolute center of a network. Networks 11:351–355
Mirchandani PB, Francis RL (eds) (1990) Discrete location theory. Wiley, New York
Moreno JA (1986) A new result on the complexity of the *p*-center problem. Technical Report, Universidad Complutense, Madrid, Spain
Odoni AR (1974) Location of facilities on a network: a survey of results. Technical Report No. TR-03-74, Operations Research Center, M.I.T., Cambridge, MA
Sforza A (1990) An algorithm for finding the absolute center of a network. Eur J Oper Res 48:376–390
Shaw DX (1999) A unified limited column generation approach for facility location problems on trees. Ann Oper Res 87:363–382
Shier DR (1977) A min-max theorem for *p*-center problems on a tree. Transp Sci 11:243–252
Shier DR, Dearing PM (1983) Optimal locations for a class of nonlinear, single-facility location problems on a network. Oper Res 31:292–302
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 BC, Francis RL, Lowe TJ, Chen ML (1982) Duality and distance constraints for the nonlinear *p*-center problem and covering problem on a tree network. Oper Res 30:725–744
Toregas C, Swain R, ReVelle C, Bergman L (1971) The location of emergency service facilities. Oper Res 19:1363–1373

Part IV
Covering Problems

Chapter 6
Covering Problems

Lawrence V. Snyder

6.1 Introduction

The mail-order DVD rental company Netflix chooses distribution center locations so that most of its customers receive their DVDs within one business day via first-class U.S. Mail. Similarly, many municipalities aim to have fire crews reach 911 callers within a specified time, such as four minutes. Both of these are examples of the notion of *coverage*, a concept central to several classes of facility location models; it indicates whether a demand location is within a pre-specified radius (measured by distance, travel time, cost, or another metric) of its assigned facility. Homeowners are covered if they are within four minutes of the nearest fire station, and Netflix customers are covered if they are within one mailing day of a distribution center. Note that in the fire-station example, municipalities typically want to cover *all* residents (while minimizing the number of service stations to open), whereas Netflix wants to cover as many customers as possible (subject to a limit on the number of warehouses it may operate at any time, as specified by its capital budget). The fire-station problem is an example of the *set covering location problem* (SCLP), while Netflix's problem is an example of the *maximal covering location problem* (MCLP). This chapter discusses both problems.

The set covering location problem was first introduced by Hakimi (1965) and was later formulated as an integer programming problem by Toregas et al. (1971). The maximal covering location problem was introduced by Church and ReVelle (1974). Both models, and their variants, have been applied extensively to public-sector facility location problems, such as the location of emergency medical service vehicles (Eaton et al. 1985), fire stations (Schilling et al. 1980), bus stops (Gleason 1975), wildlife reserves (Church et al. 1996), and emergency air services (Flynn and Ratick 1988). They have been applied in a much more limited extent in the private sector; see, e.g., Nozick and Turnquist (2001).

L. V. Snyder (✉)
Department of Industrial and Systems Engineering, Lehigh University, 200 West Packer Ave., Mohler Lab, Bethlehem, PA 18015, USA
e-mail: larry.snyder@lehigh.edu

H. A. Eiselt, V. Marianov (eds.), *Foundations of Location Analysis,*
International Series in Operations Research & Management Science 155,
DOI 10.1007/978-1-4419-7572-0_6,

The set covering location problem and the maximal covering location problem are closely related to the *p*-center problem, which aims to locate at most *p* facilities to minimize the maximum distance, among all customers, between the customer and its assigned facility. In the *p*-center problem, the coverage radius itself constitutes the objective function. The Introduction of this book provides a more thorough discussion of the relationships among these classical models.

Like most location problems, the SCLP and MCLP may be defined as continuous problems (in which facilities may be located anywhere on the plane) or as discrete problems (in which they may be located only at the nodes of a network). In this chapter we consider the latter approach.

The remainder of this chapter is organized as follows. In Sect. 6.2, we discuss classical papers on the set covering location problem (in Sect. 6.2.1) and the maximal covering location problem (in Sect. 6.2.2), present the results of computational experiments, and discuss more recent variations. Section 6.3 discusses the impact that these models have had and the bodies of research they have inspired, focusing on generalized notions of coverage. Finally, we conclude with Sect. 6.4, suggesting some possible future research directions.

6.2 Historical Contributions

This section first presents the classical models for the set covering location problem by Hakimi (1965) and Toregas et al. (1971). It then continues with a discussion of the maximal covering location problem by Church and ReVelle (1974) in Sect. 6.2.2.

6.2.1 The Set Covering Location Problem

Although the generic (non-location) set-covering problem had been formulated prior to Hakimi's (1965) seminal paper on the set covering location problem, Hakimi's work is important for, among other things, introducing the notion of coverage into facility location models. Hakimi's proposed solution method, which involved the use of Boolean functions, never proved to be efficient enough to warrant its use in practice; rather, the set covering location problem is generally solved using integer programming techniques, first proposed by Toregas et al. (1971). We discuss Hakimi's model and briefly outline the Boolean-function approach in Sect. 6.2.1.1. Section 6.2.1.2 presents the integer programming method of Toregas et al.

6.2.1.1 The Contribution by Hakimi (1965)

We consider a graph $G = (V, A)$ and assume that every node in V is both a customer (demand) node and a potential site for a facility. (However, one can easily extend

the models below to handle the cases in which some customers are not facilities or some facilities are not customers and therefore do not need to be covered. Below, we use terms like "customer i" and "facility j" as shorthand for "the customer located at node i" and "the facility located at node j.") Let $n = |V|$. The distance between nodes i and j is given by d_{ij}, and the maximum allowable distance between a customer and its nearest opened facility—the "coverage distance"—is given by s. If $(i, j) \in A$, then d_{ij} is the length of the arc (i, j), and otherwise it is the shortest distance from i to j on the graph. (We use the term "distance" throughout, but the parameters d_{ij} and s may just as well represent travel times, costs, or another measure of proximity.) Therefore, facility j covers customer i if $d_{ji} \leq s$. We define

$$V_i = \{j \in V : d_{ji} \leq s\},$$

that is, V_i is the set of nodes that cover customer i. Note that every V_i is nonempty, assuming that $d_{ii} = 0$ for all i.

The objective of the set covering location problem is to find the minimum-cost (or minimum-cardinality) set of locations such that every node in V is covered by some node in the set. The application that Hakimi cites for the set covering location problem is that of locating policemen along a highway network so that every intersection (vertex of the graph) is within one distance unit of a policeman. Subsequently, the problem has found a much broader range of applications, as discussed earlier.

We will assume that facilities may be located only at the nodes of the network, not along the edges. Note that it may be optimal to locate along edges, since the well known "Hakimi property"—which states that an optimal solution always exists in which facilities are located at the nodes, rather than along the edges, of the network—does *not* apply to the set covering location problem. (Hakimi introduced his famous property in an earlier paper (Hakimi 1964) in the context of the p-median problem, not of the SCLP.) A very simple counterexample consists of two nodes connected by a single edge of length 1 and a coverage distance of 0.5. If facilities are allowed on the edges, the unique optimal solution consists of one facility (located in the middle of the edge), whereas the optimal nodal solution consists of two facilities, one at each node. On the other hand, a problem in which facilities may be located on edges may be converted to a node-only problem by inserting dummy nodes onto the edges, taking advantage of the fact that there are only a finite number of possible optimal locations along edges. Readers are referred to Church and Meadows (1979) for details.

In some applications, it is desirable to use a different coverage distance for each customer—for example, if customers have service agreements that specify different response times. In this case, the coverage distance is customer dependent, s_i, and the set V_i is given by $V_i = \{j \in V : d_{ji} \leq s_i\}$. The analysis below changes in only minor ways.

The set covering location problem is closely related to the graph-theoretic *vertex cover problem*, whose objective is to find a subset of nodes in the graph such that every node in the graph is adjacent to some node in the set *and* such that no strict subset of the set has the same property. Such a set of nodes is called a *cover*. The optimization version of the vertex cover problem seeks the minimum-cardinality cover, and this problem is a special case of the SCLP in which $s = 1$ and $d_{ij} = 1$ for

all $(i, j) \in A$. Indeed, although he is usually credited for introducing the more general SCLP, this special case is the problem considered by Hakimi (1965), since he presented the problem explicitly in a facility location context. In this section we will assume, following Hakimi, that $s = d_{ij} = 1$, though in subsequent sections we will allow s and d_{ij} to be arbitrary. Hakimi notes that the assumption that $d_{ij} = 1$ is not too restrictive, since if the arc lengths are greater than 1, one could simply introduce dummy nodes along the arcs one unit apart, assuming that the arc lengths are integers. Of course, this modeling trick comes at considerable computational expense, especially since Hakimi's method relies on an enumerative approach whose computational complexity increases exponentially with the number of nodes.

In the remainder of this section, we describe Hakimi's (1965) approach to solving the set covering location problem. As noted earlier, this method is not commonly used today and is discussed here primarily for its historical interest.

Recall that V_i is the set of nodes that cover node i; given the assumption of unit arc-lengths and unit coverage distance, V_i is simply the set of nodes that are adjacent to i, plus i itself. Let S be a subset of the node set V. For each node i, we define a Boolean (binary) variable x_i that equals 1 if $i \in S$ and 0 otherwise. With a slight abuse of notation, we can write

$$S = \bigcup_{i \in V} x_i i,$$

where $x_i i$ is taken to equal the set $\{i\}$ if $x_i = 1$ and the null set otherwise. We also define X_i as the sum of the Boolean variables for the nodes in V_i; that is,

$$X_i = \sum_{j \in V_i} x_j.$$

Here, $\sum$ represents Boolean summation, analogous to the "or" operator, in which $1 + 1 = 1$. Then $X_i = 1$ if and only if S contains a node that covers node i. Finally, we define the Boolean function f, which takes as inputs the vector of Boolean variables for the nodes and returns a single Boolean value:

$$f(x_1, \ldots, x_n) = \prod_{i \in V} X_i.$$

Since node i is covered if and only if $X_i = 1$, we have the following theorem:

Theorem 1: *S contains a covering of V if and only if $f(x_1,\ldots,x_n) = 1$.*

The advantage of using the function f is that it allows us to use Boolean algebra to construct coverings of V. Although this approach still involves enumerating all coverings, it allows us to do so without enumerating all subsets of V to identify them. In particular, we will create a "minimum sum of products," i.e., the smallest possible sum of products of x_i variables that is logically equivalent to $f(x_1,\ldots,x_n)$. This method involves eliminating terms that are implied by others, then using Boolean algebra to simplify the resulting formula until we have an expression consisting

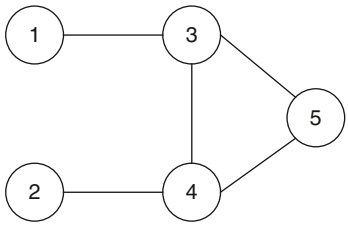


Fig. 6.1 Sample network

of the sum of simple products of variables such that no product is implied by (contains) any other. The method is best explained by use of an example.

Example 1: We illustrate the method using the sample network in Fig. 6.1.

Using the adjacencies depicted in Fig. 6.1, $X_1 = x_1 + x_3$, $X_2 = x_2 + x_4$, and so on. Therefore,

$$f(x_1, \ldots, x_n) = (x_1 + x_3)(x_2 + x_4)(x_3 + x_1 + x_2 + x_4 + x_5)$$
$$(x_4 + x_2 + x_3 + x_5)(x_5 + x_3 + x_4).$$

Using Theorem 1 to find all coverings of the graph, we need to find all possible values of $\{x_1,\ldots,x_5\}$ that make $f(x_1,\ldots,x_n) = 1$, meaning that all terms in the above product equal 1.

To begin, note that the first term is contained in the third. Since we need each term to equal 1, the third term equals 1 if the first does; we can therefore eliminate the third term. Similarly, the fourth term contains the fifth, so we can eliminate the fourth term. The resulting expression is

$$f(x_1, \ldots, x_n) = (x_1 + x_3)(x_2 + x_4)(x_3 + x_4 + x_5).$$

Boolean algebra contains two distributive laws. One says that, for any Boolean variables x, y, and z,

$$x + (yz) = (x + y)(x + z).$$

Applying this law to the last two terms, we get

$$f(x_1, \ldots, x_n) = (x_1 + x_3)(x_4 + x_2x_3 + x_2x_5).$$

The other Boolean distributive law says that

$$x(y + z) = xy + xz.$$

Applying this law to multiply the two terms, and repeatedly applying both Boolean identity laws (which say that $x+x = x$ and that $xx = x$), we obtain

$$f(x_1, \ldots, x_n) = x_1x_4 + x_1x_2x_3 + x_1x_2x_5 + x_3x_4 + x_2x_3 + x_2x_3x_5.$$

Finally, by the Boolean redundancy law ($x+xy=x$), we can remove the second and last terms:

$$f(x_1,\ldots,x_n) = x_1x_4 + x_1x_2x_5 + x_3x_4 + x_2x_3.$$

Therefore, the covers for the graph in Fig. 6.1 are

$$\{1,4\},\{1,2,5\},\{3,4\},\{2,3\}.$$

All but {1, 2, 5} are minimum covers.

Hakimi was optimistic that this enumerative approach would prove to be practical: "…since the subject of simplification of Boolean functions has been widely studied and there are efficient digital computer programs for such a purpose, the above formulation is feasible." Twenty-first century readers, however, will recognize that the enumerative approach is impractical for large instances. Moreover, since the vertex cover problem is **NP**-complete (Garey and Johnson 1979), no polynomial-time exact algorithm for the set covering location problem is known to exist. However, more efficient approaches than Hakimi's exist; we discuss a mathematical-programming-based approach in the next section.

6.2.1.2 The Contribution by Toregas et al. (1971)

Toregas et al. (1971) formulate the set covering location problem as an integer programming problem and use standard mathematical programming methods to solve it. We discuss their approach next.

The integer programming problem has one set of decision variables:

$$x_j = \begin{cases} 1, & \text{if a facility is opened at node } j \\ 0, & \text{otherwise} \end{cases}$$

for $j \in V$. Note that variable x_j has no relation to the Boolean variables x_i defined in Sect. 6.2.1.1.

The integer programming problem is formulated as follows:

$$\text{SCLP: min } z = \sum_{j\in V} x_j \tag{6.1}$$

$$\text{s.t.} \sum_{j\in V_i} x_j \geq 1 \quad \forall i \in V \tag{6.2}$$

$$x_j \in \{0,1\} \quad \forall j \in J \tag{6.3}$$

The objective function (6.1) computes the total number of facilities opened. Constraints (6.2) require at least one node from the coverage set V_i to be opened for

each node i. Constraints (6.3) are standard integrality constraints. Here, we do not assume that $s = d_{ij} = 1$ (as we did in Sect. 6.2.1.1); any values for these parameters may be used in determining the coverage sets V_i.

This formulation is virtually identical to that of the classical set covering problem; here it is discussed in the context of location theory in particular. It is well known that the set-covering problem typically has a small integrality gap; that is, the optimal objective value of the linear programming relaxation (denoted by z_{LP}) is close to that of the integer program itself (Bramel and Simchi-Levi 1997), and often the linear programming relaxation even has all-integer solutions. In fact, ReVelle (1993) argues that many facility location problems have this property and discusses "integer-friendly programming" techniques for several classical problems. However, there do exist instances of the set covering location problem whose linear programming relaxations do not have all-integer optimal solutions (otherwise the problem would not be **NP**-hard). An example follows.

Example 2: Consider the network depicted in Fig. 6.2. In this example, $s = 1$. An optimal solution to the linear programming relaxation of SCLP is given by $x_1 = x_2 = x_3 = 0.5$, $x_4 = 0$, with an objective value of $z_{LP} = 1.5$.

Since the coefficient of each x_j is 1 in the objective function of SCLP, it is clear that the objective function value is integer for any solution to the integer program. Since z_{LP} is a lower bound on z^*, the optimal objective function value for the integer program, and since z^* must be integer, we can assert that

$$z^* \geq \lceil z_{LP} \rceil ,$$

where $\lceil a \rceil$ denotes the smallest integer greater than or equal to a. Therefore, Toregas et al. propose adding the following cut to SCLP:

$$\sum_{j \in J} x \geq \lceil z_{LP} \rceil . \tag{6.4}$$

We denote the resulting problem SCLP-C. The new cut may eliminate some fractional solutions, and the linear programming relaxation to SCLP-C may have an all-integer solution as a result.

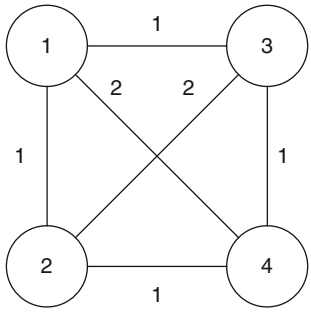


Fig. 6.2 Network for Example 2

For the example in Fig. 6.2, the problem SCLP-C does indeed have an integer solution: $x_1 = x_2 = 1, x_3 = x_4 = 0$, for example, with $z^* = 2$. (It also has optimal fractional solutions, e.g., $x_j = 0.5$ for all j, but the simplex method would find integer solutions since these represent extreme points of the feasible region.)

Toregas et al. therefore propose a two-step solution procedure for the set covering location problem.

Step 1: Solve the linear programming relaxation of SCLP. If the optimal solution is integer, STOP.

Step 2: Otherwise, solve the linear programming relaxation of SCLP-C using the optimal objective value from step 1 in the right-hand side of (6.4).

Even with constraint (6.4), the linear programming relaxation may not have an integer solution. Toregas et al. report that they found no such instance in their computational experiments, though we found several such instances in ours, see Sect. 6.2.1.3 "Computational Experiment". In fact, Rao (1974) gives two counterexamples: in one, the addition of cut (6.4) results in a fractional solution; in the other, the addition of cut (6.4) results in an integer but non-optimal solution. (See also the reply to Rao's note by Toregas et al. 1974).

Toregas et al. also discuss the relationship between the set covering location problem and a variant of the p-median problem in which each customer may only be served by facilities that are within a distance of s. The formulation is obtained simply by forcing the assignment variable to be 0 for facility–customer pairs that are more than s units apart, or, alternately, by indexing the assignment variables for each customer i over facilities j in V_i, as opposed to all facilities j in V. (We omit the formulation here.)

The optimal objective value of this p-median variant changes with s. For sufficiently large s, the objective function value is no different from the p-median without distance constraints; as s decreases, the objective function value increases as a step function; and for sufficiently small s, the problem is infeasible. Toregas et al. argue that the solution to the set covering location problem provides some information about the feasibility of this problem. In particular, for a given value of p, the smallest value of s for which the p-median variant is feasible is equal to the smallest value of s for which SCLP has an optimal objective value of p. On the other hand, the solution to a set covering location problem does not provide any information about the breakpoints of the step function that relates the p-median objective to s.

6.2.1.3 Experiments and Variants

In the "Computational Experiment" section below, we discuss the results of our computational experiment related to SCLP. In "Row and Column Reduction", we discuss a technique for reducing the problem size of the set covering location problem, and in "Facility Fixed Costs", we discuss a variant involving fixed costs.

Computational Experiment

We performed a computational experiment to confirm the results reported by Toregas et al.—namely, that the linear programming gap for SCLP is small, and that cut (6.4) produces integer solutions. For each value of $n = 50, 100, 200, 400, 800$, we generated 100 random instances of the set covering location problem. Parameters were generated as follows:

- x- and y-coordinates were drawn from $U[0,100]$,
- distances were calculated using the Euclidean metric, and
- the coverage distance s was drawn from $U[0,140]$ ($140 \approx$ maximum possible distance between two points in 100×100 grid).

For each instance, we solved the linear programming relaxation of SCLP using CPLEX v. 10.2.0 to obtain z_{LP}. If the optimal solution to the linear program was not integer, we added cut (6.4) and solved the linear programming relaxation to SCLP-C to obtain $z_{LP\text{-}C}$. If the optimal solution was still not integer, we solved SCLP as an integer program to obtain z^*. (If either of the linear programming relaxations resulted in integer solutions, their objective values give us z^*.)

The results are shown in Table 6.1. The columns labeled "% Integer" list the percentage of instances for which the linear programming relaxation produced an integer optimal solution. The columns labeled "Avg LP Gap" and "Max LP Gap" list the average and maximum, respectively, of the linear programming gap, measured as $(z_{LP} - z^*)/z^*$ for SCLP and $(z_{LP\text{-}C} - z^*)/z^*$ for SCLP-C.

The linear programming gap for SCLP is small and tends to decrease with lower values of n. The largest gap we found was 33.5% for a problem with $n = 800$. The addition of cut (6.4) reduces the linear programming gap substantially (from 0.0132 to 0.0004, on average), but does not guarantee integer solutions—even with the cut, 11.2% of instances had fractional optimal solutions. Several of these

Table 6.1 Performance of linear programming relaxations of SCLP and SCLP-C

n	% Integer	Avg LP Gap	Max LP Gap
SCLP			
50	94.0	0.0068	0.2500
100	87.0	0.0104	0.1667
200	88.0	0.0074	0.2500
400	73.0	0.0217	0.3350
800	76.0	0.0195	0.2500
Total	83.6	0.0132	0.3350
SCLP-C			
50	98.0	0.0000	0.0000
100	92.0	0.0000	0.0000
200	90.0	0.0000	0.0000
400	82.0	0.0003	0.0250
800	82.0	0.0017	0.0714
Total	88.8	0.0004	0.0714

instances also had integer optimal solutions, though CPLEX did not find these. In general, CPLEX solved the integer programming problem in well under one minute on a laptop computer, even for the largest problems.

Row and Column Reduction

The size of SCLP can often be reduced substantially by using row- and column-reduction techniques. These methods exploit the coverage structure by eliminating rows and columns that are dominated by others. In particular:

- A facility j_1 dominates another facility j_2 if it covers all of the customers that j_2 does; that is, if $j_2 \in V_i$ implies $j_1 \in V_i$ for all $i \in V$. In this case, there is no reason to open facility j_2 since j_1 covers all the same customers and possibly more. Therefore we can set $x_{j2} = 0$, or equivalently, eliminate the column corresponding to j_2.
- A customer i_1 dominates another customer i_2 if every facility that covers i_1 also covers i_2; that is, if $V_{i_1} \subseteq V_{i_2}$. In this case, if constraint (6.2) holds for i_1 it also holds for i_2, and therefore we can eliminate the row corresponding to i_2.

Row and column reduction techniques are appropriate for the SCLP because of the binary nature of coverage. Most facility location problems with distance objectives cannot generally accommodate these techniques, except heuristically, since under most metrics it is impossible for a facility to dominate another, i.e., to be closer to every customer than another facility is.

These techniques were proposed by Toregas and ReVelle (1972). See also Daskin (1995) and Eiselt and Sandblom (2004) for thorough discussions and examples of row- and column-reduction techniques.

Facility Fixed Costs

If the facilities each have a different fixed cost f_j, then the problem becomes choosing facilities to cover all demands at minimum possible cost. This problem can be formulated simply by replacing the objective function (6.1) with the objective

$$\text{Minimize } z = \sum_{j \in V} f_j x_j.$$

The set covering location problem as formulated above is a special case in which $f_j = 1$ for all j. The linear-programming-based solution methods described in Sect. 6.2.3 can easily accommodate this variation. So can the Boolean-function approach: at the final step, we simply choose the cover that has the smallest total fixed cost.

6.2.2 The Maximal Covering Location Problem

Whereas the set covering location problem has the form

SCLP: minimize the number of facilities opened,
s.t. cover all demand,

the *Maximal Covering Location Problem* MCLP has the inverse form:

MCLP: maximize the demand covered,
s.t. a limit on number of facilities opened.

The set covering location problem treats all demand nodes as equivalent since the coverage constraint applies equally to all. In contrast, in the maximal covering location problem nodes are weighted by the demand that they generate, and the objective favors coverage of larger demands over smaller ones. As the number of allowable facilities increases, the demand covered naturally increases as well. The modeler can plot a tradeoff curve depicting the performance of a range of solutions along these two dimensions; the decision maker can then choose a solution based on this tradeoff.

In Sect. 6.2.2.1, we discuss the maximal covering location problem as formulated by Church and ReVelle (1974). Section 6.2.2.2 then describes some computational experiments and several variants of the model.

6.2.2.1 Church and ReVelle (1974)

This section commences with a formal statement of the maximal covering location problem as a mathematical programming model. The next section discusses heuristics, followed by an exact algorithm in "Linear Programming Approach". "Mandatory Closeness Constraints" investigates the effects of a constraint that enforces an additional level of coverage.

Introduction and Formulation

Our notation in this section is identical to that in Sect. 6.2.1, with the addition of two new parameters: a_i is the demand at node i per unit time, and p is the maximum allowable number of facilities. We also introduce a new set of decision variables:

$$y_i = \begin{cases} 1, & \text{if customer } i \text{ is covered by some facility} \\ 0, & \text{otherwise} \end{cases}$$

The maximal covering location problem is formulated by Church and ReVelle (1974) as follows:

$$\text{MCLP: Max } z = \sum_{i \in V} a_i y_i \tag{6.5}$$

$$\text{s.t.} \sum_{j \in V_i} x_j \geq y_i \quad \forall i \in V \tag{6.6}$$

$$\sum_{j \in V} x_j = p \tag{6.7}$$

$$x_j \in \{0,1\} \quad \forall j \in V \tag{6.8}$$

$$y_i \in \{0,1\} \quad \forall i \in V \tag{6.9}$$

The objective function (6.5) computes the total demand covered. Constraints (6.6) prohibit a customer from counting as "covered" unless some facility that covers it has been opened. Constraint (6.7) requires exactly p facilities to be opened. Constraints (6.8) and (6.9) are standard integrality constraints. (In fact, it suffices to relax constraints (6.9) to $0 \leq y_i \leq 1$, since integer values for the y_i are optimal if the x_j are integer.)

Church and ReVelle cite White and Case (1973) as formulating a similar model to MCLP that maximizes the number of demand nodes covered, rather than the total demand. Case and White's model is therefore a special case of the maximal covering location problem in which $a_i = 1$ for all i.

Church and ReVelle also present an alternate formulation that uses a new decision variable $\bar{y}_i$ defined as $\bar{y}_i = 1 - y_i$; that is,

$$\bar{y}_i = \begin{cases} 1, & \text{if customer } i \text{ is not covered by any facility} \\ 0, & \text{otherwise} \end{cases}$$

In the alternate formulation, constraints (6.6) are replaced by

$$\sum_{j \in V_i} x_j + \bar{y}_i \geq 1 \quad \forall i \in V.$$

The revised constraints state that if node i is not covered by any facility (i.e., $\sum_{j \in V_i} x_j = 0$), then $\bar{y}_i$ must equal 1. The objective function (6.5) can be rewritten as

$$\text{maximize} \sum_{i \in V} a_i (1 - \bar{y}_i) = \sum_{i \in V} a_i - \sum_{i \in V} a_i \bar{y}_i, \tag{6.10}$$

or equivalently,

$$\text{minimize} \sum_{i \in V} a_i \bar{y}_i, \tag{6.11}$$

since the first term in (6.10) is a constant. The revised objective (6.11) minimizes the uncovered demand. The revised formulation is then given by

$$\text{MCLP2: Min } z = \sum_{i \in V} a_i \bar{y}_i \tag{6.12}$$

$$\text{s.t.} \sum_{j \in V_i} x_j + \bar{y}_i \geq 1 \quad \forall i \in V \tag{6.13}$$

$$\sum_{j \in V} x_j = p \tag{6.14}$$

$$x_j \in \{0, 1\} \quad \forall j \in V \tag{6.15}$$

$$y_i \in \{0, 1\} \quad \forall i \in V. \tag{6.16}$$

The two formulations are mathematically equivalent, as are their linear programming relaxations.

Megiddo et al. (1983) proved that the maximal covering location problem is **NP**-hard. The next two sections describe heuristic and exact approaches to solving the problem, all of which are discussed by Church and ReVelle (1974).

Heuristic Solution Methods

Like many facility location problems, the maximal covering location problem lends itself nicely to greedy heuristics such as the *Greedy Adding* heuristic, which Church and ReVelle (1974) credit to Church's (1974) doctoral dissertation. The Greedy Adding heuristic begins with all facilities closed, then opens p facilities in sequence, choosing at each iteration the facility that increases coverage the most. For a discussion of greedy and other heuristics for facility location problems, see Current et al. (2002).

Solutions obtained with the Greedy Adding heuristic are nested in the sense that all of the facilities in the solution to the p-facility problem are also opened in the solution to the $(p+1)$-facility problem. Optimal solutions to the maximal covering location problem are not, in general, nested in this way. Therefore, Church and ReVelle also suggest an alternate heuristic, called the *Greedy Adding with Substitution* heuristic, which attempts to rectify this problem by allowing an open facility to be closed and a closed facility to be opened at each iteration. Like any heuristic, Greedy Adding and the Greedy Adding with Substitution are not guaranteed to find the optimal solution. The latter, however, tends to perform well in practice, and both heuristics execute very quickly.

Linear Programming Approach

Church and ReVelle propose solving MCLP2 directly using linear programming and branch-and-bound. Like the set covering location problem, the linear programming relaxation of the maximal covering location problem often yields all-integer solutions: Church and ReVelle report that approximately 80% of their test instances had integer solutions; we found an even higher percentage in our computational tests (Sect. 6.2.2.2 "Computational Experiment"). Branch-and-bound may be applied to resolve fractional solutions to the linear programming relaxation, but Church and ReVelle also suggest a method that is effective when solving the same problem for consecutive values of p.

The method takes as input a fractional solution to the p-facility problem and an integer solution to the $(p-1)$-facility problem. It is effective when the $(p-1)$-facility solution covers all but a few nodes. We illustrate the method using an example.

Example 3: Consider an instance of the maximal covering location problem for which the total demand across all nodes is 100 units. Suppose we have found an integer solution to the 4-facility problem and that it covers all but two nodes, for a total of 91 demand units covered. These two uncovered nodes (we will call them 1 and 2) have demands of 3 and 6, respectively. Suppose further that the linear programming relaxation to the 5-facility problem is fractional and covers 98 demand units. Finally, suppose that the minimum a_i among all nodes i is 3.

The optimal integer solution with $p = 5$ cannot cover all of the nodes, since the linear programming relaxation has an objective value of 98. In fact, the integer solution may cover at most 97 demand units, since at best it leaves the 3-demand node uncovered. We can create an integer solution to the $p = 5$ problem by adding node 2 to the $p = 4$ solution. Since the $p = 4$ solution covered 91 demands, not including node 2, this new solution covers $91 + 6 = 97$ demands. This solution must be optimal for $p = 5$ since 97 is an upper bound on the objective value. An optimal solution for the problem with $p = 6$ can now be found by adding node 1 to the $p = 5$ solution; the resulting solution covers all demands.

Church and ReVelle refer to this method as the "inspection" method. It can be summarized as follows. Let $z_{\mathrm{IP}}(p)$ be the optimal p-facility objective value of MCLP, that is, the optimal demand covered by p facilities, and let $z_{\mathrm{LP}}(p)$ be the optimal p-facility objective value of the linear programming relaxation of MCLP. We assume that we know the integer optimal solution with $p-1$ facilities and that the optimal solution to the linear programming relaxation with p facilities is not integer. Let $a_{\min} = \min\{a_i : i \in V\}$ and $a_\Sigma = \sum_{i \in V} a_i$. We summarize the inspection method in the following theorem. (Church and ReVelle illustrate this method with an example, rather than stating it formally as a theorem.)

Theorem 2: *Suppose the following conditions hold:*

1. $Z_{LP}(p) < a_\Sigma$, *and*
2. $Z_{IP}(p-1) + a_i = a_\Sigma - a_{min}$ *for some node i that is not covered in the optimal solution to the (p − 1)-facility problem,*

then an optimal solution to the p-facility problem consists of the optimal solution to the (p − 1)-facility problem plus node i.

Church and ReVelle report that, of the 20% of their test instances where linear programming relaxations did not have integer solutions, half could be solved using the inspection method. The other half was solved via branch-and-bound.

Mandatory Closeness Constraints

Church and ReVelle discuss a variant of the maximal covering location problem in which we require that *all* customers be covered within a secondary coverage distance t ($t \geq s$). For example, we might want to maximize the demand covered within 50 miles but require all demands to be covered within 100 miles. This model, known as the MCLP with Mandatory Closeness Constraints, can be viewed as a hybrid between the maximal covering location problem and the set covering location problem, since it has a max-coverage objective plus a hard coverage constraint.

The problem can be formulated simply by adding the following constraint to either formulation of the MCLP:

$$\sum_{j \in U_i} x_j \geq 1 \quad \forall i \in V,$$

where $U_i = \{j \in V: d_{ji} \leq t\}$. The resulting model can be solved using linear programming and branch-and-bound.

Suppose we solve SCLP and find that, for a given instance, the minimum number of facilities that covers all demand nodes with a coverage distance of t is p^*. Generally there are many optimal solutions to this problem. The maximal covering location problem with mandatory closeness constraints gives us a mechanism for choosing among these, by selecting the solution that also maximizes the demands covered within some distance s. In particular, we solve MCLP with mandatory closeness constraints using p^* as the number of facilities to open and t as the secondary coverage distance.

6.2.2.2 Experiments and Variants

Computational Experiment

We performed a computational experiment to verify Church and ReVelle's claim that the MCLP often results in all-integer solutions. We set n = 50, 100, 200, 400, 800. For each value of n, we generated 100 random instances and tested three different values of p. The random instances were generated as described in Sect. 6.2.1.3 "Computational Experiment", with one additional parameter: Demands a_i were drawn from $U[0,100]$.

Table 6.2 Performance of linear programming relaxation of MCLP2

n	p	% Integer	Avg LP Gap	Avg LP Gap > 0	Max LP Gap
50	2	95.0	0.0011	0.0542	0.0646
	5	96.0	0.0019	0.0635	0.1109
	8	99.0	0.0000	–	0.0000
100	2	100.0	0.0000	–	0.0000
	5	98.0	0.0002	0.0232	0.0232
	8	98.0	0.0000	–	0.0000
200	4	96.0	0.0016	0.0540	0.1293
	10	93.0	0.0092	0.1308	0.3957
	16	92.0	0.0028	0.0699	0.1296
400	4	98.0	0.0000	–	0.0000
	10	92.0	0.0006	0.0190	0.0280
	16	92.0	0.0158	0.5254	0.9632
800	4	100.0	0.0000	–	0.0000
	10	91.0	0.0002	0.0089	0.0089
	16	89.0	0.0195	0.4865	0.9704
Total		95.3	0.0035		0.9704

We solved the linear programming relaxation of MCLP2 using CPLEX v. 10.2.0 and, if the solution was not all integer, we solved the integer program. The results are displayed in Table 6.2. The column labeled "p" gives the value of p in MCLP2. The column labeled "Avg LP Gap > 0" gives the average integrality gap among only those instances with a positive integrality gap, or "—" if there were no such instances. All other columns are interpreted as in Table 6.1.

The linear programming relaxation of MCLP seems to generate integer solutions even more frequently than the relaxation of SCLP (at least for our test instances): an average of 95.3% of the time. When it fails to do so, the integrality gap can be quite large, though this is partly a function of the minimization objective, which may have optimal values near zero and hence any suboptimal solution may have a large error on a percentage basis.

Note that for some instances the linear programming relaxation had fractional solutions but an integrality gap of 0, as evidenced by the fact that some rows have "% Integer" <100% but an average linear programming gap of 0. For these instances, an optimal integer solution exists for the linear programming relaxation but CPLEX returned a fractional optimal solution instead.

Tradeoff Curve

Figure 6.3 displays the optimal objective function value of MCLP2—the number of demand units uncovered—as p varies for a particular random instance with $n = 100$ and $s = 15$. As expected, the uncovered demand decreases as p increases. For $p \geq 18$, all demands are covered. The convex shape is typical of tradeoff curves for the maximal covering location problem, meaning that additional facilities provide decreasing marginal returns in terms of additional coverage.

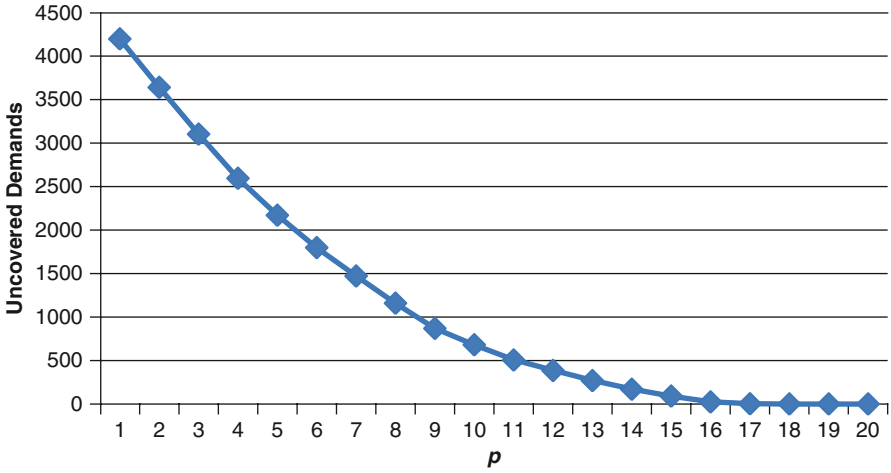


Fig. 6.3 Tradeoff curve: demands uncovered vs. p

Lagrangian Relaxation Approach

The maximal covering location problem can also be solved using Lagrangian relaxation. The key idea is to remove a set of constraints and add a penalty to the objective function for violating the constraints. The resulting problem is easier to solve but may produce solutions that are infeasible for MCLP. By adjusting the objective-function penalties iteratively, the solutions found approach the optimal solution for the maximal cover location problem. The use of Lagrangian relaxation for MCLP was detailed by Galvão and ReVelle (1996), although Daskin et al. (1989) also report computational results from a similar method without providing details. See Fisher (1981, 1985) for an excellent overview of Lagrangian relaxation.

We illustrate the Lagrangian relaxation method using formulation MCLP, though it can also be applied to MCLP2. We relax constraints (6.6) using Lagrangian multipliers λ_i to obtain the following Lagrangian subproblem:

$$\begin{aligned} \text{MCLP-LR: Max } z &= \sum_{i \in V} a_i y_i + \sum_{i \in V} \lambda_i \left(\sum_{j \in V_i} x_j - y_i \right) \\ &= \sum_{i \in V} (a_i - \lambda_i) y_i + \sum_{j \in V} \left(\sum_{i \in V : j \in V_i} \lambda_i \right) x_j \end{aligned} \tag{6.17}$$

$$\text{s.t.} \sum_{j \in V} x_j = p \tag{6.18}$$

$$x_j \in \{0, 1\} \quad \forall j \in V \tag{6.19}$$

$$y_i \in \{0,1\} \quad \forall i \in V \tag{6.20}$$

This problem decouples by x and y since there are no constraints involving both sets of variables. As a result, it can be solved easily. The optimal y-values are given by

$$y_i = \begin{cases} 1, & \text{if } a_i - \lambda_i > 0, \\ 0, & \text{otherwise.} \end{cases}$$

To find the optimal x-values, we set $x_j = 1$ for the p facilities with the largest values of $\sum_{i \in V: j \in V_i} \lambda_i$. The optimal objective value of MCLP-LR provides an upper bound on that of MCLP. Feasible (lower bound) solutions can be found by setting $x_j = 1$ for the p facilities that are opened in the upper-bound solution and setting $y_i = 1$ for each customer i that is covered by some existing facility. Lagrange multipliers can be updated using subgradient optimization, and branch-and-bound can be used if the Lagrangian procedure fails to yield a suitably small optimality gap; see Daskin (1995) for more details. Daskin et al. (1989) report that the procedure works quite well, especially when the lower-bound heuristic is supplemented by a substitution heuristic.

Budget Constraints

We can incorporate fixed costs into the model in a similar manner as we did for the set covering location problem in "Facility Fixed Costs". Here, the fixed cost appears in the constraints rather than the objective function. In particular, we replace constraint (6.7) or (6.14) with

$$\sum_{j \in V} f_j x_j \leq B,$$

where B is a budget imposed exogenously on the total fixed costs. This constraint can be easily handled by the linear programming approach discussed in Sect. 6.2.2.1, but it somewhat complicates the Lagrangian approach in Sect. 6.2.2.2 since determining the optimal x values now requires us to solve the following knapsack problem:

$$\text{Max} \sum_{j \in V} \left(\sum_{i \in V: j \in V_i} \lambda_i \right) x_j$$

$$\text{s.t.} \sum_{j \in V} f_j x_j \leq B$$

$$x_j \in \{0, 1\} \quad \forall j \in V.$$

Although this problem can be solved quite quickly using modern codes, it is still **NP**-hard, and it may slow the Lagrangian procedure significantly.

Relationship to *p*-Median Problem

The maximal covering location problem can be formulated as a special case of the *p*-median problem through a simple transformation of the distance matrix. In particular, we set

$$d_{ji} = \begin{cases} 0, & \text{if } j \in N_i \\ 1, & \text{otherwise.} \end{cases}$$

That is, we redefine the distance metric so that the distance from node j to node i is 0 if j covers i and 1 otherwise. The *p*-median problem is then formulated as usual (see, e.g., Daskin 1995). The optimal solution will cover as many demand units as possible using p facilities. Any algorithm for the *p*-median problem can then be applied to solve the maximal covering location problem.

6.3 Extensions

The literature contains many enhancements to the set covering and maximal covering location problems. In this section, we focus in particular on generalizations of the notion of coverage. One common criticism of the two types of problems is that they assume that all customers within a facility's coverage radius can be served by the facility, and served equally. In practice, facilities are not always available when needed, especially in the public-sector arena where facilities may represent such essential services as ambulances and fire crews. One approach to this issue is *backup coverage*, in which customers are required or encouraged to be covered by more than one open facility. Another approach is *expected coverage*, which accounts for probabilistic information. Moreover, in many cases the coverage benefit changes as the distance between a customer and its assigned facility changes. This dependency is captured by the notion of *gradual coverage*. We briefly discuss models for backup, expected, and gradual coverage in the next three subsections. For thorough reviews of backup and expected coverage models, see Daskin et al. (1988) or Berman and Krass (2002).

6.3.1 Backup Coverage Models

Both the set covering location problem and the maximal covering location problem have been extended to consider solutions in which customers are covered by more

than one facility. One may *require* backup coverage in order for a customer to count as "covered," or one may simply *reward* solutions for backup coverage.

6.3.1.1 Required Backup Coverage

It is simple to formulate a required-backup version of either covering problem. In the set covering location problem, we simply modify constraints (6.2) to read

$$\sum_{j \in V_i} x_j \geq m \quad \forall i \in V,$$

where m is the desired number of times that each customer is to be covered. In the maximal covering location problem, we can replace constraints (6.6) with

$$\sum_{j \in V_i} x_j \geq m y_i \quad \forall i \in V,$$

where y_i must equal 0 unless at least m facilities that cover customer i are open. This constraint is likely to weaken the linear programming relaxation of MCLP, however.

6.3.1.2 Rewards for Backup Coverage

We focus on models in which $m = 2$. Extensions to these models to consider $m > 2$ are straightforward but often yield weaker linear programming relaxations, as discussed above. Let

$$w_i = \begin{cases} 1, & \text{if customer } i \text{ is covered by two or more facilities} \\ 0, & \text{otherwise.} \end{cases}$$

The models formulated below contain a reward in the objective function for each customer who is covered twice. However, the backup coverage reward is strictly a secondary objective; in no case should a solution with more facilities have a better objective than one with fewer facilities, even if it has better backup coverage.

Daskin and Stern (1981) propose the following model for the set covering location problem with backup coverage:

$$\text{SCLP-BC: Min } z = (|V| + 1) \sum_{j \in V} x_j - \sum_{i \in V} w_i \tag{6.21}$$

$$\text{s.t.} \sum_{j \in V_i} x_j - w_i \geq 1 \quad \forall i \in V \tag{6.22}$$

$$x_j \in \{0,1\} \quad \forall j \in V \tag{6.23}$$

$$w_i \in \{0,1\} \quad \forall i \in V. \tag{6.24}$$

The objective function (6.21) enforces the hierarchical nature of the primary objective (minimizing the number of facilities) and the secondary one (maximizing twice-covered customers). It does so by multiplying the primary objective by a constant large enough that even if the primary objective is as small as possible (equal to 1), the secondary objective can never exceed it. Therefore, the solution will never open more facilities than necessary solely to improve the secondary objective. Constraints (6.22) require each customer to be covered at least once, and prohibit w_i from equaling 1 unless customer i is covered at least twice.

Another advantage of this formulation is that its solutions avoid facilities that are dominated by others in the sense described in "Row and Column Reduction". As a result, the linear programming relaxation to SCLP-BC is more likely to have all-integer solutions than that of SCLP is. Readers are referred to Daskin and Stern (1981) for justifications for both of these claims.

A similar hierarchical version of the maximal covering location problem was introduced by Storbeck (1982) and reformulated by Daskin et al. (1988). We modify their formulation somewhat in what follows.

$$\text{MCLP-BC: Max } z = (|V|+1)\sum_{i\in V} a_i y_i + \sum_{i\in V} w_i \tag{6.25}$$

$$\text{s.t.} \sum_{j\in V_i} x_j - y_i - w_i \geq 0 \quad \forall i \in V \tag{6.26}$$

$$\sum_{j\in V} x_j = p \tag{6.27}$$

$$x_j \in \{0,1\} \quad \forall j \in V \tag{6.28}$$

$$y_j \in \{0,1\} \quad \forall i \in V \tag{6.29}$$

$$w_j \in \{0,1\} \quad \forall i \in V. \tag{6.30}$$

The objective function (6.25) maximizes a sum of the primary coverage (first term) and backup coverage (second term); the weight on the first term ensures that primary coverage will never be sacrificed in order to achieve backup coverage. Note that the secondary coverage objective considers *nodes* covered, rather than *demand units* covered. This is required in order for the weighting to achieve the desired hierarchy. Constraints (6.26) stipulate that customer i may be considered covered ($y_i = 1$) only if at least one facility in V_i is open, and may be considered twice covered ($w_i = 1$) only if two such facilities are open. Since the objective function coef-

ficient for y_i is greater than that for w_i, the model will always set $y_i = 1$ before it sets $w_i = 1$, thus ensuring the desired coverage hierarchy.

6.3.2 Expected Coverage Models

The class of expected coverage models is descended primarily from the Maximum Expected Covering Location Problem (*MEXCLP*) introduced by Daskin (1982). Daskin's primary application is in the siting of emergency medical service vehicles. The *MEXCLP* maximizes the *expected* coverage of each node, defined using probabilistic information about facility availability, subject to a constraint on the number of facilities.

The *MEXCLP* assumes that the average system-wide probability that a given facility (vehicle) is busy is given by q. If a customer is covered by k facilities, then the probability that all those facilities are busy at a given point in time is given by q^k, and the probability that at least one facility is available is $1-q^k$. The maximum expected covering location problem defines new variables to keep track of the number of covering facilities for each customer. Define variables

$$y_{im} = \begin{cases} 1, & \text{if customer } i \text{ is covered by at least } m \text{ facilities} \\ 0, & \text{otherwise} \end{cases}$$

for all $i \in V$ and $m = 1,\ldots,p$. Note that if customer i is covered by *exactly* k facilities, then $y_{im} = 1$ for $m = 1,\ldots,k$ and $y_{im} = 0$ for $m = k+1,\ldots,p$. Then

$$\sum_{m=1}^{p} (1-q)q^{m-1} y_{im} = \sum_{m=0}^{k-1} (1-q)q^m = 1 - q^k$$

using a standard formula for geometric sums. In other words, the first summation in the equation above expresses the probability that customer i is covered by an available facility in terms of the decision variables y_{im}. Using this approach, Daskin formulates the *MEXCLP* as follows:

$$MEXCLP: \text{Max } z = \sum_{i \in V} \sum_{m=1}^{p} (1-q)q^{m-1} a_i y_{im} \tag{6.31}$$

$$\text{s.t.} \sum_{m=1}^{p} y_{im} - \sum_{j \in V_i} x_j \leq 0 \quad \forall i \in V \tag{6.32}$$

$$\sum_{j \in V} x_j = p \tag{6.33}$$

$$x_j \in \{0,1\} \quad \forall j \in V \tag{6.34}$$

$$y_j \in \{0,1\} \quad \forall i \in V. \tag{6.35}$$

The objective function (6.31) calculates the expected coverage. Constraints (6.32) allow the total number of y_{im} variables, for fixed i, to be no more than the total number of opened facilities that cover i. At first it may seem that the model needs a constraint of the form

$$y_{im} \leq y_{i,m+1} \quad \forall i \in V, m = 1, \ldots, p-1$$

in order to ensure that y_{im} is set to 1 for the correct values of m; that is, for the k smallest values of m, where k is the number of opened facilities that cover i. However, such a constraint is not necessary since the objective function coefficient is larger for smaller values of m; the model will automatically set $y_{im} = 1$ for the k smallest values of m.

Daskin (1983) proposes a heuristic for *MEXCLP* based on node exchanges, and several metaheuristics have been proposed subsequently; see, e.g., Aytug and Saydam (2002), and Rajagopalan et al. (2007).

The primary criticism that has been leveled at the *MEXCLP* concerns the assumption of a uniform system-wide availability probability, since availability might vary based on geographic area or on the demand assigned to each facility. ReVelle and Hogan (1989) address this concern in the *Maximum Availability Location Problem (MALP)*, a chance-constrained version of MCLP. They formulate two versions of the model, one in which the availability probability is assumed to be the same throughout the system; the main difference between this model and *MEXCLP* is that *MALP* maximizes the number of demand units that are covered with at least a certain probability, whereas *MEXCLP* includes the expected coverage in the objective. ReVelle and Hogan's second *MALP* model estimates the busy probability separately for each customer by assuming that facilities within the coverage radius of a given customer are available only to that customer. Obviously this assumption is not true, but it provides an easy, and fairly accurate, estimate of the availability probability. The two models are nearly identical once the availability probabilities are calculated. Galvão et al. (2005) present a framework that attempts to unify *MEXCLP* and *MALP*.

Batta et al. (1989) embed Larson's (1974, 1975) hypercube queuing model into *MEXCLP* to compute the availability probabilities endogenously. They find that their model disagrees substantially with *MEXCLP* in terms of the expected coverage predicted, but nevertheless results in similar sets of facilities chosen. Marianov and ReVelle (1996) formulate a version of the *MEXCLP* that endogenously calculates the availability property using a queuing model at each facility. The region around each customer node is treated as an $M/M/s/s$ queue, where s is the number of servers located within the coverage radius. Their model implicitly assumes that that the call rate in the neighborhood is not substantially different from that in adjacent neighborhoods. The resulting model is structurally similar to the *MALP* but uses different (but pre-computable) values for the coverage radius.

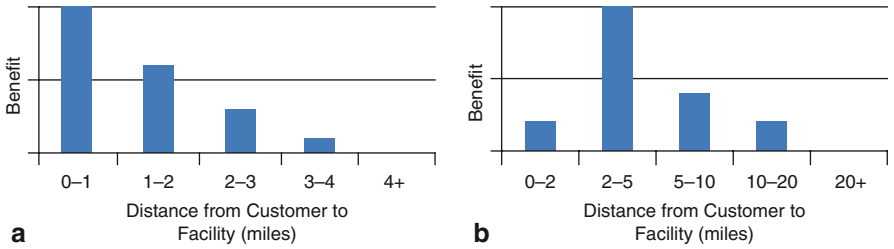


Fig. 6.4 Benefit of coverage versus distance: **a** strictly decreasing, **b** non-monotonic

6.3.3 Gradual Covering Models

The models discussed in this chapter so far all assume that coverage is a binary concept: either a customer is covered or it is not, and the distance from the customer to the covering facility is irrelevant. In practice, though, customers who are located very close to a facility such as a fire station may be served better than those located farther away, even if both customers are within the nominal coverage radius. In this case, the benefit from coverage decreases with the customer–facility distance, as illustrated in Fig. 6.4a. Moreover, some facilities such as garbage dumps are most beneficial when they are close (to reduce transportation costs) but not too close (to reduce odors and truck traffic), as illustrated in Fig. 6.4b.

Church and Roberts (1983) introduce the Weighted Benefit Maximal Coverage (*WBMC*) Model, which extends the maximal covering location problem to accommodate non-binary coverage benefits. The objective is to maximize the sum of all customers' coverage benefits (defined as the benefit per unit of demand times the demand of that customer) subject to a constraint on the number of facilities located. The formulation is a relatively straightforward modification of MCLP and includes a coverage variable (y) and a constraint for each customer–distance pair. (Each "distance" is really a range of distances, as in Fig. 6.4.) The number of variables and constraints therefore grows linearly with the number of distance ranges. If the benefits are not monotonically decreasing with the distance, as in Fig. 6.4b, then an additional set of constraints is required to ensure that customers are assigned to their nearest opened facilities, a property that is automatic if benefits are monotonically decreasing. The resulting formulations are more complex than MCLP, but Church and Roberts find that they still retain their "integer-friendliness:" the linear programming relaxation is generally very tight and often all-integer.

6.4 Conclusions and Future Research Directions

In this chapter we have discussed two classical models for locating facilities to ensure coverage of customer nodes. One model, the set covering location problem, requires *every* customer to be covered and does so with the minimum

number of facilities, while the other, the maximal covering location problem maximizes the demand covered subject to a limit on the number of facilities. Both models have garnered considerable attention in the location theory literature, and both models (and their extensions) have been widely applied in practice, especially in public-sector applications such as the location of emergency medical services.

Both covering problems are reasonably easy to solve, in the sense that modern general-purpose integer programming solvers such as CPLEX can solve problems with hundreds or thousands of nodes to optimality in a few minutes on a desktop computer. This stems in part from the fact that the linear programming relaxations of both problems tend to be tight, and even yield integer optimal solutions for a large percentage of instances. Therefore, although these problems are **NP**-hard, they are among the easiest problems in that class.

On the other hand, many of the extensions of these models are much more computationally challenging. Daskin's (1982) *MEXCLP* model, for example, or the queuing-based congestion models discussed by Berman and Krass (2002), have more complex structures than SCLP or MCLP and therefore cannot be solved using off-the-shelf solvers, except for small instances. One important direction for future research, therefore, is the development of effective, accurate algorithms and heuristics for extensions of SCLP and MCLP.

Of particular interest are stochastic and robust variants of coverage models. Although the literature on stochastic facility location models is extensive (see, e.g., Snyder 2006 for a review), most such models consider cost-based objectives rather than coverage-based ones. (Notable exceptions are the expected-coverage models described in Sect. 6.3.2, and their variants.) An important topic for future study is therefore the incorporation of stochastic elements—such as demands, travel times, server availabilities, and supply disruptions—into coverage models. The resulting models are likely to be significantly more complex than their deterministic counterparts, but the stochastic programming and robust optimization literatures are vast, and many of their more sophisticated tools have yet to be tapped by the location science community.

The distinction between cost- and coverage-based models made in the previous paragraph is an important one since it is often equivalent to the distinction between private- and public-sector applications—the former is primarily concerned with cost minimization while the latter is often encouraged or mandated to provide adequate coverage to all demand locations (ReVelle et al. 1970). Public-sector and humanitarian applications have gained increased attention in the operations research community in recent years—for example, the 2008 INFORMS Annual Meeting featured "Doing Good with OR" as a central theme, as did the February 2008 issue of *OR/MS Today*. The application of coverage models to emergency medical services and other services has been a success story in public operations research for decades, and recent renewed interest provides an opportunity for existing and new coverage models to be applied for the public good.

References

Aytug H, Saydam C (2002) Solving large-scale maximum expected covering location problems by genetic algorithms: a comparative study. Eur J Oper Res 141:480–494

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

Berman O, Krass D (2002) Facility location problems with stochastic demands and congestion. In: Drezner Z, Hamacher HW (eds) Facility location: applications and theory. Springer, New York (Chapter 11)

Bramel J, Simchi-Levi D (1997) On the effectiveness of set covering formulations for the vehicle routing problem with time windows. Oper Res 45:295–301

Church R (1974) Synthesis of a class of public facilities location models. PhD thesis, The Johns Hopkins University, Baltimore

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

Church RL, Meadows B (1979) Location modelling using maximum service distance criteria. Geogr Anal 11:358–373

Church RL, Roberts KL (1983) Generalized coverage models and public facility location. Pap Reg Sci 53:117–135

Church RL, Stoms DM, Davis FW (1996) Reserve selection as a maximal covering location problem. Biol Conserv 76:105–112

Current J, Daskin MS, Schilling D (2002) Discrete network location models. In: Drezner Z, Hamacher HW (eds) Facility location: applications and theory. Springer, New York (Chapter 3)

Daskin MS (1982) Application of an expected covering model to emergency medical service system design. Decis Sci 13:416–439

Daskin MS (1983) A maximum expected covering location model: formulation, properties and heuristic solution. Transp Sci 17:48–70

Daskin MS (1995) Network and discrete location: models, algorithms, and applications. Wiley, New York

Daskin MS, Stern EH (1981) A hierarchical objective set covering model for emergency medical service vehicle deployment. Transp Sci 15:137–152

Daskin MS, Hogan K, ReVelle C (1988) Integration of multiple, excess, backup, and expected covering models. Environ Plan B 15:15–35

Daskin MS, Haghani AE, Khanal M, Malandraki C (1989) Aggregation effects in maximum covering models. Ann Oper Res 18:115–140

Eaton DJ, Daskin MS, Simmons D, Bulloch B, Jansma G (1985) Determining emergency medical service vehicle deployment in Austin, Texas. Interfaces 15:96–108

Eiselt HA, Sandblom C-L (2004) Decision analysis, location models, and scheduling problems. Springer, New York

Fisher ML (1981) The Lagrangian relaxation method for solving integer programming problems. Manag Sci 27:1–18

Fisher ML (1985) An applications oriented guide to Lagrangian relaxation. Interfaces 15:10–21

Flynn J, Ratick S (1988) A multiobjective hierarchical covering model for the essential air services program. Transp Sci 22:139–147

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, ReVelle C (1996) A Lagrangean heuristic for the maximal covering location problem. Eur J Oper Res 88:114–123

Garey MR, Johnson DS (1979) Computers and intractability: a guide to the theory of NP-completeness. Freeman, New York

Gleason JM (1975) A set covering approach to bus stop location. Omega 3:605–608

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
Larson RC (1974) A hypercube queuing model for facility location and redistricting in urban emergency services. Comput Oper Res 1:67–95
Larson RC (1975) Approximating the performance of urban emergency service systems. Oper Res 23:845–868
Marianov V, ReVelle C (1996) The queueing maximal availability location problem: a model for the siting of emergency vehicles. Eur J Oper Res 93:110–120
Megiddo N, Zemel E, Hakimi SL (1983) The maximum coverage location problem. SIAM J Algebra Discrete Method 4:253–261
Nozick LK, Turnquist MA (2001) Inventory, transportation, service quality and the location of distribution centers. Eur J Oper Res 129:362–371
Rajagopalan HK, Vergara FE, Saydam C, Xiao J (2007) Developing effective meta-heuristics for a probabilistic location model via experimental design. Eur J Oper Res 177:83–101
Rao A (1974) Counterexamples for the location of emergency service facilities. Oper Res 22:1259–1261
ReVelle C (1993) Facility siting and integer-friendly programming. Eur J Oper Res 65:147–158
ReVelle C, Hogan K (1989) The maximum availability location problem. Transp Sci 23:192–200
ReVelle C, Marks D, Liebman JC (1970) An analysis of private and public sector location models. Manag Sci 16:692–707
Schilling DA, ReVelle C, Cohon J, Elzinga DJ (1980) Some models for fire protection locational decisions. Eur J Oper Res 5:1–7
Snyder LV (2006) Facility location under uncertainty: a review. IIE Transactions 38:537–554
Storbeck JE (1982) Slack, natural slack and location covering. Socioecon Plan Sci 16:99–105
Toregas C, ReVelle C (1972) Optimal location under time or distance constraints. Pap Reg Sci Assoc 28:133–143
Toregas C, Swain R, ReVelle C, Bergman L (1971) The location of emergency service facilities. Oper Res 19:1363–1373
Toregas C, ReVelle C, Swain R (1974) Reply to Rao's note on the location of emergency service facilities. Oper Res 22:1262–1267
White J, Case K (1973) On covering problems and the central facilities location problem. Unpublished paper, Virginia Polytechnic Institute and State University, Blacksburg

Part V
Other Location Models

Chapter 7
Equilibria in Competitive Location Models

H. A. Eiselt

7.1 Introduction

Whereas the usual location models locate facilities based on the wishes and objectives of a single decision maker, competitive location models consider the location of facilities that are under the jurisdiction of more than one decision maker. The economist Hotelling (1929) was the first to introduce competition into location models. His results stood unchallenged for fifty years, until d'Aspremont et al. (1979) corrected an inconsistency that invalidated Hotelling's main result. Nonetheless, this has not diminished the originality and importance of the original contribution, and it is also the reason why the present paper reviews Hotelling's contribution and its impact on location models with multiple decision makers.

Arguably, the best way to deal with competitive location models is to assess their components. Most prominent among them are the number of decision makers involved, the pricing policy, the rules of the game, and the behavior of the customers. Eiselt et al. (1993) provide a taxonomy and annotated bibliography that includes these features. Rather than restating their description, I will only very briefly summarize the main features. The most prominent *pricing policies* include *mill pricing*, where the price at each branch is fixed by the decision maker and customers provide for their own transportation, *spatial price discrimination*, where the firm sets the price a customer will be charged for the goods that are delivered to his place, and *uniform delivered pricing*, in which case all customers will receive the good for the same price (which typically means that customers located closer to a branch of the firm will subsidize those farther away). Other policies such as *zone pricing* may also be investigated.

The rules of the game are more complex. They essentially include rules that govern the process of decision making. In particular, they specify whether the firms' decisions are made sequentially or simultaneously. In case of pure location competi-

H. A. Eiselt (✉)
Faculty of Business Administration, University of New Brunswick,
Fredericton, NB E3B 5A3, Canada
e-mail: haeiselt@unb.ca

H. A. Eiselt, V. Marianov (eds.), *Foundations of Location Analysis,*
International Series in Operations Research & Management Science 155,
DOI 10.1007/978-1-4419-7572-0_7,

tion, i.e., the case in which all firms compete only in terms of locations, a sequential process would indicate that, say, firm *A* locates first, followed by firm *B*, then firm *C*, and so forth. This location with *foresight* is discussed in Chap. 8 of this volume. The distinguishing feature of the sequential location process is its asymmetry. The first firm, being aware of the fact that other firms will locate after it has chosen locations for its own branches, will take this knowledge into account and use what Teitz (1968) called "conservative maximization." Subsequent firms will also attempt to guard themselves against firms that follow but, at the same time, will take the locations of already existing firms into account. This chapter deals exclusively with simultaneous location.

The situation becomes more complex when variables other than location exist. Many authors, including Hotelling (1929) in his seminal work, allow the firms to not only choose locations for their branches, but also to determine their prices. One possibility is to require that all firms make their choices simultaneously. Most authors, however (including Hotelling) use a two-stage process: in the first stage, all firms simultaneously choose their respective locations. Once these choices have been made, they are revealed to all firms. In the second stage, all firms then simultaneously determine the prices they want to charge. This sequence has been chosen as the much more permanent location decision comes first, followed by the price decision, which can easily be adjusted or modified later on. Furthermore, when making a decision in Stage 1, firms will anticipate the price competition in Stage 2. Such a game will be solved by backward recursion: for each pair of locations, the two firms will independently determine their optimal prices. Given those prices, firms will then—again independently—determine their optimal locations.

One question that arises rather naturally in all of these models is whether or not the set of locations that arises from such a process is stable. The concept applied here is the Nash (sometimes also referred to as Cournot-Nash) equilibrium. Loosely speaking, a Nash equilibrium is a situation in which none of the firms has an incentive (meaning can improve its objective) by unilaterally changing any of its parameters, be it location, price, quantity, or any of the other variables in the model. Most papers, especially in the economic literature, investigate whether such an equilibrium exists in the model under considerations, and, if so, if it is unique. While simple Nash equilibria can be determined in pure location competition, the two-stage "first location, then price" game requires a refinement of the equilibrium concept. The optimality concept that applies in such a procedure is Selten's (1975) subgame perfect Nash equilibrium.

In addition to locating facilities such as warehouses, retail stores, fast food outlets, gas stations, or other facilities of this nature, it has also been suggested to use location models for seemingly unrelated problems such as the design of brands, the determination of positions for political candidates, or the allocation of tasks to employees. The main features of these nonphysical location models are described below.

First consider the design of products, which is typically referred to as the *brand positioning problem*. In this application, we first define a continuous "feature space," in which each dimension represents a specific feature of the class of prod-

ucts under consideration. For example, in the case of automobiles these features could include horsepower, maximal speed (or, alternatively, acceleration), and gas mileage. Clearly, it is required that each feature under consideration be quantitative. Also note the correlation between some of the factors, e.g., horsepower and gas mileage. The products are then also mapped into space according to their features. This is followed by the mapping of (potential) customers, who are also mapped into the feature space by their respective ideal points, i.e., the product features they would like best. It then stands to reason that a customer will evaluate a product based on the distance between his own ideal point and the location of the product. The reason is that, just like physical distances, the distance between potential customer and product in a feature space expresses the disutility of a customer for that product. And, continuing that line of argument, a potential customer will choose the product that is closest to his own ideal point. One problem associated with this model is the existence of features such as price and gas consumption, which have an ideal point that is zero (or, if you will, negative infinity). Anderson et al. (1982) suggest an "outside game," a construct that allows the meaningful inclusion of such features in the model.

Another somewhat similar application is found in the area of political science. While the spatial analysis of political scenarios is not at all new—consider the classical contributions by Downs (1957) and Black (1958)—advances in location analysis helped tremendously to improve modeling and the solution of political models. Models of this nature first construct an "issue space," an n-dimensional space in which each dimension represents a political issue that is deemed relevant in an election. One of the key problems of the analysis is the quantification and measurability of issues, such as domestic policies, economic policies, etc. Candidates and likely voters are then mapped into this space by way of their ideal point (for the voters) and their stand on the issues (for the candidates) respectively, and assuming that—following some metric—voters will vote for the candidate closest to their own ideal point. That way, it is possible to determine the number of voters that will vote for each of the candidates and, more importantly, how each of the candidates should redefine his stand on the issues so as to maximize the number of votes he will obtain. In addition to the aforementioned difficulty of measurability there is also the determination of the ideal points of millions of voters. In their seminal contribution on the subject, Rusk and Weisberg (1976) used more or less well-defined groups such as "policemen," "urban rioters," "Republicans," "Democrats," and others to determine their average ideal point and, with the help of the variance determined by a sample, define a "cloud" around this ideal point that will then represent the voters in this group. The authors get around the problem of measurability of the axes by applying a multidimensional scaling technique (see, for example, Kruskal 1964). Additional contributions can be found in the other papers in the edited volume by Niemi and Weisberg (1976). It is also worth pointing out that one of the few features of this model that makes the political positioning simpler than the Hotelling's original scenario is the absence of prices in the model.

The workload allocation problem follows a similar logic. Here, tasks and employees are mapped into an ability space that expresses their requirements and abili-

ties, respectively. The idea is to allocate tasks to employees so as to minimize the distance between employee and task, matching the requirements of the tasks and the employees' abilities as closely as possible. A close match may be desired to increase job satisfaction and hence avoid high job turnovers, absenteeism, and other work-related problems. Again, some of the main problems related to these applications are the quantifications of the abilities and the determination of an appropriate distance function. Readers are referred to Schmalensee and Thisse (1988) for their survey on applications in feature spaces and ability spaces. For a recent reference, see Eiselt and Marianov (2008a).

The contributions surveyed in this paper all have one feature in common: they all emphasize the analysis of equilibria in competitive location models. Other aspects of competitive location models are dealt with in Chaps. 8 and 9 of this volume.

7.2 Hotelling (1929): Competitive Location on a Linear Market

Hotelling starts his paper with a critical evaluation of past contributions. Of interest are particularly the embedding of his own work into the framework provided by Bertrand and Cournot. The discussion of a duopoly dates back to Cournot (1838). In his model, Cournot considers a duopoly with both firms competing on the same market with the same product. The variable costs have been normalized to zero (we may assume that they have been deducted from the price that the firms charge), and the two firms face a common demand function. The duopolists compete in quantities and the resulting solution is a Cournot-Nash equilibrium. Bertrand (1883), on the other hand, has duopolists competing in prices. Such competition is very intense, as even a slight undercutting will revert the entire market to the cheaper firm. While Hotelling's contribution is in the footsteps of these two (and other) predecessors, its novelty is that he includes competition in space, while his predecessors' models were set in a spaceless economy.

Hotelling's basic model includes a space in the form of a closed line segment of length ℓ. It is worth noting that Hotelling justified the choice of a line segment by referring to it as "main street" or a stretch of a transcontinental railroad. Later authors have claimed that Hotelling's "justification" of the "linear market" was based on "two ice cream vendors on a beach," an example never envisaged by Hotelling but put forth by later contributors.

Customer demand is distributed uniformly along the line at a unit density, so that the total demand equals ℓ. The demand is assumed to be completely inelastic. Two competing firms face the task of simultaneously locating one facility each and setting the price for a homogeneous product. Both firms use mill pricing, so that customers have to drive to the facility of their choice, pay for the product at the facility, and then ship it home: their full price includes the mill price charged at the facility and the transportation costs for shipping the good from the facility to

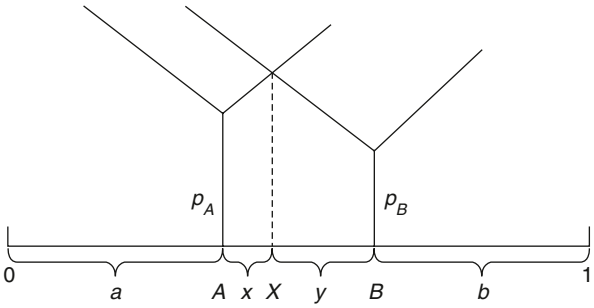


Fig. 7.1 Price functions of duopolists on a line segment

their home. Given a homogeneous (standardized) good, customers are indifferent between purchasing the good from either facility, so that they will choose the facility from which they can obtain the good for the lower full price, regardless of how distant the closest facility is. The transportation costs are assumed to be linear in the distance. The two firms are assumed to have equal cost functions, which have been normalized to zero.

Formally, define the market as a line segment of length ℓ and assume that firm A is located a units from the left end of the market, while firm B is located at a distance of b from the right end of the market. The only condition is that firm A is located to the left of firm B (which does not restrict generality, as this situation, if violated, can always been achieved by exchanging the names of the facilities). The facilities charge mill prices of p_A and p_B, respectively, and the unit transportation costs are c. Figure 7.1 shows the present situation. Each of the Y-shaped functions shows the full price (the mill price plus transportation costs) customers have to pay if they purchase from the facility in question: the stem of the "Y" is the mill price, and the slope of the two branches of the "Y" is the unit transportation cost c. Given that the good is homogeneous, customers will purchase from the source with the lower full price, i.e., the lower envelope of the branches of the two "Ys." This results in a marginal customer X (Hotelling did not use the expression), who is defined as a customer indifferent between purchasing from firm A or from firm B. Clearly, all customers to the left of the marginal customer can buy the good more cheaply from firm A, while those to the right of X can purchase the good more cheaply from firm B. This will define firm A's market area from the left end of the market to the marginal customer, while firm B's market area extends from the marginal customer to the right end of the market.

Authors who followed Hotelling usually refer to the region to the left of A as "A's hinterland," the region to the right of B as "B's hinterland," and the area between firms A and B as the "competitive region." (It appears that Smithies (1941) was the first author to use these terms.) The two hinterlands are of length a and b, and the competitive region is divided by the marginal customer X into pieces of lengths x and y, respectively. Formally, we have

$$a + x + y + b = \ell, \tag{7.1}$$

and the marginal customer X is defined as a place at which prices are equal, i.e., $p_A + cx = p_B + cy$ with unit transportation costs c. Solving this system of two equations for x and y, we obtain

$$x = \tfrac{1}{2}[\ell - a - b + \tfrac{1}{c}(p_B - p_A)] \tag{7.2}$$

and

$$y = \tfrac{1}{2}[\ell - a - b + \tfrac{1}{c}(p_A - p_B)], \tag{7.3}$$

so that the profits are

$$\pi_A = p_A q_A = p_A(a + x) = \tfrac{1}{2}(\ell + a - b)p_A + \tfrac{p_B}{2c} p_A - \tfrac{1}{2c} p_A^2 \tag{7.4}$$

and

$$\pi_B = p_B q_B = p_B(b + y) = \tfrac{1}{2}(\ell - a + b)p_B + \tfrac{p_A}{2c} p_B - \tfrac{1}{2c} p_B^2. \tag{7.5}$$

For any given values of ℓ, a, b, and π_i, $i = A, B$, iso-profit lines can be plotted in p_A, p_B space as hyperbolas. Since each duopolist will adjust his own price so as to maximize his profit, we can take partial derivatives

$$\frac{\partial \pi_A}{\partial p_A} = \tfrac{1}{2}(\ell + a - b) + \frac{p_B}{2c} - \tfrac{1}{c} p_A = 0 \tag{7.6a}$$

or

$$p_A^* = \tfrac{1}{2}c(\ell + a - b) + \tfrac{1}{2}p_B \tag{7.6b}$$

and

$$\frac{\partial \pi_B}{\partial p_B} = \tfrac{1}{2}(\ell - a + b) + \frac{p_A}{2c} - \tfrac{1}{c} p_B = 0 \tag{7.7a}$$

or

$$p_B^* = \tfrac{1}{2}c(\ell - a + b) + \tfrac{1}{2}p_A. \tag{7.7b}$$

(Note that $\frac{\partial^2 \pi_A}{\partial p_A^2} < 0$ and $\frac{\partial^2 \pi_B}{\partial p_B^2} < 0$, so that these conditions determine a local maximum).

The expressions (7.6b) and (7.7b) are usually (although not by Hotelling) referred to as reaction functions of the two firms. In particular, if firm B were to set any price p_B, then firm A would react by setting its price to a level specified by relation (7.6b). Similarly, firm B will react by using relation (7.7b) to any price p_A set by its competitor A.

Solving for the prices p_A and p_B results in the equilibrium prices

$$\bar{p}_A = c\left(\ell + \frac{a-b}{3}\right) \tag{7.8a}$$

and

$$\bar{p}_B = c\left(\ell - \frac{a-b}{3}\right), \tag{7.8b}$$

and the quantities at equilibrium are

$$\bar{q}_A = a + x = \tfrac{1}{2}\left(\ell + \frac{a-b}{3}\right) \tag{7.9a}$$

and

$$\bar{q}_B = b + y = \tfrac{1}{2}\left(\ell - \frac{a-b}{3}\right). \tag{7.9b}$$

This can best be explained graphically. Hotelling's original example involves values of $\ell = 35$, $a = 4$, $b = 1$, $c = 1$, and it is shown in Fig. 7.2. Given his numerical example, the optimality conditions result in the reaction functions $p_A^* = 19 + \tfrac{1}{2}p_B$ and $p_B^* = 16 + \tfrac{1}{2}p_A$, respectively. Solving the two linear equations results in the equilibrium prices $\bar{p}_A = 36$ and $\bar{p}_B = 34$.

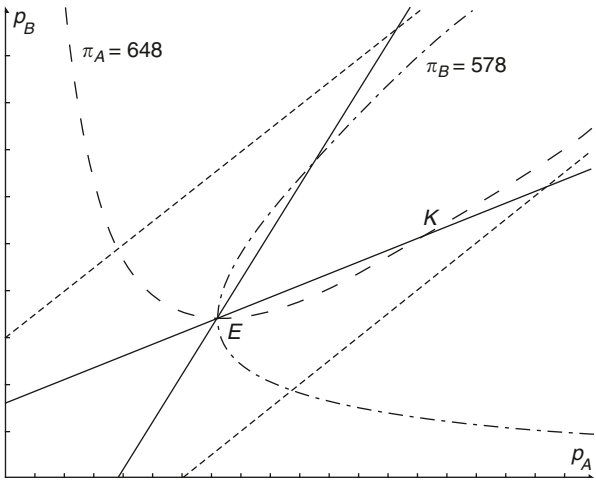


Fig. 7.2 Reaction functions of the duopolists and equilibrium

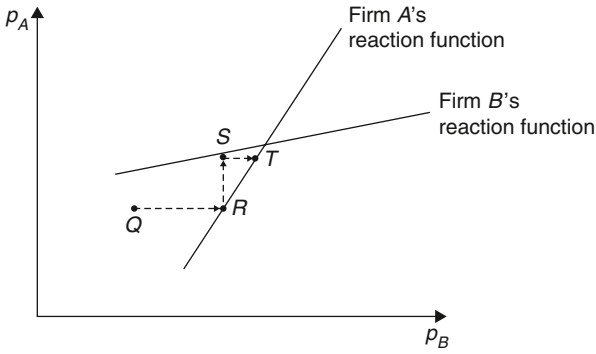


Fig. 7.3 Price adjustments over time

The lines with short dashes define a "corridor" between the lines $p_B \leq p_A + 30$ and $p_B \geq p_A - 30$. This corridor is the set of price combinations in which the price difference is no larger than the cost of shipping one unit from one facility to the other. In other words, it is the area within which neither competitor cuts out its opponent. The solid lines represent the reaction functions that result from the optimality conditions. The steeper line is firm A's reaction function, while the flatter line is firm B's reaction function. The broken line with long dashes denotes the set of price combinations that result in $\pi_A = 648$ (the profit that results from the equilibrium prices at point E, *viz*., $p_A = 36$ and $p_B = 34$). Finally, the broken and dotted line is the set of price combinations that result in $\pi_B = 578$.

Hotelling then describes a procedure in which the two firms start with non-equilibrium prices that they subsequently adjust in sequential fashion. For simplicity, the two reaction functions are shown again in Fig. 7.3, where E again denotes the equilibrium point. Suppose now that the two firms charge prices so as to realize point Q. Given this combination of (below equilibrium) prices, either of the firms has an incentive to change (here: raise) its price. Suppose that firm A will react first. Firm A will assume that, at least for some time, its competitor will not react. This assumption was later referred to as "zero conjectural variation" by Eaton and Lipsey (1975). Furthermore, firm A will act without any foresight and consequently move from point Q to the point on firm A's reaction function, which is point R. Once this has been accomplished, firm B will react and move to the point on its reaction function, *viz*., point S. Then firm A reacts again by moving to point T, and so on. The price adjustment from points in any of the three other cones is similar. Note also the similarity of the adjustment process here to that in the famed cobweb theorem in economic theory.

At this point, Hotelling remarks in a footnote that the above conclusions are true only as long as the difference in price does not exceed the cost of shipping one unit from A to B or vice versa. Formally, the condition is

$$|p_A - p_B| \leq c(\ell - a - b). \tag{7.10}$$

If this condition is not satisfied, the equilibrium is not point E but some other point. It is important to note that Hotelling does indeed realize that his computations are

valid only for a certain range of prices (price differences, to be exact). However, he does not elaborate. Hotelling's result, the clustering of the duopolists at the center of the market has also been referred to as the *principle of minimal differentiation* (in reference to product design and the political model introduced in the beginning of this paper) or *Hotelling's law*.

An interesting case of cooperation results. Starting again at the equilibrium point E in Fig. 7.2, assume that firm A is willing to forego profits in the near future and moves out of point E by raising its price, and moving to the right. Behaving optimally, firm B will again move towards its point on its reaction function by increasing its price as well. As long as firm A's price increase was modest, the point that will be realized will be located on firm B's reaction function to the left of point K. This point does provide both firm A *and* firm B with higher profits than at equilibrium. However, the solution is inherently unstable (similar to the well-known *Prisoner's dilemma*), as firm A has an incentive to increase its profit even more by moving onto its own reaction function. Such a move will, however, result in sequential price adjustments that ultimately lead back to the equilibrium solution E.

Part II of Hotelling's paper deals with a variety of extensions of his basic model, as well as alternative explanations. He first notes that the profits at equilibrium are

$$\bar{\pi}_A = {}^1\!/_2 c\left(\ell + \frac{a-b}{3}\right)^2 \tag{7.11a}$$

and

$$\bar{\pi}_B = {}^1\!/_2 c\left(\ell - \frac{a-b}{3}\right)^2. \tag{7.11b}$$

Given that, it is apparent that the profit of both firms increases with increasing unit transportation costs c. In other words, rather than promoting better means of transportation, the two firms would fare better if transportation were to be made more difficult. The reason is that if transportation were very difficult, each firm could behave as a local monopolist and charge monopolist's prices. It is important to point out that while higher transportation costs as applied to shipments from the firms to their customers do, in fact, increase profits, they will have a detrimental effect on the variable costs as they also apply to shipments from the firms' suppliers to the firms. These costs were neglected in the model. This means that the argument regarding the parameter c is better explained by the existence of tariffs.

The paper then examines the case in which one firm's location (without loss of generality assume this is firm A) has fixed its location and firm B now chooses its own location. Given its profit at equilibrium as shown in relation (7.11b), it is apparent that firm B's profit increases with increasing value of b. In other words, it will pay firm B to locate as close to its competitor as possible. This is again the "agglomeration result" (or "principle of minimum differentiation" as it became known later). However, Hotelling again notes the problem that occurs when the two facilities are sufficiently close so that one firm can cut out its opponent.

Another interesting result relates to the total profit of the two firms, which are, say, governed by a central planner. Formally, we have

$$\pi_A + \pi_B = c\left[\ell^2\left(\frac{a-b}{3}\right)^2\right], \tag{7.12}$$

indicating that it would benefit the planner to have the two facilities locate at sites that are as different from each other as possible, i.e., maximizing $a - b$.

The next few paragraphs of the paper examine the relationship between the solution arrived at by profit maximization as opposed to the solution that optimizes some social objective. The social objective chosen is the minimization of total transportation costs. For simplicity, consider the left end of the market between 0 and A an interval of length a. The transportation costs in this interval for all shipments to the facility at point A are $\int_{t=0}^{a} ct\,dt = \frac{1}{2}ca^2$. Applying this result to all intervals, *viz*., those from *0* to A (an interval of length a), from A to the marginal customer X (an interval of length x), from the marginal customer X to facility B (an interval of length y), and finally the interval from facility B to the end of the market (an interval of length b), results in total transportation costs

$$TTC = \tfrac{1}{2}c\left(a^2 + b^2 + x^2 + y^2\right). \tag{7.13}$$

Given fixed locations of the facilities A and B, the values of a and b are fixed as well, and so is $x + y$. Then $x^2 + y^2$ is minimized, if $x = y$. This, in turn, is only satisfied, if $p_A = p_B$, which, while entirely possible under the direction of a central planner or commissar, is an outcome that is highly unlikely under competition. It does, however, indicate that social planners will prefer equal prices charged at the facilities. Assume now that $a \neq b$. Without loss of generality, let $a > b$, which, given individual profit maximization, implies that at equilibrium, $\bar{p}_A > \bar{p}_B$, see relation (7.8a). This means that some customers in the competitive region, although they are located closer to facility A, will make their purchases and resulting shipments from facility B. This results in higher transportation costs as if they were to make their purchases at facility A, which renders this solution not "socially optimal." In fact, Hotelling states, "Consequently some buyers will ship their purchases from B's store, though they are closer to A's and socially it would be more economical for them to buy from A." This clearly indicates Hotelling's allocation rule assumes that customers purchase their goods from the source that offers the lowest full price (even though he may not advocate this practice). This is worth pointing out since some authors use the term "Hotelling's allocation" to mean the allocation of a customer to his closest facility, which is not correct.

If the facilities can be moved at will, the social optimum again minimizes the function shown in (7.13) with a, b, x, and y all variable and the single constraint that $a + b + x + y = \ell$ plus the nonnegativity constraints. At optimum, all variables assume equal values ($a = b = x = y = \frac{1}{4}\ell$), so that the two facilities are located at the quartiles of the market. The highest transportation cost paid by any customer in this

arrangement is then ¼ℓ. In contrast, competition will have the two facilities cluster at the center of the market (Hotelling again notes the "unimportant qualification" that deals with the possibility of one competitor cutting out its opponent), so the highest possible full price is ½ℓ. The author uses this as an example of "wastefulness of private profit-seeking management."

Another extension deals with additional firms. In the case of individual profit maximization, Hotelling notes that the third firm will locate "close to *A* and *B*, but not between them." In some sense, this anticipates the analyses performed later by Lerner and Singer (1937) and subsequently by Eaton and Lipsey (1975). For more facilities, Hotelling asserts that clustering will occur, but no specifics are given. The case of social optimization for three facilities is again easy: the facilities locate symmetrically at $\frac{1}{6}\ell$, $\frac{3}{6}\ell$, and $\frac{5}{6}\ell$, respectively.

Hotelling then extends the range of applicability of his model from scenarios that involve the physical transportation of items to multidimensional spaces (today referred to as *feature spaces*), in which each dimension symbolizes a (quantifiable) feature of (a class of) products. He uses one dimension to distinguish between different brands of cider, and the attribute of the cider that identifies the particular brand is its sweetness. What used to be facilities in the competitive location model discussed above now represents brands of cider. Customers are again distributed along the line segment, such that each customer is represented by its "most preferred point" (or "ideal point") on the line, such as the point that represents the sweetness of cider that this customer desires most. The distance between a customer's ideal point and a brand is then a measure that expresses the customer's disutility associated with buying and consuming that particular brand of cider.

The results of the preceding analysis, *viz*., the clustering in case of individual profit maximization, then imply "excessive sameness." Hotelling credits this in part to standardization and economies-of-scale in the production process, but also to the results derived in this study. The main lesson for a firm that intends to enter the market with a new product is *not* to make the product identical to existing products (in which case Bertrand price competition would ensue, driving down prices), but design a product that differs slightly from existing products by locating the brand in feature space close, but not too close, to existing brands. Hotelling's assertion that the similarities of political platforms of Republican and Democratic parties (which are again represented by their main issues in *issue space*) can also be attributed to the effects studied here are not valid *per se*, as political models do not involve prices, thus reducing the model to a much simpler version. Some remarks regarding political models are provided in the next section in this chapter as well as Chap. 19 in this volume.

Some further generalization and extensions are discussed. First, Hotelling affirms that demand densities other than the uniform demand distribution used in his analysis provide "no essential change in conclusion." In the case of buyers being located in a two-dimensional plane, the market areas of the two firms are divided by a hyperbola. In case of more than two facilities, the market areas will be bounded by arcs of hyperbolas. In multidimensional spaces (such as feature spaces), the demand density is typically not uniform and it occurs within a finite bounded region. Here,

not all facilities need to belong to the same firm. There is a general tendency among outsiders to move inward and approach the cluster, which is again the agglomeration result of this paper. This result is asserted, but not proven. For more on market areas and their use in location planning, see the Chaps. 18 and 19 in this volume.

An important extension concerns the elasticity of demand. So far, it has been assumed that firms offer a product for which the demand is fixed, i.e., completely inelastic. While this may occur in the case of essential goods, it is highly unlikely for most products. One of the central questions is whether the price or the quantity should be a variable. So far in the analysis, the quantities have been restricted to the constant ℓ. Given elastic demand, this limitation no longer applies and prices or quantities can be used as independent variables. Hotelling asserts that even with elastic demand, the results derived above will remain "qualitatively true," even though there will be less of a tendency to cluster.

7.3 The Impact of Hotelling's Contribution

Hotelling's original paper has sparked controversy, as well as a flurry of papers written about his model and similar scenarios. In their survey and taxonomy, Eiselt et al. (1993) already list about a hundred papers on the subject. Since then, at least another hundred contributions have been published. It is possible to broadly distinguish between two types of contributions: those that deal with the existence of Nash equilibria, and those that examine von Stackelberg solutions. There is no doubt that the impact of Hotelling's paper has been felt by both streams. However, this chapter will only survey those papers that deal with Nash equilibria; von Stackelberg solutions are examined in detail in Chaps. 8 and 9 of this volume. This chapter will follow the developments of those works that can be considered continuations and refinements of Hotelling's work. Most contributions in this area are made by economists, and their tool of choice is game theory.

Those who followed in Hotelling's footsteps generalized his model in various directions. These directions include (but are by no means limited to)

- different spaces
- $n > 2$ facilities,
- different assumptions about competitors' behavior
- different transport cost functions and different pricing policies,
- different assumptions concerning customer behavior,

and other generalizations. A few of the many milestones are highlighted below.

Probably the earliest contribution to deal with Hotelling models is put forward by Lerner and Singer (1937). The authors point to Hotelling's assumption of fixed demand and the customers' willingness to pay any amount to satisfy their demand as one of the main deficiencies of his model, particularly when applying his argument to favor a social/socialist solution as more efficient than a capitalist solution. The authors thus introduce a "demand price," defined as the highest amount customers

are prepared to pay to satisfy their demand. The authors also criticize Hotelling's assumption that stipulates that a facility planner uses more information when choosing a location than when setting a price. The reason is that the location decision in stage 1 of the two-stage "first location, then price" game is made with the assumption that the opponent's price will be what results from a long line of price adaptations. However, in stage 2 this knowledge is no longer assumed to exist. In contrast, Lerner and Singer assumed that a firm's planner will not react when his opponent moves closer and takes a part of his customers, but he will react when undercut so that all of his customers are not supplied by his opponent. This is a concept, a variant of which Eaton and Lipsey (1975) referred to as "zero conjectural variation." This assumption leads to locations at about 3/8 away from the respective ends of the market. A further analysis in the paper assumed again that a firm, whose competitor is in the process of relocating, does not react except if undercut. The last part of the paper dealt with a Hotelling model with fixed and equal prices, resulting in pure location competition. The authors identified a large number of equilibrium locations for $n \geq 2$ facilities. Two competing firms will have a unique equilibrium solution by clustering at the center of the market; this is the "minimum differentiation" result Hotelling envisaged for his own model. The case of three firms is interesting: the two peripheral firms crowd in on the firm between them in order to gain a higher market share until the central firm has no market share left. It then "leapfrogs" to the outside, becomes a peripheral facility itself, and starts moving inwards as well. Teitz (1968) referred to this later as "dancing equilibria," which really means that this case has no equilibrium. For four or more firms locating on the linear market, their locations are at $\frac{1}{2\lceil n \rceil}, \frac{3}{2\lceil n \rceil}, \ldots, \frac{2\lceil n \rceil - 1}{2\lceil n \rceil}$. Finally, in their analysis of the model with price discrimination, the equilibrium locations are at $\frac{1}{2n}, \frac{3}{2n}, \ldots, \frac{2n-1}{2n}$, which happens to be socially optimal in that it minimizes the total transportation costs. An interesting feature of this result is that a customer closer to a facility will have to pay more than one that is more remote from a firm. The reason is that the level of competition close to a firm is fairly low, which increases the price.

Smithies (1941) continued where Lerner and Singer (1937) left off. His particular interest were the assumptions concerning the behavior of the competitors. In particular, Smithies did not believe that competitive price cutting was a reasonable policy, as it would lead to an all-out price war. Given a price-quantity relation, his model included three cases that exhibited different levels of cooperation. In the first case, facilities would charge the same price and would locate symmetrically. This "full quasi-cooperation," as the author called it. This case includes little, if any competition, and it is not surprising that the results would be the same as if a monopolist were to locate two plants. The second behavioral assumption was for both firms to charge identical prices but compete in locations. Finally, case 3 exhibited "full competition" in the sense that both firms independently optimized their prices and locations. The results were examined according to their dependence on freight rates and changes in marginal costs. Kohlberg and Novshek's (1982) contribution followed Smithies in many respects in that each relocating facility would assume that its competitors would keep their locations and prices at the present level, except

if undercutting occurred, in which case the firm that was undercut would reduce its price to its marginal cost. The main result was that there exists a certain length of market below which there is no equilibrium, while in case of longer markets, there exists a unique location-price Nash equilibrium for which the authors provide a necessary and sufficient condition. Along similar lines is the analysis by Stevens (1961), who was probably the first author to use matrix games for a discretized version of Hotelling's game. Given elastic demand similar to Smithies, the result was still central agglomeration.

Another generalization concerns locations on a circle. While the space may appear somewhat contrived, the results indicated not only the fragility of Hotelling equilibria, but also some of the special features of the linear market that are lost on a circle: hinterlands, for instance, are specific to linear markets (and tree networks, for that matter), but they do not exist on circles or on general networks. On a circle, multiple equilibria exist for all cases with two or more facilities, given rectangular demand density functions. Finally, some locational patterns on a disk were investigated regarding their equilibrium status. Based on simulation attempts, the authors conjectured that there is no equilibrium for $n > 2$ facilities.

The aforementioned contribution by Eaton and Lipsey is one of the papers most frequently referred to in the context of Hotelling's result, even though their model is quite different from Hotelling's contribution. Their work first restated the results obtained by Lerner and Singer (1937) before performing a variety of sensitivity analyses on the problem. Their first model was Hotelling's linear market with uniform demand density and the zero conjectural variation, i.e., no foresight. Model 2 was the same as Model 1, but with no zero conjectural variation. It results in minimax strategies, and as such anticipates the results by Prescott and Visscher (1977) that are presented in Chap. 8 of this volume. Finally, their third model is again similar to Model 1, but with the assumption of uniform demand density relaxed. The result for two firms was similar (the facilities will cluster at the median of the density function), and there was no equilibrium for three firms, and there may not be equilibria for more than three firms either, given a condition on the demand function. In particular, the authors proved that for an equilibrium to exist, it is necessary that the number of firms on the market is no more than twice the number of modes in the demand distribution.

The authors then tackled the much more complex problem of equilibria in two-dimensional space. Again, they avoided boundary problems by considering a disk. Due to the difficulty of the problem even with fixed and equal prices, they investigated a number of patterns that are potential candidates for equilibria and determine whether or not they are indeed equilibria. The first pattern has facilities located on a circle around the center of the disk. This pattern self-destructs immediately as soon as individual firms (re-) optimize their location. Pattern 2 is similar, except with one facility at the center of the disk. This pattern also turn out to be unstable. Finally, pattern 3 is the Löschian honeycomb pattern that consists of hexagons. (Details concerning Lösch's work are found in Chap. 20 of this volume.) This pattern also self-destructs immediately as individual firms optimize their locations, thus the authors conjectured that there exists no equilibrium pattern on a disk with $n > 2$

facilities. (The case of $n = 2$ facilities is easily dispensed of: the two facilities cluster at the center of the market, a simple pairing observed on the linear market for $n \geq 4$ facilities.) For a discussion of the case of competition in bounded two-dimensional space, readers are referred to Chap. 19 of this volume.

Probably the most important contribution following Hotelling's work is the short paper by d'Aspremont et al. (1979), published fifty years after the original work appeared. It first pointed out an error in Hotelling's original work that resulted in the wrong conclusion: not only does the duopoly model described by Hotelling not have an equilibrium at the center of the market (central agglomeration), but the model does not have an equilibrium anywhere. Hotelling was aware that his results would need some refinements (see his footnote referred to above), but he was not aware of the severity of the consequences. Actually, the equilibrium he computed for facilities that are located closely together is wrong. However, d'Aspremont et al. (1979) were not the first to recognize that there were problems with Hotelling's analysis. To quote the earlier work by Prescott and Visscher (1977):

> The difficulty with this solution concept, as others have noted (Smithies 1941, Eaton 1976, and Salop 1979) is that when locations in Nash are sufficiently close, Nash equilibrium prices will not exist.

Without resorting to formalities, the lack of an equilibrium can readily be seen by the following arguments. Consider any locational arrangement that has the two facilities not clustered together. First of all, there is an incentive for firm A to move closer to its opponent until the right branch of its Y-shaped full price function coincides with that of firm B. Similarly, firm B has an incentive to move to the left until the left arm of its Y-shaped full price function coincides with that of firm A. Once that has been achieved (note that there is no clustering of the facilities yet), the firm with the lower mill price could lower its price by an arbitrarily small amount and, in doing so, be cheaper on the entire market. In doing so, its profit would jump up, meaning that the cheaper facility certainly has an incentive to undercut its opponent. The more expensive facility could now react by lowering its price so as to undercut its opponent (which is Bertrand's price competition). Once prices have reached a very low level, it would benefit either of the two facilities to move significantly far away from its opponent so as to enjoy a local monopoly and the associated positive profits.

D'Aspremont et al. (1979) used a more formal argument. The authors first proved that any equilibrium if it exists at all, it either has $a + b = \ell$ (both facilities locate at the center at the market), in which case both prices are equal to zero (the Bertrand solution), or $a + b < \ell$, in which case the price difference must satisfy

$$|\bar{p}_A - \bar{p}_B| < c(\ell - a - b). \tag{7.14}$$

Condition (7.14) expresses the requirement that the difference in prices is less than the cost required to ship one unit from one facility to another. If this condition were violated, it would imply that the lower-price facility is able to cut out its opponent and capture the entire market. Clearly, this cannot be an equilibrium solution as the higher-priced facility would be left without a zero profit that it could increase by undercutting its opponent in turn.

We are now able to present a formal expression for the existence of an equilibrium. Recall that the equilibrium profits were determined in relations (7.11a) and (7.11b) as

$$\bar{\pi}_A = \tfrac{1}{2}c\left(\ell + \frac{a-b}{3}\right)^2 \quad \text{and} \quad \bar{\pi}_B = \tfrac{1}{2}c\left(\ell - \frac{a-b}{3}\right)^2.$$

An equilibrium can then only exist if and only if a firm's equilibrium profit is larger than the profit it would obtain if it were to slightly undercut its opponent by some small value ε. If for instance, firm A were to undercut firm B, then its profit would be $p_A\ell$, as it captures the entire market. Assuming that firm A undercuts firm B by setting its price to $p_A = p_B - c(\ell - a - b) - \varepsilon$ with some $\varepsilon > 0$, while firm B charges its equilibrium price $\bar{p}_B$ specified in relation (7.8b), firm A's profit would be $\pi_A = [\bar{p}_B - c(\ell - a - b) - \varepsilon]\ell$. Clearly, an equilibrium can only exist if undercutting does not result in a higher profit than the equilibrium profit. Formally, an equilibrium will exist, if $\bar{\pi}_A \geq \pi_A$, or, equivalently,

$$\tfrac{1}{2}c\left(\ell + \frac{a-b}{3}\right)^2 \geq [\bar{p}_B - c(\ell - a - b) - \varepsilon]\ell.$$

Applying some standard algebraic transformations and repeating the process for firm B, undercutting firm A, we obtain the necessary and sufficient existence conditions for equilibria as

$$\left(\ell + \frac{a-b}{3}\right)^2 \geq \tfrac{4}{3}\ell(a + 2b) \tag{7.15a}$$

and

$$\left(\ell + \frac{a-b}{3}\right)^2 \geq \tfrac{4}{3}\ell(2a + b). \tag{7.15b}$$

Note that for symmetric equilibria $a = b$, so that the conditions (7.15a) and (7.15b) reduce to $a = b \leq \frac{1}{4}\ell$. This means that the condition requires the two facilities being located outside the first and third quartiles, which is, of course, not satisfied by Hotelling's "central agglomeration" result.

The authors continued to examine a model that is identical to that investigated by Hotelling, except that it uses quadratic transportation costs of the type $c(\text{distance})^2$. While physical transportation is unlikely to exhibit such cost function, models with nonphysical spaces very well may. The result is not only that this model does have a unique equilibrium, but that at equilibrium, we have maximum (rather than minimum) differentiation with both firms locating at the respective ends of the market. This is but one indication of the instability of Hotelling models in general. This point was driven home even further by Anderson (1988), who considered a

Hotelling model with a linear quadratic transportation cost function of the type c_1(distance) + c_2(distance)^2. This type of cost function was first introduced by Gabszewicz and Thisse (1986). With this cost function, there exists an equilibrium only if $c_1 = 0$, i.e., the function has no linear part at all, regardless how small. However, for certain pairs of locations with the duopolists located close together, there is a price equilibrium. In case only pure strategies are allowed in stage 1 but mixed strategies are permitted in stage 2, an equilibrium exists only if the transport costs are "sufficiently" convex as expressed by the relation of parameters a and b. The Hotelling model with linear-quadratic transportation costs was picked up again by Hamoudi and Moral (2005).

Shaked (1982) considered a mixed strategy version the Hotelling model with fixed and equal prices and three competitors. Customers were uniformly distributed on the line. Following the result by Dasgupta and Maskin (1986), the solution would be doubly symmetric: both firms use the same mixed strategies, and the strategy is symmetric about ½ℓ. In particular, firms avoid locations in the extreme quartiles and choose locations instead in the central half of the market with equal probability. Osborne and Pitchik (1986) followed this line of investigation. Their model has fixed and equal prices, allows nonuniform demand distributions, and let the firms use mixed strategies. The authors first noted the well-known sensitivity of the model. For instance, for $n \geq 5$ facilities, the model does not have an equilibrium if the customer distribution is either strictly convex or strictly concave, regardless how close the distribution is to uniformity. The main results are: for $n \geq 3$, the game has a symmetric mixed strategy equilibrium and if the customer distribution is symmetric about the center of the market, so is the mixed strategy equilibrium; for $n = 3$, a unique equilibrium exists with one firm at the center of the market and the other two firms using mixed strategies for their locations.

The contribution by Kohlberg (1983) is different, as this appears to be the first paper that includes factors other than price and location. In particular, Kohlberg's model included not only the transportation cost (here interpreted as travel time), but also the time spent waiting at a facility. The waiting time is assumed to be increasing with the facility's market share. The author then proved that, while there is a unique equilibrium in the case of duopolists with both of them locating at the center of the market, there exist no equilibria for $n > 2$ facilities. Silva and Serra (2007) picked up the model but solved an optimization problem in discrete space; however, they do not investigate equilibria.

De Palma et al. (1985) took a different route. In their analysis, they employed Hotelling's original model with locations and prices variable, a linear market of length ℓ, and a uniform demand, but their model included n facilities and a random utility function that expresses the customers' evaluation of customer preferences. The authors put their model in the context of product placement with n products to be located on a line segment that determines the products' feature. A customer's (dis-) like of a product is expressed as a function of the distance between the customer's ideal point on the line and the product's location. The main assumption of their paper was that products and customers are heterogeneous. In particular, customers value purchasing a product according to the function

$$\text{(random utility)} = \text{(valuation of product)} - \text{(price)} - \text{(unit disutility cost } c\text{)} \times \text{(distance)} + \mu\varepsilon_i,$$

where $\mu > 0$ denotes the degree of heterogeneity of customer tastes (so that $\mu = 0$ equals homogeneous tastes), and the random variable ε that has a zero mean and unit variance. It turns out that heterogeneity in the logit function removes discontinuities in the products' profit functions.

After first considering only the location and then only the price model, the authors proved that in the location-price model, for $n > 2$ products and a degree of heterogeneity of $\mu < ½c\ell(1–2/n)$, there is no agglomerated Nash equilibrium, meaning an equilibrium with all facilities locating at the same point. However, if $\mu \geq c\ell$, central agglomeration with equal prices is a Nash equilibrium. In other words, large values of $\mu/c\ell$ lead to clustering, whereas small values of $\mu/c\ell$ result in dispersion. There are no results regarding the existence and the nature of other equilibria. Some tests revealed that equilibria may exist for $n = 3$. In summary, if all customers have very similar tastes, then there exists no equilibrium with similar products, while in case of very diversified customer tastes, products will tend to be the same. One may look at the result from the following angle: if tastes are very similar, then the firms have to diversify the products to appeal to different segments of the customer base, while in case of significantly diverse tastes, all products can occupy a similar position in feature space.

A follow-up of their 1985 paper was provided by De Palma et al. (1987a). The assumptions were again a linear market, fixed and equal prices, a linear transportation cost function, and the same random utility function shown above with μ again denoting the degree of heterogeneity in customers' tastes. Numerical computations reveal the following results: for $\mu/c < 0.157$, no symmetric equilibria exist; for $\mu/c \in [0.157; 1/6]$, only symmetric dispersed equilibria exist; for $\mu/c \in [1/6; 0.27]$, agglomerated and symmetric dispersed equilibria exist; for $\mu/c \geq 0.27$, only agglomerated equilibria exist. As far as an interpretation goes, consider a competitive location model in product (or feature) space. Here, less wealthy customers tend to be nondiscriminating, meaning that they tend not to care that much if a product is not exactly as they would like it to be. This implies that the value of c is small for this group, implying more heterogeneity and a larger value of μ. We can therefore associate a large value of μ/c for less affluent groups, while wealthier groups may be characterized by a small value of μ/c. The results of this study then indicate that less affluent customers with a large μ/c value will end up with products that are very similar to each other, while wealthy customers will face a market segment whose products are significantly different. This can, for instance, be observed in the automobile market, though to a much lesser extent today than ten or twenty years ago.

De Palma et al. (1987b) considered a competitive location model on a linear market that uses uniform delivered pricing. Apart from this feature, the usual Hotelling assumptions apply. Given the reasonable assumption that consumers purchase the product from the firm that offers the lowest delivered price and assuming that the products are perfectly homogeneous, the analysis indicates that there is no location—price equilibrium. The authors then changed the assumption concerning

customer behavior. First of all, they assumed that customer tastes are homogeneous with a degree μ, which is taken as the standard deviation of the distribution of consumer tastes. This is the same assumption made in their earlier papers. The authors then proved that the model has indeed an equilibrium, as long as the degree of heterogeneity is sufficiently large, *viz*., $\mu \geq c\ell/8$. At that equilibrium, central agglomeration occurs. It was also shown that the result generalizes to n firms, in which case the existence condition is $\mu \geq [(n-1)/n](c\ell/4)$. At equilibrium, all firms are clustered at the center of the market and the equilibrium prices are independent of the number of facilities. Comparing the results with those obtained by De Palma et al. (1985) for mill pricing, it turns out that the mill price charged at the facility plus the transportation cost equals the uniform delivered price in this model, and that customers close to the facilities (in particular those inside the first and third quartiles) prefer mill pricing over uniform delivered pricing. The firms' profits are identical in both cases.

The paper by Labbé and Hakimi (1991) considered a network with customers located at the nodes. The delivered prices charged by the firms and paid at the nodes depend on the total quantity of the homogeneous good supplied by the duopolists at the node. The demand-price function is linear and has a negative slope. The authors use a two-stage procedure: in the first stage, firms choose their locations; in the second stage, they determine their production quantities. This feature was quite distinct from other contributions that use locations and prices as variables, whereas this work considers competition in locations and quantities (which is thus much closer to Cournot's original work, rather than Bertrand's unstable price competition). Employing the usual recursion, the authors prove that for any fixed pair of locations, the quantity game has an equilibrium. Under a condition that requires that it is always profitable to supply any market on the graph with a positive quantity of goods, a locational Nash equilibrium exists at the nodes of the graph. If this condition is not satisfied, the authors provided examples demonstrating that a locational Nash equilibrium either does not exist at all, or may exist on the edges of the graph.

The competitive location model investigated by Eiselt and Laporte (1993) included three firms, each attempting to maximize its own market share. The demand is located at the vertices of a tree. Contrary to the linear market, in which three market-share maximizing facilities end up without ever finding an equilibrium, the paper outlined under what conditions equilibria exist. In particular, there may be an agglomerated equilibrium with all facilities locating at the median of the tree, a semi-agglomerated equilibrium with two facilities locating at the median, while the third facility chooses an adjacent site, a dispersed equilibrium, in which the three facilities locate at three mutually adjacent vertices (one of which is the median), or no equilibrium. Loosely speaking, the more evenly the weights are distributed on the tree, the more likely it is that an equilibrium exists.

The focus of the contribution by Bhadury and Eiselt (1995) was the usual equilibrium—no equilibrium dichotomy. The paper proposed a measure that indicates not only whether or not an equilibrium exists, but how stable or unstable the solution is. While the paper demonstrated the computation of the measure in a tree network, it applies to all competitive location models. There are two cases to be

considered. In the first case, at least one Nash equilibrium exists. The measure then determines the effort that is required to convince at least one of the firms to move out of its equilibrium location. Clearly, if it takes a large subsidy to make a firm move out of its present location, the situation can be considered very stable. On the other hand, in case no equilibria exist, a tax for any moves (or, alternatively, moving costs) will indicate how much it takes to stop a firm from relocating. If this amount is substantial, it indicates that much effort is needed to stop the firms from relocating, so that the situation is far from an equilibrium and as such is highly unstable. A continuous measure of this nature contains much more information than the usual existence/non-existence analysis.

Eiselt and Bhadury (1998) considered the problem of reachability of equilibria, given that they actually exist. Their space is a tree network with demand occurring at the nodes. Two competing firms locate one branch each at the nodes of a tree. They charge fixed, but not necessarily equal, mill prices. The authors developed necessary and sufficient criteria for the existence of equilibrium locations on a tree. Given that equilibrium locations exist, the paper then examined whether or not a sequential and repeated relocation procedure that starts at an arbitrary location will eventually lead to the equilibrium. The authors first demonstrated that, in general bimatrix games with an arbitrary starting point, a Nash equilibrium, even if its exists, may not be reached. They then described a "reasonable" optimization procedure. In this process, one of the duopolists optimizes his own location, given his opponents present location. The assumption is that his opponent does not react, at least not for some time, so that the planner can reap the benefit of his own relocation. In the next step, the firm that relocated is now fixed at the site it chose and its opponent optimizes his location. This sequential process terminates when repeated reoptimization does not change the locations. The main result of the paper was that an equilibrium will be reached in this process, provided a proper tie-breaking rule is used. Table 7.1 summarizes some of the highlights in the analysis of Hotelling models.

Table 7.1 Some of the major contributions to Hotelling's model

Authors	Year	Major aspect of the model
Hotelling	1929	The basic model
Lerner and Singer	1937	Hotelling results for $n > 2$
Smithies	1941	Different behavioral assumptions
Eaton and Lipsey	1975	Equilibria with $n > 2$, 2-D results
d'Aspremont et al.	1979	Hotelling was wrong, quadratic cost function
Shaked	1982	Firms use mixed strategies
Kohlberg	1983	A model with waiting time
De Palma et al.	1985, 1987a	Customers use probabilistic choice rule
De Palma et al.	1987b	The model with uniform delivered pricing
Andersson	1988	Linear-quadratic transportation costs
Labbé and Hakimi	1991	Equilibria on networks
Eiselt and Laporte	1993	Three facilities on a tree
Bhadury and Eiselt	1995	Stability of equilibria
Eiselt and Bhadury	1998	Reachability of equilibria

In summary then, what has Hotelling's contribution done for location science? First and foremost, it has alerted the location science community (by which I include all interested parties from regional scientists to mathematicians, engineers, and computer scientists) to the interdependencies of different factors of location planning, and it has provided insight into location models. While, for instance, it will be virtually impossible to compute Nash equilibria for any real location scenarios, the decision makers now know which factors are required to stabilize a solution and which will lead to instability. Similarly, decision makers know that competition means having to look over their shoulders and anticipate a reaction, and the hundreds of contributions that have followed Hotelling's original analysis have enabled decision makers to know what to look for: adaptations of prices, quantities, attractiveness of their facilities, and many others. Another area in which Hotelling's work has impacted the field is in the—still somewhat underdeveloped—area of nonphysical location. Much more work is needed to develop brand positioning models, the assignment of tasks to employees in ability space, and the positioning of political candidates in issue space to a point where they become viable tools for practical location problems.

7.4 Future Work

As highlighted in the above sections, much work has been done in the field of competitive location models. Below, I will list a few of the areas that appear to offer promising research leads.

1. *Models with additional parameters*. While in the original contributions firms were competing in location and price, additional factors exist that may be taken into consideration. One such possibility is weights that symbolize the attractiveness of firms or brands. In the retail context, the attractiveness of a store may be expressed in terms of floor space, opening hours, (perceived) friendliness of staff, and similar factors. Attraction functions have been used for a long time, such as in the original work by Huff (1964). In the locational context, models with attraction functions are also not new, as witnessed by the contributions by Eiselt and Laporte (1988, 1989), Drezner (1994), and Eiselt and Marianov (2008b). Another recent contribution that uses repeated optimization with a Huff-style attraction function is put forward by Fernández et al. (2007). However, none of these models discusses equilibrium issues. Another feature that may be included is the choice of technology.
2. An interesting aspect is *asymmetric models*, i.e., models in which competing firms have either different objective functions, use different pricing policies, or have different perceptions of existing demand structures. The paper by Thisse and Wildasin (1995) is a step in this direction, as it includes not only competing duopolists, but also a public facility. A model with different pricing policies on a linear market has been put forward by Eiselt (1991).

3. An obvious extension concerns the discussion of competitive location in 2- or *higher-dimensional spaces*. It is questionable, though, if this is a promising route: experience with two-dimensional models, even if price competition is ignored altogether, has shown it to be very difficult. Some results with have been obtained by Irmen and Thisse (1998). More details concerning pure location competition can be found in Chap. 19 of this volume.
4. A different angle concerns the product with market segmentation. It refers to firms competing in different markets. Again, these markets could either be separated in physical space or in abstract feature or issue spaces in nonphysical applications. Especially in the context of product design, it would be very interesting to see whether or not there are instances in which a firm will decide not to compete in some of market.
5. The issue of data *aggregation* in the context of competitive location models has recently been put forward by Plastria and Vanhaverbeke (2007). The discussion is still in its infancy and it remains to be seen if conclusive results can be obtained.

Acknowledgments This work was in part supported by a grant from the Natural Sciences and Engineering Research Council of Canada. This support is gratefully acknowledged. Thanks are also due to an anonymous referee whose comments helped to streamline the paper.

References

Anderson SP (1988) Equilibrium existence in a linear model of spatial competition. Economica 55:479–491

Anderson SP, DePalma A, Thisse J-F (1992) Discrete choice theory of product differentiation. The MIT Press, Cambridge

Bertrand J (1883) Theorie mathematique de la richesse sociale. J savants 67:499–508

Bhadury J, Eiselt HA (1995) Stability of Nash equilibria in locational games. Oper Res 29:19–33 (Recherche opérationnelle)

Black D (1958) The theory of committees and elections. Cambridge University Press, Cambridge

Cournot AA (1838) Recherches sur les principes mathématiques de la théorie des richesses. Hachette, Paris (English translation by N.T. Bacon in 1897)

d'Aspremont C, Gabszewicz JJ, Thisse J-F (1979) On Hotelling's 'stability in competition.' Econometrica 47:1145–1150

Dasgupta P, Maskin E (1986) The existence of equilibrium in discontinuous economic games, I: theory. Rev of Econ Stud 53(1):1–26

De Palma A, Ginsburgh V, Papageorgiou YY, Thisse J-F (1985) The principle of minimum differentiation holds under sufficient heterogeneity. Econometrica 53:767–781

De Palma A, Ginsburgh V, Thisse J-F (1987a) On existence of location 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

Downs A (1957) An economic theory of democracy. Harper & Row, New York

Drezner T (1994) Locating a single new facility among existing, unequally attractive facilities. J Reg Sci 34:237–252

Eaton BC (1976) Free entry in one-dimensional models: pure profits and multiple equilibria. J Reg Sci 16:31–33

Eaton BC, Lipsey RG (1975) The principle of minimum differentiation reconsidered: some new developments in the theory of spatial competition. Rev Econ Stud 42:27–49
Eiselt HA (1991) Different pricing policies in Hotelling's duopoly model. Cahiers du C.E.R.O. 33:195–205
Eiselt HA, Bhadury J (1998) Reachability of locational Nash equilibria. Oper Res Spektrum 20:101–107
Eiselt HA, Laporte G (1988) Location of a new facility on a linear market in the presence of weights. Asia-Pac J Oper Res 5:160–165
Eiselt HA, Laporte G (1989) The maximum capture problem in a weighted network. J Reg Sci 29:433–439
Eiselt HA, Laporte G (1993) The existence of equilibria in the 3-facility Hotelling model in a tree. Transp Sci 27:39–43
Eiselt HA, Marianov V (2008a) Workload assignment with training, hiring, and firing. Eng Optim 40:1051–1066
Eiselt HA, Marianov V (2008b) A conditional *p*-hub location problem with attraction functions. Comp Oper Res 36:3128–3135
Eiselt HA, Laporte G, Thisse J-F (1993) Competitive location models: a framework and bibliography. Transp Sci 27:44–54
Fernández J, Pelegrín B, Plastria F, Tóth B (2007) Solving a Huff-like competitive location and design model for profit maximization in the plane. Eur J Oper Res 179:1274–1287
Gabszewicz JJ, Thisse J-F (1986) Spatial competition and the location of firms. Fundam Pure Appl Econ 5:1–71
Hamoudi H, Moral MJ (2005) Equilibrium existence in the linear model: concave versus convex transportation costs. Pap Reg Sci 84:201–219
Hotelling H (1929) Stability in competition. Econ J 39:41–57
Huff DL (1964) Defining and estimating a trade area. J Mark 28:34–38
Irmen A, Thisse J-F (1998) Competition in multi-characteristic spaces: Hotelling was almost right. J Econ Theory 78:76–102
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
Kruskal JB (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrica 29:1–27
Labbé M, Hakimi SL (1991) Market and locational equilibrium for two competitors. Oper Res 39:749–756
Lerner AP, Singer HW (1937) Some notes on duopoly and spatial competition. J Polit Econ 45:145–186
Niemi RG, Weisberg HF (1976) Controversies in American voting behavior. WH Freeman, San Francisco
Osborne MJ, Pitchik C (1986) The nature of equilibrium in a location model. Int Econ Rev 27:223–237
Plastria F, Vanhaverbeke L (2007) Aggregation without loss of optimality in competitive location models. Netw Spat Econ 7:3–18
Prescott E, Visscher M (1977) Sequential location among firms with foresight. Bell J Econ 8:378–393
Rusk JG, Weisberg HF (1976) Perceptions of presidential candidates: implications for electoral change. In: Niemi RG, Weisberg HF (eds) Controversies in American voting behavior. WH Freeman, San Francisco
Salop SC (1979) Monopolistic competition with outside goods. Bell J Econ 10:141–156
Schmalensee R, Thisse J-F (1988) Perceptual maps and the optimal location of new products. Int J Res Mark 5:225–249
Selten R (1975) Re-examination of the perfectness concept for equilibrium points in extensive games. Int J Game Theory 4:25–55

Shaked A (1982) Existence and computation of mixed strategy Nash equilibrium for 3-firms location problems. J Ind Econ 31:93–96
Silva F, Serra D (2007) Incorporating waiting time in competitive location models. Netw Spat Econ 7:63–76
Smithies A (1941) Optimum location in spatial competition. J Polit Econ 49:423–439
Stevens BH (1961) An application of game theory to a problem in location strategy. Pap Proc Reg Sci Assoc 7:143–157
Teitz MB (1968) Locational strategies for competitive systems. J Reg Sci 8:135–148
Thisse J-F, Wildasin DE (1995) Optimal transportation policy with strategic locational choice. Reg Sci Urban Econ 25:395–410

Chapter 8
Sequential Location Models

Hassan Younies and H. A. Eiselt

8.1 Introduction

Competitive location models have been discussed in the location literature since Hotelling's (1929) seminal paper. As other location contributions, his model includes customers, who are located in some metric space and who have a demand for some good. This demand may be satisfied by firms that offer the product, given some pricing policy. The difference between standard location problems and competitive location models is that in the competitive case, there are at least two competing firms, who offer the same product. Depending on the complexity of the model under consideration, the differences between the firms may include their different locations, prices, pricing policies, or the attractiveness of their respective facilities.

In their simplest form, competitive location models are based on the assumption that customers will patronize the firm that offers them the best value, in terms of price, transportation costs, and general attractiveness. Given some objective function, each firm will then attempt to determine the optimal value of the variables that are under their respective jurisdictions, such as its location, price, and possibly other features. Models including some of these features can be found in literature, see, e.g., Eiselt and Laporte (1996) and Plastria (2001).

There are two main types of analyses that have been performed on competitive location models. The first asks whether or not there exists a stable situation for the model, i.e., an equilibrium. Depending on the tools available to the decision makers of the firms, we may have location equilibria, price equilibria, etc. In the context of location, the equilibrium question was first addressed by Hotelling (1929) and a summary of his contribution can be found in Chap. 7 in this volume. His analysis assumes that the competitors play a simultaneous game, in the sense that they

H. Younies (✉)
School of Management, New York Institute of Technology, Abu Dhabi, United Arab Emirates
e-mail: hassan.younies@gmail.com

H. A. Eiselt
Faculty of Business Administration, University of New Brunswick,
Fredericton, NB E3B 5A3, Canada
e-mail: haeiselt@unb.ca

H. A. Eiselt, V. Marianov (eds.), *Foundations of Location Analysis,*
International Series in Operations Research & Management Science 155,
DOI 10.1007/978-1-4419-7572-0_8,

choose their strategies at the same time. Another type of analysis involves sequential moves, i.e., the competitors make their choices one after the other in a prescribed sequence. For simplicity, this chapter will concentrate on location choices, assuming that the firms' prices are set at a fixed and equal level.

One type of analysis starts with an arbitrary locational arrangement of the firms on the market and applies short-term optimization by allowing them to relocate one by one so as to maximize their profit. The main question is what location pattern will result from such a process, and whether or not it is stable. The first to investigate such process appears to be Teitz (1968), which is the first paper reviewed in this chapter.

Another type of analysis that employs a sequential location process was first proposed by the economist Freiherr von Stackelberg (1943). His analysis assumed that in any industry, there are firms that lead (in innovation, product development, or in any other way), while there are others that follow. This concept was later extensively used in marketing, where leaders and followers were referred to as "first movers" and "second movers." In our analysis, we will consider a firm the leader, if it acts (most prominently: locates) first, while a follower is a firm that acts after the leader has chosen his strategy.

Note the asymmetry in the decision making processes of leaders and followers: The follower faces a situation in which the values of his opponent's decision variables are known to him, so that he faces possibly a number of restrictions, but deals with certainty, at least in regard to his competitor's key decisions. On the other hand, taking into account his opponent's decision, the leader faces uncertainty, as he usually does not know what his competitor's objectives are. Even if he did, he first has to establish what is known as a *reaction function*, i.e., for each of his own potential decisions, the leader must establish the reaction of his competitor and determine the outcome based on this pair of decisions. Given the complete reaction function, i.e., having established his competitor's reactions to each of his own possible actions, the leader can then choose the course of action that benefits him most. The setting here is a straightforward application of bilevel programming problems, (see, e.g., Dempe 2002), in which the follower's solution becomes the input in the leader's problem. If the model setting is simple, there may be a (closed-form) description of the reaction function. However, in most practical cases, the reaction function consists of solutions that are much more complex, e.g., solutions of integer programming problems, making the leader's problem very difficult, to say the last.

Another aspect of von Stackelberg solutions is that firms are not necessarily designed to be leaders or followers. As a matter of fact, this choice may be up to the firm as part of the decision-making process. In order to be a leader, there are essentially two requirements: First, a firm must have the *capability* to be a leader, and secondly, it must have an *incentive* to become a leader. For instance, the capability could require a firm to have a large research and development lab, to have a foothold in a country they want to compete in, or similar advantages in the industry. Typically, only firms that have significant capital can possibly be leaders. The second requirement has nothing to do with the firms themselves, but with the way the process is structured. For instance, if the system does not protect the leader, it

may not be beneficial to become a leader. As an example, take the pharmaceutical industry. A leader would be a firm that develops new drugs for certain illnesses. The requirement of capability is clear. Consider now the need for appropriate protection. In this example, protection is provided in the form of patenting laws. In case a patent lasts for a very long time, then there is a strong incentive for a capable firm to develop all sorts of new medicines, as they will be able to reap the benefits for a very long time. On the other hand, if the time of a patent elapses after only a few years, the firm will have little time to introduce the drug, publicize it, and recover some of its costs before the patent runs out, allowing other firms to produce generic versions of the drug, thus dramatically cutting down the leader's market share. Knowing this in advance, the leader may not consider the time sufficient to recover costs and make a profit, so that he may not conduct the research and consequently will not introduce the product. In other words, the leader-follower game will not be played. Another aspect concerns the existence of more than two firms in the market. It has been suggested that in such a case, there will be a waiting line of firms, the first being the leader, the second will follow thereafter, and so forth. However, the question is why any firm would accept to take an specific place in line rather than choose what is most beneficial for his firm (other than may be first in line, which requires special capabilities). It is much rather likely that the firm will group into leaders and followers, depending on their abilities and the system's incentives.

The major assumptions of the sequential location model are that

1. Location decisions are costly and are made once and for all. Relocating is considered prohibitively costly and is not permitted.
2. Firms enter in sequence, one after another.
3. The leader and the follower have full and complete knowledge about the system, and the follower will have complete knowledge about the leader's decisions, once they have been made.

Among the first to propose the sequential entry of firms to the market are Teitz (1968), Rothschild (1976) and Prescott and Visscher (1977). The paper by Prescott and Visscher (1977) introduced the sequential entry of firms in a competitive location model from the perspective of operations research. Their ideas are illustrated through a set of examples, which are covered in this chapter.

Prior to von Stackelberg's work, other theories regarding market competition were known, mainly the one by Cournot. In his analysis two firms A and B are competing to supply the market with a homogeneous product at the same price. The two firms compete in the amounts of the product that they will put on the market. Each firm's objective is to determine the amount of the product it will make and sell in order to maximize its profit. In order to do so, each firm will determine its own supply reaction to the other firm's supply. Cournot stability assumes that each firm will move along its reaction curve. Cournot asserts that if each supplier takes the amount offered by his rival as a parameter of action, then the two firms can reach a point of equilibrium as the point of intersection of the firm's reaction curve to its competitor's supply. While these contributions provide the basic ideas for sequential location problems, their main emphasis is in economics, which is why we have chosen not to review them in detail.

The remainder of this contribution is organized as follows. Section 8.2 will review two classic contributions that use sequential location processes: one by Teitz (1968), and the other by Prescott and Visscher (1977). Section 8.3 will then assess the major impact of these contribution and outlines some directions of future research.

8.2 Classic Contributions

In this paper, we have chosen to survey two of the major contributions to the field. The first is a paper by Teitz (1968), in which he discusses a sequential relocation problem for two firms, one of which locates a single facility, while the other locates multiple facilities. The stability of the locational arrangement is investigated. The second paper is by Prescott and Visscher (1977). Its main contribution is the description of the sequential location of three facilities with foresight. This paper was the first to use von Stackelberg solutions in the context of competitive location models. Many contributions have used the basic ideas put forward in this work and extended them.

8.2.1 Teitz (1968): Competition of Two Firms on a Linear Market

While Prescott and Visscher (1977) are usually credited as the pioneers of sequential location, the contribution by Teitz (1968) predated their work by more than a decade. Teitz's paper considers the usual competitive system on a linear market, but with fixed and equal prices. In contrast to other contributors, the author does not consider simple competition between firms that locate one branch or facility each, but competition, in which each firm locates a given number of branches. The main thrust of the paper deals with repeated short-term optimization of the facilities' locations. For simplicity, the space customers and firms are to be located in a "linear market" of length 1, i.e. a line segment, on which customers are uniformly distributed.

The paper starts with the simple case of each firm locating a single branch each and, starting with initial random locations, use "short-term optimization" to relocate. This is done in order to maximize the firm's maximal profit, which, given fixed prices and fixed demand, reduces to the maximization of market share. The author uses sequential and repeated optimization by the two firms. In each step, the relocating firm takes the location of its competitor as fixed and optimizes. Given short-term maximization of market share, the relocation rule is to locate next to the competitor on the "longer" side of the market. Once this is accomplished, the other firm relocates in the same fashion. In this way, the two firms will cluster in each step and slowly move towards the center of the market, where neither of them has

any more incentive to relocate further. This central agglomeration solution recreates Hotelling's "central agglomeration."

The paper then investigates the case of firm A locating two facilities A^1 and A^2 and firm B locating a single facility by the same name. Either firm has two choices: either locate directly next to one of the other two branches on the outside and thus capture the hinterland of that branch, or locate between the two branches and capture half of what is called the "competitive region." Clearly, if a branch were to relocate to the outside, it would chose the branch with the larger hinterland and locate right next to it.

In our example shown in Fig. 8.1, the branches relocate in the sequence B, A^1, A^2, B, and so forth. At first, the two branches of firm A are located arbitrarily on the market. Then firm B locates its firm. It does so directly to the left of A^2, as its hinterland is larger than that of A^1 or half the competitive region. In the next step,

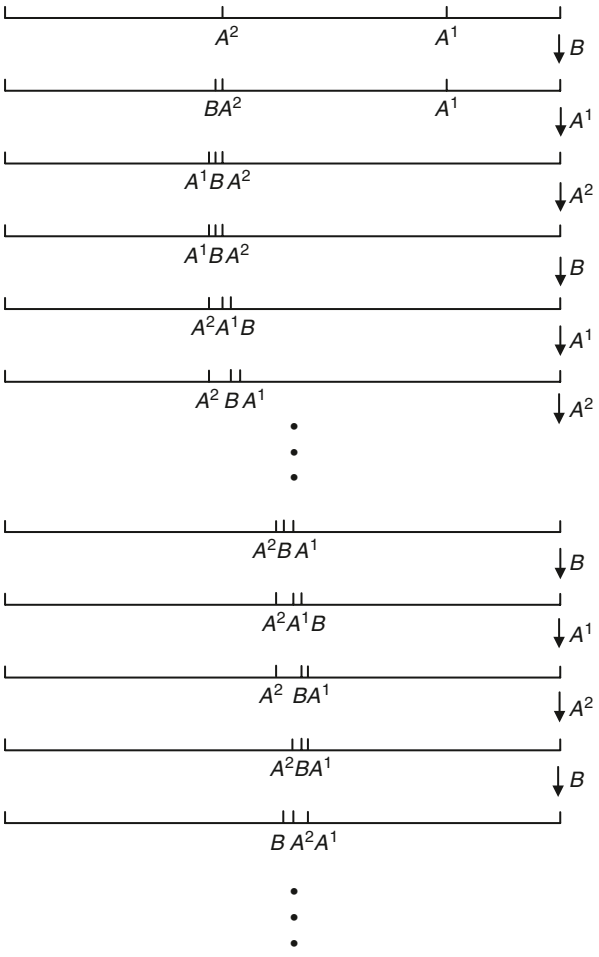


Fig. 8.1 Repeated relocation of two firms

branch A^1 relocates by moving directly to the left of B, even though A^1 has a larger hinterland. The reason is that both branches belong to the same firm, and by locating next to B on its outside, firm A will capture almost the entire market with both of its branches. Thus, when it is branch A^2's turn, it will not move as it is already located in its optimal place.

The next round of relocations starts again with B. It will move to the right of A^1, as this is the longer side of the two outside branches. Branch A^1 will then counter by moving just outside of B, reducing its market share again to a very small segment of the market. Branch A^2 will then move just a bit to the right towards B, so as to almost reduce its market share to zero again. The relocation process will continue in this fashion until all three branches are clustered at or near the center of the market.

At the center, the branch at the center will move to the outside; if it is branch B, it moves to the longer of the two hinterlands, if it is one of firm A's branches, it will move next to branch B and locate on its outside. Teitz referred to this relocation process as a "dancing equilibrium." The market shares of the two firms can be determined as follows. Firm A captures the entire market after one of its branches relocated, while it gets half the market after firm B relocates. Assuming equal relocation speed, firm A captures an average of ¾ of the market, while B obtains an average of ¼ of the market.

The instability of the solution leads to the author's conclusion that short-term optimization may not be the best solution. Instead, he suggests "long-term or conservative maximization." This can be explained as follows. Suppose that firm A locates both of its facilities at the first and third quartiles. Firm B can then either locate adjacent to either of A's branches on the outside (thus capturing the entire hinterland), or anywhere in between A's branches and capture half of the central region. Either way, firm B will capture ¼ of the market. Once this has happened, the author suggests that A does not relocate his branches (although such a relocation would benefit firm A in the short run), but stay put, thus ending the location process. That way, firm A will capture ¾ of the market, while firm B obtains the remaining ¼. This, incidentally, is the same market share that the two firms had obtained if they engaged in short-term optimization, giving the two firms an incentive to behave in this manner (especially when relocation costs are introduced, which are ignored in this discussion). While the author mentions that firm A uses a minimax objective, there is no mention of von Stackelberg and his leader-follower model. There is also no mention of what would happen if firm B were to locate first (which will always result in firm A capturing the entire market, as the two branches of A would "sandwich" firm B regardless of its location).

The analysis is then extended to include one facility of firm B, but 3 branches of firm A. As long as firm A knows that its competitor will locate only a single branch, it is aware that the branches will either follow the pattern *BAAA* or *ABAA*, all other location patterns reduce to these two based on symmetry. If firm A locates its branches at the first, third, and fifth sextiles, firm B will again either locate adjacent to either of the two outside facilities and capture the hinterland of length ⅙, or anywhere between any of firm A's branches and capture also ⅙.

This result can be further generalized to the case in which firm B locates a single facility, while firm A locates n_A branches. Firm A will then subdivide the market, so that the two hinterlands are half as long each as the region between any two of its branches, so that it will locate at $1/2n_A$, $3/2n_A$, $5/2n_A$, …, $(2n_A-1)/2n_A$, while firm B will locate its single facility again either adjacent to A's outside facilities in one of the two hinterlands, or anywhere between A's facilities. With that locational arrangement, firm B will capture $1/2n_A$, while firm A will capture the remaining $1-1/2n_A$ of the market.

The next step in the analysis is to allow firm B to locate more than one facility. In general, we now allow firm B to locate n_B branches, so that $n_B<n_A$. Following a reasoning similar to that above, we find that firm A locates again at the odd $2n_A$-tile points, while firm B locates its branches in the same manner prescribed above. The results are market shares of $M(B)=1/2n_A$ for firm B and $M(A)=1-n_B/2n_A$, for firm A. Figure 8.2 plots Firm A's market share against the number of facilities it locates. Even though n_A must obviously be an integer, we plot for all values for improved visibility. The solid, broken, dashed-and-dotted, dashed-and-double-dotted, and dotted lines show firm A's capture for $n_B=1$, 2, 3, 4, and 10.

A bit of elementary algebra leads to another interesting result. We can rewrite firm A's capture function $M(A)=1-n_B/2n_A$ as $n_A = \frac{n_B}{2[1-M(A)]}$ and then determine the number of branches firm A must locate in order to obtain the desired market share. For instance, for $M(A)=0.5$, we obtain $n_A=n_B$ (an obvious result), for $M(A)=0.75$, we obtain $n_A=2n_B$, for $M(A)=0.99$, $n_A=50n_B$, and so forth. One of the author's conclusions of this process is that the results do not exhibit agglomeration, but are quite similar to the social optima that minimize overall transportation costs. As Teitz put it, "Even a small gadfly can keep the big operator 'honest'." The remainder of the contribution deals with an investigation into equilibria for models for firms with fixed locations and variables prices, which is not of interest in this context.

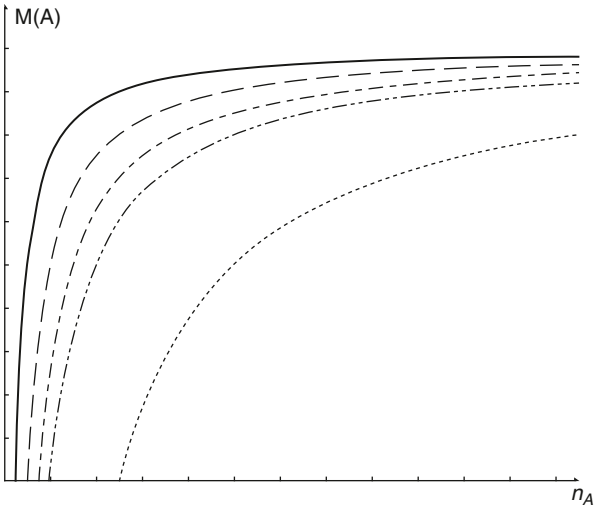


Fig. 8.2 Firm A's market share against the number of facilities

8.2.2 Prescott and Visscher (1977): Extensions of the Model on a Linear Market

The paper by Prescott and Visscher examines a number of different scenarios by way of five examples. Each such "example" relates to a competitive location model. Of specific interest in the context of sequential location models is Example 1, which demonstrates the complexities of the process for the case of three facilities that enter the market sequentially.

The novelty of Prescott and Visscher's approach is the use of foresight. In other words, the leader of the location game locates first, knowing that the follower will locate, so that its profit will be optimized. Such sequential location problems are typically solved in recursive fashion. If, for instance, n facilities are to be located by n independent decision makers, we first assume that $n-1$ facilities already are located and we are to locate the last facility. This will result in some general location rules. These rules, commonly called "reaction function," will then be used as input by the $(n-2)$nd facility. In particular, the decision maker at that facility will consider all possible location configurations of the first $(n-3)$ facilities and plan his location, taking into account the reaction function of the n-th facility. It is apparent that this process will get exceedingly tedious once the number of facilities increases.

For now, suppose there are two firms located somewhere on the market. Without loss of generality, assume that firm A is located to the left of firm B. As in our discussion of Teitz's paper, the area to the left of A is called A's hinterland, the area to the right of B is referred to B's hinterland, and the region between A and B is called the competitive region. Finally, that part of the market that is closer to a facility is said to be *captured* by that facility. (Note that this phrase was coined later by ReVelle (1986).

Consider first the simple case of two firms that locate a single branch each. Suppose that firm A is the leader who locates at some point x_A, while firm B is the follower who locates at x_B. The recursive argument assumes for the time being that x_A is fixed and that firm B's task is to optimally locate its facility. Then there are two cases: either firm B (the follower) will now locate to the left of A (i.e., $x_B<x_A$), or it will locate to the right of A (i.e., $x_B>x_A$). In the former case, firm B will capture the hinterland on its left in its entirety and half of the competitive region between itself and its competitor, i.e., $x_B + ½(x_A-x_B)=½(x_A + x_B)$. Since its capture depends positively on x_B, firm B will choose the largest possible value of x_B. Since its location is only limited by x_A, it will choose $x_B=x_A-\varepsilon$ for some arbitrarily small $\varepsilon>0$. In other words, firm B will locate directly to the left of the leader A. In doing so, firm B will capture x_A, while firm A will capture the remaining $1-x_A$ of the market. A similar argument applies to the case, in which firm B locates directly to the right of its competitor A. In this case, firm B locates at $x_B=x_A + \varepsilon$ and captures $1-x_A$, while firm A captures the remaining x_A.

In summary, firm B will now locate directly to the left of A, if $x_A>1-x_A$ or, equivalently, $x_A>½$, while B will locate directly to the right of A, if $x_A<1-x_A$ or

$x_A < ½$. Or, even shorter, if $x_A < ½$, then B locates at $x_A + \varepsilon$ and A will receive x_A, while if $x_A > ½$, then B will locate at $x_A - \varepsilon$ and A will receive $1 - x_A$. Knowing this to be firm B's reaction function, firm A will then decide as follows. In the former case, firm A's capture depends positively on its location, so that it will locate at $x_A = ½ - \varepsilon$, while its capture in the latter case depends negatively on its location, so that it will locate at $x_A = ½ + \varepsilon$. this means that the leader's location is best chosen at the center of the market, and the follower will then locate to either side, so that both capture about half of the market each.

This is the type of argument employed by Prescott and Visscher in their contribution. Below, we discuss two examples that constitute the major contribution of their paper to sequential location theory.

Example 1: Sequential location of three firms

This is example is a straightforward (albeit tedious) extension of the argument put forward above for two firms. Here, three facilities locate in sequential fashion. The facilities are A, B, and C, and their respective locations are x_A, x_B, and x_C. The facilities are going to locate in the order A, B, and C. Again, the length of the market has been generalized to 1. All facilities charge fixed and equal prices, so that we deal with pure location competition. Without loss of generality, we assume that $x_A < ½$. Starting with firm C, we note the C will either locate in one of the two hinterlands or in the competitive region, created by the already existing locations of firms A and B.

We now consider the following four cases.

Case 1: Facility B is located in the left half of the market, i.e., $x_B < ½$. This case allows two Subcases 1a and 1b: either B is located to the left of A, or B is located to the right of A.

Subcase 1a: If B is located at some point to the left of A (i.e., $x_B < x_A$), then C can either locate directly to the left of B (and capture x_B, which, by assumption, satisfies $x_B < x_A \leq ½$), or C can locate between A and B (thus capturing $½(x_B - x_A)$, which, by virtue of the assumptions concerning x_A and x_B, is less than ¼), or locate immediately to the right of A, which results in C capturing $1 - x_A > ½$. Clearly, this option dominates, so that C will locate immediately to the right of A at $x_A + \varepsilon$ for some arbitrarily small $\varepsilon > 0$.

Subcase 1b: Suppose now that B is located to the right of A at some point $x_B > x_A$, while still maintaining that $x_B \leq ½$. Again, facility C's best option is to locate directly to the right of firm B at $x_B + \varepsilon$, thus capturing $1 - x_B \geq ½$.

Summarizing Case 1, we find that Firm C will always locate at $x_C = \max\{x_A, x_B\} + \varepsilon$ and capture about $1 - \max\{x_A, x_B\}$ of the market. Incidentally, given firm C's behavior, firms A and B capture $½(x_A - x_B)$ and $½(x_A + x_B)$ in Subcase 1a, and $½(x_A + x_B)$ and $½(x_B - x_A)$ in Subcase 1b, respectively.

In the remaining three cases, we assume that firm B has located to the right of the mid-market point at $x_B > ½$. The cases differ in that in Case 2, firm C best locates in A's hinterland just to the left of firm A; in Case 3, Firm C best locates in B's hinterland just to the right of firm B; and in Case 4, Firm C's best option is to locate between the two firms A and B. The three cases establish the condi-

tions that the chosen solution provides a larger capture to Firm C than the other two options.

Case 2: Firm B locates at $x_B > ½$, and $x_A > \max\{1 - x_B, ½(x_B - x_A\}$. In this case, Firm C will locate at $x_A - \varepsilon$, capturing somewhat less than x_A. The three firms then capture:

Firm A captures $½(x_B - x_A)$,
firm B captures $1 - ½(x_A + x_B)$, and
firm C captures x_A.

Case 3: Firm B locates at $x_B > ½$, and $1 - x_B > \max\{x_A, ½(x_B - x_A)\}$. Here, Firm C will locate at $x_B + \varepsilon$ and capture somewhat less than $1 - x_B$. The captures of the firms in this case are:

Firm A captures $½(x_A + x_B)$,
firm B captures $½(x_B - x_A)$, and
firm C captures $1 - x_B$.

Case 4: Firm B locates at $x_B > ½$, and $½(x_B - x_A) > \max\{x_A, 1 - x_B\}$. In this case, Firm C can locate anywhere between its competitors A and B and capture about half of the competitive region. Prescott and Visscher assume that Firm C will locate in the middle of the competitive region at $x_C = x_A + ½(x_B - x_A) = ½(x_A + x_B)$ and capture $½(x_B - x_A)$. The captures of the three firms are then:

Firm A captures $¾\, x_A + ¼ x_B$,
firm B captures $1 - ¼ x_A - ¾ x_B$, and
firm C captures $½(x_B - x_A)$.

This completes the reaction of firm C. Consider now the reaction of firm B, which will depend on what firm A has done (something that firm B can observe) and the anticipated reaction of firm C derived above. Note for firm B in Case 1, Subcase 1a dominates Subcase 1b. Table 8.1 shows firm B's options, where LB and UB denote the bounds derived from the conditions imposed in the four cases.

Table 8.1 Summary of cases in example 1

Case #	Conditions (in addition to $x_A < ½$)	Firm B's capture	Strongest condition for x_B
1	$x_B < x_A < ½$	$½(x_A + x_B)$	$x_B < x_A$
2	$x_A > 1 - x_B$ or $x_B > 1 - x_A$ $x_A > ½(x_B - x_A)$ or $x_B < 3x_A$	$1 - ½(x_A + x_B)$	$x_B \in [1 - x_A, 3x_A]$ $(x_A \geq ¼$, as $3x_A \geq 1 - x_A)$
3	$1 - x_B > x_A$ or $x_B < 1 - x_A$ $1 - x_B > ½(x_B - x_A)$ or $x_B < ⅔ + ⅓\, x_A$	$½(x_B - x_A)$	$x_B < ⅔ + ⅓\, x_A$, if $x_A \leq ¼$ $x_B < 1 - x_A$, if $x_A \geq ¼$ $x_B > ½$ in both cases
4	$½(x_B - x_A) > x_A$ or $x_B > 3x_A$ $½(x_B - x_A) > 1 - x_B$ or $x_B > ⅔ + ⅓\, x_A$	$1 - ¼ x_A - ¾ x_B$	$x_B > ⅔ + ⅓\, x_A$, if $x_A \leq ¼$ $x_B > 3x_A$, if $x_A \geq ¼$

In Case 1, firm B's capture is positively correlated with its location, so that B will choose as large a value of x_B, which is achieved at $x_B = x_A - \varepsilon$. Firm B's gain is then about x_A, while firm A is wedged in between B and C and will get nothing.

Note that in Case 2, firm B's capture is negatively correlated with its location, so that B will attempt to decrease x_B as much as possible. The same argument applies in Case 4, while firm B will increase x_B as much as possible in Case 3.

Assume now that $x_A < ¼$. Then firm B's choice is to either locate at $x_B = x_A - \varepsilon$ and capture $½(x_A + x_B) \approx x_A$ (Case 1), at $x_B = ⅔ + ⅓ x_A$ and capture $⅓(1 - x_A)$ (Case 3), or at $x_B = x_B = ⅔ + ⅓x_A$ and capture $½(1 - x_A)$ (Case 4). Note that (Case 2) does not apply. Clearly, Case 4 dominates, so that

- if $x_A < ¼$, firm B will locate at $x_B = ⅓(2 + x_A)$ and capture $½(1 - x_A)$.

In case $x_A > ¼$, firm B can either locate at $x_B = x_A - \varepsilon$ and capture x_A (Case 1), locate at $x_B = 1 - x_A$ and capture ½ (Case 2), locate at $1 - x_A$ and capture $½ - x_A$ (Case 3), or locate at $3x_A$ and capture $1 - 2½x_A$ (Case 4). As $x_A \geq ¼$ in case 4, firm B's capture in that case cannot exceed ⅜, so that Case 2 dominates. This results in the location rule for firm B:

- If $x_A > ¼$, firm B will locate at $x_B = 1 - x_A$ and capture ½.

This now completely describes the reaction function of firm B.

On the last, and highest, level, consider now firm A's planning. Note that firm A knows exactly how its two competitors will react to any of its own actions. In particular, our above discussion reveals that if firm A locates somewhere at $x_A < ¼$, then firm B will locate at $x_B = ⅓(2 + x_A)$ and firm C will locate at $x_B + \varepsilon$ or at $⅓(1 + 2x_A)$. As for firm C, Cases 3 and 4 dominate. In both cases, firm A will maximize its own capture by choosing as large a value of x_A as possible, so that $x_A = ¼$ (resulting in $x_B = ¾$ and $x_C = ¾ + \varepsilon$ or $x_C = ½$).

As the former case requires some distance between firms B and C, firm C's capture will be somewhat less than in the latter case, so that we assume that firm C locates at the center of the market. Given locations at ¼, ¾, and ½ for the firms A, B, and C, their captures are ⅜, ⅜, and ¼, respectively.

Suppose now that firm A locates at some point $x_A > ¼$. As derived above, firm B will then locate at $x_B = 1 - x_A$, while firm C will either apply Case 2 and locate at $x_A - \varepsilon$, or apply Case 3 and locate at $1 - x_B + \varepsilon$. Each of these two cases result in firm C capturing x_A. Given that, firm A will capture $½(x_B - x_A) = ½(1 - 2x_A)$ or $½(x_A + x_B)$. In the former case, firm A's best option is to choose x_A as small as possible, so that $x_A = ¼$ (resulting in $x_B = ¾$ and $x_C = ½$), given the same argument used above), while in the latter case, firms A and B will cluster about the center with firm C locating next to either one of them, cutting out one of the firms. This outcome is unlikely, leaving again the symmetric locations of the firms at the first and third quartile and the center respectively. As demonstrated above, the captures of the three firms are then ⅜, ⅜, and ¼ for the firms A, B, and C. One question not addressed by Prescott and Visscher is why firm C would agree to be the third firm to locate, given that its capture is one third less than that of the second firm.

Example 2: Sequential location of infinite number of firms

Prescott and Visscher's second example considers the case in which an infinite number of firms, with a fixed cost of locating, can be potential entrants. Given that α is the market share needed to cover the fixed costs, then the largest number of firms that can enter the market with positive profit is $1/\alpha$, assuming again a market of length one. The authors describe three basic rules for the location. The first rule considers the case, in which two facilities are located at x_A and x_B, respectively, where, without loss of generality, x_A is located to the left of x_B. It is assumed that the two facilities are direct neighbors, i.e., there exists no facility between them.

Then, if the two facilities are no more than 2α apart, no new facility will ever locate between them, as the space is not sufficient to make a positive profit. If the space is more than 2α but no more than 4α, then a facility may profitably locate between x_A and x_B and, as in their previous example, the authors claim that the new facility would locate halfway between the two existing facilities. Finally, if there is more than 4α between the two existing facilities, the authors assert that a new facility would locate at a distance of 2α to the right of A or to the left of B with equal probability.

The second rule (and, by virtue of symmetry, the third rule) considers the situation that a facility exists at some point x and no other facility is located to its left. Clearly, if the space to the left of x is less than α, no facility will be able to profitably locate to the left of x. On the other hand, if the space left of x is larger than α, a new facility can locate there, which, as the authors assert, will happen at point α. This point, of course, guarantees that there will be no additional facility locating at any point in time to the left of the newly entering facility.

After some algebra, the authors determine that the model will locate facilities, so that the outside firms are α distance units from the two ends of the market, and each subsequent firm is located at a distance of 2α from its neighbor. The only disruption of the uniformity is the last interval that is either too short for an additional facility to locate in, or in which a facility will locate in the center.

Example 3: Competitive location model with product quality as "location"

This example introduces a revised Hotelling problem, in which firms choose a level of product quality in addition to the price. The product quality characteristic in this example is waiting time. The introduction of such a product characteristic enables the authors to formulate the classical Hotelling problem without discontinuities in the reaction function, thus avoiding the disequilibrium problem that is inherent to Hotelling's classical model. The solution of the duopoly model is made by numerical models, in which the equilibrium is unique and, as opposed to Hotelling's assertion of "minimal differentiation," the locations are widely dispersed. Equilibria with more than two competitors cannot be guaranteed.

Example 4: An extension of a competitive location model with product quality as "location"

This example expands the model introduced in Example 3. In particular, it is assumed that once the facilities have chosen their location, they can no longer be moved. First, the authors observe that if the duopolists were to choose location and

price simultaneously and irreversibly, then the follower firm always has an advantage, as it can locate at the same site as the leader, but charge a slightly lower price, thus being guaranteed higher profits than the leader. This raises the questions if any location will actually take place at all, as both firms may wait for the other to lead, so they can have the advantage to follow.

However, prices are not very likely to be as inflexible as locations, so that while locations (waiting times) are chosen once and for all, prices can be determined subsequently, so that they constitute a Nash equilibrium. The authors describe a recursive procedure that includes the possibility that the leader firm decides not to enter the market. Given some specific parameters, the authors then compute equilibrium solutions. The authors obtain some interesting results.

- Fixed costs are a barrier to the entry of additional facilities.
- Increasing fixed costs allows duopoly firms to locate farther apart, thus realizing local monopolies, so that the firms' profits actually increase.
- Earlier entrants hive higher profits.
- If the first firm to enter is allowed and has the resources to locate multiple branches, it will locate branches at all profitable locations, resulting in a monopoly.

Example 5: A competitive location model with plant capacity as "location"

This final example assumes that the price of a good in an industry is determined by the total capacity of all firms in the industry. If the number of firms that enter the market is not set in advance, the first entrant will build just as much capacity so as to ensure that no subsequent firms enter the market, thus resulting in a monopoly. This case is reminiscent of the last observation in the previous example. The results are very different, if a predetermined number of firms will enter the market. In such a case, early entrants will chose smaller capacities, so that subsequent entrants increase the total capacity of the industry to a level that is beneficial to the early entrants.

8.3 Impact of the Classic Contributions and Future Research

Competitive location models can be and have been applied to a variety of problems in marketing, political science, product positioning, and others. Sequential location procedures are appealing for many of these applications, so that it is not surprising that many researchers have discussed different aspects of sequential location models. In what he called the von Stackelberg equilibrium problem, Drezner (1982) introduced the planar sequential location problem and offered a polynomial time algorithm for this problem. Macias and Perez (1995) used rectilinear distance for planar competitive problem with an $O(n^5)$ algorithm. The case of asymmetric distance was first studied by Nilssen (1997) and more recently by Lai (2001). Lai's results show that equilibrium results cannot be attained in continuous location. Teraoka et al. (2003) considered the case of the two-firm planar von Stackelberg problem

with customers distributed continuously according to a random distribution with a probability that a customer would patronize a certain facility. Bhadury et al. (2003) describe heuristic solution methods for centroid and medianoid problems in the plane. Eiselt and Laporte (1996) studied the case where the facilities have different levels of attractiveness based on certain characteristics. Plastria (1997) introduced a competitive model based on location and attractiveness level.

Neven (1987) and Anderson (1987) investigate sequential location models from an economist's perspective. Both authors determine—as Prescott and Visscher did before them—that locations are much more difficult to change than prices and are therefore much more likely to be permanent. Prices, on the other hand, can easily be changed without cost. This lead them to a "first location, then price" game. The contribution by Ghosh and Buchanan (1988) allows duopoly firms to locate multiple facilities each. The authors also introduce the marketing concept of "first mover advantage" into the discussion.

Eiselt and Laporte's review (1996) of the sequential location problem listed the major contributions that employ a linear market or two-dimensional real space. The authors identify three main research issues: different objectives for firms; endogenizing the leader/follower choice; and the position of firms in a queue for entrance to the market.

One of the major contributions that uses the concept of sequential location choice is by Hakimi (1983). In his paper, the author defines centroid problems and medianoid problems, the former pertaining to the leader in the location game, while the latter is the decision problem by the follower. While his paper deals with the location of facilities on a network, the concepts easily translate to other spaces. In Hakimi (1990), the author further develops specific results given different customer choice rules. For further details on Hakimi's results and an in-depth discussion thereof, readers are referred to Chap. 9 of this volume.

One major assumption of the sequential location model is that the firms in the model enter the market at different points of time, an issue closely related to that of a firm's position in the entry queue. The time between entries enables leader firms to gain more profit and market share for a certain period, while the followers decide on the timing of their entry. Important issues for future studies include the effects of changes of the cost of entry over time due to different factors such as fixed cost change, inflation or a reduction in the cost of technology.

Other factors to be included could be market penetration costs for followers and customer retention costs for leader firms. Models that take such factors into account should allow for uncertainty of these factors. A stochastic approach to the competitive environment was introduced by Choi et al. (1990), who produced a model with one leader and multiple followers and customers with a stochastic utility function.

Another open area of research is the incorporation of the concepts of competitive location in the context of supply chain models. A model based on competition for customers can be considered as a model which is looking forward in the supply chain, while a model looking backward would consider competing for suppliers, e.g., manufacturers for retailers or natural resources such as oil. Sequential loca-

tion problems will include location decisions with respect to both suppliers and customers.

An interesting aspect of competitive location models concerns cases, in which customers cannot arbitrarily switch between competitors without incurring an early termination fee. As an example, this situation applies to the cell phone industry. Other examples involve suppliers of mineral water or heating gas, where customers are bound by annual contracts with a supplier. Adding switching costs as well as binding contracts time to the models would create more realistic models for certain industries.

A sign of globalization is the tendency of competing firms to form bigger companies through consolidations, acquisitions and mergers. Questions in this context include: what location factors would lead a firm to decide to consolidate with firm *A* and not firm *B*? What are the impacts of such mergers and acquisitions on the market and on present and future competitors?

Finally, an issue that could be included in competitive location models includes the privatization of services such as of waste management and disposal. While the location of undesirable facilities (for details, see Chap. 10 of this volume) is a well-studied area of location theory, it has not been investigated in the context of competitive location.

Acknowledgments The work of the second author was in part supported by a grant from the Natural Sciences and Engineering Council of Canada. This support is much appreciated. Thanks are also due to Professor Vladimir Marianov for his careful reading of the manuscript and for many helpful suggestions.

References

Anderson S (1987) Spatial competition and price leadership. Int J Ind Organ 5:369–398

Bhadury J, Eiselt HA, Jaramillo JH (2003) An alternating heuristic for medianod and centroid problems in the plane. Comput Oper Res 30:553–565

Choi SC, DeSarbo WS, Harker PT (1990) Product positioning under price competition. Manag Sci 36:175–199

Dempe S (2002) Foundations of bilevel programming. Springer, Berlin

Drezner Z (1982) Competitive location strategies for two facilities. Reg Sci Urban Econ 12:485–493

Eiselt HA, Laporte G (1996) Sequential location problems. Eur J Oper Res 96:217–231

Ghosh A, Buchanan B (1988) Multiple outlets in a duopoly: a first entry paradox. Geogr Anal 20:111–121

Hakimi SL (1983) On locating new facilities in a competitive environment. Eur J Oper Res 12:29–35

Hakimi SL (1990) Locations with spatial interactions: competitive locations and games. In: Mirchandani PB, Francis RL (eds) Discrete location theory. Wiley, New York, pp 439–478

Hotelling H (1929) Stability in competition. Econ J 39:41–57

Macias MR, Perez MJ (1995) Competitive location with rectilinear distances. Eur J Oper Res 80:77–85

Lai FC (2001) Sequential locations in directional markets. Reg Sci Urban Econ 31:535–546

Neven DJ (1987) Endogenous sequential entry in a spatial model. Int J Ind Organ 5:419–434

Nilssen T (1997) Sequential location when transportation costs are asymmetric. Econ Lett 54:191–201

Plastria F (1997) Profit maximising single competitive facility location in the plane. Stud Locat Anal 11:115–126

Plastria F (2001) Static competitive facility location: an overview of optimization approaches. Eur J Oper Res 129:461–470

Prescott E, Visscher M (1977) Sequential location among firms with foresight. Bell J Econ 8:378–393

ReVelle C (1986) The maximum capture or sphere of influence problem: Hotelling revisited on a network. J Reg Sci 26:343–357

Rothschild R (1976) A note on the effect of sequential entry on choice of location. J Ind Econ 24:313–320

von Stackelberg H (1943) Grundlagen der theoretischen Volkswirtschaftslehre. Kohlhammer, Berlin. English edition: von Stackelberg (1952) The Theory of the Market Economy (trans. Peacock AT). W. Hodge, London

Teitz MB (1968) Locational strategies for competitive systems. J Reg Sci 8:135–148

Teraoka Y, Osumi S, Hohjo H (2003) Some Stackelberg type location game. Comput Math Applications 46:1147–1154

Chapter 9
Conditional Location Problems on Networks and in the Plane

Abdullah Dasci

9.1 Introduction

Location decisions are of critical importance to all firms. Opening, closing, and relocating facilities require careful planning due to the strategic nature of these decisions. When customers do not have physical contact with the facilities (such as plants, distribution centers, or call centers), demand for the products or services can be assumed to be relatively independent of location. However, location choices of some stores (such as coffee shops, supermarkets, bank branches, and restaurants) do have a direct impact on demand. Therefore, such decisions should not be made without consideration of consumer behavior and market conditions.

In this chapter I present a class of competitive location models that are inspired by three streams of historical developments in location analysis: retail location from marketing, spatial competition from economics, and location theory from operations research. Estimating trade areas and market shares of retail facilities are central to strategic management of retail networks that involves decisions such as store location, relocation, or dismantling. Simple models such as Voronoi tessellations and Reilly's law of retail gravitation have existed for much of the twentieth century to assist marketing managers (see, e.g., Ghosh and McLafferty 1987). These methods and their various extensions have been used to delineate trade areas and locations for retail outlets but they usually lack formal optimization or equilibrium approaches. The paper by Hotelling (1929), which pioneered the field of spatial competition, considers the location and pricing decisions of two firms competing on a linear market. In his model, each firm simultaneously locates one facility and, after observing each other's choices, they simultaneously decide on their prices. Hotelling's model and many of its extensions are, however, usually stylized models whose primary purpose is to provide insights rather than to prescribe solutions to more realistic problems. The 1960s witnessed important strides towards modeling

A. Dasci (✉)
School of Administrative Studies, York University, 4700 Keele Street,
Toronto, ON M3J 1P3, Canada
e-mail: dasci@yorku.ca

H. A. Eiselt, V. Marianov (eds.), *Foundations of Location Analysis,*
International Series in Operations Research & Management Science 155,
DOI 10.1007/978-1-4419-7572-0_9,

and solving progressively more realistic location problems as a result of advances in operations research techniques. Much of the early works focus on plant or warehouse location issues where competitive interactions are naturally neglected (see, e.g., Daskin 1995). The late 1970s and early 1980s witnessed a convergence of these historical developments towards competitive location models that consider more realistic market spaces and customer characteristics.

One might expect that the competitive framework proposed by Hotelling could be generalized and equilibria could be found, at least numerically. However, despite its simple formulation, the analysis of even Hotelling's original simplistic problem is rather involved. Therefore, early works needed to make two key assumptions to obtain manageable models. First, they neglected the pricing decisions altogether, since the second stage competition in prices immensely complicated the analysis of the Hotelling's model. Unfortunately, the problem could still be difficult or poorly defined because the equilibrium might be quite difficult to identify or it might not exist in general. Therefore, as a second simplification, the games are defined in a leader-follower framework, in which von Stackelberg solutions are sought, rather than scenarios with simultaneous moves and Nash equilibria.

Individually and separately, Zvi Drezner and Louis Hakimi presented a series of competitive location models along these lines at the International Symposium of Location Decisions (*ISOLDE*) in Denmark in 1981. Subsequently, their works have appeared as Drezner (1982) and Hakimi (1983), which have become among the most influential works in the location literature. Although there are some contributions that predate these papers (such as those by Wendell and Thorson 1974; Slater 1975; Wendell and McKelvey 1981; and Hansen and Thisse 1981), Drezner and Hakimi study these problems under fairly general conditions for the time and present a number of important constructs that have become a framework for substantial further study.

Their problems can briefly be described as follows: a number of customers, each endowed with a certain buying power, will purchase a homogeneous good at the closest facility. First, the leader establishes all of her facilities in the market and then, after observing her choices, the follower sets up his facilities. Under these assumptions, both Hakimi and Drezner define two "conditional" problems.

The *Follower's problem* [$(r|X_p)$-*medianoid*]: Given the locations $X_p = \{x_1, x_2, \ldots, x_p\}$ of p existing (i.e., the leader's) facilities serving the customers, find the locations of r new facilities that will capture the most buying power from the customers.

The *Leader's problem* [$(r|p)$ *centroid*]: Find the locations for p facilities such that they will retain the most buying power against the best possible locations of r competing facilities.

Their definitions are virtually identical, except for one fundamental difference: Drezner assumes that the location space is a plane while Hakimi assumes that it is a network. This difference leads to somewhat different models and solution approaches. The purpose here is to re-introduce the conditional locational models of Drezner and Hakimi and review subsequent developments. It is, however, impossible to review all of the contributions in this growing literature, and therefore I refer the interested reader to a series of survey papers published over the years such as

Hakimi (1990), Hansen et al. (1990), Eiselt et al. (1993), Eiselt and Laporte (1996), Plastria (2001), and Santos-Penate et al. (2007).

The remainder of this chapter is organized as follows. In the next section, I formally define the medianoid and centroid problems, and then present Drezner's (1982) and Hakimi's (1983) results in the following two sections. The three subsequent sections are devoted to a review of extensions of these works. First, I look at the extensions of medianoid problems along some important dimensions such as variable expenditure and unequal store attractiveness levels. I then review the centroid and some related problems such as those from voting theory and Voronoi games. Finally, the works that consider pricing will be reviewed. The chapter concludes with possible future research directions.

9.2 The Classical Contributions

In order to describe the main contributions, we will use the following conventions. Let there be n customers at locations $V=\{v_1, v_2, \ldots, v_n\}$, each endowed with a buying power or demand, $w(v_i)$. For any set Z of points on the space (either the plane or a network $G(V, E)$ with the node set V and edge set E) let $D(v, Z)=\min\{d(v, z)|z \in Z\}$, where $d(v, z)$ is the distance between points v and z. If the problem is defined in the plane the distance will be Euclidean; if it is defined on a network, the distance will be the length of the shortest path from one point to another.

Let $X_p=\{x_1, x_2, \ldots, x_p\}$ and $Y_r=\{y_1, y_2, \ldots, y_r\}$ be the locations of the leader's and the follower's facilities, respectively. Customers will buy from the closest follower facility if the Euclidean distance between this follower's facility and the customer is less than the closest leader's facility. Hence, the ties are assumed to be broken in favor of the leader. The customers that are captured by the new facilities are defined as $V(Y_r|X_p)=\{v \in V|D(v, Y_r)<D(v, X_p)\}$ and the total buying power is $W(Y_r|X_p)=\sum w(v)|v \in V(Y_r|X_p)$.

Therefore, given V and X_p, Y_r^* is called an $(r|X_p)$ medianoid if $W(Y_r^*|X_p) \geq W(Y_r|X_p)$ for all feasible Y_r in the space. Also, given V, X_p^* is called an $(r|p)$ centroid if $W(Y_r^*(X_p^*)|X_p^*) \leq W(Y_r^*(X_p)|X_p)$ for all sets of points X_p in the space. The implicit assumption in these definitions is that the demand is essential or inelastic, and hence minimizing the follower's payoff is equivalent to maximizing the leader's payoff. It is rather straightforward to extend these definitions to profit maximization under more general conditions.

9.2.1 Drezner (1982)

Drezner begins his paper with an illustrative example in which both firms open a single facility. He then proposes solution algorithms for both problems and analyzes

several generalizations. Here, I follow his steps but also add a section for immediate generalizations that follow his work.

9.2.1.1 An Illustrative Example

Consider an instance of six demand points with equal buying power located at the vertices of a hexagon, as depicted in Fig. 9.1. If the leader opens her facility at *O*, the center of the hexagon, she can retain three customers. To see this, consider any pairs of points that are diametrically opposite to each other. For example, if the follower opens his facility at *G*, he can attract *F* but the leader retains *C*. Similarly *A* and *B* are captured by the follower, but *D* and *E* are retained by the leader. Hence, the follower captures a buying power of three when the leader locates at the center. If the leader locates anywhere else in the hexagon, the follower will capture at least four demand points. For example, if the leader locates her facility at *G*, the best location for the follower can be found by drawing a perpendicular line from *G* to one of the diagonals connecting opposite demand points. Location *H* captures demand points *C*, *D*, *E*, and *F*. As a result, the unique solution to the leader's problem in this instance is to locate at the center of the hexagon.

Note also that wherever the leader locates her facility in the polygon, she retains at least one customer. Put differently, the follower can never capture all the buying power as long as the leader does not locate outside the polygon. This observation will later be utilized in the analysis of the problem.

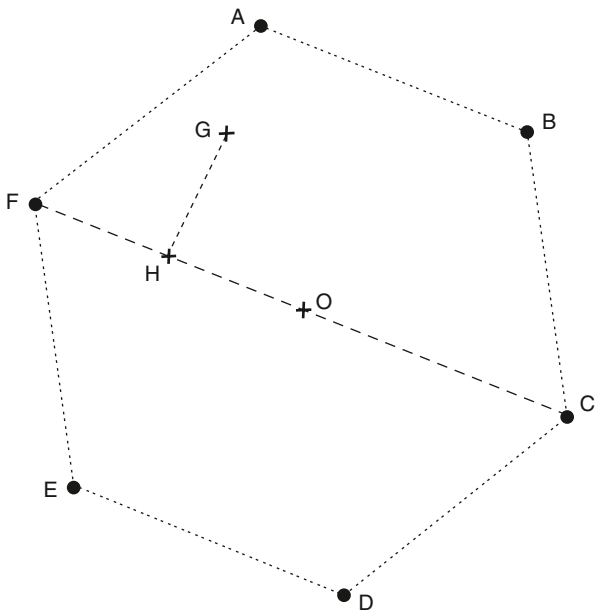


Fig. 9.1 An example in the plane

9.2.1.2 The $(1|X_1)$ Medianoid

I now proceed with Drezner's treatment of the follower's problem when both firms locate a single facility. For the sake of notational simplicity, here I denote the facility locations of the leader and the follower as X and Y, respectively. Suppose Y is a candidate location in the plane to compete against the facility located at X as shown in Fig. 9.2a. The market can easily be divided between the firms by the perpendicular bisector of the line connecting X and Y. The open-half space containing Y is the follower's market area and the closed half-space is the leader's market area. Note that the follower can do no worse by moving his facility arbitrarily close to the leader's, see Fig. 9.2b. However, he will avoid co-locating as the leader is assumed to have the advantage in case of a tie.

The above discussion shows that the optimal solution for the follower is to locate infinitesimally close to the leader. The only problem for the follower is to find the angle at which his location divides the market. This angle determines the set of customers and hence the buying power captured by each firm. The follower's objective function will be a piecewise step function of the angle, at which discontinuities happen at angles where a customer is captured or lost. Therefore, all one needs to do is to search over a finite number of angles. The calculation of captured buying power can be performed in linear time, but the sorting part takes $O(n \log n)$, which determines the complexity of the overall procedure.

9.2.1.3 The (1 | 1) Centroid

Having solved the follower's problem, Drezner turns to the leader's problem: Consider a subset V_0 of demand points, whose total buying power is at least w_0. Unless the leader locates her facility outside the convex hull defined by these demand points, the follower cannot attract *all* the buying power (recall the earlier discussion). As a result, if $X=X_1(w_0)$ is inside the intersection of all convex hulls of the sets with the buying power of at least w_0, then $W(Y_1|X=X_1(w_0))<w_0$. This is because no new facility can attract all the buying powers from those sets. While providing a starting point, this observation is unlikely to lead to an efficient algorithm, as the

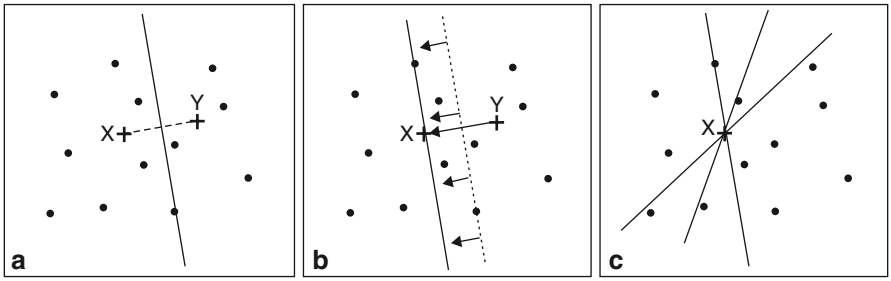


Fig. 9.2 An illustration of the treatment of $(1|X_1)$ medianoid problem

number of convex hulls to be considered grows exponentially with the number of customers. The following theorem presents a powerful result that greatly simplifies the analysis.

Theorem 1: *The intersection of all convex hulls for the sets with buying power of at least w_0 is identical to the intersection of all halfplanes whose buying power is at least w_0.*

Proof Suppose that a point z is not in the intersection of such convex hulls. Then there must be at least one convex hull to which this point does not belong. One can easily divide the plane into two parts: one containing the convex hull, the other containing the point z. The halfplane containing the convex hull will have a buying power of at least w_0. Therefore, this point cannot be a member of the intersection of such half planes. Proving the converse is similar; if the point is not in the intersection of half-planes, then there is a convex hull whose buying power is at least w_0 to which this point does not belong. □

The following result then follows immediately.

Corollary 2: *The condition $w(Y_1|X_1(w_0))<w_0$ holds for all $Y_1 \in \mathbb{R}^2$ if and only if $X_1(w_0)$ belongs to $\cap H$, where $\cap H$ is the intersection of all closed half-planes H for which $\sum\{w(v)|v \in H\} \geq w_0$.*

In order to find a point $X_1(w_0)$ such that $W(Y_1|X_1(w_0))<w_0$ for all $y_1 \in \mathbb{R}^2$, therefore, one needs to look at the intersection of all such closed half-planes. Fortunately, one can restrict the set of halfplanes to be considered to those that are defined by lines passing through at least two demand points. To see this, consider a closed half-plane generated by a line L where the total buying power is at least w_0. If there is no demand point on the line, it can be shifted back until it touches a demand point. Similarly, it later can be rotated once clockwise and once counter-clockwise until it touches two other points. All of the closed halfplanes defined by these lines is at least w_0 and the intersection of them is a subset of the original half-plane defined by L.

After these observations, Drezner presents an algorithm, which is essentially a bisection on the buying power w_0. The procedure is shown as *Algorithm* 1.

Algorithm 1: A Bisection Algorithm for the (1|1) Centroid Problem in the Plane

Step 1: Find the lines passing through all pairs of points and calculate all w_0 for all halfplanes.

Step 2: Sort w_0 in decreasing order. Find w_{min} and w_{max}.

Step 3: Set w_0 to the median value among all w_0 such that $w_{min}<w_0<w_{max}$. If there is no such w_0, go to *Step 7*.

Step 4: Find if there is feasible point at the intersection of all halfplanes with buying power w_0.

Step 5: If there is a feasible solution X_1 in *Step 4*, then $W(Y_1^*(X_1)|X_1) < w_0$. Set w_{max} to w_0 and go to *Step 3*.

Step 6: Otherwise, $W(Y_1^*(X_1^*)|X_1^*) \geq w_0$. Set w_{max} to w_0 and go to *Step 3*.
Step 7: The feasible point for the last w_{max} is an optimal solution and the value of the objective is $W(Y_1^*(X_1^*)|X_1^*) < w_{min}$.

Step 4 can be formulated as a linear program with two variables and $O(n^2)$ constraints. While numerous results on the complexity of linear programs are now known, Drezner provides a complexity result independent of them: he notes that the dual problem can solved in $O(n^4)$ and is repeated at each bisection step, which itself takes $O(\log n)$. Hence, the worst-case performance of this algorithm is bounded by $O(n^4 \log n)$.

Drezner also presents an alternative algorithm. Since the solution must be at the intersection of half-planes, there must be a feasible point that is a vertex of that intersection. Obviously, such a vertex is at the intersection of two lines. Since there are $O(n^2)$ lines, the number of intersection points is bounded by $O(n^4)$. One then needs to solve a $(1|X_1)$ medianoid for each point, which takes $O(n \log n)$. Hence the overall complexity of the second algorithm is $O(n^5 \log n)$.

9.2.1.4 Extensions

The centroid problem when $r>1$ and $p \geq 1$ is quite difficult to solve and no analysis is performed by Drezner, who concentrates on the other cases. When $p=1$ and $r>1$, the solution is rather simple. All the leader can retain is one customer since the follower can sandwich the leader's store between his facilities and capture the rest of the buying power. Therefore, the leader locates at the customer with the most buying power. The medianoid problem with $p>1$ and $r>1$ is also not available, but Drezner provides an algorithm of complexity $O(n^2 \log n)$ to solve the $(1|X_p)$ medianoid problem.

Let $B(v, X_p)=\{z \in \mathbb{R}^2 | d(v, z) < D(v, X_p)\}$ be the circle containing the set of locations for the follower's facility that captures the customer at v. For a given $V' \subset V$, let $B(V', X_p) = \cap\{(Bv, X_p) | v \in V'\}$. If $B(V', X_p) \neq \varnothing$, then any $Y_1 \in B(V', X_p)$ captures all the buying power of V'. Now consider all the circles $B(v, X_p)$ centered at $v=v_1, v_2, \ldots, v_n$. Since any two of these circles intersect in at most two points, a circle may have at most $2(n-1)$ intersection points. Over each circle then, there are at most $2(n-1)$ intervals with location on one of the intervals (actually infinitesimally inside the circle) that yields the maximum capture. One can easily sort the intersection points and compute the capture when the follower's facility is located on different segments. This requires $O(n \log n)$ operations. Since this step has to be performed for each circle, the overall complexity of the algorithm is $O(n^2 \log n)$.

Finally, Drezner analyzes the version $p=r=1$ with a minimal distance requirement, whereby the follower is not allowed locate within a distance R from the leader. One can replicate the analysis of the $(1|X_1)$ medianoid problem almost exactly

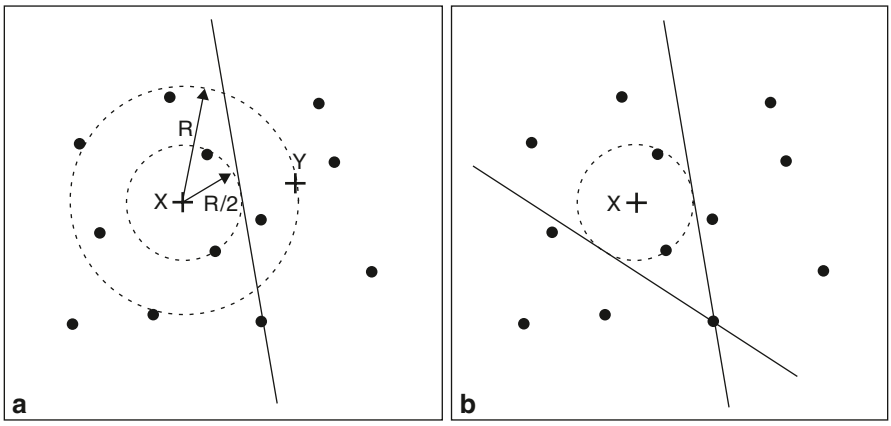


Fig. 9.3 An illustration of the $(1|X_1)$-Medianoid with minimum distance requirement

in this case. The follower's best response will be to locate on the circle of radius R centered at $X=X_1$. Let $Y=Y_1$ be such a point. Then, as usual, the perpendicular bisector of the line segment connecting X and Y separates the market between the facilities. This bisector is tangent to the circle centered at X with radius ½R as shown in Fig. 9.3a. Hence, all the follower needs to do is to find the best tangent line, which is the one that leads to the highest buying power for himself. Doing this is almost as easy as solving the original problem. If a demand point is inside the smaller circle, where it will never be captured by the follower. For all points outside a circle, two directions are calculated using the tangent lines passing through them as shown in Fig. 9.3b. Finally, one can go over all possible tangent lines and find the half-space with the highest buying power, which in turn gives the solution to the follower's problem. There will be at most $2n$ lines, but due to sorting of the directions as in the original algorithm, this procedure has a complexity of $O(n \log n)$.

One may try to extend the original algorithm for the centroid problem to the modified version. The earlier convex hull idea can also be used here. However, now the convex hulls are smooth at the corners. Drezner notes that it seems difficult to generalize the earlier algorithm due to this problem. He, however, provides an immediate generalization of his second approach. That is, if the intersection of all convex hulls with buying power of at least w_0 is a nonempty set, it must contain a vertex. Such a vertex is either an intersection of two lines, two circles, or a line and a circle. Since there are $O(n^2)$ such lines and n circles, we have at most $O(n^4)$ possible intersection points. Similarly, since a medianoid problem must be solved for each point, the overall complexity of this algorithm is also $O(n^5 \log n)$.

9.2.1.5 Immediate Generalizations and Algorithmic Developments

Two of the earlier works, Lee and Wu (1986) and Hakimi (1990), examine Drezner's results for the $(1|X_1)$ medianoid and $(1|1)$ centroid problems, respectively. Lee and

Wu show that $O(n \log n)$ is indeed the lowest complexity bound to solve the $(1|X_1)$ medianoid problem. They prove this by showing the equivalence of the medianoid problem to the ε-Closeness problem, which is to determine whether any two of n real numbers $\{x_1, x_2, \ldots, x_n\}$ are within a fixed ε of each other, i.e., if $|x_i - x_j| < \varepsilon$. The ε-Closeness problem is shown to have an established lower bound of $O(n \log n)$. Lee and Wu show that one can transform an instance of the $(1|X_1)$ medianoid problem to ε-Closeness in linear time, thereby establishing the lower-bound complexity.

Hakimi (1990), on the other hand, provides a more promising result on the $(1\,|\,1)$ centroid problem. Recall that Drezner's first algorithm for this problem requires solving a linear program with two variables and $O(n^2)$ constraints. Although Drezner points out that the dual of this problem can be solved in $O(n^4)$, subsequent results considerably improve this bound to $O(n^2)$. Since finding the set of constraints actually takes $O(n^2 \log n)$ time and there is a search over w_0, which takes $O(\log n)$, Hakimi shows that the worst-case complexity of the $(1\,|\,1)$ centroid problem in the plane is $O(n^2 \log^2 n)$.

9.2.2 *Hakimi (1983)*

Hakimi's paper consists of mainly two parts. First, he presents a series of stylized examples to investigate properties of medianoid and centroid problems, such as node-optimality, or relationships to classical concepts such as p-median or p-center of networks. He then presents the complexity results to these problems. Here, I also add some comments on some immediate works.

9.2.2.1 Illustrative Examples

One would expect that some of the earlier classical location results on medians or centers could be generalized or somehow related to the medianoid and the centroid concepts. Through a number of designed instances, Hakimi shows that no such result can be generalized in the conditional location problems. Figures 9.4, 9.5 and 9.6

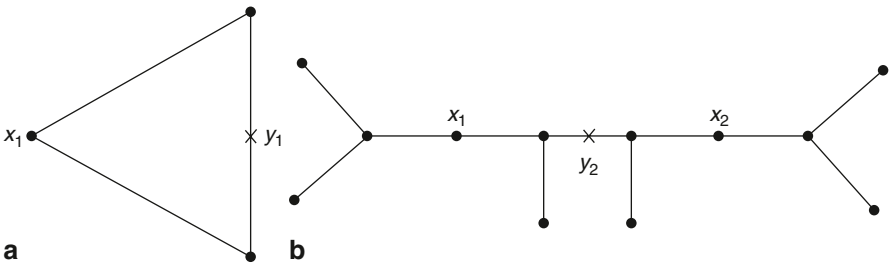


Fig. 9.4 **a** A $(1|X_1)$-Medianoid that is not a vertex, and **b** a $(1|X_2)$-medianoid that is not a vertex

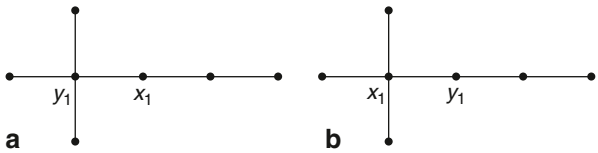


Fig. 9.5 **a** x_1 is the 1-Center but not a (1|1)-Centroid, while in **b** x_1 is a (1|1)-Centroid but not a 1-Center

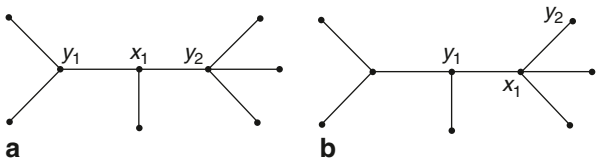


Fig. 9.6 **a** x_1 is the 1-Median but not a (1|2)-Centroid, while in **b** x_1 is a (2|1)-Centroid but not a 1-Median

depict some of the examples discussed in his paper. In all the networks in these figures, each edge has a unit length and each node has a unit weight or a buying power.

Figure 9.4a depicts an example where the solution to a $(1|X_1)$ medianoid problem might not be on a vertex. It is however easy to see that there exists an $(r|X_1)$ medianoid on vertices if the network is a tree and $p=1$. Unfortunately, this result does not generalize to $p>1$. The $(1|X_2)$ medianoid defined on the tree depicted in Fig. 9.4b, clearly does not possess a node solution that can capture any more than the four obtained by the current location.

Since the leader's problem can be cast as a minmax problem, a p-center of the network could be a reasonably good choice for the leader's facilities. Hakimi again provides a counterexample for the simplest case of (1|1) centroid, as depicted in Fig. 9.5. The solution x_1 in (a) is the 1-center, and the leader retains three buying power after the follower's best response, y_1. It is, however, clear that the leader can retain four if she locates as in (b), which is indeed the (1|1) centroid solution. If the network is a tree, though, it is shown earlier by Slater (1975) that the 1-median and the (1|1) centroid coincide. Unfortunately, this result does not generalize either. Hakimi constructs another example, as depicted in Fig. 9.6, to show that an $(r|1)$ centroid of a tree network with $r>1$ is not necessarily at a 1-median. It is easy to see that at 1-median the leader retains two customers, while she retains three customers at the (2|1) centroid. Hakimi provides a number of other interesting and important instances and counterexamples to convincingly demonstrate that not much can be said for the general cases with $p>1$ and $r>1$ even for tree networks. All of this groundwork suggests that these problems are indeed very difficult, which is proved in the remainder of Hakimi's paper.

9.2.2.2 Complexity Results

Hakimi presents the complexity results of both medianoid and centroid problems and shows that these problems are **NP**-Hard. He reduces two problems from graph theory with known complexity, *Dominating Set* and *Vertex Cover*, to instances of

medianoid and centroid problems. I first define these problems and then present two of the theorems proved by Hakimi.

Dominating Set (DS) Problem: Given a graph $G(V, E)$ and an integer $r<|V|$, is there a subset $V' \subset V$, such that $|V'| \leq r$ and $d(v, V') \leq 1$ for all $v \in V$? In plain language, is there a subset of nodes, such that all nodes not in this subset are directly connected to at least one of the selected nodes?

Vertex Cover (VC) Problem: Given a graph $G(V, E)$ and an integer $p<|V|$, is there a subset $V' \subset$ V, such that $|V'| \leq p$ and each edge $e \in E$ has at least one end vertex on V'? In other words, is there a subset of nodes, such that all edges have at least one adjacent node among those selected?

Theorem 3: *The problem of finding an* $(r|X_1)$ *medianoid of a network is* ***NP-****hard*.

Proof: Hakimi proves this theorem by reducing the dominating set problem to an instance of the $(r|X_p)$ medianoid problem. Given an instance of the dominating set problem, construct a network G_1 with vertex set $V \cup \{x_1\}$ and edge set $E \cup \{(x_1, v)|v \in V\}$. The vertex weights, i.e., buying powers, are all equal to one and if an edge $e \in E$, then $\ell(e) = 1.5$, while if $e=(x_1, v)$, then $\ell(e) = 2$. Hakimi shows that there exist a set of points $Y_r(x_1)$ on G_1, such that $W(Y_r(x_1)|x_1) \geq |V|$, if and only if the dominating set problem has a feasible solution. Roughly speaking, if it does, then all the original points are adjacent to a vertex in V', i.e., all the customers except the one at x_1 are closer to one of the follower facilities on G_1. Therefore it is easy to see that $W(Y_r(x_1|x_1)=|V|$. On the other hand, suppose that there is an $(r|X_1)$ medianoid on G_1, such that $W(Y_r(x_1|x_1)=|V|$ (keep in mind that Y_r can be anywhere on the network). In this case, for each node except x_1 the distance from a customer to the nearest follower location must be less than 2. If indeed all Y_r are at vertices, then this would be a feasible solution to the dominating set problem in G. It is an easy matter to show that if a location is on the edges, an equivalent solution can be obtained by changing these locations with one of the adjacent nodes. This process eventually yields a feasible solution to the *DS* problem. □

The above proof also shows that even if the locations are restricted to the nodes, the problem remains **NP-**hard.

Theorem 4: *The problem of finding a* $(1|p)$ *centroid of a network is* ***NP-****hard*.

Proof: Hakimi proves this theorem by reducing the Vertex Cover (VC) problem to the $(1|p)$ centroid problem. Given an instance of the vertex cover problem, we construct a network $G_1(V_1, E_1)$ from G by replacing each edge $e_i=(u, v)$ in G with a diamond structure depicted in Fig. 9.7a. The lengths and weights in G_1 are all equal to one. Note that $|E_1|=5|E|$, and $|V_1|=|V|+2|E|$. Hakimi then proceeds to show that there exist p points X_p on G_1, such that $W(Y_1(X_p)|X_p) \leq 3$ for every point $Y_1(X_p)$ on G_1 if and only if the vertex cover problem on G has a feasible solution.

Suppose V' with $|V'| = p$ is a solution to the vertex cover problem in G. Set now $X_p=V'$ on G_1. Then, for any diamond, the leader has a facility at either u or v, and $W(Y_1(X_p)|X_p) \leq 3$ for any point $Y_1(X_p)$ on G_1. The follower can achieve the

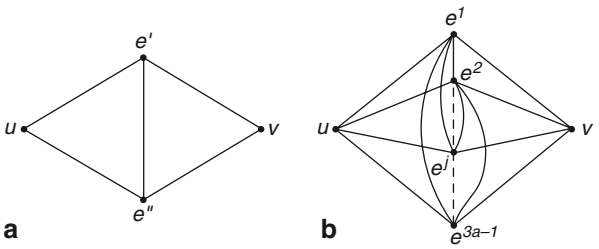


Fig. 9.7 The diamond structures used in Hakimi's proofs

upper bound if he locates anywhere inside the edge (e', e'') in any diamond structure depicted in Fig. 9.7a. While proving the converse is rather lengthy, one can replicate the steps of the earlier proof. For a given set of p points X_p on G_1, such that $W(Y_1(X_p)|X_p) \leq 3$ for every point $Y_1(X_p)$ on G_1, there must be a leader facility located on each diamond; and these locations could be shifted either to u or v without violating feasibility, thus resulting in a feasible solution to the vertex cover problem. □

Similarly, this problem also remains **NP**-hard even if the locations are restricted to the nodes. The centroid problem has proven to be substantially more difficult than the medianoid problem. In fact, Hakimi shows that even obtaining an approximate $(1|p)$ centroid of a network is **NP**-hard. Here, I do not formally state the proof, which is similar to the earlier proof except the diamond structure shown in Fig. 9.7b is used.

9.2.2.3 Immediate Generalizations and Algorithmic Developments

Hakimi's paper has led to a number of extensions and generalizations in the literature. A solution algorithm for the $(r|X_p)$ medianoid problem is provided by Megiddo et al. (1983). Their solution approach is based upon the fact that there are a finite number of points at which facilities can be located in an optimal solution. Let $G(V, E)$ be the network with $|V|=n$, $|E|=m$, and locations X_p already known. For each $v \in V$, let $B(v, X_p) = \{z \in G | d(v, z) < D(v, X_p)\}$. The customer at v will use one of the new facilities, provided one of them is located in $B(v, X_p)$. Therefore, for a $V' \subset V$, if $B(V', X_p) = \cap \{B(v, X_p) | v \in V'$ is nonempty, then there is a location that can capture all the customers in V'. The endpoints of these sets $B(V', X_p)$ are therefore attractive points to locate the new facilities. One, however, need not consider all the subsets V', since an attractive location in any $B(V', X_p)$ is defined by one of $B(v, X_p)$ for $v \in V'$. All one has to consider, for each vertex in the graph, are those points that are just at the capture distance. Megiddo et al. show that there are $O(nm)$ such points on networks (a number that reduces to $O(n)$, if the network is a tree), which are subsequently used to develop an algorithm that runs in $O(n^r m^r / r!)$ time and $O(n^2 r)$ time, if the network is a tree). This implies that if r is fixed, these algorithms run in polynomial time. If r is a problem input then the algorithm runs in exponential time for general networks. Perhaps one of the most influential approaches to solv-

ing the $(r|X_p)$ medianoid problem was given by ReVelle (1986). In his maximum capture model he formulates a version of the problem with the pre-determined set of alternative location sites, similar to plant location problems. ReVelle's model is technically more restrictive than the original problem, but one should keep in mind that despite nice discretization results, the general problem is still difficult to solve and this formulation provides a promising starting point for the solution process. As will be reviewed later, it is also one of the most extended models in the literature. ReVelle himself describes some of the extensions along pre-existing follower facilities, elastic demand, and multiple objectives in his paper.

9.3 Impact on Subsequent Work

In this section, I present an overview of the literature that follows these two influential works. Much of these developments are concerned with increasing the realism of the competitive models by relaxing some of the basic assumptions. Regardless of the settings however, centroid problems continue to be far more challenging than the medianoid problems and addressed by much fewer works. Therefore, in what follows, a major part is devoted to a systematic review of the medianoid problems. The section is followed by a review of the centroid problems and concludes with a short review of models that consider pricing issues.

9.3.1 Extended Medianoid Problems

While the classical works and their immediate generalizations expand the boundaries of solvable problems under basic assumptions, much of the subsequent contributions focus on more realistic consumer choice processes and differential store characteristics. A considerable amount of works published especially within the past two decades focus on these extensions and therefore, they will be reviewed in the first part of this section. The medianoid problems have also been extended along other dimensions. The remaining parts are devoted to such models that consider threshold market and franchising issues, congestion and queuing effects, and flow interception.

9.3.1.1 Varying Consumer Preferences and Store Characteristics

Both Drezner (1982) and Hakimi (1983) assume that customers spend their entire buying power at the closest facility. Hence, these models fall into the class of models where customer preferences are *binary* and the demand is essential, i.e., inelastic. Hakimi (1986) is also among the first to consider cases where consumer preferences are not fixed and the demands may be inessential or elastic. In addition to the

binary preferences, he also considers *partially binary* and *proportional* preferences. In the former, consumers select their most preferred locations from both firms and spend their buying power proportionally between two rival facilities. Let $w_x(v)$ and $w_y(v)$ denote the buying powers the leader and the follower captures, respectively. Under partially binary rule, consumers split their buying power according to

$$\frac{w_y(v)}{w_x(v)} = \frac{\max\limits_{y_i}\{a(v, y_i)\}}{\max\limits_{x_j}\{a(v, x_j)\}},$$

where $a(v, x_j) = 1/f_v(d(v, x_j))$ and $f_v(0) > 0$ is a concave increasing function. Under proportional rule, the customers spread their purchases to all facilities, hence

$$\frac{w_{z_i}(v)}{w_{z_j}(v)} = \frac{a(v, z_i)}{a(v, z_j)},$$

where z_i and z_j could denote facilities of either firms. The partially binary case would be more suitable for firms with a high degree of standardization between its branches, while the proportional rule would be more suitable when facilities display differences.

Note that in Hakimi's model, the only differentiation among the facilities is the proximity to customers. It is well established in the marketing literature that customers usually consider distance and some other attributes of the facilities such as size, quality of products and services, and parking space when they make a decision. One of the most popular extensions is to consider more sophisticated customer preference schemes such as Huff-like preferences. In the model proposed by Huff (1964), there is an attraction function that each customer evaluates for each facility. In his model, the attraction is directly proportional to the size of the facility and inversely proportional to a power of the distance, that is, a customer at node v will have an attraction towards a facility at x_j.

$$a(v, x_j) = \frac{A_j}{d(v, x_j)^{\beta}},$$

where A_j is the size of facility j and $d(v, x_j)$ is the shortest distance between customer v and the facility x_j. Later, I assume that A_j is a scalar measure of important facility characteristics such as service quality, size, product variety. Although this model has also been used in a few works, the attraction function

$$a(v, x_j) = \frac{A_j}{f_v(d(v, x_j))},$$

has been used more commonly due to its generality. Nakanishi and Cooper (1974) present a major improvement in their Multiplicative Competitive Interaction (*MCI*) model, which replaces the floor area by a product of factors, each a component of

attractiveness. For most of the models that are solved through numerical means, however, Huff attraction can easily be extended to the multiplicative competitive interaction model. Finally, few works consider these problems under random utility. Here, for each customer v, a utility for each open facility x_i is defined as

$$U_v(i) = a_i - bd(v, x_i) + \varepsilon_{iv},$$

where ε_{iv} is independently and identically distributed random variable. If the distribution is Normal, then the model is called probit and may be numerically quite demanding. Therefore, the most common form of this utility is the logit model, which assumes that the distribution of ε_{iv} is a double-exponential (Gumbel). Under the logit model, the probability that a customer will patronize a store at x_j is given as

$$P_v(j) = \frac{\exp(a_j - bd(v, x_j))}{\sum_i \exp(a_i - bd(v, x_i))}.$$

These extensions have been studied quite extensively in the past. To effectively review this voluminous literature, I present them in three subsections as planar models, network models, and discrete models.

Planar Models

Much of the extensions of the early work on planar models is conducted by Tammy Drezner and her colleagues. Even the simplest case in which a single new facility with fixed attractiveness is located can be a difficult problem since a large number of local optima are observed in these problems. Therefore, much of the early work proposes heuristics. Recent years have witnessed a renewed interest in these problems with a bolder algorithmic agenda that aims to solve larger instances to optimality.

A series of papers, Drezner (1994a, b), Drezner and Drezner (1996), and Drezner et al. (1998), all study models that consider $(1|X_p)$ medianoid under unequal store attractiveness levels. Drezner (1994a) considers a model under binary preferences and extends Drezner's (1982) algorithm to solve the problem. Drezner (1994b) considers the Huff-like preferences, for which she proposes a Weiszfeld-like fixed-point iteration solution procedure. Through computational experience, it is found that many local optimal solutions typically exist. Therefore, the solution obtained here is a local optimum solution. Finally, Drezner and Drezner (1996) study the problem under a specific random utility, i.e., the generalized probit. Dealing with this type of utilities might be computationally prohibitive because the objective function evaluation requires computing a p-dimensional integral. In a follow-up paper, Drezner et al. (1998) show that a simple logit model can accurately be used in place of a probit model. Again, an iterative Weiszfeld-type procedure is proposed for the model with the logit model, and a set of computational experiments are carried out.

An extension to multiple entering facilities, an $(r|X_p)$ medianoid problem with existing company owned facilities, is discussed by Drezner et al. (2002a). They propose and test five heuristic solution methods. They also observe that the accuracy of the market share estimate is enhanced by replacing the distance with the "corrected" version $\sqrt{d^2+\varepsilon}$. This distance correction is needed to avoid market shares being artificially inflated when the facilities locate at a demand point. It also assists the numerical methods, as it maintains the continuity of the objective function at customer locations. McGarvey and Cavalier (2005) present a gravity-based utility model, in which a facility's capacity is used as its measure of its attractiveness. A new formulation of the $(r|X_p)$ medianoid in the plane with elastic and gravity-based demand, including capacity, forbidden region and budget constraints, is given and solved by an exact branch-and-bound and a heuristic method.

Some models consider both locations and attraction levels as decision variables. Usually, attraction levels are modeled through a budgeted formulation or in a profit maximization framework. Although Drezner (1998) presents a fairly general version of the problem, no special algorithm is provided. Recent years have witnessed a renewed interest in such problems. However, the problems are of considerable difficulty; only $(1|X_p)$ and $(2|X_p)$ medianoid problems that involve the location and the attractiveness criteria have been solved to optimality.

Fernandez et al. (2007a, b) have reported the application of interval analysis tools in a branch and bound solution procedure for a series of single facility location and attractiveness problems under Huff-like preferences. In Fernandez et al. (2007a), a profit maximization version of the problem is modeled and solved by a Weiszfeld-type heuristic algorithm and the interval analysis based branch and bound procedure. Their computational experiments suggest that the heuristic method does not yield consistent results (a further testament to the large number of local optimal solutions), whereas the global optimization method yields a solution within a guaranteed optimality after an acceptable computational effort. In a companion paper, Fernandez et al. (2007b) consider the same problem for a chain that likes to maximize its share as a primary objective and to minimize the cannibalization of the existing stores as a secondary objective.

More recently Toth et al. (2008) and Redondo et al. (2009a, b) study a $(2|X_p)$ medianoid problem that involves the location as well as the attractiveness criteria. In the former, the problem is solved by sequential and simultaneous approaches based on the exact interval analysis based branch-and-bound algorithm. The running time of this algorithm, however, might be impractical for a further generalization as it can take up from six to 140 hours for one instance. Later, Redondo et al., in two papers, investigate the effectiveness of several heuristic procedures for this problem.

Network Models

Most network models focus on obtaining discretization and node optimality results, which may subsequently be used in various search techniques such as those demonstrated in Megiddo et al. (1983). The general conclusion that can be drawn from

these works is that most models under partially binary and proportional preferences show node optimality property. However, models under binary preferences do not posses this property, but instead display certain discretization results. The earlier works assume that all facilities have equal attractiveness. I first review those models, then shift the focus to more general cases.

As mentioned earlier, Hakimi (1986) studies a series of problems on networks under binary, partially binary, and proportional preferences, as well as under essential and nonessential demands. Hakimi's results for these problems are rather promising, as he proves node-optimality in all $(r|X_p)$ medianoid problems except those with binary preferences. His proofs use the convexity of the objective function of the points along edges, which depends on the properties of the distance function, i.e., concave (linear, in case of proportional preferences and nonessential demand), nondecreasing, and $f_v(0)>0$ for all nodes. Suarez-Vega et al. (2004a) provide an alternative treatment of the problems studied in Hakimi (1986). First, they extend the node optimality result of proportional preferences and nonessential demand for the general concave distance functions. Second, they provide a comprehensive computational study where they test three solution heuristic procedures.

An earlier attempt to include Huff-like preferences with unequal attractiveness is made by Peeters and Plastria (1998). First, they extend Hakimi's node optimality results and then they analyze a further extension in which customers purchase from only those stores that lie in the Pareto frontier of the distance functions and the facility attractiveness. For this Pareto-Huff model, they show that the set of nodes can further be augmented to look for the best set of locations. Subsequently, Suarez-Vega et al. (2004b) unify and extend the essential demand case. They show that respective discretization and node-optimality results continue to hold. Similarly, they provide discretization results for the facility attraction levels, if the locations are assumed to be given. Finally, Santos-Penate et al. (2007) report that respective node-optimality results extend to cases with elastic demand under all customer preferences, and Suarez-Vega et al. (2007) extend the earlier node-optimality results to cases where customers proportionally spend their buying power at the facilities which exceed a specified attraction-distance threshold.

Discrete Models

As noted earlier, ReVelle's (1986) maximum capture (*MAXCAP*) model provides a workable integer programming formulation that is quite amenable to various extensions. In fact, discrete competitive location problems with Huff-like preferences predate even ReVelle's work. For example, Achabal et al. (1982) present a problem of chain expansion under Multiplicative Competitive Interactions (*MCI*) model. They formulate the problem as a nonlinear integer program and propose an interchange heuristic. Similarly, Ghosh and Craig (1983) present an earlier work in which they formulate both medianoid and centroid problems under proportional preferences, and propose an iterative heuristic solution procedure.

Subsequently, a number of works followed. Eiselt and Laporte (1989) extend the maximum capture formulation by including attraction functions. They then provide the solution method for $(1|X_p)$ medianoid, as well as the facility's optimal attraction level. Karkazis (1989) includes two criteria that customers consider when deciding which facility to patronize. Serra et al. (1999b) and Serra and Colome (2001) formulate and solve maximum capture problems under partially binary and proportional preferences with essential demand. They have solved the problems exactly and approximately and present their numerical experiments. Benati (1999) and Benati and Hansen (2002) present a medianoid problem under a random utility model. Benati (1999) proves two important theoretical features of the problem; first he shows the submodularity of the objective function; second, he demonstrates that the problem can be formulated as a *p*-median type problem. He uses these properties to develop two exact solution methods. Benati and Hansen (2002) present further results and also discuss heuristic solution methods. Finally, Benati (2003) presents a version of *MAXCAP* in which the number of new facilities are endogenized through the introduction of location dependent fixed costs.

Finally, Berman and Krass (2002) present a general modeling framework for the follower's problem. In their framework, customers have Huff-like preferences and general expenditure functions, which imply generalized elastic demand functions. It is shown that the form of the total expenditure function profoundly effects the difficulty of the problem. For example, while cases with linear expenditure functions are easy to solve those with the bounded ones are rather difficult. Subsequently, Aboolian et al. (2007a) analyze a special case of the expenditure function, where the demand function satisfies certain concavity assumptions. The problem is then formulated as a nonlinear knapsack problem and solved by an approach based on a piecewise linear approximation scheme for the objective function. This leads to exact and near-optimal solution approaches capable of dealing with relatively large scale instances of the model. In a companion paper, Aboolian et al. (2007b) extend the problem to simultaneously optimize the locations and designs, i.e., attractiveness, of a set of new facilities under a budget constraint.

9.3.1.2 Threshold Market and Franchising Issues

Most large retail firms give their franchisees some kind of territorial exclusivity. Even if contracts may vary in specifics, there are provisions that guarantee a level of revenue opportunity, i.e., the threshold market, to franchisees. This concept is also relevant for planning public services such as hospitals, schools, and post offices. Policymakers may need to consider the threshold market to have economically viable facilities. There are a number of works that incorporate threshold market concept into medianoid and capture models under varying assumptions.

Balakrishnan and Storbeck (1991) present a model in which franchise locations and individual franchise market areas selected so that market coverage is maximized and the required threshold level of demands are maintained for all sites. Serra et al. (1999a) provide a mixed integer formulation and propose a metaheuristic.

Ghosh and Craig (1991) and Current and Storbeck (1994) realize that total market coverage and threshold market objectives might be in conflict, and therefore study the multi-objective version. Ghosh and Craig consider the effects of one new outlet on demand and on competition, and then develop a network optimization model to evaluate potential sites for their impact on total revenues and the revenues of the existing outlets. Current and Storbeck, on the other hand, formulate a multiobjective model that guarantees that all franchise locations were assigned at least a minimal threshold market area with sufficient demand to ensure economic survival.

The above works consider deterministic demand in which the threshold markets are also expressed deterministically. Drezner et al. (2002b) and Colome et al. (2003) present extensions along the stochastic demand. In both papers, customers are assumed to have buying power distributed according to some statistical distribution. Drezner et al. present a single facility planar location problem with the gravitational attraction function and stochastic threshold. Their objective is to minimize the probability of falling short of the required threshold. Colome et al., on the other hand, present a model based on the discrete *MAXCAP* model by introducing a stochastic threshold constraint.

9.3.1.3 Congestion Effects

In the majority of competitive location problems, the consumer utility (explicitly or implicitly) depends on exogenously given data such as distance or facility attractiveness level. When dealing with location decisions in practice, most firms need to consider quality of service and congestion effects, which are partly determined by the capacity of the facility and its market area. These issues are addressed by a family of work that incorporate queuing aspects with location decisions.

Marianov and Serra (1998) present several probabilistic maximal covering models with constrained waiting time and queue length. Their basic model addresses the issue of the location of a given number of single-server centers, such that the maximum population is served within a standard distance and satisfy chance-constraints on average queue-length and waiting time. Marianov et al. (2004) later extend the model for public service centers competing with private companies. Recently, Silva and Serra (2007) present an extension of the maximum capture problem, in which customer's choice depends on the traveling time plus the waiting time at the facilities. Finally, Zhang and Rushton (2008) present a more comprehensive discrete model that, in addition to the queuing effects, considers simultaneous facility opening and closing as well as sizing decisions. They then present` a mathematical formulation and propose a genetic algorithm to heuristically solve the problem.

9.3.1.4 Flow Interception or Capture

In all models reviewed so far, the distance between a customer and a facility plays a primary role in competitive location models. The main premise of all these papers

is that customers make special trips to stores to make purchases, and therefore proximity to facilities is important. However, in many cases, such as ATM machines, gasoline stations, and convenience stores, customers make purchases as part of their routine trips, such as those from home to work and return. These concerns are addressed in models with flow-interception, which considers not only distance concerns of the customers but also their trip profiles.

Berman and Krass (1998) appear to be first to include flow capture in a competitive location framework. Their model represents a direct generalization of the traditional spatial interaction models, and thus all results in the paper are immediately applicable to cases without flow interception. The new model is formulated as a nonlinear integer program, for which they develop both exact branch-and-bound and heuristic solution algorithms. Their computational results suggest that the heuristic method typically yields a solution within 1–2% of the optimal solution. Subsequently, Wu and Lin (2003) propose a new formulation and a greedy heuristic for the problem. They have tested their approach and model on some small problems and as well as on a network of the city of Yuanlin in Taiwan.

9.3.2 Centroid and Related Problems

The general version of the centroid problem, whether defined in the plane or on networks, is substantially more difficult than its medianoid counterpart. To date, only (1|1) centroid problem is solved to optimality with an algorithm of known complexity. On networks, Hansen and Labbe (1988) propose an algorithm that runs in $O(n^2m^2 \log mn \log D)$, where n is the number of nodes, m is the number of edges, and D is the total buying power in the network. As mentioned earlier, Drezner (1982) originally proposed two polynomial algorithms for the planar version of the problem.

In this section, I review the progress on centroid problems, which is rather limited, but I also cover some related problems such as those from voting theory and Voronoi games, as they are related to the centroid concept. The next section reviews algorithms to solve centroid problems on networks and in the plane, followed by problems from voting theory and Voronoi games. Finally, I review the location games with endogenous number of facilities and those with asymmetric information.

9.3.2.1 Algorithms for Centroid Problems

Among the earlier approaches, Serra and ReVelle (1994) propose two heuristics to solve these problems. While one of the heuristics is based on a search on leader's locations and the optimal solution to the follower's problem, the other is a heuristic at both levels. Around the same time, Benati and Laporte (1994) develop a tabu

search heuristic algorithm for the classical centroid and medianoid problems, which uses the lower bound computed for the capture function. They also show that the capture function is submodular and nondecreasing. This result is later replicated by Santos-Penate and Suarez-Vega (2003) for all cases of centroid and medianoid problems described in Hakimi (1986).

Not surprisingly, solving these problems in the plane is also difficult; to our knowledge, only one study attempts this. Bhadury et al. (2003) propose an alternating heuristic scheme where the leader and the follower repeatedly solve medianoid problems. They test two heuristic decision models by the follower: the first one is a greedy heuristic based on Drezner's algorithm for solving the $(1|X_p)$ medianoid problem. In the second, which is found to be better overall, the players try to locate as close to the leader facilities as possible. They also indicate, however, that solutions appear to be very sensitive with respect to the assumptions of the model and the relocation process.

Drezner and Drezner (1998) consider a $(1|1)$ centroid problem under Huff-like consumer preferences, for which they propose three solution procedures. Recently, Saiz et al. (2007) consider a version of the problem where both the leader and the follower have pre-existing stores with different qualities. They propose a branch-and-bound algorithm that solves both problems up to a desired accuracy.

9.3.2.2 Voting Theory

While the focus in voting theory is somewhat different, there are intimate connections between the solutions of competitive location models and those of voting models. In voting theory, the main purpose is to locate one facility (such as a school or a hospital) as a result of a voting process. The same problem could be cast for a politician who likes to position on a platform that would win the election. One of the important issues in voting theory is to compare the quality of the voting solution to other plausible solutions, such as those that arise out of competition or, more commonly, to the benevolent dictator's solution (the Weber solution), which minimizes the average distance that must be traveled by users.

A point is called a Condorcet point if no strict majority prefers another point to this point. Therefore, a Condorcet point gets at least 50% of the votes. It is an attractive definition, however, a Condorcet point does not necessarily exist under general conditions. In the plane, a considerably earlier work by Wendell and Thorson (1974) shows that every Weber point is also a Condorcet point, if the distances are rectilinear. On networks, Hansen and Thisse (1981) as well as Wendell and McKelvey (1981) show that every Condorcet point on a tree network is a Weber point. Hansen and Thisse also demonstrate that, on general networks, the average distance between users and a Condorcet point may be almost three times larger than that of a Weber location. This indeed indicates that sometimes voting solutions may lead to poor outcomes. Labbe (1985) and Bandelt (1985) describe other networks for which Condorcet points always exist or coincide with the Weber points.

When a Condorcet point fails to exist, a natural alternative to look for is the least objectionable solution: a point minimizing the maximum number of users preferring another point. This is called the Simpson point in voting theory and it is precisely the (1 | 1) centroid of a network. Maximum relative rejection of a point is defined as the proportion of customers preferring another location, which is equivalent to the follower's market share. Bandelt and Labbe (1986) study the properties of the Simpson point and compare it to the Weber point. They show that in general networks, a Simpson point could indeed be very bad: the maximum relative rejection of a Simpson point could be made arbitrarily close to one, and a Simpson point could be as bad as any point on the network when average distance between the users and the facility are considered.

Subsequent works in voting theory consider various extensions. A recent paper by Campos-Rodriguez and Moreno-Perez (2003) relaxes the conditions for a Condorcet point in two ways: first, by considering two locations as indifferent for every user if the difference of the distances to them is within a positive threshold; second, by considering that the proportion of users needed to reject a location could be more than one half. Noltemeier et al. (2007) study a multiple voting location problem on networks. Apart from a very efficient algorithm for tree networks, they prove a result towards lower bounds on the complexity of the single voting location problem.

9.3.2.3 Voronoi Games

The Voronoi game is played on a bounded continuous area by two players who put p points alternately, and the continuous field is subdivided according to the nearest neighbor rule. The player who gets the larger area at the final step wins the game. It is a simplified model for a competitive facility location.

There are potentially large numbers of different games that can defined as Voronoi games. For example, the classical ($p|p$) centroid problem is a one-round Voronoi game, in which the leader locates p facilities followed by the follower's p facilities. However, Voronoi games may be very difficult to solve in general and few special cases have been dealt with so far. Among those, Ahn et al. (2004) investigate the case where the market is a line segment with uniformly distributed customers. They have shown that the follower always has a winning strategy that would capture $½+\varepsilon$, $\varepsilon>0$ of the market. But the leader also has a strategy that could make ε arbitrarily small.

On the other hand, Cheong et al. (2004) and Fekete and Meijer (2005) deal with a 2-dimensional cases, but they restrict themselves to the one-round game. The former paper extends Ahn et al.'s result in the following way: the follower still does have an advantage, but the leader does not have a strategy that would make ε arbitrarily small. In fact, after a threshold value of p, there is an ε that is independent of the number of stores. Fekete and Meijer contributes towards the complexity of the problem, showing that even the follower's problem is **NP-**hard.

9.3.2.4 Endogenous Number of Facilities

Both medianoid and centroid problems are defined by prespecified numbers of facilities to be opened. In noncompetitive location models, the implicit understanding is that a decision maker solves a family of problems and then makes a decision on the number and locations of facilities. Its implication in a competitive environment however is not clear. Some works consider budgeted version but, in reality, firms could open facilities as long as it is profitable to do so, as they usually invest with borrowed money. Hence, fixed costs are introduced, which also act to endogenize the number of facilities that firms open.

The literature is rather scant in this area. Dobson and Karmarkar (1987) and Dasci and Laporte (2005) consider these problems with fixed costs, but this is where their similarities end. Dobson and Karmarkar consider the problem on networks, where firms have equal and identical fixed costs, and co-location by the follower is not permitted. Under these assumptions, the follower can never enter the market profitably, because if there were any profitable location left to the follower, the leader would have taken it already as a worst case action. Hence, the authors focus on various stable set definitions, i.e., the sets of leader store locations that deter entry, and algorithms to find those sets. Dasci and Laporte (2005), on the other hand, treat the consumers as entities continuously spread over the plane according to a given density and consider any point as an alternative for location. They then transform the problem to a Voronoi-like game and analyze entry and entry deterrence conditions under various fixed costs scenarios. Unlike Dobson and Karmarkar's model, here firms can co-exist in the market.

9.3.2.5 Uncertainty and Information

These topics have not attracted enough interest from the literature, partly because problems with different informational structures or uncertainty are quite involved. Eiselt (1998) studies the case of an information asymmetry on a tree network, where both the leader and the follower locate a single facility. He analyzes the problem under the assumption that firms locate at the nodes and investigates the impact of information asymmetry on firms' relative advantages. For uncertainty, Shiode and Drezner (2003) consider a (1 | 1) centroid problem on a tree network, where customers are located at the nodes and are assumed to have stochastic demand. They have proven that the leader's optimal solution is to locate at one of the nodes. Once the leader's decision is made, the follower's solution is characterized rather easily.

9.3.3 Competitive Location with Pricing Issues

The discussion so far confirms that these problems are difficult even without pricing decisions. Although the results are still scant, progressively more manageable

problems have been identified and presented. There are relatively more diverse sets of models that include pricing and therefore, I present these works in three broad classes: those that consider only the follower's location and pricing decisions, those that deal with Hotelling's problem on trees, and those of network and discrete models that consider location and some form of pricing decisions.

In the plane, Zhang (2001) extends the original $(1|X_p)$ medianoid problem with a price decision to be applied at the facility and presents a procedure with a worst-case complexity of $O(n^3 \log n)$. There are also works that incorporate pricing decisions in discrete domain: Serra and ReVelle (1999) extend the maximum capture model with a price decision and study both the inelastic and elastic demand cases. Recently, Plastria and Vanhaverbeke (2009) introduce a version of a maximum capture model with inelastic demand. They have explored various properties of the optimal solution, which are later used in an efficient enumeration method.

Eiselt (1992) is among the first to consider the Hotelling's problem on a tree network. He shows that if both competitors have price and location as decision variables, no equilibrium exists in pure strategies. If, however, prices are fixed in advance, equilibria may exist under certain conditions. Subsequently, Eiselt and Bhadury (1998) have shown that if equilibria exist, they could be reached as a result of a sequential location game played by the firms. Garcia-Perez and Pelegrin (2003) study the same problem and prove a few interesting node-optimality results. Under different prices, an optimal location for the follower at a node always exists, while for the leader optimal locations could only be in the interior of an edge. Under equal prices, an optimal location for the leader at a node always exists, but for the follower it might be on the interior of an edge.

A part of this literature also considers multi stage games that involve location and/or pricing decisions. For example, some earlier works such as Tobin and Friesz (1986), Friesz et al. (1989), and Miller et al. (1991) study a set of leader-follower problems where an entering firm(s) wishing to establish production facilities to maximize its profit, taking into account the changes in prices at each market that would result from the increase in supply provided by its new facilities and from the response of competing firms that are already established. More recently, Fischer (2002) and Pelegrin et al. (2006) study these problems on networks. In the latter, leader's locations are fixed and once the follower sets up his facility, firms engage in a price competition. They study the equilibrium and optimization problems under mill and spatial delivered pricing policies, and present extensive discretization results. Fischer, on the other hand, presents a series of leader-follower models with spatial discriminatory pricing, and develops bilevel programming formulations.

9.4 Conclusion

This chapter provides an overview of conditional location problems with a special emphasis on the works that pioneered this field. Both Drezner's and Hakimi's works are so central to this field that there are still open research issues that are

rather basic extensions of these works. I would like to point out that many of the works that are reviewed in earlier sections such as those with threshold market concepts, congestion and queuing effects, and flow-interception could be extended in various ways.

Perhaps two important weaknesses of conditional location problems are the designations of the leader and the follower and the exogenously set number of facilities. While in some settings these choices could arise naturally or these models could very well be useful in decision-making, it is not clear what makes one firm a leader and the other one a follower, or what really determines the number of facilities they plan to open. Therefore, an important future research avenue is to develop models in which the leader and follower designations and/or the number of facilities arise endogenously as a result of competition and strategic interactions.

Most of the existing models assume that firms are identical in terms of the objectives they pursue. However, there are a number of possible scenarios where they are not the same. For example, one of the firms could be a public or a quasi-public provider, or it might not have the ownership of all stores, or it might choose to pursue multiple objectives. Therefore, considering alternative objectives in these models is an important research avenue. Furthermore, including franchising contracts, such as those related to pricing and service delivery, may also be a related area of study.

Finally, today it is not uncommon for firms to own and operate a number of chains. For example, consider KFC, Pizza Hut, and Taco Bell, or KMart and Sears. Therefore, an integrated location planning approach is needed for the management of such intertwined retail networks. There are a number issues, ranging from market segmentation and cannibalization to brand management to economies of co-location, that could be explored in a competitive location framework.

References

Aboolian R, Berman O, Krass D (2007a) Competitive facility location model with concave demand. Eur J Oper Res 181:598–619

Aboolian R, Berman O, Krass D (2007b) Competitive facility location and design problem. Eur J Oper Res 182:40–62

Achabal D, Gorr WL, Mahajan V (1982) MULTILOCL: a multiple store location decision model. J Retail 58:5–25

Ahn H-K, Cheng S-W, Cheong O, Golin M, van Oostrum R (2004) Competitive facility location: the Voronoi game. Theor Comput Sci 310:457–467

Balakrishnan P, Storbeck J (1991) Mctresh: modeling maximum coverage with threshold constraints. Environ Plan B 18:459–472

Bandelt H-J (1985) Networks with Condorcet solutions. Eur J Oper Res 20:314–326

Bandelt H-J, Labbe M (1986) How bad can a voting location be? Soc Choice Welf 3:125–145

Benati S (1999) The maximum capture problem with heterogeneous customers. Comput Oper Res 26:1351–1367

Benati S (2003) An improved branch and bound method for the uncapacitated competitive location problem. Ann Oper Res 122:43–58

Benati S, Hansen P (2002) The maximum capture problem with random utilities: problem formulation and algorithms. Eur J Oper Res 143:518–530

Benati S, Laporte G (1994) Tabu search algorithms for the ($r|Xp$)—medianoid and ($r|p$)—centroid problems. Locat Sci 2:193–204
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, Krass D (2002) Locating multiple competitive facilities: spatial interaction models with variable expenditures. Ann Oper Res 111:197–225
Bhadury J, Eiselt HA, Jaramillo JH (2003) An alternating heuristic for medianoid and centroid problems in the plane. Comput Oper Res 30:553–565
Campos-Rodriguez CM, Moreno-Perez JA (2003) Relaxation of the Condorcet and Simpson conditions in voting location. Eur J Oper Res 145:673–683
Cheong O, Har-Peled S, Linial N, Matousek J (2004) The one-round Voronoi game. Discrete Comput Geom 31:125–138
Colome R, Lourenco HR, Serra D (2003) A new chance-constrained maximum capture location problem. Ann Oper Res 122:121–139
Current J, Storbeck J (1994) A multiobjective approach to design franchise outlet networks. J Oper Res Soc 45:71–81
Dasci A, Laporte G (2005) A continuous model for multistore competitive location. Oper Res 53:263–280
Daskin MS (1995) Network and discrete location: models, algorithms, and applications. Wiley, New York
Dobson G, Karmarkar US (1987) Competitive location on a network. Oper Res 35:565–574
Drezner Z (1982) Competitive location strategies for two facilities. Reg Sci Econ 12:485–493
Drezner T (1994a) Locating a single new facility among existing unequally attractive facilities. J Reg Sci 34:237–252
Drezner T (1994b) Optimal continuous location of a retail facility, facility attractiveness, and market share: an interactive model. J Retail 70:49–64
Drezner T (1998) Location of multiple retail facilities with limited budget constraints—in continuous space. J Retail Consum Serv 5:173–184
Drezner T, Drezner Z (1996) Competitive facilities: market share and location with random utility. J Reg Sci 36:1–15
Drezner T, Drezner Z (1998) Facility location in anticipation of future competition. Locat Sci 6:155–173
Drezner Z, Wesolowsky GO, Drezner T (1998) On the logit approach to competitive facility location. J Reg Sci 38:313–327
Drezner T, Drezner Z, Salhi S (2002a) Solving the multiple competitive facilities location problem. Eur J Oper Res 142:138–151
Drezner T, Drezner Z, Shiode S (2002b) A threshold satisfying competitive location model. J Reg Sci 42:287–299
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, Bhadury J (1998) Reachability of locational Nash equilibria. OR Spectrum 20:101–107
Eiselt HA, Laporte G (1989) The maximum capture problem in a weighted network. J Reg Sci 29:433–439
Eiselt HA, Laporte G (1996) Sequential location problems. Eur J Oper Res 96:217–231
Eiselt HA, Laporte G, Thisse J-F (1993) Competitive location models: a framework and bibliography. Transp Sci 27:44–54
Fernandez J, Pelegrin B, Plastria F, Toth B (2007a) Solving a Huff-like competitive location and design model for profit maximization in the plane. Eur J Oper Res 179:1274–1287
Fernandez J, Pelegrin B, Plastria F, Toth B (2007b) Planar location and design of a new facility with inner and outer competition: an interval lexicographical-like solution procedure. Netw Sp Econ 7:19–44
Fekete SP, Meijer H (2005) The one-round Voronoi game replayed. Comput Geom Theory Appl 30:81–94

Fischer K (2002) Sequential discrete *p*-facility models for competitive location planning. Ann Oper Res 111:253–270
Friesz T, Tobin L, Miller T (1989) Existence theory for spatially competitive network facility location models. Ann Oper Res 18:267–276
Garcia-Perez MD, Pelegrin BP (2003) All Stackelberg location equilibria in the Hotelling's duopoly model on a tree with parametric prices. Ann Oper Res 122:177–192
Ghosh A, Craig CS (1983) Formulating retail location strategy in a changing environment. J Mark 47:56–68
Ghosh A, Craig CS (1991) FRANSYS: a franchise location model. J Retail 67:212–234
Ghosh A, McLafferty SL (1987) Location strategies for retail and service firms. Lexington Books, Lexington
Hakimi SL (1983) On locating new facilities in a competitive environment. Eur J Oper Res 12:29–35
Hakimi SL (1986) *p*-median theorems for competitive location. Ann Oper Res 5:79–88
Hakimi SL (1990) Locations with spatial interactions: competitive locations and games. In: Mirchandani PB, Francis RL (eds) Discrete location theory. Wiley, New York, pp 439–478
Hansen P, Thisse J-F (1981) Outcomes of voting and planning: Condorcet; Weber and Rawls locations. J Public Econ 16:1–15
Hansen P, Labbe M (1988) Algorithms for voting and competitive location on a network. Transp Sci 22:278–288
Hansen P, Thisse J-F, Wendell RW (1990) Equilibrium analysis for voting and competitive location problems. In: Mirchandani PB, Francis RL (eds) Discrete location theory. Wiley, New York, pp 479–501
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
Karkazis J (1989) Facilities location in a competitive environment: a Promethee-based multiple criteria analysis. Eur J Oper Res 42:294–304
Labbe M (1985) Outcomes of voting and planning in single facility location problems. Eur J Oper Res 20:299–313
Lee DT, Wu YF (1986) Geometric complexity of some location problems. Algorithmica 1:193–211
Marianov V, Serra D (1998) Probabilistic maximal covering location-allocation for congested systems. J Reg Sci 38:401–424
Marianov V, Rios M, Taborga P (2004) Finding locations for public service centres that compete with private centres: effects of congestion. Pap Reg Sci 83:631–648
McGarvey RG, Cavalier TM (2005) Constrained location of competitive facilities in the plane. Comput Oper Res 32:359–378
Megiddo N, Zemel E, Hakimi SL (1983) The maximum coverage location problem. SIAM J Algebra Discrete Method 4:253–261
Miller T, Tobin R, Friez T (1991) Stackelberg games on a network with Cournot-Nash oligopolistic competitors. J Reg Sci 31:435–454
Nakanishi M, Cooper LG (1974) Parameter estimate for multiplicative interactive choice model: least squares approach. J Mark Res 11:303–311
Noltemeier H, Spoerhase J, Wirth H-C (2007) Multiple voting location and single voting location on trees. Eur J Oper Res 181:654–667
Peeters PH, Plastria F (1998) Discretization results for the Huff and Pareto-Huff competitive location models on networks. Top 6:247–260
Pelegrin B, Fernandez J, Suarez R, Garcia-Perez MD (2006) Single facility location on a network under mill and delivered pricing. IMA J Manag Math 17:373–385
Plastria F (2001) Static competitive facility location: an overview of optimization approaches. Eur J Oper Res 129:461–470
Plastria F, Vanhaverbeke L (2009) Maximal covering location problem with price decision for revenue maximization in a competitive environment. OR Spectrum 31:555–571
Redondo JL, Fernández J, García I, Ortigosa PM (2009a) Solving multiple competitive facilities location and design problem on the plane. Evolutionary Computation 17:21–53

Redondo JL, Fernández J, García I, Ortigosa PM (2009b) A robust and efficient algorithm for planar competitive location problems. Ann Oper Res 167:87–105
ReVelle C (1986) The maximum capture or "sphere of influence" location problem: Hotelling revisited on a network. J Reg Sci 26:343–358
Saiz ME, Hendrix EMT, Fernandez J, Pelegrin B (2007) On a branch-and-bound approach doe a Huff-like Stackelberg location problem. Discussion paper No. 37, Mansolt Graduate School
Santos-Penate DR, Suarez-Vega R (2003) Submodular capture functions in competitive location: the (*r*|*Xp*)—medianoid and the (*r*|*p*)—centroid problems. 27 Congreso SEIO, Lerida
Santos-Penate DR, Suarez-Vega R, Dorta-Gonzalez P (2007) The leader--follower location model. Netw Sp Econ 7:45–61
Serra D, Colome R (2001) Consumer choice and optimal location 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 34:549–561
Serra D, ReVelle C (1999) Competitive location and pricing on networks. Geogr Anal 31:109–129
Serra D, ReVelle C, Rosing K (1999a) Surviving in a competitive spatial market: the threshold capture model. J Reg Sci 39:637–650
Serra D, Eiselt HA, Laporte G, ReVelle C (1999b) Market capture models under various customer choice rules. Environ Plan B 26:741–750
Shiode S, Drezner Z (2003) A competitive facility location problem on a tree network with stochastic weights. Eur J Oper Res 149:47–52
Silva F, Serra D (2007) Incorporating waiting time in competitive location models. Netw Sp Econ 7:63–76
Slater PJ (1975) Maximin facility location. J Nat Bureau Stand B 79:107–115
Suarez-Vega R, Santos-Penate DR, Dorta-Gonzalez P (2004a) Competitive multi-facility location on networks: the (*r*|*Xp*)-medianoid problem. J Reg Sci 44:569–588
Suarez-Vega R, Santos-Penate DR, Dorta-Gonzalez P (2004b) Discretization and resolution of the (*r*|*Xp*)—medianoid problem involving quality criteria. Top 12:111–133
Suarez-Vega R, Santos-Penate DR, Dorta-Gonzalez P (2007) The follower location problem with attraction thresholds. Pap Reg Sci 86:123–137
Tobin R, Friesz T (1986) Spatial competition facility location models: definition, formulation and solution approach. Ann Oper Res 6:49–74
Toth B, Fernandez J, Pelegrin B, Plastria F (2008) Sequential versus simultaneous approach in the location and design of two new facilities using planar Huff-like model. Comput Oper Res (to appear)
Wendell RE, Thorson SJ (1974) Some generalizations of social decisions under majority rule. Econometrica 42:893–912
Wendell RE, McKelvey R (1981) New perspectives in competitive location theory. Eur J Oper Res 6:174–182
Wu TH, JN Lin (2003) Solving the competitive discretionary service facility location problem. Eur J Oper Res 144:366–378
Zhang S (2001) On a profit maximizing location model. Ann Oper Res 103:251–260
Zhang L, Rushton G (2008) Optimizing the size and locations of facilities in competitive multi-site service systems. Comput Oper Res 35:327–338

Chapter 10
The Location of Undesirable Facilities

Emanuel Melachrinoudis

10.1 Introduction

Undesirable facilities are those facilities that have adverse effects on people or the environment. They generate some form of pollution, nuisance, potential health hazard, or danger to nearby residents; they also may harm nearby ecosystems. Examples are incinerators, landfills or sewage plants, airports, stadia, repositories of hazardous wastes, nuclear or chemical plants, prisons, and military installations. Although they provide some disservice to nearby residents, these facilities are necessary to society. In addition, there is often some travel involved to and from these facilities and an associated transportation cost that increases with distance from the population, which in turn suggests that they should be placed away but not very far away. The terms *semi-obnoxious* and *semi-desirable* have also been used for some of these facilities, but the undesirable features (perceived or real) of these facilities dominate the desirable ones. Since the analytical models used for locating these facilities do not change much with their degree of undesirability, as Erkut and Neuman (1989) suggested, we will use the term *undesirable* for all of them.

Since its inception, location theory has been dominated by models and methods for locating desirable facilities, such as warehouses, hospitals, and firehouses, which need to be placed close to the population centers receiving their services. This changed in the 1970s with the launching of undesirable facility location research. Several reasons are attributed to this late entry in the location literature, notably that most of the aforementioned undesirable facilities, such as airports, mega-stadia, and sewage, chemical, and nuclear plants, are the byproducts of the technological advances and industrialization of the second half of the twentieth century. In addition, both industrial waste and municipal waste increase with world population and economic development while the waste generated by some of these facilities is toxic and has to be disposed of safely.

E. Melachrinoudis (✉)
Department of Mechanical and Industrial Engineering, Northeastern University,
360 Huntington Avenue, Boston, MA 02115, USA
e-mail: emelas@coe.neu.edu

H. A. Eiselt, V. Marianov (eds.), *Foundations of Location Analysis,*
International Series in Operations Research & Management Science 155,
DOI 10.1007/978-1-4419-7572-0_10,

In the early 1970s, public environmental concerns triggered federal legislation in the United States, which, in turn, enhanced awareness of the potential hazards and generated a need for the systematic placement of these facilities to minimize their undesirable effects. Prior to the 1970s, the protection of basic air and water supplies was a matter mainly left to each state. During the 1970s, responsibility for clean air and water shifted to the federal government. The Environmental Protection Agency was created in 1970, and during the ensuing decade several regulatory laws were passed to protect human health and the environment from potential hazards of pollution and waste disposal. These included the *Clean Air Act* of 1970 and the *Safe Drinking Water Act* of 1974 for enforcing clean air and drinking water standards, the *Resource Conservation and Recovery Act* of 1976 for regulating the disposal of solid and hazardous wastes, the *Clean Water Act* of 1977 for eliminating releases of toxic substances into the water, and the *Comprehensive Environmental Response, Compensation, and Liability Act* of 1980 for protecting people from abandoned heavily contaminated toxic waste sites.

A suitable location objective for locating an undesirable facility is the maximization of some increasing function of the distance between the facility and the affected customers. Analogous to the *minisum* and *minimax* objectives, most popular for locating desirable facilities, the *maxisum* and *maximin* objectives are established for locating undesirable facilities; see, e.g., Eiselt and Laporte (1995). The maxisum objective maximizes the sum of distances (or average distance) between the facility and the customers, while the maximin objective maximizes the distance between the facility and the closest customer to it. Sometimes weights are assigned to customers to represent the relative incompatibility between a customer and the facility and *weighted distances* are used. The objectives for undesirable facilities are frequently referred to as *push objectives*, since they push the undesirable facility away from the customers, while the objectives for desirable (attractive) facilities are referred to as *pull objectives*, since they pull the facility closer to the customers. To avoid pushing the undesirable facility to an infinite distance from the customers, which does not make sense in a real life problem, the objectives for undesirable facilities have to be optimized within a bounded region, a distinct difference from the desirable facilities objectives. In addition, the optimization models for undesirable facility location are more difficult to solve. Unlike the desirable facility location models, undesirable facility location models are nonconvex, typically having many local optima.

Although Goldman and Dearing (1975) are credited with first discussing the concept of optimally locating "semi-desirable" or "partially noxious facilities" in a conference paper, Church and Garfinkel's (1978) paper was the first published work on undesirable facility location. Their paper dealt with the maxisum location problem on a general network: they found a point of the network such that the sum of weighted shortest path distances from the nodes is maximized. They reduced the network solution space to a finite set of candidate points for optimality, consisting of the set of *bottleneck points* of the network and the leaf nodes of the network. Church and Garfinkel first showed that the maxisum objective renders

a nonconvex problem having many local optima, and that Hakimi's principle of optimality at a node does not hold. Exploiting the structure of the problem and utilizing bounds, their algorithm found the global optimum by partially enumerating local maxima. Their algorithm was adapted later for other undesirable facility location problems. This pioneering work stimulated a large body of research in undesirable facility location that complemented the desirable facility location literature.

The maximin location problem first appeared in the works of Shamos (1975) and Shamos and Hoey (1975) who studied the complexity of several fundamental problems in computational geometry. One of these problems is finding the *largest empty circle* of a given set of points in the plane, i.e., the circle that contains no points of the set, yet whose center is in the convex hull of the given points. The center of that circle is the maximin point as it maximizes the Euclidean distance to the closest point in the set. The solution is found by constructing the Voronoi diagram for the set of points. The first papers on the maximin location problem were published five years later by Dasarathy and White (1980) and Drezner and Wesolowsky (1980).

Building on their earlier work on pattern recognition, Dasarathy and White (1980) first formulated and solved the maximin problem with Euclidean distances for a feasible region, which is a convex polyhedron in k-dimensional space. They delineated their general algorithm for a 3-dimensional space. For the 2-dimensional space, they expanded Shamos and Hoey's Voronoi construction to account for optimality at the boundary of the feasible region. Their principal contributions are the characterization of the problem as nonlinear and nonconvex, the establishment of the properties of local optima using the Karush-Kuhn-Tucker optimality conditions, and the development of a general algorithm for solving the problem.

Drezner and Wesolowsky (1980) considered the same problem but with customer weights and a convex planar region defined by maximum distance constraints, one for each point (customer). Equivalently, the customers want the facility as far away as possible but within certain distance from them, which in turn signifies the semi-obnoxious character of the facility. Their optimization procedure was different from the one in Dasarathy and White (1980). They used a bisection search based on a graphical approach to approximate the optimal solution.

The above classical contributions, which cast the foundation of undesirable facility location theory during the late 1970s, are presented in this chapter. The detail and illustrative examples are helpful to introduce a beginner into the basic concepts of undesirable facility location research but also there is sufficient depth for the veteran researcher in the field to review and appreciate the classical contributions. An effort has been made to include major theoretical results, the thought process of the authors at the time, and the impact their work had on location literature. Although this is not a survey paper, major works that followed the classical contributions are surveyed.

This chapter is organized as follows: The classical contributions are presented in Sect. 10.2 and their impact is assessed in Sect. 10.3. The chapter ends with a summary and outlook of undesirable facility location research in Sect. 10.4.

10.2 The Classical Contributions

The classical contributions on the location of undesirable facilities can be classified according to the objective functions used in the respective optimization problems: *maxisum* and *maximin*. In an effort to minimize the adverse effects of the facility to be located, both of these objectives maximize some increasing function of the distance between the facility and the affected customers, namely the sum of distances and the minimum distance. The original papers appeared within a five-year period in the second half of the 1970s. We first present the classical contribution of Church and Garfinkel (1978) that utilizes the maxisum objective on a network. This work is followed by contributions that consider the maximin objective in continuous space: Shamos (1975) and Shamos and Hoey (1975), Dasarathy and White (1980), and Drezner and Wesolowsky (1980).

10.2.1 The Maxisum Problem on a Network: Church and Garfinkel (1978)

Let $G = (N, A)$ be a connected and undirected graph with no loops or multiple arcs, where N is the set of n nodes and A is the set of m arcs. The nodes represent customers that exhibit some adverse interaction with the new facility. We want to find a point x, $x \in G$, for locating the facility that maximizes

$$T(x) = \sum_{i \in N} w_i d(i, x), \tag{10.1}$$

where $w_i \geq 0$ is the weight of node i and $d(i, x)$ is the length of the shortest path between node i and $x \in G$.

Church and Garfinkel first formulated the above problem and named it *maxian* as it is identical to the *median* problem except that the objective is maximizing instead of minimizing. Thus, the solutions of these two problems find the two extreme values of $T(x)$. As they remark, this may help in evaluating how bad a given solution is with respect to any one of the two objectives.

Whereas the median problem attempts to find a location that is close to a given set of points, the maxian problem attempts to find a location that is as far as possible from these points. It should be noted that in the maxian problem, which is more often called maxisum problem, the type of facility to be located is not as important as is the adverse interaction between the given points and the new facility. In fact, Church and Garfinkel gave as an example of application the location of a house or a business—by no means undesirable—in a city among pockets of high crime incidence concentrated at the nodes of the network. The interaction between the nodes and the facility results in danger or potential harm to the facility that decreases with its distance from the nodes. In this example, the weight w_i represents the relative danger of node i to the facility.

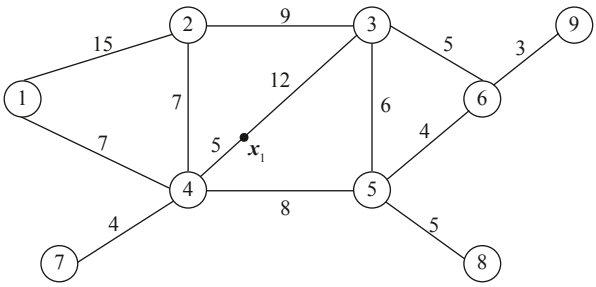


Fig. 10.1 A nine node network

For median problems, Hakimi (1964) proved that there exists a node which is optimal. This suggests a straightforward procedure for finding the optimal solution: evaluate objective function (10.1) at all nodes of the network and select the node(s) with the minimum value. Changing the optimization operator to "max" results in a surprisingly more complicated problem, in which the optimal point cannot be only at the nodes but also on the arcs of the network. Moreover, the maxisum is a non-convex problem, thus possessing many local optima. One of the key results of this work is that the search for the optimal solution is reduced from the infinite number of points of network G to a finite set of candidate points, often referred to as a *Finite Dominating Set* of points (*FDS*). After a brief notation, the points of interest are introduced below and optimality properties are established.

An interior point x on an arc (i, j) divides it into two arc segments (i, x) and (j, x). Denote the lengths of the two segments $c(i, x)$ and $c(j, x)$, respectively, and denote point x by the arc (i, j) is on and its distance from node i: $[(i, j); c(i, x)]$. For example, in Fig. 10.1 (slightly modified from Church and Garfinkel 1978), $x_1 = [(3, 4); 12)]$.

It is assumed that the shortest path distances between nodes are known. Ahuja et al. (1993) demonstrated that they can be effectively computed by several algorithms. Table 10.1 contains the shortest path distances between the nodes of the above network, $d(i, k)$, and the sum of distances from a node k to all nodes,

$$T(k) = \sum_{i \in N} d(i, k).$$

Let x be an interior point of arc (i, j). If there exists a node k such that

$$d(k, i) + c(i, x) = d(k, j) + c(j, x), \tag{10.2}$$

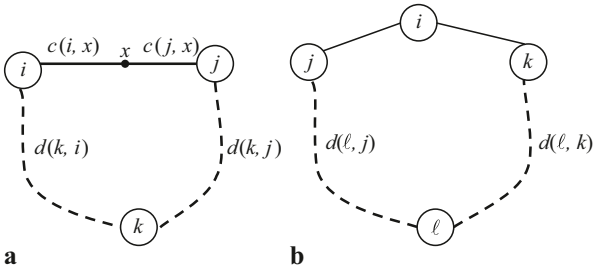
x is called *arc bottleneck point* with respect to node k and $B_A(k)$ denotes the set of bottleneck points generated by node k on A. This is illustrated in Fig. 10.2a, where the shortest paths from node k to i and j are shown as broken lines. Since $c(i, x)$, $c(j, x) > 0$ and $c(i, x) + c(j, x) = c(i, j)$, it follows from (10.2) that arc (i, j) contains an arc bottleneck point with respect to node k if and only if $|d(k, i) - d(k, j)| < c(i, j)$. By letting

$$p(k) = d(i, k) - d(j, k), \tag{10.3}$$

Table 10.1 Node-to-node shortest path distance matrix $d(i, k)$ and $T(k)$

Node i	Node k									$T(k)$
	1	2	3	4	5	6	7	8	9	
1	0	14	21	7	15	19	11	20	22	129
2	14	0	9	7	15	14	11	20	17	107
3	21	9	0	14	6	5	18	11	8	92
4	7	7	14	0	8	12	4	13	15	80
5	15	15	6	8	0	4	12	5	7	72
6	19	14	5	12	4	0	16	9	3	82
7	11	11	18	4	12	16	0	17	19	108
8	20	20	11	13	5	9	17	0	12	107
9	22	17	8	15	7	3	19	12	0	103

Fig. 10.2 Bottleneck points



the condition for arc (i, j) to contain an arc bottleneck point generated by node k can be written as

$$|p(k)| < c(i, j). \tag{10.4}$$

Note that $p(k)$ measures how much farther away node i is from k than from j. The range of values of $p(k)$ can be found by the triangle inequality for shortest path distances. Since $d(i, k) \le d(i, j) + d(j, k)$ and $d(j, k) \le d(j, i) + d(i, k)$, substituting in (10.3) we obtain $-d(i, j) \le p(k) \le d(i, j)$, or

$$|p(k)| \le d(i, j). \tag{10.5}$$

If $d(i, j) < c(i, j)$, (10.4) and (10.5) imply that every node $k \in N$ has an arc bottleneck point on arc (i, j). This is later illustrated for arc (1, 2) of Fig. 10.1.

Inequality (10.4) implies that no shortest path from k to i or from k to j contains arc (i, j). Clearly, a bottleneck point on arc (i, j), with respect to node k, is associated with a cycle formed by the shortest path from node k to node i, arc (i, j), and the shortest path from node j back to node k, as shown in Fig. 10.2a. The bottleneck point is the point in a cycle that is the farthest away from node k. For example, $x_1 \in B_A(6)$ in Fig. 10.1 as node 6 and point x_1 are the endpoints of two equidistant paths, $\{(6, 3), (3, x_1)\}$ and $\{(x_1, 4), (4, 5), (5, 6)\}$, and inequality (10.4) is satisfied as $|5 - 12| < 17$. By considering all cycles that contain node 6, in Fig. 10.1, and identifying arcs on those cycles containing bottleneck points using inequality

(10.4), the complete set $B_A(6)$ can be derived: $B_A(6) = \{[(3, 5); 2.5], [(3, 4); 12], [(2, 4); 2.5], [(1, 2); 5]\}$. Since there is a unique path between leaf node 9 and node 6, $B_A(9) = B_A(6)$. Note that it is possible for arc (i, j) to contain a bottleneck point with respect to each one of its vertices i and j. This happens if and only if $d(i,j) < c(i,j)$, i.e., the shortest path between the end nodes of an arc is not that same arc. In Fig. 10.1, arc (1, 2) satisfies that condition, as $d(1, 2) = 14 < c(1, 2) = 15$ and the associated cycle is $\{(1, 2), (2, 4), (4, 1)\}$. Therefore, point [(1, 2); 14.5] is in $B_A(1)$ and point [(1, 2); 0.5] is in $B_A(2)$.

Bottleneck points can also appear on nodes. If there exist distinct arcs (i, j) and (i, k) incident to node i and a node $\ell \neq i$ such that $d(\ell, j) + c(j, i) = d(\ell, k) + c(k, i)$, then node i is a *node bottleneck point* with respect to node ℓ, denoted by $i \in B_N(\ell)$, and illustrated in Fig. 10.2b. For example, node 2 is in $B_N(5)$ in Fig. 10.1. There are two equidistant paths from node 5 to node 2, one containing arc (2, 3) and the other (2, 4), whose union forms a cycle. Each arc or node bottleneck point is associated with a cycle in G that contains the point. Conversely, Church and Garfinkel show that every cycle in G contains a bottleneck point with respect to every node in the cycle. This result suggests a method for finding all bottleneck points of a network: find all cycles in G and for every node in a cycle identify the corresponding bottleneck point.

Let $B_A = \bigcup_{k \in N} B_A(k)$ denote the set of all arc bottleneck points, $B_N = \bigcup_{k \in N} B_N(k)$ the set of all node bottleneck points, and $B = B_A \cup B_N$ the set of all bottleneck points of G. Let D denote the set of dangling (leaf) nodes of G. In the network of Fig. 10.1, for example, $D = \{7, 8, 9\}$. Since bottleneck points are defined by cycles, D and B have no elements in common. The following theorem reduces the solution space from an infinite set (network G) to a finite set of candidate points for optimality consisting of the set of leaf nodes and the set of bottleneck points of G.

Theorem 1: *There exists a point* $\hat{x}$ *which maximizes* (10.1) *such that* $\hat{x} \in X^* = D \cup B$.

Proof: Church and Garfinkel show that for every point $x \in G$, $x \notin D \cup B$, there exists an $x \in D \cup B$ with a better objective value. Consider first an interior point x of arc (i, j), $x \notin B_A$. Then, within the ε-neighborhood of x, the sum in (10.1) can be decomposed into two, one over nodes $k \in N_i(x)$ and one over $k \in N_j(x)$, where $N_i(x)$ and $N_j(x)$ are nodes $k \in N$ whose shortest path to x includes segment (i, x) and (j, x), respectively:

$$T(x) = \sum_{k \in N_i(x)} w_k d(k, x) + \sum_{k \in N_j(x)} w_k d(k, x).$$

For a point x_e in the interior of arc (i, j), such that $c(j, x_e) = c(j, x) + \varepsilon$, for $\varepsilon > 0$ and infinitesimal, $N_i(x)$ and $N_j(x)$ remain unchanged and therefore,

$$T(x_e) - T(x) = \left[\sum_{k \in N_j(x)} w_k - \sum_{k \in N_i(x)} w_k \right] \varepsilon = q(x)\varepsilon.$$

Assuming without loss of generality that $q(x) \geq 0$, it follows that $T(x_e) - T(x)$ increases with ε until x_e reaches an arc bottleneck point or node i.

Consider now a point x that is on a node and $x \notin D \cup B_N$. Similarly, in this case a path from x can be found on G of increasing objective value until a point of $D \cup B$ is encountered. □

On a given arc $(i, j) \in G$, there exist at most n bottleneck points identified by cycles containing arc (i, j) and each node $k \in G$. Therefore, there exist at most mn bottleneck points. Since the number of leaves in a network is at most n, the size of the set containing the optimal solution is $O(mn)$.

Since bottleneck points occur only on cycles of G and a tree network has no cycles, the following corollary follows from Theorem 1.

Corollary 1: *If the network is a tree, there exists an optimal point which is a leaf node.*

A straightforward approach for solving (10.1) is to find the best point on each edge, and then compare these points and select the optimal point in G.

The shortest path distance between a node $k \in N$ and a point $y \in (i, j)$, is $d(k, y) = \min\{d(i, k) + c(i, y), d(j, k) + c(j, y)\}$. It is maximized when $d(i, k) + c(i, y) = d(j, k) + c(j, y)$. Substituting $c(i, y) = c(i, j) - c(j, y)$, the point on arc (i, j) with the maximum distance from node k, denoted by $y(k)$, is at a distance from node j, $c(j, y) = \frac{1}{2}[d(i, k) - d(j, k) + c(i, j)]$. After it is simplified using (10.3), it becomes:

$$c(j, y) = \tfrac{1}{2}\left[p(k) + c(i, j)\right]. \tag{10.6}$$

In other words, for a given arc (i, j) the length of (j, y) is increasing with $p(k)$. The greater the value of $p(k)$, the further $y(k)$ is from node j. If $p(k) = c(i, j)$, $y(k) = i$, while if $p(k) = -c(i, j)$, $y(k) = j$. Therefore, if we reorder nodes $k \in N$ in increasing magnitude of $p(k)$ they will map to $y(k)$ points in the same order on arc (j, i), according to (10.6). To reorder the nodes $k \in N$ for arc (i, j) we re-index them by $r(k)$ in terms of increasing $p(k)$, i.e., $r(k_2) > r(k_1) \rightarrow p(k_2) \geq p(k_1)$. Clearly, $p(i) = -d(i, j)$ and $p(j) = d(i, j)$, and we can let $r(i) = 1$ and $r(j) = n$.

Table 10.2 contains $p(k)$ and $r(k)$, $k \in N$, for arc $(i, j) = (1, 2)$ of Fig. 10.1. The distance of point $y(k)$ from node 2, $c(2, y)$, is also computed according to (10.6). Note that $d(1, 2) < c(1, 2)$ and therefore $|p(k)| < c(1, 2) = 15$ for every $k \in N$. Based on an earlier observation, every node has a bottleneck point on arc (1, 2) although not all of them are distinct.

We want to express the objective function (10.1) at some point $y \in (i, j)$. Consider two consecutive $y(k_1)$ and $y(k_2)$ points, i.e., $r(k_1) = t$ and $r(k_2) = t + 1$, $t = 0, \ldots, n$,

Table 10.2 Bottleneck points on arc (1, 2)

k	1	2	3	4	5	6	7	8	9
$p(k)$	-14	14	12	0	0	5	0	0	5
$r(k)$	1	9	8	2	3	6	4	5	7
$c(2, y)$	.5	14.5	13.5	7.5	7.5	10	7.5	7.5	10
$T(y(k))$	111.5	133.5	140.5	160.5	160.5	158	160.5	160.5	158

where $r(k) = 0$ and $r(k) = n + 1$ are associated with node j and i, respectively. For notational simplicity, use $y = c(j, y)$. In other words, y represents both a point $y \in (i,j)$ and its distance from j on arc (i, j). For $y \in (y(k_1), y(k_2))$,

$$d(k,y) = \begin{cases} d(k,i) + c(i,j) - y, & \text{for } \{k | r(k) \le t\} \\ d(k,j) + y, & \text{for } \{k | r(k) \ge t+1\} \end{cases} \tag{10.7}$$

and the objective function $T(y)$ can now be expressed as

$$T(y) = \sum_{k|r(k)\ge t+1} w_k d(k,j) + \sum_{k|r(k)\le t} w_k [d(k,i) + c(i,j)] + W(t)y, \tag{10.8}$$

where

$$W(t) = \sum_{k|r(k)\ge t+1} w_k - \sum_{k|r(k)\le t} w_k, \quad t = 0, \ldots, n \tag{10.9}$$

is the gradient of $T(y)$. When $y(k_1)$ and $y(k_2)$ are distinct points, $T(y)$ is a line segment with slope $W(t)$. When $y(k_1) = y(k_2)$, the line segment becomes a degenerate point. The number of different line segments of $T(y)$ for $y \in (i, j)$ depends on the number of distinct numbers $y(k)$. It can be as low as 1 (whole arc), when $|p(k)| = c(i, j)$ for all $k \in N$, and as high as $n + 1$, when all values $y(k)$ are distinct bottleneck points. The former happens for arc (4, 5) of Fig. 10.1. Clearly, the slope $W(t)$ is nonincreasing with increasing t, as one scans consecutive arc segments $(y(k_1), y(k_2))$ from node j to node i on arc (j, i). Since it is also continuous, $T(y)$ is piecewise linear and concave, and the theorem below follows.

Theorem 2: *A best point* $y^*(k)$ *on arc* (i, j) *satisfies* $r^*(k) = \min\{r(k) | W(r(k)) \le 0, r(k) = 1, \ldots, n\}$, *and its distance from node j is given by* (10.6).

It is possible that $W(r(k)) = 0$ at the best point $y^*(k)$. Then every point on the arc segment between $y(k^*)$ and $\{y(\ell) | r(\ell) = r^*(k) + 1\}$ maximizes $T(y)$ on arc (i, j).

Although Church and Garfinkel allude to the concavity property of the objective function on an arc, they do not explicitly state it. To illustrate the "piecewise linear and concave" property of the objective function $T(y)$ over an arc, objective values at all $y(k) \in (2, 1)$ are calculated using (10.8) and (10.9). Equal weights, $w_k = 1$, $k \in N$, are assumed. The objective values $T(y(k))$ are displayed in the last row of Table 10.2. In Fig. 10.3, $T(y)$ is plotted for arc (1, 2) using the data of Table 10.2. Point $y = 0$ and point $y = 15$ correspond to nodes 2 and 1, respectively. The objective value at node 2 and 1 is 107 and 129, respectively, taken from the last column of Table 10.1.

The *unweighted maxian problem* is the maxian problem in which all weights w_i are equal to 1. As is shown below, the procedure for finding the best point on an arc can be greatly simplified when weights are equal.

Note in (10.9) that as point y on arc (i, j) moves from one interval $[y(k_1), y(k_2)]$ to the next $[y(k_2), y(k_3)]$, such that $r(k_1) = t$, $r(k_2) = t + 1$ and $r(k_3) = t + 2$, the gradient

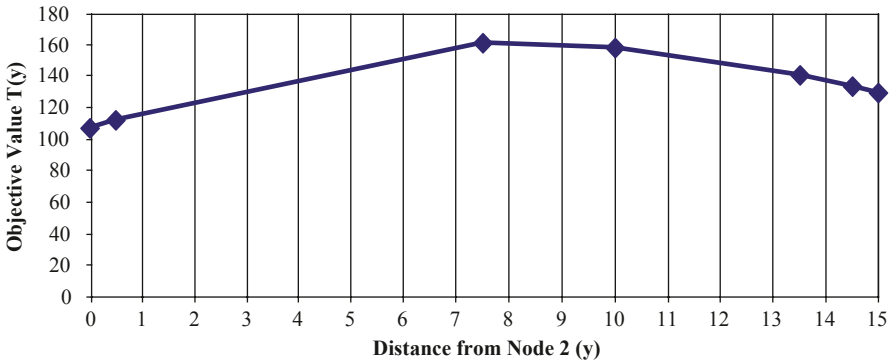


Fig. 10.3 Objective function value over arc (2, 1)

$W(t)$ decreases by $2w_{k_2}$. If all weights are equal to 1 the gradient decreases by 2. For our 9-node example, $W(t) = 9 - 2t$, $t = 0, \ldots, 9$. The following corollary simplifies the procedure for finding the best point on an arc.

Corollary 2: *A best point* $y^*(k)$ *on arc* (i, j) *satisfies*

$$r^*(k) = \begin{cases} \frac{1}{2}n, & \text{if } n \text{ is even} \\ \frac{1}{2}(n+1), & \text{if } n \text{ is odd} \end{cases}$$

and its distance from node j is given by (10.6).

The best point on an arc can be found in $O(n \log n)$ time by sorting the n points with respect to increasing values of $p(k)$. Therefore, the total time required to find the optimal maxisum point in an unweighted network is $O(mn \log n)$.

Consider again arc (1, 2) of the network of Fig. 10.1. Since n is odd, the best point $y^*(k)$ on arc (1, 2) is associated with $r^*(k) = \frac{1}{2}(n+1) = 5$. From Table 10.2 we find that the arc bottleneck point $y^*(k)$ is generated by node $k = 8$ and is at distance $y = 7.5$ units from node 2 (see also Fig. 10.3). Note that, in addition to node 8, the best point on arc (1, 2) is the bottleneck point of nodes 4, 5 and 7.

Bounds of the objective function over an arc can be found as follows. Since $T(y)$ is concave over each arc (i, j), its minimum occurs at an endpoint, i.e., at one of the two nodes, i or j, or both. Therefore, a lower bound of $T(y)$ over arc (i, j), $\underline{T}(i, j)$, is specified in the following relation

$$\underline{T}(i, j) = \min \left\{ \sum_{k \in N} w_k d(k, i), \sum_{k \in N} w_k d(k, j) \right\}. \tag{10.10}$$

A good lower bound of the optimal objective value over G is $\underline{T}$. It is obtained by comparing the objective values of the nodes of G:

$$\underline{T} = \max_{i \in N} \sum_{k \in N} w_k d(k, i) \tag{10.11}$$

To find an upper bound of $T(y)$ over an arc (i, j), $\overline{T}(i, j)$, we consider the upper bound of $d(k, y), y \in (i, j)$. The maximum distance point on arc (i, j) from node k is point $y(k)$, found earlier. Its distance from node k is $d(k, j) + c(j, y)$, where $c(j, y)$ is given by (10.6), such that

$$d(k, y) \leq \tfrac{1}{2}[c(i, j) + d(k, i) + d(k, j)].$$

Multiplying both sides of this expression by w_k, and taking the summation for all $k \in N$, we obtain

$$\begin{aligned} T(y) \leq \overline{T}(i, j) &= \tfrac{1}{2}\left[c(i, j)\sum_{k\in N} w_k + \sum_{k\in N} w_k d(k, i) + \sum_{k\in N} w_k d(k, j)\right] \\ &= \tfrac{1}{2}\left[c(i, j)\sum_{k\in N} w_k + T(i) + T(j)\right]. \end{aligned} \tag{10.12}$$

For the unweighted maxian problem (with $w_k = 1, k \in N$), the upper bound reduces to

$$\overline{T}(i, j) = \tfrac{1}{2}\left[nc(i, j) + T(i) + T(j)\right]. \tag{10.13}$$

The search for the optimal solution starts with the node $\hat{x}$ associated with $\underline{T}$. This is the incumbent solution. Upper bounds on all arcs, $\overline{T}(i,j)$, are used to eliminate at the outset as many arcs as possible. After finding the best point on an arc, the lower bound is updated and is used to eliminate additional arcs. When all remaining arcs have been considered, the incumbent is the optimal solution.

Algorithm 1: Maxisum Problem on a Network

Step 1: Compute a lower bound on $T(x^*)$, $\underline{T}$, using (10.11) and identify the point $\hat{x} \in G$ at which it occurs.

Step 2: Compute the upper bounds on all arcs $(i, j) \in A$, $\overline{T}(i, j)$, using (10.12).

Step 3: Eliminate every arc $(i, j) \in A$ for which $\overline{T}(i, j) < \underline{T}$ from further consideration. Let the set of remaining arcs be A'.

Step 4: Until $A' = \varnothing$, repeat.

Step 4.1: Let the arc $(i, j) \in A'$ with the largest $\overline{T}(i, j)$ be (u, v). Find the best point on arc (u, v) and denote it by y^* and its objective value by $T(y^*)$. Set $A' \leftarrow A' - (u, v)$.

Step 4.2: If $T(y^*) > \underline{T}$, update $\underline{T} \leftarrow T(y^*)$ and $\hat{x} \leftarrow y^*$, and eliminate remaining arcs $(i, j) \in A'$ for which $\overline{T}(i, j) < \underline{T}$.

Step 5: The optimal solution is $x^* = \hat{x}$ and $T(x^*) = \underline{T}$.

Table 10.3 Objective value upper bounds on arcs

Arc (i,j)	(1, 2)	(1, 4)	(2, 3)	(2, 4)	(3, 4)	(3, 5)
$\overline{T}(i,j)$	185.5	136	140	125	162.5	109
Arc (i,j)	(3, 6)	(4, 5)	(4, 7)	(5, 6)	(5, 8)	(6, 9)
$\overline{T}(i,j)$	109.5	112	112	95	112	106

Table 10.4 Bottleneck points on arc (3, 4)

K	1	2	3	4	5	6	7	8	9
$p(k)$	14	2	−14	14	−2	−7	14	−2	−7
$r(k)$	7	6	1	8	4	2	9	5	3
$c(4, y)$	15.5	9.5	1.5	15.5	7.5	5	15.5	7.5	5
$T(y(k))$								125.5	

Algorithm 1 is illustrated for the unweighted maxian of the network of Fig. 10.1. Node 1 provides a lower bound $\underline{T} = T(1) = 129$ from the last column of Table 10.1. Upper bounds on the arcs $\overline{T}(i,j)$ have been calculated according to (10.13) and are shown in Table 10.3. According to Step 3 of the algorithm, all arcs except $A' = \{(1, 2), (1, 4), (2, 3), (3, 4)\}$ are eliminated. Of the remaining, arc (1, 2) is selected for its largest upper bound (185.5). The best point on arc (1, 2) was found earlier at its midpoint with objective value 160.5. The new $\underline{T} = 160.5$ allows us to eliminate all remaining arcs except (3, 4).

To find the best point on arc (3, 4) we construct Table 10.4, from which $r^*(k) = 5$ occurs for $k = 8$. The best point is [(4, 3); 7.5] with objective value 125.5, calculated directly from (10.8) and (10.9). This is inferior to the incumbent point, which is optimal because there are no other remaining arcs to consider. Therefore, the optimal point is [(1, 2); 7.5] with objective value 160.5.

10.2.2 The Maximin Location Problem in Continuous Space with Euclidean Distances

Analogous to *minimax* problem, the *maximin* objective attempts to find a location for an undesirable facility that minimizes the adverse impact on the most affected customer, which is the one closest to the facility. The maximin objective was first used in continuous space with Euclidean distances. Three original contributions for undesirable facility location are analyzed in this subsection. The earliest work by Shamos (1975) and Shamos and Hoey (1975) characterized the unweighted maximin problem in the plane as an interesting problem of computational geometry, whose solution is the byproduct of the construction of a Voronoi diagram. Although they suggested the use of the maximin objective for undesirable facility location, Dasarathy and White (1980) and Drezner and Wesolowsky (1980) first formulated the maximin problem with practical feasible regions making it a suitable location model for undesirable facilities.

10.2.2.1 Shamos (1975) and Shamos and Hoey (1975): The Origins of the Maximin Problem

The maximin location problem first appeared in the works of Shamos and Hoey, who studied the complexity of several fundamental problems in computational geometry. The maximin problem is stated as follows: Given a set N of n points a_i in $\mathbb{R}^2$, find the largest empty circle that contains no points of the set yet whose center x is interior to the convex hull of the points, $CH(N)$. Equivalently,

$$\max_{x \in CH(N)} \min_{i \in N} d(i, x), \tag{10.14}$$

where $d(i, x) = \|a_i - x\|$ is the Euclidean distance between point a_i and x. The center of such a circle is the unweighted maximin point. Since such a point is farthest away from the closest customer point, it is suitable for the location of an undesirable facility, such as a source of pollution. For the same reason, the maximin point is suitable for locating a new business—albeit a desirable facility—that does not wish to compete for territory with established outlets represented by existing points. The solution point is restricted to a bounded feasible region, $CH(N)$, because otherwise it is going to be at an infinite distance from the customers. Moreover, Shamos and Hoey characterized the new problem as the dual of the (unweighted) minimax problem, posed much earlier by Sylvester (1857), which found the smallest circle enclosing all points of set N. The minimax objective was thoroughly investigated during the 1970s as an alternative to the minisum objective for the location of "desirable" facilities. Shamos and Hoey solved the maximin problem by constructing the Voronoi diagram of the a_i points.

Associated with each point a_i, $1 \leq i \leq n$, there exists a polygon V_i, called a Voronoi polygon, with the following property: if $x \in V_i$, then $\|a_i - x\| \leq \|a_j - x\|$, $1 \leq j \leq n$. The polygon V_i is the intersection of halfplanes containing a_i, where the halfplanes are determined by the perpendicular bisectors of the line segments joining a_i and a_j, $j \neq i$. The edges of the Voronoi polygons, some of which are unbounded halflines, are called Voronoi edges and their vertices Voronoi vertices. A Voronoi vertex is the common point of at least three Voronoi polygons, i.e., is equidistant from at least three a_i points. The circle drawn with its center at a Voronoi vertex and its radius the distance to its equidistant points contains no a_i points in its interior; it is an empty circle. The Voronoi vertex associated with the largest empty circle is the optimal solution to (10.14). The interior points of a Voronoi edge are equidistant from exactly two (neighboring) points a_i. The union of the boundaries of the Voronoi polygons is called a Voronoi diagram. The union of the Voronoi diagram and the interior sets of all Voronoi polygons constitute $\mathbb{R}^2$. The Voronoi diagram uses all relevant proximity information and is constructed very efficiently in $O(n \log n)$ time. The maximin problem in one dimension reduces to finding a pair of two consecutive points on a line that are farthest apart. Shamos and Hoey observe that this problem is also solved in $O(n \log n)$ time.

10.2.2.2 Dasarathy and White (1980): The Unweighted Maximin Problem in a Bounded Convex Region

Dasarathy and White (1980) first formulated the unweighted maximin problem for a general feasible region S that is a bounded and convex polyhedron in $\mathbb{R}^k$,

$$\max_{x \in S} \min_{i \in N} d(i, x), \tag{10.15}$$

where $d(i,x) = \|a_i - x\|$ is the Euclidean distance between point a_i and x in $\mathbb{R}^k$ and S is described by a set of m linear constraints, so that $S = \{x | c_j x \le b_j, 1 \le j \le m\}$.

The authors described a number of applications of this problem, not necessarily all in location. Viewing (10.15) as the problem of finding the largest hypersphere centered in S, whose interior is free of points a_i, they put forward some applications in information theory and in pattern recognition. It appears that these applications influenced the authors to cast the maximin problem in a higher dimensional space and not in the 2-dimensional space where most location applications are found. Needless to say, the location application of the maximin problem (10.15) had the greatest impact in future undesirable location literature. If the a_i points represent n cities in a region S and a highly polluting industry is to be located within S, the maximin problem will find its location such that the amount of pollutants reaching any city is minimized. It is assumed that the pollutant dispersion is uniform in all directions and the amount of pollutants reaching each city is a monotonically decreasing function of the distance between the city and that industry. Modeling the spread of pollutants in conjunction with the facility that generates them was studied later by Karkazis and Papadimitriou (1992) and Karkazis and Boffey (1994). Note that unlike the maxisum objective, which attempts to minimize the unpleasant collective impact to all customers, the maximin objective attempts to minimize the impact to the most adversely affected customer, making it an equity measure relative to that customer.

Dasarathy and White view the maximin problem also as a covering problem. Consider, for example, the a_i points being the locations of n radar stations and the convex set S the region monitored by these stations. Then (10.15) finds the minimum (of the maximum) power required by the stations such that each point in S is monitored by one or more of the stations. It is assumed that the required power of a station is a monotonically increasing function of the distance over which it can receive or send signals.

Letting z represent the square of the objective function in (10.15), the maximin problem can be converted to a standard nonlinear programming formulation:

$$\text{Max } z \tag{10.16}$$

$$\text{s.t. } z - \|a_i - x\|^2 \le 0, \quad 1 \le i \le n \tag{10.17}$$

$$c_j x \le b_j, \quad 1 \le j \le m \tag{10.18}$$

The above problem described by (10.16)–(10.18) is clearly not a convex programming problem due to constraints (10.17). Therefore, it may have several local optima one has to enumerate explicitly or implicitly to find the global optimum. The properties of a local optimum can be explored by constructing the necessary Karush-Kuhn-Tucker conditions for a local optimum at (x^*, z^*). Let the Lagrangian multipliers for constraints (10.17) and (10.18) be $v_i^* \geq 0$, $1 \leq i \leq n$, and $u_j^* \geq 0$, $1 \leq j \leq m$, respectively. Then, in addition to the feasibility conditions (10.17) and (10.18) the following conditions should be satisfied at (x^*, z^*):

$$\sum_{i=1}^{n} v_i^* = 1, \tag{10.19}$$

$$\text{s.t.} \sum_{i=1}^{n} 2v_i^*(a_i - x^*) - \sum_{j=1}^{m} u_j^* c_j = 0, \tag{10.20}$$

$$v_i^* \left(\left\| a_i - x^* \right\|^2 - z^* \right) = 0, \quad 1 \leq i \leq n, \tag{10.21}$$

$$u_j^*(c_j x^* - b_j) = 0, \quad 1 \leq j \leq m. \tag{10.22}$$

In addition, a local optimum either lies on the boundary of the feasible region (Case b) or not (Case a). These two cases are analyzed below to reveal the properties of local optima of (10.15).

Case a: If x^* does not lie on the boundary of S, none of constraints (10.18) are binding at x^*, which in turn forces all $u_j^* = 0$ by (10.22). In that case, (10.19) and (10.20) indicate that x^* lies in the convex hull of the a_i points, CH(N). Furthermore, in expressing the convex combination, only the multipliers v_i^* that are associated with points that are equidistant from x^* need to be positive, due to (10.21). Equivalently, x^* can be expressed as a convex combination of the points a_i that lie on the surface of the optimal hypersphere. Since $k + 1$ or fewer points suffice to express the convex combination of more than k points in $\mathbb{R}^k$ (Caratheodory's theorem restated in Hadley 1964), a local optimum in CH(N) is equidistant from at least $k + 1$ points a_i.

Case b: If x^* lies on the boundary of S, one or more (ignoring degenerate cases, up to k) of constraints (10.18) are binding at x^*. Let d, $0 \leq d \leq k - 1$, be the dimension of the smallest facet $F, F \subset S,$ on which x^* lies. Assume now that at most d of constraints (10.17) are binding at (x^*, z^*), or equivalently, at most d of the points a_i are equidistant from x^*. Draw the projections of these equidistant points a_i on the affine space A of F (a hyperplane of the same dimension that includes F). Since the number of such projections on F results in at most d points in A, there exists a hyperplane H of A of a dimension lower than d that passes through them. If point x^* is moved away from H by an infinitesimal distance, still lying on F, the distance from the equidistant points increases and therefore the objective function z^* increases,

thus contradicting the optimality of (x^*, z^*). Therefore x^* should be equidistant from at least $d + 1$ nearest a_i points.

The results of the above two cases are summarized in the following theorem:

Theorem 3: *The optimal solution x^* of the maximin problem* (10.15) *either lies on the boundary of the convex polyhedron S or in the convex hull of the a_i points CH(N). If it lies in CH(N), x^* is equidistant from at least* $(k + 1)$ *nearest a_i. If it lies on the boundary of S, x^* is equidistant from at least* $(d + 1)$ *nearest a_i, where d is the dimension of the smallest facet on which x^* lies.*

Similar to the case of the maxisum problem on a network as formulated in relation (10.1), the above theorem reduces the feasible region to a finite candidate set of solutions containing the optimal point of the maximin problem. These candidate points are either within $CH(N)$, analogous to Church and Garfinkel's bottleneck points, or remote points of the boundary of the feasible region, analogous to the leaves of a network.

The above theorem suggests a method for identifying candidate points on $CH(N)$ and on the boundary of S as follows:

1. The point that is equidistant from every combination of $k + 1$ points a_i is found and checked for feasibility using (10.17) and (10.18). Similarly, the center and radius of the hypersphere that passes through these $k + 1$ points is found and checked if the center is in S and there are no points in the interior of the hypersphere. $\binom{n}{k+1}$ combinations of points a_i are considered and for each one of them a system of k simultaneous linear equations is solved for the k components of x.
2. The point of each facet F of the boundary of S that is equidistant from every combination of $d + 1$ points a_i is found, where d is the dimensionality of F. For each facet F of dimensionality d, $\binom{n}{d+1}$ combinations of points a_i are considered, and for each one of them a system of k simultaneous linear equations are solved, of which d linear equations stipulate that x is equidistant from $d + 1$ points a_i and $k - d$ equations define the facet F.

As in the algorithm for the maxisum problem, bounds are used so that not all candidate points are explicitly generated. Dasarathy and White used a lower bound L and an upper bound U on the global z^* to eliminate facets from further consideration and to forgo the feasibility test if a generated point in $CH(N)$ has an objective value z that falls outside these bounds. As lower bound on the global value of the objective z^*, the objective value of the current best solution is used. A good initial lower bound L_0 can be obtained by evaluating all extreme points e_j of S, and at the same time taking care of the examination of 0-dimensionality facets:

$$L_0 = \max_j \min_i \left\| a_i - e_j \right\|^2. \tag{10.23}$$

Dasarathy and White computed an upper bound U on the global z^* by maximizing the Lagrangian of the problem for any nonnegative multipliers v_i,

$$U = \max_{(x \in S, z)} \left\{ z + \sum_i v_i \left(\|a_i - x\|^2 - z \right) \right\}. \tag{10.24}$$

By letting $\sum v_i = 1$, they developed an efficient algorithm for computing U. They also used upper bounds on the objective value z on facets in an effort to eliminate them. A facet F is eliminated from further consideration if the square distances between some a_i and all the extreme points of F are smaller than the current best objective value L. Similarly, an upper bound on the objective value z on facet F is

$$\min_{i \in N} \left\{ U_i(F) = \max_{e_j \in F} \left\| a_i - e_j \right\|^2 \right\}. \tag{10.25}$$

Although Dasarathy and White provided an algorithm for a convex polyhedron S in three dimensions ($k = 3$), a general algorithm is presented below for any $k \geq 2$. This maximin algorithm below can be easily modified for other distance metrics and for weighted distances.

Algorithm 2: Maximin Problem in a Convex Polyhedron

Step 1: Find the lower bound L_0 with the corresponding extreme point e_l of S and the upper bound U on the global z^* using (10.23) and (10.24), respectively. $L \leftarrow L_0, x^* \leftarrow e_l$. ($L$ keeps track of the current best local optimum, $L \leq z^* \leq U$).

Step 2: (Search for the best local optimum interior to $CH(N)$.) Consider all combinations of the points a_i taken $k + 1$ at a time. For each one of these combinations find the point x that is equidistant from the points a_i and its square distance z. If $L < z \leq U$ and x is feasible $L \leftarrow z, x^* \leftarrow x$. Repeat Step 2 for all $\binom{n}{k+1}$ combinations of the a_i points.

Step 3: (Search for local optima on the boundary of S.) Set $d \leftarrow k - 1$.

Step 3.1: For every facet F of dimensionality d, repeat: If a point a_i exists, $1 \leq i \leq n$, such that $U_i(F) \leq L$, eliminate F from further consideration. If F is not eliminated, consider all combinations of the a_i points taken $d + 1$ at a time. For each one of these combinations, find the point $x \in A$ (A is the affine space of $F, F \subset A$) that is equidistant from the points a_i and its square distance z. If $L < z \leq U$ and $x \in F$, set $L \leftarrow z, x^* \leftarrow x$.

Step 3.2: Set $d \leftarrow d - 1$. If $d > 1$, go to Step 3.1.

Step 4: The optimal solution is (x^*, L).

Algorithm 2 assumes that the facet structure of S is known. For a 3-dimensional space, Dasarathy and White described an $O(m^2 \log m)$ algorithm to obtain the face structure. The worst-case complexity of Algorithm 2 is $O(n^{k+2})$. For a 2-dimensional feasible region, there is a lower worst-case complexity algorithm that utilizes the Voronoi diagram of the a_i points.

For $S \subset \mathbb{R}^2$ (i.e., $k = 2$), Algorithm 2 can be simplified as follows. Step 1 considers all the 0-dimensional facets of S and selects the best extreme point of S as a starting solution, and its z-value as the starting lower bound on the global z^*. Since the nearest a_i point to a vertex of a convex polygon having m edges can be determined in $O(\log^2 n)$ time after $O(n \log n)$ preprocessing (Shamos and Hoey 1975), Step 1 can be executed in $O(m \log^2 n + n \log n)$ time. The local optima sought in Step 2 are among the Voronoi vertices of the Voronoi diagram of the a_i points. If a Voronoi vertex is in S it is a local optimum. Shamos and Hoey (1975) provided an $O(\log m)$ algorithm for determining if a given point is within an m-edge polygon. The $O(n)$ vertices of the Voronoi diagram can be generated in $O(n \log n)$ time and tested for feasibility in $O(n \log m)$ time. In Step 3, only the edges of the polygon S ($d = 1$) have to be searched for local optima. The points of the edges that are equidistant from a_i points taken two at a time are the intersections of the Voronoi edges with the edges of S. A Voronoi edge can intersect the boundary of S at most twice. Using a binary search, Dasarathy and White found the intersections of the Voronoi edges with the edges of S in $O(n \log m)$ time.

In summary, the optimum point of the maximin problem in a convex polygon S can be a vertex of S, a Voronoi vertex, or an intersection of a Voronoi edge with an edge of S. The required effort to solve it is $O(m \log^2 n + n \log n + n \log m)$.

Figure 10.4 shows a set N of 10 points, $a_1, \ldots, a_{10}$, within a square region S. The boundary of $CH(N)$ is displayed by dotted line segments. The Voronoi diagram of the points consists of 10 vertices and 19 edges. There are 8 intersections of Voronoi edges with the edges of S. The solution to (10.14), i.e., the maximin point in $CH(N)$, is vertex v of the Voronoi diagram, which is the center of the largest empty circle, shown in Fig. 10.4. The extreme point e_4 is the maximin point in S, solution to (10.15), with objective value $\|a_7 - e_4\|$.

10.2.2.3 Drezner and Wesolowsky (1980): Weighted Maximin Problem with Maximum Distance Constraints

The difference in this contribution compared to the previous work is that positive weights w_i are assigned to customers and the solution method does not search the feasible region for local optima to find the best one(s), but instead progressively reduces the feasible space to trap the global optimum in an infinitesimal area. The feasible region is a convex bounded planar area defined by the intersection of circles, each having at their center a customer point a_i, and radius r_i representing the maximum distance the facility can be from customer i with $i \in N$. Clearly, the authors had in mind the location of a semi-obnoxious facility, which is pushed away by each customer to a different degree depending on its weight, but at the same time is wanted within certain distance from each customer. The problem can be formulated as

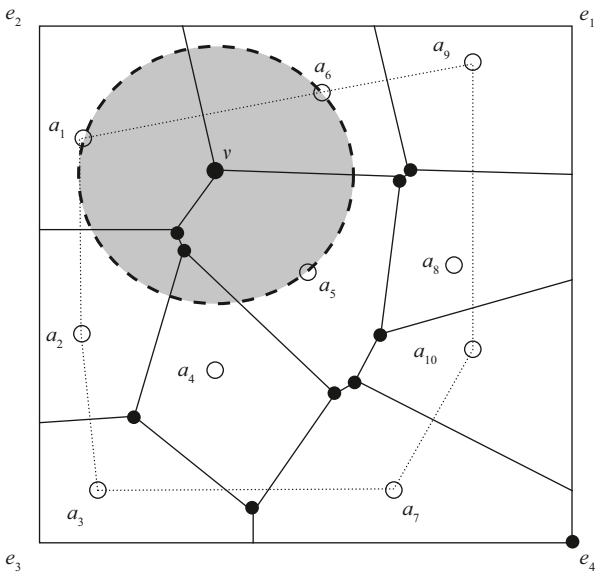


Fig. 10.4 Example problem using the Voronoi diagram

$$\max_{x \in S} \min_{i \in N} w_i d(i, x), \tag{10.26}$$

where $d(i,x) = \| a_i - x \|$ is the Euclidean distance between point a_i and $x \in \mathbb{R}^2$ and S is a set of n maximum distance constraints, $S = \{x | d(i,x) \le r_i, 1 \le i \le n\}$.

The solution methodology is graphical in nature and is speeded up by a bisection search. Consider some objective value z, $z = \min_{i \in N} w_i d(i,x)$. The points of the plane with better objective value than z are outside of the union of circles having centers a_i and radii z/w_i, or $C(z) = \{x | d(i,x) \ge \frac{z}{w_i}\}$. Starting with a relatively small value of z, one can solve the problem interactively by increasing z until the last point in S is covered, or $S \cap C(z)$ is an infinitesimally small area. In fact, Brady and Rosenthal (1980) used this interactive graphical approach on the computer to solve constrained minimax problems. Instead of an interactive approach, Drezner and Wesolowsky used an efficient bisection search as follows. At some iteration, let be the objective value of the best solution found so far (lower bound on z*) and be an upper bound on z*. A new objective value is generated and a procedure is used to find out if a point x exists in $S \cap C(z)$ with that z-value. If it does, $\leftarrow z$, otherwise, $\leftarrow z$. The iterations continue until (-) becomes smaller than a small preset constant. The solution x associated with is close to the best point x* within an approximation.

10.3 Impact of the Original Papers

The above classical contributions stimulated a large body of research in undesirable facility location that complemented the existing (desirable) location literature. Up to that time, pull objective location models, such as minisum (median) and minimax

(center), dominated the location literature. The introduction of the push objective location models leveled the field of location science and opened it to new methods, applications and location problems. An outstanding example has been the launch of a new class of location problems that utilize a combination of push and pull objectives to find locations that best trade off the conflicting objectives. This section describes the immediate impact of the original papers on the location literature in the period of 10–15 years that followed as well as the major works that were afterwards influenced by the classical works and contributed to the undesirable location literature.

10.3.1 The Impact of Church and Garfinkel's Contribution

The pioneering work of Church and Garfinkel initiated the field of undesirable facility location by introducing the *maxisum* location problem and distinguishing it from the existing (desirable) location problems of the time. As Goldman (2006) notes, such a "three letter change" (substitution of max for min) might seem innocuous, but in fact substantially increases the difficulty to carry out the optimization. Church and Garfinkel showed that the new problem is nonconvex and thus may have many local optima, so that it is necessary to generate all or at least a subset of them by improving bounds on the optimal objective value to find the global optimum. In fact, many algorithms that were developed later for variations of the maxisum and the maximin objectives resemble Church and Garfinkel's algorithm. Similarly, the "existence of a finite candidate set of points containing the global optimum" that resonates throughout the undesirable location literature originated in this paper. Among the points in that set are local optima that arise due to the nonconvex property of the undesirable facility location problem which, in turn, render the "Hakimi property" of a network invalid. Finally, new terms were coined in their paper to enrich the location lexicon: "obnoxious" and "semi-obnoxious" facilities and "bottleneck points" of a network. In the remainder of this subsection we will include early contributions that built on the work of Church and Garfinkel.

Ting (1984) dealt with the maxisum problem on trees and developed an $O(n)$ algorithm by using a special data structure. This is an improvement over Church and Garfinkel's $O(n^2)$ algorithm for trees. Minieka (1983) addressed the unweighted maxisum problem and essentially developed the same algorithm as Church and Garfinkel to find the *antimedian* of the network, as he named the solution of the maxisum problem. In the same paper, Minieka studied another max-type problem, $\max_{x \in G} \max_{i \in N} d(i, x)$, whose solution named the *anticenter* of the network.

Hansen et al. (1981) considered a more general maxisum problem on a continuous and bounded feasible region S with $S \in \mathbb{R}^2$. By modeling the nuisance from the obnoxious facility located at x to a population center i as decreasing and continuous function of their distance, $D_i[d(i, x)]$, they actually formulated a minisum model, named the *anti-Weber problem*, as the counterpart of the Weber problem with the objective

$$\min_{x \in S} T(x) = \sum_{i \in N} D_i[d(i,x)], \tag{10.27}$$

where $d(i, x)$ can be any distance metric, of which the most commonly used are the Euclidean, rectilinear and Tchebycheff metrics.

Similar to Church and Garfinkel's (1978) Theorem 1 above, Hansen et al. established a theorem that reduces the feasible region that contains the optimal location to the union of two sets. The first set, analogous to the set of bottleneck points of a network, consists of the points of S that are in the convex hull $CH(N)$ of the points, i.e., $S \cap CH(N)$. The second set, analogous to the leaf nodes of a network, consists of the points of $S - CH(N)$ that are remote from $CH(N)$. A point $y \in Y$ is said to be remote from set X if there exists $x \in X$ such that the straight halfline starting from x and passing through y contains no point of Y beyond y. The results of this theorem are used by Hansen et al. to rationalize the locational pattern of nuclear power plants in France. Some power plants are at interior locations while many others are located at the border of France with Germany and Belgium or on the Atlantic coast. They solved the anti-Weber problem by a branch-and-bound method, similar to the *Big Square-Small Square* algorithm developed earlier by the same authors for the generalized Weber problem.

For the special case where D_i is a linear function of distance $d(i, x)$, relation (10.27) reduces to the (ordinary) maxisum objective on the plane:

$$\max_{x \in S} T(x) = \sum_{i \in N} w_i d(i,x). \tag{10.28}$$

For the maxisum problem (10.28), Hansen et al. reduced the set containing the optimal solution even further by excluding all interior points of S:

Theorem 4: *For the maxisum problem in the plane there exists an extreme point of the convex hull of the feasible region S that is optimal.*

The proof follows directly from the convexity property of the objective function. When the feasible region is approximated by a nonconvex polygon S with m vertices, Melachrinoudis and Cullinane (1986a) described a simple $O(mn)$ algorithm for finding the weighted maxisum point by evaluating the vertices of $CH(S)$.

Theorem 4 states that the maxisum point is at remote points of the boundary of the feasible region. This result is analogous to Church and Garfinkel's result for trees where the optimal point is one of the leaves of the tree. The feasible region S therefore has to be bounded, otherwise the optimal solution of the maxisum problem is "out at infinity." It is even possible that the optimal location is at an existing facility point as Eiselt and Laporte (1995) illustrated in the following example. Consider the case of a square feasible region with four equally weighted customer points at the corners of the square. According to Theorem 4, the optimal maxisum locations are at the extreme points of the feasible region which coincide with the customer locations. Prescribing always a boundary solution and sometimes even a customer's location for the undesirable facility does not make the maxisum model very attractive for use in the plane. However, the maxisum objective is very useful

in a multiobjective setting when it is combined with a pull objective as described later in this section.

10.3.2 The Impact of the Original Maximin Location Papers

The original papers on the maximin location problem, directly or indirectly, had an impact on the undesirable facility location works that followed during the 1980s. A number of variations of the maximin problem with Euclidean distances have been solved using a solution approach similar to Dasarathy and White's for generating local optima by using the Karush-Kuhn-Tucker optimality conditions. Melachrinoudis (1985) and Melachrinoudis and Cullinane (1985) extend the weighted maximin problem to nonconvex regions and to regions that enclose forbidden areas, respectively. They provided an example for locating a toxic dump in the state of Massachusetts, which was represented by a nonconvex bounded planar region with forbidden areas for the facility around cities, wetlands, rivers, lakes, and ecosystems. The forbidden areas were approximated by the union of circles. Weights assigned to the customers, such as cities and towns, reflected the population size. The most important customer point, the city of Boston ($i = 6$), was assigned a weight of 1, while the weight of the population center i was calculated relative to Boston by the formula $w_i = (N_i/N_6)^{-1}$, where N_i denotes the population of city i.

Since it has not been elaborated in the location literature, it is important to note here that unlike the weight of a desirable facility, the maximin weight is a decreasing function of the degree of incompatibility between a customer and the facility. For example, consider the simple case of a one-dimensional feasible region in the interval [0, 1], with customer A at point 0 having weight 1 and customer B at point 1 having weight 3. The maximin point is at 0.75, meaning that the customer with the lower weight (A) pushes the facility further away than the customer with the higher weight. By the way, the minimax point happens to be the same in this very small example, thus the customer with the higher weight pulls the desirable facility closer to it.

To explain this counterintuitive property of the maximin weights, consider the generalization of the weighted maximin problem of (10.26), where $S \subset \mathbb{R}^k$, $k \geq 1$. Let the optimal point be x^* and the optimal objective value be z^*. Theorem 3, generalized for the weighted maximin problem, states that x^* is equidistant (in a weighted sense) from a subset of customers, N', and $|N'|$ depends on the dimensionality of S and on whether x^* lies in $CH(N)$ or on the boundary of S. Theorem 3 and (10.26) imply that $d(i, x) = z^*/w_i,\ i \in N'$, and $d(i, x) > z^*/w_i,\ i \in N - N'$. Therefore, the distance of the maximin point from every point $i \in N'$ is inversely proportional to its weight w_i, while the distance from each of the remaining points $i \in N - N'$ is greater than a lower bound that is inversely proportional to its weight w_i.

The above property of the maximin weights is probably the reason some authors do not consider weights with the maximin problem. Karkazis (1988) studied an un-

weighted Euclidean maximin problem in which the facility was to be located within a polygonal region S but as far away as possible from any point of the boundary of protected areas. These were more generally defined forbidden regions than in Melachrinoudis and Cullinane (1985). Although there were no apparent customers, the optimization approach—similar to the geometrical approach of Shamos and Hoey (1975)—suggests that the customers constitute an infinite set represented by the boundaries of the protected areas. The solution amounts to finding the largest (empty) circle that contains no points of the protected areas yet whose center is in S.

Melachrinoudis and Smith (1995) extended the Voronoi method of Dasarathy and White (1980) and developed an $O(mn^2)$ algorithm for the weighted maximin problem. For two points a_k, a_ℓ having weights w_k, w_ℓ such that $w_k > w_\ell$, the loci of weighted equidistant points is the Apollonius circle. This circle has center on the line connecting a_k and a_l, at point $o_{k\ell}$, and radius $\gamma_{k\ell}$, both expressed in terms of the weights ratio, $r_{k\ell} = w_\ell / w_k$, in (10.29). The edges of the weighted Voronoi diagram are therefore circular segments or whole circles.

$$o_{k\ell} = \frac{a_k - r_{k\ell}^2 a_\ell}{1 - r_{k\ell}^2}, \quad \text{and} \quad \gamma_{k\ell} = \frac{r_{k\ell} \, \|a_k - a_\ell\|}{1 - r_{k\ell}^2}. \tag{10.29}$$

Melachrinoudis and Cullinane (1986b) developed a minimax model for undesirable facility location. The model seeks a facility location that minimizes the maximum weighted inverse square distance over all customers, or

$$\min_{x \in S} \max_{i \in N} \{w_i / d^2(i, x)\}. \tag{10.30}$$

The objective is justified in many situations since the concentration of pollutants such as noise or radiation follows the inverse square law, see Poynting's Theorem in Lipscomb and Taylor (1978). A customer weight represents the degree of incompatibility between the customer and the facility and unlike with the maximin objective, the higher the weight of the customer, the further away the facility is pushed. Similar to the one for the maximin problem, an $O(n^4)$ algorithm was developed for a convex polygonal region S, while for a nonconvex feasible region, composed of many disjointed nonconvex sets representing irregular land and islands, a graphical computer procedure was suggested as in Drezner and Wesolowsky (1980). The minimax problem in (10.30) was shown by Erkut and Öncü (1991) to be equivalent to the maximin problem with weights $w_i^{-1/2}$, implying the above mentioned inverse relationship between the magnitude of a maximin weight and the degree of incompatibility it represents. Their proof used a more general formulation with an arbitrary exponent, i.e., $d^q(i, x)$, in which case the weights in the equivalent maximin problem were $w_i^{-1/q}$. The negative exponent explains the opposite behavior of weights in the two problems.

A minimax objective was also developed by Hansen et al. (1981) for the location of an undesirable facility in which, however, a general continuous and decreasing function of distances was used. The authors named the problem the *anti-Rawls*

problem, since the objective can be characterized as an equity measure to the worst-off customer. When the function of distances is linear, the minimax reduces to the maximin problem. The authors used a simple method called *Black and White*, which is similar to the ε-approximation method of Drezner and Wesolowsky (1980).

For the rectilinear maximin location problem in a convex polygon S, Melachrinoudis and Cullinane (1986a) and Melachrinoudis (1988) developed optimality properties similar to those described by Dasarathy and White, except that the convex hull $CH(N)$ is replaced by the smallest rectangle H whose sides are parallel to the two coordinate directions and encases all customers. Thus, local optima exist in the union of two sets, the boundary of S and $S \cap H$.

10.3.3 Major Contributions on Undesirable Facility Location that Followed the Classical Works

Following the classical contributions, numerous papers on undesirable facility location problems have been published in the last thirty years. A few of them, which built on the classical contributions and those on which the classical contributions had a direct or indirect impact, were reviewed in the previous two subsections. In this subsection, a short survey of major works that followed the classical contributions is presented. This short and by no means all-inclusive survey includes representative works with similar distance metrics and solution spaces as well as multiobjective approaches. A comprehensive survey of undesirable facility location models, though less contemporary, can be found in Erkut and Neuman (1989) Eiselt and Laporte (1995) and Plastria (1996).

The classical contribution of Church and Garfinkel (1978) itself was followed by only an algorithmic refinement. Their algorithm requires $O(mn^2)$ time to find the weighted maxisum point on a general network. By making use of the observation that $T(x)$ in (10.1) is a piecewise linear and concave function of x on a given arc, Tamir (1991) briefly suggested an improvement leading to an $O(mn)$ algorithm. Colebrook et al. (2005) described a complete algorithm of this improved complexity by making use of the above concavity property and by computing efficiently in $O(n)$ time a new upper bound of $T(x)$ over an arc. Their experimental results showed that the improved algorithm, compared to Church and Garfinkel's, ran in about half the time and processed about 25% fewer arcs due to tighter upper bounds on the arcs.

The unweighted maximin problem on a network admits a trivial solution, in that the optimal is the midpoint of the longest arc of the network. The weighted maximin problem on a network, $\max_{x \in G} \min_{i \in N} w_i d(i, x)$, has similar properties to the maxisum problem. It is nonconvex and it has a unique local optimum on each arc (Melachrinoudis and Zhang 1999). In addition to the set of arc bottleneck points, the finite set of candidates includes the set of center bottleneck points. For a complete coverage of finite dominating sets to the maximin and other location problems on

networks with general monotone or non-monotone distance functions, see Hooker et al. (1991). The algorithm for solving the maximin problem on networks is similar to Algorithm 1: searching arcs for local maxima, updating the lower bound and eliminating arcs using upper bounds on arcs. For each unfathomed arc, a linear program with two variables can be constructed which can be solved very efficiently by an $O(n)$ algorithm. Melachrinoudis and Zhang (1999) and Berman and Drezner (2000) independently provided $O(mn)$ algorithms by using $O(n)$ algorithms for linear programming problems of Dyer (1984) and Megiddo (1982), respectively.

The first paper on the maximin problem using the rectilinear metric was published by Drezner and Wesolowsky (1983). Since the rectilinear distance is piecewise linear, the problem can be linearized. The feasible region is divided into rectangular segments by drawing horizontal and vertical lines through each customer point and a linear optimization problem is solved for each one of the $O(n^2)$ linear programming problems. Upper bounds for each region are used to reduce the number of linear programs that need to be solved. Mehrez et al. (1986) proposed a new upper bound for that purpose which was further improved by Appa and Giannikos (1994). Sayin (2000) formulated the rectilinear maximin location problem as a mixed integer program that can be solved by any standard MIP solver. Nadirler and Karasakal (2008) simplified the mixed integer programming formulation and improved further the bounds to increase the computational performance of a branch and bound algorithm very similar to the *Big Square-Small Square* algorithm of Hansen et al. (1981) and the generalized Big Square-Small Square method of Plastria (1992).

As was mentioned earlier, undesirable facility location problems provide some service to the community and some travel may be required to and from it. Therefore, in addition to minimizing the undesirable effects on populations, the minimization of transportation cost is of interest. This gives rise to a bi-objective problem for locating undesirable facilities. Depending on the application, either the minimax or the minisum (desirable facility) objective can be combined with the maximin or maxisum (undesirable facility) objective. An advantage of solving a bi-objective undesirable facility location problem is that one can obtain the whole efficient frontier, i.e., the set of points that exhibit the complete tradeoff between the two objectives, including the two points that optimize the individual objectives. For a formal definition of efficient points and other concepts in multicriteria optimization, see Steuer (1989).

The first bi-objective problem for undesirable facility location was formulated by Mehrez et al. (1985). The authors combined the minimax and maximin unweighted objectives using rectilinear distances. They generated the whole efficient set by examining the intersections of any two lines forming the equirectilinear distances between every pair of customer points or boundary edges of the feasible region. Also using rectilinear distances, Melachrinoudis (1999) combined the minisum and maximin objectives and solved the problem by generating a series of $O(n^2)$ linear programs as in Drezner and Wesolowsky (1983), but instead of solving the linear programs by the simplex method, he constructed the whole efficient frontier by reducing each linear program to simple variable ranges using the Fourier-Motzkin

elimination process. Brimberg and Juel (1998) used Euclidean distances in a bi-objective problem that combined the minisum objective and a second minisum objective (undesirable objective), which had the Euclidean distances raised to a negative power. They outlined an algorithm for generating the efficient set by solving the weighted-sum of the two objectives with varying weights. Skriver and Andersen (2001) solved the same problem using the Big Square-Small Square method and generated an approximation of the efficient set. A similar bi-objective model was developed by Yapicioglu et al. (2007) where the second minisum objective was modified further to model undesirable effects with distance. They approximated the effects at a distance $d(i, x)$ from the facility as a piecewise linear and decreasing function of $d(i, x)$; up to a certain distance, they argued, the obnoxious effects are constant, then decreasing with distance in a piecewise fashion, while beyond a certain distance the effects are nonexistent. Particle Swarm Optimization is used to approximate the efficient set.

Melachrinoudis and Xanthopulos (2003) solved the Euclidean distance location problem with the minisum and maximin objectives. They developed the whole trajectory of the efficient frontier by a combination of a problem that optimizes the weighted sum of the objectives and the Voronoi diagram of the customer points. Using Karush-Kuhn-Tucker optimality conditions showed that this trajectory is not necessarily continuous and may consist of (a) a parametric curve of the weighted-sum of the objectives starting at the minisum point and ending at the boundary of its Voronoi polygon, (b) segments of the Voronoi edges as the weight of the maximin objective is increasing while the weight of the minisum objective is decreasing, and (c) segments of the boundary until the maximin point is reached.

A different type of undesirable facility location problem is the *minimal covering problem* in which the undesirable effects of a facility are evident only within certain distance from it, referred to as the circle of influence. Given n populations of size w_i, $i = 1, \ldots, n$, that are concentrated in n points on the plane, the location of an undesirable facility is to be found within a feasible planar region S to minimize the population covered within a certain distance r from the facility. This problem was introduced by Drezner and Wesolowsky (1994) who, in addition to the circle, determined the rectangle that contains the minimum total population. Berman et al. (1996) extended the problem to the network space. By considering the radius of the circle as a continuous variable and second objective, Plastria and Carrizosa (1999) solved the problem with two objectives. First, to maximize the radius r of a circle with center the point x at which the facility is to be located, and second, to minimize the population covered in that circle, $\sum_{d(i,x)<r} w_i$. They developed polynomial algorithms for generating all efficient discs (x, r) whose number they show is finite. The trade-off information of efficient solutions can provide answers to interesting coverage questions, such as finding the facility location that minimizes the population covered within a given radius (previously defined as minimal covering problem) or finding the largest circle not covering more than a given total population. They considered a feasible region of any shape in the plane and the results can be extended to a planar network.

A more recent approach for locating an undesirable facility is with expropriation. The rationale is that in certain cases there is no point in the feasible region that is far enough from all customers to locate the undesirable facility. One possibility to resolve this issue is by buying (or compensating) some of the customers. Berman et al. (2003) introduced two models for the *location problem with expropriation*. In the first model, a location on a network was sought that maximizes the minimum distance (maximin) from the facility to the non-expropriated customer points, subject to a given expropriation budget. In the second model, the expropriation cost was minimized while ensuring that the facility is located at least certain distance away from all non-expropriated customer points. Berman and Wang (2007) added a second objective to the last model: the minimization of transportation cost. The two cost objectives were added into one, so the resulting problem is treated as a single objective problem. For a planar feasible region and rectilinear metric, they identified a finite dominating set that contains the optimal solution.

There are not many papers on undesirable facility location on networks using two objectives. Zhang and Melachrinoudis (2001) formulated the first bi-objective problem on a network by combining the maxisum objective with the maximin objective. Using the piecewise linear and concave property of both objectives on an arc, they developed fathoming rules for eliminating inefficient arcs and arc segments. Unfathomed arc segments were mapped onto the 2-dimensional objective space and a direct search was undertaken to construct the nondominated set, followed by the efficient set, which was shown to consist of discontinuous arc segments. Hamacher et al. (2002) developed several multicriteria models for undesirable facility location problems on a network with minisum and center objectives, and proposed methods for solving them.

A general model for the undesirable facility location problem with Euclidean distances in a polygonal feasible region S was presented by Saameno et al. (2006). By setting a parameter to certain values the model reduces to problems already defined: maximin, maxisum and bi-objective maximin/maxisum problems. In addition, the model reduces to the r-anticentrum problem, which maximizes the weighted sum of distances between the undesirable facility and its r closest customers. The maximin and maxisum problems are special cases of the r-anticentrum problem for $r = 1$ and $r = n$, respectively. The authors generalized the properties of local optima developed by Dasarathy and White (1980) and Melachrinoudis and Smith (1995) for the whole class of objectives. They identified a finite dominating set consisting of the set of vertices of S, V, the set of intersections of weighted bisectors (10.29) of customer points with the edges of S, W, and the set of intersections of the weighted bisectors taken two at a time, I. The finite dominating set was obtained in $O(nm^2 + m^4)$. In their algorithm they generated all candidate points in the set, eliminated some of them using the Lipschitzian property of the objective function, and evaluated the remaining points to obtain the optimal solution.

In addition to the above classes of models for undesirable facility location there are other models, such as multifacility, discrete, and location and routing models, which followed the classical contributions; we cannot elaborate upon these papers due to the limited space in this chapter. For the interested reader, some excellent

papers and surveys are available. For the *p-dispersion* problem, Chandrasekaran and Daughety (1981), Kuby (1987), Erkut (1990), and Pisinger (2006); for the *p-defense* problem, Moon and Chaudhry (1984), Kincaid (1992), and Klein and Kincaid (1994); for generic *discrete multifacility* undesirable location problems, Chhajed and Lowe (1994). For locating multiple undesirable facilities on graphs using maxisum and maximin objectives, Tamir (1988, 1991); using coverage objectives, Berman and Huang (2008); using expropriation, Berman and Wang (2007). A recent survey for *location and routing* problems that includes undesirable facility location and routing of hazardous wastes can be found in Nagy and Salhi (2007).

10.4 Summary and Outlook

The advent of more stringent environmental standards, the resurgence of environmental groups and a greater awareness of the public of the potential dangers of pollution in the early 1970s generated a research need for systematically locating polluting and environmentally hazardous facilities. Undesirable facility location research began with the pioneering work of Church and Garfinkel (1978). Their work on the maxian (maxisum) problem was analyzed in detail followed by the first works on the Euclidean maximin problem of Dasarathy and White (1980) and Drezner and Wesolowsky (1980).

Church and Garfinkel (1978) first formulated a model for locating an undesirable facility on a network by replacing the *min* with the *max* operator in the median model that had dominated the location literature since the seminal work of Hakimi (1964). Unlike the median model, they demonstrated that the maxian model is hard to solve because it is nonconvex and typically has many local optima, a characteristic of undesirable facility location problems. They showed that local optima occur on the cycles (bottleneck points) and on the leaves of the network and developed a simple solution procedure that decomposes the network into its arcs in the search for the global optimum. Arcs were considered for fathoming using bounds, and the local maxisum point was found on unfathomed arcs by utilizing the concavity property of the objective function. This algorithm became a standard for future algorithms in undesirable facility location. For example, instead of arcs, parts of the feasible region are considered for fathoming in the Big Square-Small Square method or individual facets of the feasible region in Dasarathy and White's (1980) algorithm that partially enumerates local maxima. The work of Church and Garfinkel had an enormous impact by stimulating research and establishing the field of undesirable facility location in the 1980s.

Undesirable facility location in the continuous space has its origins in the *largest empty circle* problem, one of several problems Shamos (1975) and Shamos and Hoey (1975) studied in computational geometry. Dasarathy and White (1980) were the first to define the maximin problem using Euclidean distances as a nonconvex and nonlinear problem, derive the properties of the local optima using the Karush-Kuhn-Tucker optimality conditions, and identify a finite dominating set. They

proved that the global optimum is either in the convex hull of the customer points or on the boundary of the feasible region and developed an algorithm for searching those parts of the feasible region. As in Church and Garfinkel (1978), they used upper bounds to fathom facets of the feasible region and updated the lower bound on the optimal objective value using the current best feasible solution. For a 2-dimensional feasible region, they extended Shamos and Hoey's Voronoi diagram approach to search for local optima at the boundary of the feasible region.

Drezner and Wesolowsky (1980) first formulated the weighted maximin location problem. The 2-dimensional feasible region is the intersection of circles each having its center at a customer point and radius equal to the maximum distance the undesirable facility can be located away from that customer, implying that the facility performs some service to the customer and has to be within reach. Their solution approach is different from previous ones. It is graphical in nature and is implemented on the computer with a bisection search of the feasible region. Although the bisection search seems very efficient for this feasible region, it has not been used generally. The contributions by Dasarathy and White and by Drezner and Wesolowsky incited a large body of research in undesirable facility location using the maximin objective with various distance metrics and solution spaces. Unlike the maxisum objective, the maximin objective does not limit the optimum to the boundary and excludes customer points for locating the undesirable facility. Its use represents obvious advantages over the maxisum objective in continuous spaces.

Important works that followed the original papers were analyzed with special attention given to location models or methods that extended the classical contributions, such as considering single facility location models on network and planar space and under multiple objectives. It was not the purpose of this chapter to survey the literature on undesirable facility location, and therefore many important papers have not been included. A complete survey of this area is important and its time is due, so therefore it is suggested that such an effort be undertaken in the near future.

Regarding future research directions, consider what has been accomplished so far, what has not and what can be accomplished given the technological advances and the changing needs of society. The location literature is full of elegant mathematical models which admit neat solution algorithms. As ReVelle and Eiselt (2005) point out, the "location field is active from a research perspective but when it comes to applications it appears to be a significant deficit, at least as compared to other, similar fields." It is known that real life problems are complex with nasty feasible regions and multiple objectives that may not be necessarily functions of straightforward distance metrics such as Euclidean or rectilinear. When it comes to undesirable facilities, pollution density or its effects often are neither symmetric nor linear functions of distance. Very often, a real feasible region is not a simple polygonal area but the union of many disjoint regions.

The parameters of the problem, such as customer weights, may change over time depending on the population size, technological developments, and legislation for hazardous wastes and facility standards. Multiple stakeholders and decision makers are usually involved in undesirable facility location decisions; therefore, more realistic, integrated and robust location models need to be developed that relate to the

practitioners' concerns. New technological tools such as geographical information systems are readily available, together with versatile optimization tools and vast computer power to make the task easier. Without discouraging the development of elegant mathematical models that admit creative solution procedures, researchers should be encouraged to tackle real-life problems with creative formulations, even if they have to solve them for a near-optimal solution by a standard optimization software package or a heuristic procedure.

Although in this chapter we reviewed single facility location on continuous or network spaces, discrete location models appear to be more realistic from a practitioner's decision making point of view, maybe because it is more natural to compare the merits of given sites rather than find one among an infinite number of possible sites. As a strategic decision, the facility location process usually involves two stages: one approach is to evaluate many candidate sites in the first stage and come up with a few using constraints and minimum requirements, and in the second stage to select a site using multiple criteria optimization, as in Min et al. (1997); another approach is to generate a small number of candidate sites in the first stage with analytical models and in the second stage to use discrete multiobjective tools to select the candidate site, as suggested by Erkut and Neuman (1989) and Plastria (1992).

An interesting non-geographical area in which undesirable facility location models could be applied is product design; see, e.g., Goldman (2006). The attributes of a product, such as physical dimensions, expected lifetime, and cost can be regarded as coordinates in the attribute or design space. Given the existing products in the market (points in space), a company may want to design a new product to differentiate from the existing ones as an alternative to purchase, yet not make it very different. The new product has the properties of a semi-desirable facility that needs to be located in the design space away from existing points but within reach.

Finally, location researchers should adapt their models to fill new needs of the society and use tools made available by new technologies. Some examples are applications in telecommunications and especially wireless networks, homeland security, environment change and global warming (Francis 2008), and use of geographical information systems (Murray and Church 2008).

References

Appa GM, Giannikos I (1994) Is linear programming necessary for single facility location with maximin of rectilinear distance? J Oper Res Soc 45:97–107

Ahuja RK, Magnanti TL, Orlin JB (1993) Network flows: theory, algorithms and applications. Prentice Hall, Englewood Cliffs

Berman O, Drezner Z (2000) A note on the location of an obnoxious facility on a network. Eur J Oper Res 120:215–217

Berman O, Huang R (2008) The minimum weighted covering location problem with distance constraints. Comput Oper Res 35:356–372

Berman O, Wang Q (2007) Locating semi-obnoxious facilities with expropriation: minisum criterion. J Oper Res Soc 58:378–390

Berman O, Drezner Z, Wesolowsky GO (1996) Minimum covering criterion for obnoxious facility location on a network. Networks 28:1–5
Berman O, Drezner Z, Wesolowsky GO (2003) The expropriation location model. J Oper Res Soc 54:769–776
Brady SD, Rosenthal RE (1980) Interactive computer graphical solutions of constrained minimax problems. AIIE Trans 12:241–248
Brimberg J, Juel H (1998) A bicriteria model for locating a semi-desirable facility in the plane. Eur J Oper Res 106:144–151
Chandrasekaran R, Daughety A (1981) Location on tree networks: *p*-centre and *n*-dispersion problems. Math Oper Res 6:50–57
Chhajed D, Lowe TJ (1994) Solving structured multifacility location problems efficiently. Transp Sci 28:104–115
Church RL, Garfinkel RS (1978) Locating an obnoxious facility on a network. Transp Sci 2:107–118
Colebrook M, Gutierrez J, Sicilia J (2005) A new bound and an $O(mn)$ algorithm for the undesirable 1-median problem (maxian) on networks. Comput Oper Res 32:309–325
Dasarathy B, White LJ (1980) A maxmin location problem. Oper Res 28:1385–1401
Drezner Z, Wesolowsky GO (1980) A maximin location problem with maximum distance constraints. AIIE Transactions 12:249–252
Drezner Z, Wesolowky GO (1983) Location of an obnoxious facility with rectangular distances. J Reg Sci 23:241–248
Drezner Z, Wesolowky GO (1994) Finding the circle or rectangle containing the minimum weight of points. Locat Sci 2:83–90
Dyer ME (1984) Linear time algorithms for two and three-variable linear programs. SIAM J Comput 13:31–45
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
Erkut E (1990) The discrete *p*-dispersion problem. Eur J Oper Res 46:48–60
Erkut E, Neuman S (1989) Analytical models for locating undesirable facilities. Eur J Oper Res 40:275–291
Erkut E, Öncü TS (1991) A parametric 1-maximin location problem. J Oper Res Soc 42:49–55
Francis RL (2008) A discussion of some location problems global warming can cause. Abstract, ISOLDE 11:139
Goldman AJ (2006) Optimal facility-location. J Res Nat Inst Stand Technol 111:97–101
Goldman AJ, Dearing PM (1975) Concepts of optimal location for partially noxious facilities. Bull Oper Res Soc Am 23(Suppl 1):B-31
Hadley G (1964) Nonlinear and dynamic programming. Addison-Wesley, Reading
Hakimi SL (1964) Optimal location of switching centers and the absolute centers and medians of a graph. Oper Res 12:450–459
Hamacher HW, Labbé M, Nickel S, Skriver AJV (2002) Multicriteria semi-obnoxious network location problems (MSNLP) with sum and center objectives. Ann Oper Res 110:33–53
Hansen P, Peeters D, Thisse JF (1981) On the location of an obnoxious facility. Sistemi Urbani 3:299–317
Hooker JN, Garfinkel RS, Chen CK (1991) Finite dominating sets for network location problems. Oper Res 3:100–118
Karkazis J (1988) The general unweighted problem of locating obnoxious facilities on the plane. Belgian J Oper Res Stat Comput Sci 28:43–49
Karkazis J, Boffey C (1994) Modeling pollution spread. Stud Locat Anal 7:91–104
Karkazis J, Papadimitriou B (1992) A branch and bound algorithm for location of facilities causing atmospheric pollution. Eur J Oper Res 58:363–373
Kincaid RK (1992) Good solutions to discrete noxious location problems via metaheuristics. Ann Oper Res 40:265–281
Klein CM, Kincaid RK (1994) The discrete anti-*p*-center problem. Transp Sci 28:77–79

Kuby MJ (1987) Programming models for facility dispersion: the *p*-dispersion and maxisum dispersion models. Geogr Anal 19:315–329

Lipscomb DM, Taylor AC Jr (1978) Noise control, handbook of principles and practices. Van Nostrand Reinhold, New York

Megiddo N (1982) Linear-time algorithms for linear programming in $\mathbb{R}^3$ and related problems. SIAM J Comput 4:759–776

Mehrez A, Sinuany-Stern Z, Stulman A (1985) A single facility location problem with a weighted maximin-minimax rectilinear distance. Comput Oper Res 12:51–60

Mehrez A, Sinuany-Stern Z, Stulman A (1986) An enhancement of the Drezner-Wesolowsky algorithm for single facility location with maximin of rectilinear distance. J Oper Res Soc 37:971–977

Melachrinoudis E (1985) Determining an optimal location for an undesirable facility in a workroom environment. Appl Math Model 9:365–369

Melachrinoudis E (1988) An efficient computational procedure for the rectilinear maximin location problem. Transp Sci 22:217–223

Melachrinoudis E (1999) Bicriteria location of a semi-obnoxious facility. Comput Ind Eng 37:581–593

Melachrinoudis E, Cullinane TP (1985) Locating an undesirable facility within a geographical region using the maximin criterion. J Reg Sci 25:115–127

Melachrinoudis E, Cullinane TP (1986a) Locating an undesirable facility within a polygonal region. Ann Oper Res 6:137–145

Melachrinoudis E, Cullinane TP (1986b) Locating an undesirable facility with a minimax criterion. Eur J Oper Res 24:239–246

Melachrinoudis E, Smith JM (1995) An O(mn^2) algorithm for the maximin problem in E^2. Oper Res Lett 18:25–30

Melachrinoudis E, Xanthopulos Z (2003) Semi-obnoxious single facility location in Euclidean space. Comput Oper Res 30:2191–2209

Melachrinoudis E, Zhang FG (1999) An $O(mn)$ algorithm for the 1-maximin problem on a network. Comput Oper Res 26:849–869

Min H, Melachrinoudis E, Wu X (1997) Dynamic expansion and location of an airport: a multiple objective approach. Transp Res—Part A 31:403–417

Minieka E (1983) Anticenters and antimedians of a network. Networks 13:359–364

Moon ID, Chaudhry SS (1984) An analysis of network location problems with distance constraints. Manag Sci 30:290–307

Murray AT, Church RL (2008) Location analysis and GIS. Abstract, ISOLDE 11:105

Nadirler D, Karasakal E (2008) Mixed integer programming-based solution procedure for single-facility location with maximin of rectilinear distance. J Oper Res Soc 59:563–570

Nagy G, Salhi S (2007) Location-routing: issues, models and methods. Eur J Oper Res 177:649–672

Pisinger D (2006) Upper bounds and exact algorithms for the *p*-dispersion problems. Comput Oper Res 33:1380–1398

Plastria F (1992) GBSS: The generalized big square small square method for planar facility location. Eur J Oper Res 62:163–174

Plastria F (1996) Optimal location of undesirable facilities: a selective overview. JORBEL 36:109–127

Plastria F, Carrizosa E (1999) Undesirable facility location with minimal covering objectives. Eur J Oper Res 121:302–319

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

Saameno JJ, Guerrero C, Munoz J, Merida E (2006) A general model for the undesirable single facility location problem. Oper Res Lett 34:427–436

Sayin S (2000) A mixed integer programming formulation for the 1-maximin problem. J Oper Res Soc 51:371–375

Shamos MI (1975) Geometric complexity. Proceedings of the seventh ACM symposium on Theory of Computing, pp 224–233

Shamos MI, Hoey D (1975) Closest-point problems. 16th annual symposium on foundations of computer science, pp 151–162
Skriver AJV, Andersen KA (2001) The bicriterion semi-obnoxious location (BSL) problem solved by an ε-approximation. Eur J Oper Res 146:517–528
Steuer ER (1989) Multiple criteria optimization: theory, computation, and application. Krieger, Malabar
Sylvester JJ (1857) A question in the geometry of situation. Q J Pure Appl Math 1:79
Tamir A (1988) Improved complexity bounds for center location problems on networks by using dynamic data structures. SIAM J Discrete Math 1:377–396
Tamir A (1991) Obnoxious facility location on graphs. SIAM J Discrete Math 4:550–567
Ting SS (1984) A linear time algorithm for maxisum facility location on tree networks. Transp Sci 18:76–84
Yapicioglu H, Smith AE, Dozier G (2007) Solving the semi-desirable facility location problem using the bi-objective particle swarm. Eur J Oper Res 177:733–749
Zhang FG, Melachrinoudis E (2001) The maximin-maxisum network location problem. Comput Optim Appl 19:209–234

Chapter 11
Stochastic Analysis in Location Research

Oded Berman, Dmitry Krass and Jiamin Wang

11.1 Introduction

Modern Location Theory includes a large and growing field studying the impact of various types of uncertainty in location models. This field of Stochastic Location Models can be traced back to the pioneering work of Frank, whose 1966 paper "Optimum Locations on a Graph with Probabilistic Demands" represents the first formal analysis of stochastic issues in location theory. This paper appeared during the "golden age" of Operations Research, when many new applications of probability theory to optimization problems were being developed, spawning fundamental contributions in a variety of fields such as inventory theory, queuing theory, and stochastic dynamic programming, among others.

The development of Stochastic Location Models parallels the trends in other fields of Operations Research. While, as evidenced by several chapters in the current volume, the roots of location theory go back to the nineteenth century, the modern beginnings (particularly of network models) can be traced to the papers by Hakimi (1964, 1965) who formalized classical median and center objectives in deterministic location models. These objectives are concerned with the interactions of three elements: customer demands (assumed to be originating at the nodes of the network), location of the facility (early papers largely dealt with single-facility models; multi-facility models, generally requiring much larger computer processing power, came later), and the travel distance between customers and the facility. Frank's paper analyzed the effect of uncertainty in the first component: the customer

O. Berman (✉) · D. Krass
Rotman School of Management, University of Toronto,
105 St. George Street, Toronto, ON M5S 3E6, Canada
e-mail: berman@rotman.utoronto.ca

D. Krass
e-mail: krass@rotman.utoronto.ca

J. Wang
College of Management, Long Island University,
720 Northern Blvd., Brookville, NY 11548, USA
e-mail: jiamin.wang@liu.edu

H. A. Eiselt, V. Marianov (eds.), *Foundations of Location Analysis,*
International Series in Operations Research & Management Science 155,
DOI 10.1007/978-1-4419-7572-0_11, © Springer Science+Business Media, LLC 2011

demands, also known as node weights. This work provided direct impetus for subsequent works in two main directions: further and more general analysis of location models with stochastic node weights (which can be thought of as "direct descendants" of Frank's original work), and the analysis of the impacts of stochasticity in all other aspects of customer-facility interactions (the "indirect descendants"), such as customer-facility travel times, the provision of service once the customer reaches the facility (or the mobile server reaches the customer), and facility breakdowns. Many aspects of these analyses have been extended to, for example, multi-facility settings, planar location, or non-nodal demands. The obvious practical importance of these models, coupled with the significant technical difficulties they represent (often combining **NP**-hard deterministic location problems with very difficult stochastic problems) has led to the continued strong interest in these models on the part of location theorists. It is fair to say that significant advances continue to be made, while many important problems remain open.

This chapter is structured as follows. We begin, in Sect. 11.2, with a detailed look at Frank's original contribution, describing both the various models he introduced and the main methodological approaches used to analyze these models. In Sect. 11.3 we describe direct generalizations of Frank's original work, i.e., new results on location models with stochastic node weights. Important open problems in this line of research are discussed in Sect. 11.3.5. Finally, in Sect. 11.4, we provide a more general overview of Stochastic Location Theory. As noted earlier, this is a large and growing field and thus we do not attempt to provide a thorough treatment here. Instead, we outline main research directions and key contributions, citing a number of references that could serve as the basis for further reading.

11.2 Frank (1966): Maximum Probability Center and Median Problems

Let $G = (N, L)$ represent the underlying network, where N is the discrete set of nodes and L is the set of links. Let $n = |N|$ and $\ell = |L|$ be the number of nodes and the number of links, respectively. We assume that customer demand is concentrated at the nodes of the network, while facilities may be located both at the nodes and along the links. The weight associated with a nodal point $i \in N$, denoted by h_i, represents the number of potential customers originating from the corresponding node, while the length of a link $(p, q) \in L$, denoted by ℓ_{pq} measures the traveling distance between the two end nodes. If a facility is located at some internal point $x \in (p, q)$, we will adopt the following common notational convention: we will let $x \in [0, \ell_{pq}]$ represent *both* the distance of the facility from the left end-point p and the actual location of the facility. Thus, $x = 0$ means that the facility is located at p, while $x = \ell_{pq}$ means it is located at q, making clear from the context which link x is located on. We will also use $d(a, b)$ to represent the shortest distance between points a and b on G, where a and b may be nodes or points along the links.

The concepts of absolute (deterministic) m-centers and m-medians of a network were rigorously defined by Hakimi (1964, 1965). In these models, the weight of

each nodal point and the length of each link are assumed to be constant and known *a priori*. Frank (1966) argued that the weight of a nodal point, e.g., the number of messages originating from a node of a communication network, may not be deterministic, but is better represented as a random variable with some probability distribution. This leads to several possible stochastic generalizations of the concepts of medians and centers: in the expected value sense ("absolute expected" median and center), in the sense of maximizing the probability of achieving some threshold ("maximum probability" median and center), or in the sense of variance minimization ("minimum variance" absolute median). These concepts, which were introduced and analyzed in Frank (1966), are discussed in Sect. 11.2 below.

Frank also provided some additional analysis for his models. He derived upper and lower bounds for most problems, developed some solution methods (under fairly stringent assumptions), and suggested several asymptotic results. Many of the ideas he introduced turned out to be quite useful and were employed by subsequent researchers to develop more general approaches that will be reviewed in Sect. 11.3. Furthermore, we note that in his paper, Frank focused on single-facility models. Though his definitions can be easily generalized to the case of multiple facilities, the resulting models are quite difficult to analyze, and remain open to a large extent. Thus, we will focus on single-facility models throughout Sects. 11.2 and 11.3, briefly discussing the multiple facility case in Sect. 11.3.5.

11.2.1 Absolute Expected Centers and Medians

Suppose node weights h_i, $i \in N$ are random variables whose distributions have finite moments. Denote the expectation operator by E. An *absolute expected center* (*AEC*) x_{0e} of G is a point such that

$$\max_{i \in N}\{E(h_i)d(i, x_{0e})\} \leq \max_{i \in N}\{E(h_i)d(i, x)\}$$

holds for every point $x \in G$. The maximum expected weighted distance at the absolute expected center, $r_0 = \max_{i \in N}\{E(h_i)d(i, x_{0e})\}$ is referred to as the *expected radius* of G.

Similarly, a point y_{0e} is an *absolute expected median* (*AEM*) if for every point x on G

$$\sum_{i \in N} E(h_i)d(i, y_{0e}) \leq \sum_{i \in N} E(h_i)d(i, x),$$

and $R_0 = \sum_{i \in N} E(h_i)d(i, y_{0e})$ is called the *expected median length* of G. Note that x_{0e} and y_{0e} are equivalent to the absolute center and median defined by Hakimi, with node weights taken to be the expected values of the random weights.

In the probabilistic setting, it may be important to evaluate the likelihood that the objective values of the absolute expected center and the absolute expected median (which represent expected maximum and total travel costs, respectively) exceed

certain pre-determined thresholds. This leads to the following expressions where r and R are the threshold values:

$$P\left(\max_{i\in N}\{h_i d(i, x_{0e})\} > r\right), \tag{11.1}$$

and

$$P\left(\sum_{i\in N} h_i d(i, y_{0e}) > R\right). \tag{11.2}$$

Suppose the threshold values satisfy $r \geq r_0$ and $R \geq R_0$. Noting that $P(Z<c) \geq 1 - E(Z)/c$ always holds for a nonnegative random variable Z, the following upper bounds for the probabilities in (11.1, 11.2) can now be established:

$$P\left(\max_{i\in N}\{h_i d(i, x_{0e})\} > r\right) \leq 1 - (1 - r_0/r)^n$$

and

$$P\left(\sum_{i\in N} h_i d(i, y_{0e}) > R\right) \leq R_0/R.$$

Note that the right-hand side of the first expression (the center objective) rapidly approaches 1 as the number of nodes $|N|$ grows. On the other hand, the right-hand side of the second expression (the median objective) is independent of $|N|$.

11.2.2 Maximum Probability Absolute Centers and Medians

Given the thresholds and probabilities defined in (11.1) and (11.2), where are these probabilities minimized? That is, if we want to minimize the probability that the maximum or total weighted distance from the facility to the demand points exceeds certain undesirable level, what is the best location for the facility? Is it necessarily the absolute expected center or median? The following example demonstrates that this may not be the case.

Example 1: Consider a network consisting of a single link 10 units long with the two end nodes denoted by A and B. Suppose that the random weight associated with node A takes two values 10 and 20 with equal probabilities, while the weight of B takes the same two values with respective probabilities of 0.4 and 0.6. It is obvious that $E(h_A) = 15$ and $E(h_B) = 16$. Applying the solution approach suggested by Hakimi (1964) for the deterministic absolute center problem with node weights of $h_A = 15$ and $h_B = 16$, we have $x_{0e} = 5.16$, so the absolute expected center is 5.16 units away from A (or, equivalently, 4.84 units away from B). The expected radius r_0 is computed as $5.16(15) = 77.40$. It is easy to verify that $P\left(\max_{i\in\{A,B\}}\{h_i d(i, x_{0e})\} > r_0\right) = 80\%$. Now con-

sider the location $x = 7.5$, i.e., a point that is 7.5 units away from A and 2.5 units away from B. It is easy to check that the value $P(\max_{i\in\{A,B\}}\{h_i d(i,x)\} > r_0) = 50\%$. A similar situation occurs for the absolute expected median of the network, where $y_{0e} = 10$ (node B is the absolute expected median) and the expected median length R_0 equals 150. Likewise, $P(\sum_{i\in\{A,B\}} h_i d(i,y_{0e}) > R_0) = 50\%$, while $P(\sum_{i\in\{A,B\}} h_i d(i,x) > R_0) = 30\%$ when $x = 5$.

The above example suggests that the absolute expected center and the absolute expected median may be suboptimal in terms of minimizing the *risk* of unacceptably high values. This leads to the concepts of maximum probability absolute centers and medians. A point x_r is a *maximum probability absolute r center* (MPArC) of G if it minimizes the probability that the maximum weighted distance to the nodal points exceeds $r > 0$, i.e.,

$$P\left(\max_{i\in N}\{h_i d(i,x_r)\} > r\right) \le P\left(\max_{i\in N}\{h_i d(i,x)\} > r\right), \tag{11.3}$$

which holds for every point x on G. Similarly, a *maximum probability absolute R median* (MPARM) of G, denoted by y_R, minimizes the probability that the total weighted distance to the nodal points is greater than R, so that for any x on G,

$$P\left(\sum_{i\in N} h_i d(i,y_R) > R\right) \le P\left(\sum_{i\in N} h_i d(i,x) > R\right). \tag{11.4}$$

Here r and R can be interpreted as "aspiration levels" of the maximum weighted distance and total weighted distance, respectively: it is desirable for the probabilistic weighted distances to stay within these targets. The MPArC and MPARM are the locations that maximize the likelihood of achieving the specified aspiration levels. Therefore, models (11.3) and (11.4) are in the same spirit as the concept of "satisficing objectives" (in contrast to "optimizing objectives") that is developed by Simon (1957) for the behavior theories.

Assuming that the random weights are independent, solving model (11.3) is equivalent to finding a point x maximizing $\prod_{i\in N} F_i(x)$, where $F_i(x) = P(h_i d(i,x) \le r)$. Furthermore, if the probability distributions of the weights h_i are discrete, the function $F_i(x)$ on a link changes its value only at a finite number of *jump points* $x = x_{i_m}$, where there exists some realization $w_{(i)}$ of the random weight h_i such that $w_{(i)} d(i, x_{i_m}) = r$. Let J be the set of distinct jump points of $F_i(x)$ for $i \in N$. It is obvious that the objective function is constant on each open interval formed by two consecutive jump points in J. Therefore, in principle, an optimum on each link may be found by enumerating all the jump points and then repeating this procedure for all links on the network. An algorithmic approach to accomplish this was suggested by subsequent researchers and will be outlined in Sect. 11.3.

Unlike the deterministic absolute median model, the MPARM model (11.4) does not have the vertex optimality property (this follows from Example 1 above). In fact, the evaluation of the objective function at a given point is usually difficult to obtain since, even under the independence assumption, it involves a convolution of

distributions. One exception occurs when the distribution of node weights is Normal, since the convolution of Normal distributions is also Normal and is available in closed form. Moreover, when the number of nodal points is large enough, the central limit theorem indicates that the distribution of the weighted sum (i.e., the objective function) will be approximately Normal. Therefore, Frank suggests that the approximate MPARM can be found by solving the model

$$\max_{y \in G} \left[R - \sum_{i \in N} E(h_i) d(i, y) \right] \Big/ \sqrt{\sum_{i \in N} \sigma^2(h_i) d(i, y)^2}, \tag{11.5}$$

where $\sigma^2(h_i)$ is the variance of h_i. We note that the nonlinear optimization model above can be hard to solve, since the objective function is not convex along each link. Thus, local optima may occur and it is not clear how the globally best location is to be found. Solution approaches for this problem will be discussed in Sect. 11.3.4.

In practice, the probability distributions of the weights may be unknown. Frank shows that the distributions estimated by sampling may give good results for the MPArC and MPARM as long as the sample sizes are sufficiently large.

11.2.3 Minimum Variance Absolute Medians

The MPARM objective discussed in the previous section was introduced to limit the risk of large weighted travel distances that may occur due to stochasticity of node weights. Another common measure of risk (or variability) is the variance, which leads to the following model that was also introduced in Frank (1966).

A point y_{0v} is a *minimum variance absolute median* (MVAM) if

$$Var\left[\sum_{i \in N} h_i d(i, y_{0v}) \right] \leq Var\left[\sum_{i \in N} h_i d(i, x) \right] \tag{11.6}$$

holds for every point x on G. This objective may be useful in situations where having a location with a low variability of total weighted travel distance is as, or more, important than having a low expected value of travel distance.

If the random demand weights h_i are independent, the variance of the total weighted distance from a point x to the nodal points can be computed as follows

$$Var\left[\sum_{i \in N} h_i d(i, x) \right] = \sum_{i \in N} \sigma^2(h_i) d(i, x)^2.$$

Under the independence assumption, analysis in Frank (1966) shows how to find a local MVAM solution on an isthmus, i.e., a link whose removal disconnects the network (also known as a "cut link" in network theory). This approach allows us to find the optimal MVAM on a tree, where every link is an isthmus. However, a gen-

eral network may have no cut links. Fortunately, Frank's approach can be extended to arbitrary links. Consider a link $(p, q) \in L$ and a node $i \in N$. The *antipode* on (p, q) with respect to i is a point $x_i \in (p, q)$ such that $d(i, p) + x_i = d(i, q) + \ell_{pq} - x_i$ (recall that x_i refers to the point on (p, q) which is at distance x_i from the left endpoint p). In other words, the shortest paths from x_i to node i through both endpoints p and q have the same length. Observe that an antipode on (p, q) exists for any node $i \in N \setminus p \setminus q$. Moreover, if the shortest path from i to q passes through p (through q), then $x_i = q$ ($x_i = p$). This must be the case when (p, q) is an isthmus—in this case all antipodes must be located at the endpoints. However, for an arbitrary link, the antipodes will often occur at an internal point. Consider all antipodes on (p, q). We define a *primary region* as a segment between two adjacent antipodes and note that if (p, q) is an isthmus, there is only one primary region consisting of the whole link. Since p is an antipode of q and vice versa, the primary regions represent a partition of (p, q).

For a primary region $\pi = [\hat{x}, \tilde{x}]$ we define two sets of nodes, denoted by L' and R', as follows: let $L' = \{i \in N | d(i, p) + \tilde{x} \leq d(i, q) + \ell_{pq} - \tilde{x}\}$ and $R' = N \setminus L'$. In words, L' represents all nodes for which the shortest path from $\tilde{x}$ goes through $\hat{x}$ and, consequently, through the left endpoint p of (p, q) as well. The shortest path from $\hat{x}$ to a node in $i \in L'$ must also pass through p; indeed the shortest path from any internal point in π to i must pass through p. Now consider a node $i \in R'$. By definition of R' the shortest path from $\tilde{x}$ to i goes through the right endpoint q. Observe that the shortest path from $\hat{x}$ to i must also pass through q, for if it does not, an internal antipode x_i must exist in π. But then x_i would also be an antipode on (p, q) that would contradict π as a primary region. It follows that the shortest path from any point in π to a node in R' must pass through q.

These observations allow us to derive first-order conditions that lead to the following expression for the unique minimum z of the objective function within the primary region π (where z is the unique MVAM with respect to π):

$$z = \begin{cases} \hat{x} & \text{if } D \leq \hat{x} \\ D & \text{if } \hat{x} \leq D \leq \tilde{x}, \\ \tilde{x} & \text{if } D \geq \tilde{x} \end{cases} \tag{11.7}$$

where $D = \left[\sum_{i \in R'} \sigma^2(h_i)(d(i, q) + \ell_{pq}) - \sum_{i \in L'} \sigma^2(h_i) d(i, p) \right] \Big/ \sum_{i \in N} \sigma^2(h_i)$. The overall minimum over the link (p, q) can now be found by enumerating the solutions over all primary regions on the link (as noted earlier, when the link is an isthmus, it contains only one primary region). By enumerating solutions over all links, the MVAM of the network G can be found.

We note that the technique introduced in this section, dividing a link into segments over which the objective function is well-behaved and the solution is easier to find, is an important approach that will be applied to many objectives other than MVAM in the next section.

11.3 The Impact of Frank's Work: Models with Random Weights

The work on extending and generalizing Frank's original results has proceeded in several directions. First, additional location models with random weights have been introduced; as in Frank's original work, these models are counterparts of the well-studied deterministic location models such as anti-median, anti-center, and maximal cover. These models will be presented in Sect. 11.3.1 below.

Second, much progress on finding efficient algorithms for the models introduced by Frank and their extensions has been achieved. The results here can be divided into four streams. The first stream identifies "easy cases," i.e., conditions under which problems with random weights admit easily identifiable solutions; we will review these results in Sect. 11.3.2. The second stream consists of identifying situations where the optimal solution can be *localized* to a discrete finite set of points *a priori*; the solution can then be obtained by simple enumeration. In most cases, the localization results require the distribution of node weights to be discrete and independent; these results are covered in Sect. 11.3.3. Some algorithmic approaches have been developed for general distributions of node weights; these are described in Sect. 11.3.4. Finally, in Sect. 11.3.5 we will present a brief discussion of problems that remain open and, in our opinion, deserve further attention.

11.3.1 Obnoxious and Covering Problems with Random Weights

The MPArC, MPARM and MVAM problems described earlier are probabilistic counterparts of two classical deterministic location problems: the center and the median. There are three additional deterministic location models that have been well-studied: anti-center, anti-median and maximal cover. The first two are used for the location of undesirable (so-called "obnoxious") facilities that should be located far away from the demand nodes. The anti-center (also known as the maximin problem) seeks to find a facility location that maximizes the minimum distance from any demand node. The anti-median model maximizes the average or the total weighted travel distance between the facility and the demand nodes. The reader is referred to Erkut and Neuman (1989) for detailed discussions of these models and to Chap. 10 of this volume.

Berman et al. (2003b) introduce the *maximum probability absolute r anti-center* (MPArAC), which they refer to as the "probabilistic maximin problem" with the objective of finding a point x_1 that maximizes the probability of the minimum weighted distance to the nodal points exceeding a threshold value r. That is, x_1 is an optimal solution to the model

$$\max_{x \in G} P\left(\min_{i \in N} \{h_i d(i, x)\} \geq r \right) \tag{11.8}$$

In a similar vein, Berman and Wang (2006) define the *maximum probability absolute R anti-median* (MPARAM) as a point y_1 that achieves

$$\max_{x \in G} P\left(\sum_{i \in N} h_i d(i, x) \geq R\right), \tag{11.9}$$

i.e., the probability that the total weighted distance to the nodal points is no less than R is maximized at y_1.

The deterministic maximal covering location problem (Church and Meadows 1978) seeks to locate a facility so that as much demand as possible is "covered." A node $i \in N$ is considered covered by a facility at x if it is within a pre-defined *coverage radius* c_i of x, i.e. if $d(i, x) \leq c_i$. Given a point $x \in G$, let the coverage set $\mathrm{COV}(x) = \{i \in N | d(i, x) \leq c_i\}$ be the set of all nodes covered from x. Assuming that the weights are random variables, Berman and Wang (2008a) introduce the *probabilistic maximal covering problem* (PMCP) with the objective of maximizing the probability that the total weight of the covered nodes is greater than or equal to a pre-selected threshold value T, i.e.,

$$\max_{x \in G} P\left(\sum_{i \in COV(x)} h_i \geq T\right) \tag{11.10}$$

Together with the three problems discussed in Sect. 11.2, we now have six different location models with random node weights. The solution approaches to these models under various assumptions are discussed in the following sections. The next two sections do not consider the MVAM model; some results for this model will be presented in Sect. 11.3.4.

11.3.2 Special Cases: Bounds on Aspiration Levels

In general, the location problems defined above are difficult to solve, even for a single facility on a network. However, under some conditions optimal solutions can be easily identified. In this section, we assume that the node weight distributions are bounded, i.e., that for any $i \in N$ there exist finite numbers a_i, b_i, $0 \leq a_i \leq b_i$, such that $P(a_i \leq h_i \leq b_i) = 1$. For discrete distributions a_i and b_i can be taken as the smallest and largest realizations, respectively.

Denote by $\hat{x}_b$ and $\hat{x}_a$ the deterministic absolute center of the network when $h_i = b_i$ and $h_i = a_i$ for all $i \in N$, respectively. Berman et al. (2003b) and Berman and Wang (2007a) observe that if $r \geq \max_{i \in N}\{b_i d(i, \hat{x}_b)\}$, then $\hat{x}_b$ is also the MPArC of the network since $P(\max_{i \in N}\{h_i d(i, \hat{x}_b)\} > r) = 0$. Similarly, suppose $r \leq \max_{i \in N}\{a_i d(i, \hat{x}_a)\}$. Then $P(\max_{i \in N}\{h_i d(i, x)\} > r) = 1$ holds for any $x \in G$, implying that any point is an equally poor solution for the maximum probabilistic absolute center model. Now let $\tilde{x}_b$ and $\tilde{x}_a$ be the respective optimal solutions to the

deterministic maximin (anti-center) problem when $h_i = b_i$ and $h_i = a_i$ for $i \in N$. Similar arguments to above show that $\tilde{x}_a$ is the MPArAC if $r \leq \min_{i \in N}\{a_i d(i, \tilde{x}_a)\}$, while all points are equally bad when $r \geq \min_{i \in N}\{b_i d(i, \tilde{x}_b)\}$.

The same reasoning can be developed for the probabilistic median and anti-median. Indeed, let $\hat{y}_b$ ($\tilde{y}_b$) and $\hat{y}_a$ ($\tilde{y}_a$) be, respectively, the deterministic median and anti-median of the network when $h_i = b_i$ and $h_i = a_i$ for $i \in N$. In Berman and Wang (2004, 2006) it is shown that (1) if $R \geq \sum_{i \in N} b_i d(i, \hat{y}_b)$, then $P(\sum_{i \in N} h_i d(i, \hat{y}_b) > R) = 0$ and thus $\hat{y}_b$ is the MPARM; and (2) if $R \leq \sum_{i \in N} a_i d(i, \tilde{y}_a)$, then $P(\sum_{i \in N} h_i d(i, \tilde{y}_a) \geq R) = 1$ and therefore $\tilde{y}_a$ is the MPARAM.

Moreover, the objective function value at any point $x \in G$ is 1 for the maximum probability absolute median problem if R is less than $\sum_{i \in N} a_i d(i, \hat{y}_a)$. Similarly, the objective function value at any point $x \in G$ is 0 for the maximum probability absolute anti-median problem if R is greater than $\sum_{i \in N} b_i d(i, \tilde{y}_b)$.

Finally, if $y^C(a)$, $y^C(b)$ are the solutions to the deterministic maximum covering problem when node weights for all $i \in N$ are equal to a_i and b_i respectively, then for $T \leq \sum_{i \in COV(y^C(a))} a_i$ we have $P(\sum_{i \in COV(y^C(a))} h_i \geq T) = 1$ and thus $y^C(a)$ solves the PMCP.

On the other hand, if $T > \sum_{i \in COV(y^C(b))} b_i$ then every solution to the probabilistic maximal covering problem is equally poor, achieving the objective function value of 0.

To summarize, recalling that the threshold values r, R and T in the models discussed above can be interpreted as the aspiration levels, the results presented in this section show that when the aspiration level is sufficiently easy to attain, the solution to the deterministic problem will be optimal for its probabilistic counterpart as well. On the other hand, if the aspiration level is too hard to attain, then all potential solutions will perform equally poorly in the probabilistic problem. Moreover, the sufficient conditions for these special cases are easy to check by evaluating the corresponding objective functions at the upper and lower bounds of the node weight distributions. These results thus provide upper and lower bounds for the 'reasonable' values of the aspiration levels. Methods for finding solutions corresponding to the values within these bounds are discussed next.

11.3.3 Localization Results for Problems with Stochastic Node Weights

As in the previous section, the results presented here apply to the aspiration-level models MPArC, MPArAC, MPARM, MPARAM, and PMCP. Unless stated other-

wise, we generally assume that the probability distributions of node weights are discrete and independent throughout this section. It is clear that under these conditions the joint distribution of node weights must also be discrete. For any link (p, q) of the network, the objective functions for the models (11.3), (11.8), (11.4), (11.9), and (11.10) are step-functions of the facility position x along the link. It follows that the optimum within a link can be found at one of the jump points of the objective function. It is also possible to develop dominance relationships between these jump points, allowing us to focus on a subset of potential solutions. The discussion below is based on Berman and Wang (2004, 2008a).

We illustrate this approach for the probabilistic center (MPArC) model (11.3). Suppose the weight h_i associated with node $i \in N$ is a discrete random variable with realizations $w_i[k]$, $k = 1, 2, \ldots, K_i$, where K_i is the number of realizations and the realizations are arranged in an increasing order, i.e., $w_i[k] < w_i[k+1]$ for $1 \le k \le K_i - 1$. Consider a link (p, q) and a location $x \in (p, q)$. Since the random weights are independent, as noted in Sect. 11.2.2, the objective function value at x can be computed as follows:

$$P\left(\max_{i \in N}\{h_i d(i,x)\} > r\right) = 1 - \prod_{i \in N} P(h_i d(i,x) \le r) \tag{11.11}$$

where $P(h_i d(i,x) \le r) = \sum_{k=1}^{K_i} P(h_i = w_i[k]) I\{w_i[k] d(i,x) \le r\}$ and $I\{\bullet\}$ is the indicator function.

Recall the definition of the primary regions in Sect. 11.2.3 and of the accompanying sets L' and R' of nodes to which the shortest path from within the primary region passes through the left and right endpoints of (p, q), respectively. Consider a primary region $[\hat{x}, \tilde{x}]$ and define x_{ik} as the jump point with respect to $w_i[k]$ if $\hat{x} \le x_{ik} \le \tilde{x}$ and

$$\begin{cases} w_i[k][d(i,p) + x_{ik}] = r & \text{if } i \in L' \\ w_i[k][d(i,q) + \ell_{pq} - x_{ik}] = r & \text{if } i \in R' \end{cases} \tag{11.12}$$

In other words, for the realization $w_i[k]$, the weighted distance $d(i, x_{ik}) w_i[k]$ equals the threshold r. Suppose now that $i \in R'$. Then for any $x \in (\hat{x}, x_{i1})$ and any $k \in \{1, \ldots, K_i\}$, we have $d(i,x) w_i[k] > d(i, x_{i1}) w_i[1] = r$, implying that $P(h_i d(i,x) \le r) = 0$. From (11.11) we see that the value of the objective function at x is 1.0, implying that x cannot possibly be the maximum probability absolute r center. Similar arguments rule out any $x \in (x_{i1}, \tilde{x})$ for $i \in L'$.

Let now $\hat{x}' = \max_{i \in R'}\{x_{i1}\}$ and $\tilde{x}' = \min_{i \in L'}\{x_{i1}\}$. As explained above, the objective function is equal to 1 for $x \in [\hat{x}, \hat{x}') \cup (\tilde{x}', \tilde{x})]$. Thus, if $\hat{x}' > \tilde{x}'$, the objective function is 1.0 at any point within primary region $[\hat{x}, \tilde{x}]$ and therefore the primary region

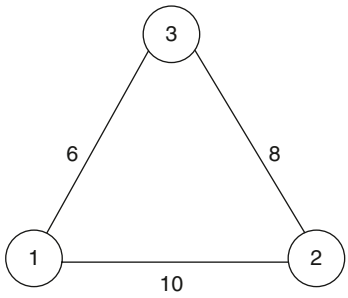


Fig. 11.1 Network for Example 2

does not contain an optimal solution. Now assume that $\hat{x}' \leq \tilde{x}'$. To find a local optimum, define the following sets of points:

$$J_1 = \{x_{ik} | \hat{x}' \leq x_{ik} \leq \tilde{x}', \quad i \in L', k = 1, 2, \ldots, K_i\},$$
$$J_2 = \{x_{ik} | \hat{x}' \leq x_{ik} \leq \tilde{x}', \quad i \in R', k = 1, 2, \ldots, K_i\}, \text{ and}$$
$$J_0 = J_1 \cup J_2 \cup \{\hat{x}'\}.$$

Berman and Wang (2004) prove that $x \in J_0$ if and only if x is a jump point of the objective function $P(\max_{i \in N}\{h_i d(i, x)\} > r)$. It follows that the best solution to MPArC problem (11.3) within the primary region $[\hat{x}, \tilde{x}]$ can be found within the set J_0, as in the following example.

Example 2: Consider the network depicted in Fig. 11.1. Suppose that $r = 30.0$ and probability distributions of the discrete random weights are given in Table 11.1. On link (1, 2), the antipode associated with node 3 is $x = 6.0$ (i.e., the point at distance 6 from node 1). Hence there are two primary regions [0, 6.0] and [6.0, 10.0]. In the primary region [0, 6.0] it is easy to verify that $L' = \{1, 3\}$ and $R' = \{2\}$. It follows that $x_{14} = 5.0$, $x_{22} = 2.5$, $x_{23} = 4.0$, $x_{32} = 4.0$, and $x_{33} = 1.5$ are valid jump points. Note that $\hat{x}' = \hat{x} = 0$ and $\tilde{x}' = \tilde{x} = 6.0$. We therefore have $J_1 = \{1.5, 4.0, 5.0\}$, $J_2 = \{2.5, 4.0\}$ and $J_0 = \{0, 1.5, 2.5, 4.0, 5.0\}$. The objective func-

Table 11.1 Probability distributions of random weights

	k_i	$w_i[k_i]$	$p_i[k_i]$
W_1	1	2	0.3
	2	3	0.2
	3	4	0.4
	4	6	0.1
W_2	1	1	0.5
	2	4	0.2
	3	5	0.1
	4	10	0.2
W_3	1	2	0.3
	2	3	0.45
	3	4	0.2
	4	6	0.05

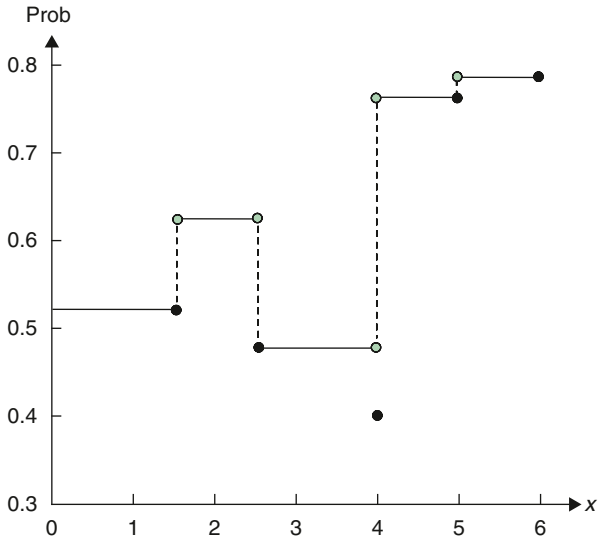


Fig. 11.2 Objective function over primary region [0, 6]

tion $P(\max_{i \in N}\{h_i d(i, x)\} > r)$ over the primary region [0, 6.0] is graphically depicted in Fig. 11.2. The optimum is reached at $x = 4$.

The procedure suggested above can be further improved by noting that many of the jump points in J_0 are dominated by other jump points and thus need not be considered. Berman and Wang (2004) demonstrate that the dominant jump points for the local optimum on the primary region $[\hat{x}, \tilde{x}]$ include $\hat{x}'$ defined earlier and any element in the set J_2, which (a) is also in the set J_1, (b) is the largest element in J_2, or (c) has a neighboring jump point to the right belonging to the set J_1 only. The reasoning behind this result is as follows. Consider $x = x_{ik} \in J_0$. If $x \in J_1$ and $x \notin J_2$, moving x by a small $\varepsilon > 0$ to the right will increase the objective function value. Similarly, if $x \in J_2$ and $x \notin J_1$, moving x by a small $\varepsilon > 0$ to the left will increase the objective function value. If $x \in J_1 \cap J_2$, moving x by a small $\varepsilon > 0$ either to the left or to the right will increase the objective function value. A local optimum on a primary region can thus be found by evaluating the objective function at only the dominant jump points. In the above example, the two dominant jump points are $x = 0$ and $x = 4.0$ with objective function values of 0.525 and 0.4, respectively. Once again we see that $x = 4.0$ is the local optimum.

The complexity of the solution procedure described above for finding a local optimum is $O(U^2n^2)$, where U is the maximum number of realizations of a random weight and n is the number of nodes. Readers are referred to Berman and Wang (2004) for a formal description of the algorithm. In order to find the global MPArC, this algorithm must be applied to all primary regions on every link. Since there are at most $n - 1$ primary regions on a link and ℓ links, in time $O(U^2n^3\ell)$ we can solve the maximum probability absolute r center problem.

A similar solution procedure, with the same worst-case complexity, can be applied to the probabilistic anti-center problem (11.8). In fact, the concepts of jump points and jump point dominance can also be applied to the probabilistic median and

anti-median models (11.4, 11.9). However, for these models the objective function involves the sum of the weighted distances and thus the jump points in Eq. (11.12) must be defined with respect to the joint, rather than marginal, distribution of node weights, see Berman and Wang (2004) for details. This implies that even the evaluation of the objective function at a given point is **NP**-hard, and thus the resulting solution procedure can only be applied to small-scale networks. As noted earlier, for networks with a large number of nodes, the sum of weighted distances can be effectively approximated by the Normal distribution due to the central limit theorem. The solution for models with Normal and other continuous weight distributions will be discussed in the next section.

For the probabilistic maximal covering problem (11.10) even stronger localization results were obtained by Berman and Wang (2008a), who prove that jump points of the objective function occur either at nodes or at *network intersect points*. The network intersect points associated with node i are all points x such that $d(i, x) = c_i$, where c_i is the coverage radius. Note that there are at most two such points along each link of the network, with one point corresponding to travel through either endpoint of the link. Thus, an optimal solution can be found by evaluating the objective function for, at most, $O(n\ell)$ points. We emphasize that this result holds irrespective of the distribution of node weights, so the assumptions that node weight distributions are both independent and discrete made elsewhere in this section are not required here.

Unfortunately, the evaluation of the objective function runs into the same difficulties as in the case of probabilistic median/anti-median problems, since the covering objective $P\left(\sum_{i\in COV(x)} h_i \geq T\right)$ also involves a sum of random variables. As noted earlier, evaluating this objective at a given point is **NP**-hard, implying that the exact solutions can be obtained only when $|COV(x)|$ is small for all $x \in G$. For the potential solution points with large $|COV(x)|$ it is necessary to use the Normal approximation. Computational experiments indicate that this approximation is very accurate for $|COV(x)| \geq 15$.

To summarize, the optimal solution is obtained by evaluating the objective function value for all jump points in the network; for jump points with $|COV(x)| < 15$, exact evaluation of the objective function (via convolution) can be used, while for points with $|COV(x)| \geq 15$, the Normal approximation is used. The jump point with the maximum value of the objective is returned as the optimal solution.

To illustrate the procedure, we refer back to the network in Example 2 to obtain the best solution on link (1, 2). Assume that the coverage radii are $c_1 = 4$, $c_2 = 5$, and $c_3 = 9$. Then the set of network intersect points on link (1, 2) is given by $NI = \{3.0, 4.0, 5.0, 9.0\}$, and the set of potential solutions on link (1, 2) is $NI \cup \{0, 10\}$ with the last two points corresponding to nodes 1 and 2, respectively. Suppose $T = 10$. For $x = 0$, the coverage set $COV(0) = \{1, 3\}$ and

$$P\left(\sum_{i\in COV(0)} h_i \geq T\right) = P(h_1 = 4)P(h_3 = 6) + P(h_1 = 6)P(h_3 = 4) + P(h_1 = 6)P(h_3 = 6) = 0.045.$$

Incidentally, the Normal approximation gives a value of $1 - \Phi(2.297) = 0.0108$ in this case. In a similar way, all other potential solutions can be evaluated. The highest objective function is achieved at $x = 9.0$ with $COV(9.0) = \{2, 3\}$ and an objective function value of $P\left(\sum_{i \in COV(9.0)} h_i \geq 10\right) = 0.215$, indicating that the optimal location on link (1, 2) is 9 units away from node 1. Observe that an alternative optimum is available at node 2. The Normal approximation of the objective function value at 9.0 is 0.188. Not surprisingly (since the cardinality of $COV(x)$ sets is small), the Normal approximation provides a relatively poor estimate of the objective function values in this example. On the other hand, the optimal location under the Normal approximation is still 9.0, indicating that it does lead to the right solution in this example.

11.3.4 Solution Approaches for General Probability Distributions of Node Weights

In this section, we consider five of the models discussed earlier: the probabilistic center, anti-center, median, anti-median, and minimum variance median. In all cases, the distributions of node weights are either continuous or general. Two principal research directions have been pursued for this case. The first one develops solution approaches when the node weights are assumed to be independent, paralleling the development in Sect. 11.3.3. The main results for this case were obtained by Berman and Wang (2006), Berman et al. (2003b), and Berman and Wang (2007). The second direction is concerned with cases, in which the distributions of node weights are not independent but follow a multivariate Normal distribution; this was investigated by Frank (1967).

11.3.4.1 Independent Node Weights

We start with the results for the probabilistic center/anti-center models for the case where the independence of node weights is assumed. Results for general probability distributions are available for this case because the evaluation of the objective function involves only marginal distributions of node weights unlike the models with median-type objectives that usually require joint distributions to evaluate.

The initial results for MPArC and MPArAC problems were obtained by Berman et al. (2003b) for the case when the node weights h_i, $i \in N$ are independent uniform random variables over $[a_i, b_i]$ with $0 \leq a_i \leq b_i$. Later, Berman and Wang (2007) extended the results to the case where the node weights follow arbitrary continuous probability distributions. The constants a_i, b_i represent the bounds of the probability distribution of h_i, if this distribution is bounded. In the case of unbounded distributions, $a_i = 0$ and $b_i = \infty$ should be used.

The common difficulty in deriving an expression for the objective function of center-type problems is the change (often a discontinuity) that occurs when the

shortest paths to some nodes shift as the potential facility location is varied over the link (p, q). In Sect. 11.3.3 these difficulties were overcome by restricting the analysis to the primary region of each link, where such shifts cannot occur. A similar strategy is followed here. For a given link (p, q) we first divide it into intervals over which the form of the objective function is invariant. The end points of such intervals are called break points. Then on each interval we find a local optimum and the best local optimum is returned as the global optimal solution. We illustrate this approach for the maximum probability absolute center problem.

Consider a primary region $[\hat{x}, \tilde{x}]$ on link $(p, q) \in L$. Similar to the definition of the jump points in the previous section, we note that the form of the objective function changes at a point x where for some node $i \in L'$, $a_i d(i, x) = r$ or $b_i d(i, x) = r$. This is because if the facility is located at $y \in (\hat{x}, x)$ and $d(i, x) = r/b_i$, then the threshold r cannot possibly be exceeded by node i, and thus this node will not be present in the expression for the objective function. On the other hand, if $y \in (x, \tilde{x})$ and $d(i, x) = r/a_i$, then the objective function is equal to 1, again changing the form. We call such points "break points."

Denote by S the collection of all break points in $[\hat{x}, \tilde{x}]$. Berman and Wang (2007) showed that $S = \overline{L}_1 \cup \overline{L}_2 \cup \overline{R}_1 \cup \overline{R}_2 \cup \{\hat{x}, \tilde{x}\}$, where

$$\overline{L}_1 = \left\{ x | \hat{x} \le x = \frac{r}{a_i} - d(i, p) \le \tilde{x}, \quad i \in L' \text{ and } a_i \ne 0 \right\},$$

$$\overline{L}_2 = \left\{ x | \hat{x} \le x = \frac{r}{b_i} - d(i, p) \le \tilde{x}, \quad i \in L' \text{ and } b_i \ne \infty \right\},$$

$$\overline{R}_1 = \left\{ x | \hat{x} \le x = d(i, q) + \ell_{pq} - \frac{r}{a_i} \le \tilde{x}, \quad i \in R' \text{ and } a_i \ne 0 \right\},$$

$$\overline{R}_1 = \left\{ x | \hat{x} \le x = d(i, q) + \ell_{pq} - \frac{r}{b_i} \le \tilde{x}, \quad i \in R' \text{ and } b_i \ne \infty \right\}.$$

Let s_m and s_{m+1} be two consecutive break points in S. If there exists $i \in L'$ such that $a_i(d(i, p) + s_m) \ge r$ or $i \in R'$ such that $a_i(d(i, q) + \ell_{pq} - s_{m+1}) \ge r$, then the objective function $P(\max_{i \in N}\{h_i d(i, x)\} \ge r) = 1$ at any $x \in [s_m, s_{m+1}]$ and the interval $[s_m, s_{m+1}]$ can be excluded from consideration. Assume that $P(\max_{i \in N}\{h_i d(i, x)\} \ge r) < 1$ at any $x \in [s_m, s_{m+1}]$. Define

$$A = \{i \in L' | a_i(d(i, p) + s_{m+1}) \le r \quad \text{and} \quad b_i(d(i, p) + s_m) \ge r\},$$

and

$$B = \{i \in R' | a_i(d(i, q) + \ell_{pq} - s_m) \le r \quad \text{and} \quad b_q(d(i, q) + \ell_{pq} - s_{m+1}) \ge r\}.$$

Denote by $Q_i(\bullet)$ the probability distribution function of random weight h_i. It is clear from (11.11) that the objective function over the interval $[s_m, s_{m+1}]$ can be written as

$$P\left(\max_{i\in N}\{h_i d(i,x)\} > r\right) = 1 - \prod_{i\in A} Q_i\left(\frac{r}{d(i,p)+x}\right) \times \prod_{i\in B} Q_i\left(\frac{r}{d(i,q)+\ell_{pq}-x}\right), \quad s_m \le x \le s_{m+1} \tag{11.13}$$

Minimizing (11.13) is equivalent to maximizing $\ln\left[\prod_{i\in A} Q_i\left(\frac{r}{d(i,p)+x}\right)\prod_{i\in B} Q_i\left(\frac{r}{d(i,q)+\ell_{pq}-x}\right)\right]$, and is done via a line search procedure.

For the case where the density of the distribution function (the derivative) is available in closed form, it is also useful to define

$$M_1(x) = \sum_{i\in A} \frac{d}{dx} \ln\left(Q_i\left(\frac{r}{d(i,\,p)+x}\right)\right)$$

and

$$M_2(x) = \sum_{i\in B} \frac{d}{dx} \ln\left(Q_i\left(\frac{r}{d(i,\,q)+\ell_{pq}-x}\right)\right)$$

If $M_1(x) + M_2(x)$ is increasing in x, the objective function has no stationary point inside the segment and therefore either s_m or s_{m+1} is optimal; we can simply compute and compare the objective values at these two points. If $M_1(x) + M_2(x)$ is decreasing in x, then the objective function is unimodal on the segment and a dichotomous search method can be used to find the optimal solution; see, e.g., Bazaraa et al. (1993). If neither of the above two conditions is met, a line search method can be applied. The procedure is illustrated by the following example.

Example 3: Refer back to the network in Fig. 11.1 and Example 2, but now assume that the weights h_1, h_2 and h_3 are uniformly distributed over the intervals [2.0, 8.0], [4.0, 5.0] and [3.0, 5.0], respectively. Let $r = 36.0$. Consider the primary region [0, 6.0] on link (1, 2). Recall from Example 2 that $L' = \{1, 3\}$ and $R' = \{2\}$. It is easy to verify that $\overline{L}_1 = \{6.0\}$, $\overline{L}_2 = \{1.2, 4.5\}$, $\overline{R}_1 = \{1.0\}$, $\overline{R}_2 = \{2.8\}$ and $S = \{0, 1, 1.2, 2.8, 4.5, 6.0\}$. Consider the segment given by the breakpoints $s_m = 1.2$ and $s_{m+1} = 2.8$. Since $b_1[d(1, 1) + s_m] < r$, $b_2[d(2, 2) + \ell_{12} - s_{m+1}] = r$, $a_2[d(2, 2) + \ell_{12} - s_m] < r$, $b_3[d(3, 1) + s_m] = r$ and $a_3[d(3, 1) + s_{m+1}] < r$ we have $A = \{3\}$, $B = \{2\}$; therefore, the objective function can be written as $1 - \dfrac{6(x-1)(6-x)}{(10-x)(6+x)}$. Similarly, we can derive the expressions for the objective function of other segments. The objective function over primary region [0, 6.0] is expressed as

$$P\left(\max_{i\in N}\{h_i d(i,x)\} > r\right) = \begin{cases} 1 & x \in [0, 1.0] \\ 1 - \dfrac{4(x-1)}{10-x} & x \in [1.0, 1.2] \\ 1 - \dfrac{6(x-1)(6-x)}{(10-x)(6+x)} & x \in [1.2, 2.8] \\ 1 - \dfrac{3(6-x)}{2(6+x)} & x \in [2.8, 4.5] \\ 1 - \dfrac{(6-x)(18-x)}{2x(6+x)} & x \in [4.5, 6.0]. \end{cases}$$

On segment [1.2, 2.8], $M_1(x) = \dfrac{d}{dx}\ln\left[\dfrac{3(6-x)}{2(6+x)}\right] = \dfrac{-8}{(6-x)(6+x)}$ and $M_2(x) = \dfrac{d}{dx}\ln\left[\dfrac{4(x-1)}{10-x}\right] = \dfrac{9}{(x-1)(10-x)}$. Since both $M_1(x)$ and $M_2(x)$ are decreasing functions of x, the objective function is unimodal on this segment and the golden section search technique (see, e.g., Bazaraa et al. 1993) is applied to find the minimum point. It turns out that $x = 2.8$ is optimal with an objective value 0.4545. The objective value is 1.0 at any point on segment [0, 1.0]. We can thus skip this segment because it cannot contain the global optimum. Local optimal solutions on other segments are presented in Table 11.2. It follows that $x = 2.8$ is the maximum probability absolute r center on primary region [0, 6.0].

To summarize, the methodology above reduces the computation of MPArC to (a) computation of the primary regions and the breakpoints, which can be done relatively efficiently, and (b) evaluation of the objective function over the subintervals defined by the breakpoints. The latter can, in principle, be accomplished via univariate line search. However, depending on the distributions of node weights, the objective function can be multi-modal even within these intervals; thus, only a local optimum can, in general, be guaranteed.

We next turn our attention to the probabilistic median/anti-median models. As noted earlier, the objective for these models involves a sum of random variables, requiring a convolution operation to evaluate, which is even more problematic in the continuous distribution case than in the discrete distribution case considered earlier. Here the exact results are only available for the cases when the weight distributions are Uniform or Normal: in these cases, the convolutions can be obtained in closed form and are thus relatively easy to evaluate. On the other hand, the Central

Table 11.2 Optimal solutions on the segments

Segment	Optimal solution	Optimal objective value
[1.0, 1.2]	1.2	0.9091
[2.8, 4.5]	2.8	0.4545
[4.5, 6.0]	4.5	0.7857

Limit Theorem ensures that Normal approximation of the distribution of the weighted sum will be quite good unless the number of nodes is very small. Thus, provided MPARM and MPARAM models can be solved efficiently for the case where node weights are Normally distributed, high-quality approximate solutions should be available for the more general distributions as well. The discussion below is based primarily on the results in Berman and Wang (2006).

When node weights are Normally distributed, the objective function of either model is unimodal in any primary region and therefore a local optimum can be obtained using the first-order conditions. This allows us to efficiently solve the problem (11.5) defined in Sect. 11.2.2. Similar approach works for the MPARAM case. An algorithm of order $O(n^5)$ is suggested for solving these models.

When the random weights are uniformly distributed, a closed form solution for the objective function is derived. A line search approach can then be applied for all links of the network to find local optima, and the best of them is an optimal solution on the entire network.

For the case of more general distributions, closed form expressions for the objective function are not available. There are two approximation approaches that can be used in this case. The first one is the Normal-based approximation discussed above and in Sect. 11.2.2. An alternative approach is to approximate node weight distributions with discrete probability distributions, and then apply the approach discussed in Sect. 11.3.3. One would expect the Normal-based approximation to perform better when the number of nodes is reasonably large, while the discrete-based approach may be better when the number of nodes is small.

Computational experiments comparing the two approaches were reported by Berman and Wang (2006). They show that the Normal approximation method outperforms the discrete approximation method in both CPU time and solution quality in the vast majority of cases; in fact, it is recommended for networks with $|N| \geq 5$. For very small networks with under five nodes, either a direct numerical evaluation of the convolutions or a discrete approximation can be used.

11.3.4.2 Correlated Node Weights

The analysis of the case where node weight distributions are correlated was carried out by Frank (1967). Our discussion is mostly based on his results. However, we extend his formulas to a general link (all of Frank's work was restricted to an isthmus) and correct some of his formulas. All results on correlated node weights assume that the joint distribution of weights is multivariate Normal. To the best of our knowledge, no results for more general distributions are available. We first present the results for the MPArC problem, followed by MPARM and MVAM. The results for MPArAC and MPARAM models are not available, but can likely be obtained along the same lines as for their MPArM and MPARM counterparts, respectively.

Suppose the random vector of node weights $\mathbf{H} = (h_1, h_2, \ldots, h_n)'$ has an n-dimensional Normal distribution with the mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)'$ and variance-covariance matrix $\mathbf{V} = [\sigma_{ij}]_{n \times n}$. For a given location $x \in G$, de-

fine the matrix $\mathbf{D}(x) = [d_{ij}]_{n \times n}$ with $d_{ij} = 0$ if $i \neq j$ and $d_{ij} = d(i, x)$ otherwise. If point x is not a node, the n-dimensional random vector of the weighted distances $\overline{\mathbf{H}}(x) = (h_1 d(1, x), h_2 d(2, x), \ldots, h_n d(n, x))'$ is Normally distributed with mean vector $\mathbf{D}(x)\boldsymbol{\mu}$ and variance-covariance matrix $\mathbf{D}(x)\mathbf{V}\mathbf{D}(x)$. If point x is a node, say node k, the distribution of $\overline{\mathbf{H}}(x)$ is singular, i.e., its kth component is zero. However the probability distribution of weighted distances of interest $\overline{\mathbf{H}}_k(x) = (h_1 d(1, x), h_2 d(2, x), \ldots, h_{k-1} d(k-1, x), h_{k+1} d(k+1, x), \ldots, h_n d(n, x))'$ can be obtained without difficulty.

We start the discussion with the maximum probability absolute center problem. The objective value at any point x that is not a node can be computed as follows, with the objective value at a node being derived in a similar way.

$$P\left(\max_{i \in N}\{h_i d(i,x)\} > r\right) = 1 - \frac{(2\pi)^{-\frac{n}{2}}}{|\mathbf{D}(x)||\mathbf{V}|^{\frac{1}{2}}} \int_{-\infty}^{r} \cdots \int_{-\infty}^{r} \exp\{-1/2[\mathbf{u} - \mathbf{D}(x)\boldsymbol{\mu}]' \times \mathbf{D}(x)^{-1}\mathbf{V}^{-1}\mathbf{D}(x)^{-1}[\mathbf{u} - \mathbf{D}(x)\boldsymbol{\mu}]\} d\mathbf{u}.$$

Let $\mathbf{u} = \mathbf{D}(x)^{-1}\mathbf{z}$. Suppose x belongs to a primary region $[\hat{x}, \tilde{x}]$ on a link $(p, q) \in L$. The expression above can then be rewritten as

$$P\left(\max_{i \in N}\{h_i d(i,x)\} > r\right) = 1 - \frac{(2\pi)^{-\frac{n}{2}}}{|\mathbf{V}|^{1/2}} \times \int_{-\infty}^{g_1} \cdots \int_{-\infty}^{g_n} \exp\{-1/2[\mathbf{z} - \boldsymbol{\mu}]'\mathbf{V}^{-1}[\mathbf{z} - \boldsymbol{\mu}]\} d\mathbf{z}, \tag{11.14}$$

where

$$g_i = \begin{cases} \dfrac{r}{d(i,p) + x}, & \text{if } i \in L' \\ \dfrac{r}{d(i,q) + \ell_{pq} - x}, & \text{if } i \in R'. \end{cases}$$

Expression (11.14) can be evaluated via numerical computation and the local optimum can be found using a line search method over $[\hat{x}, \tilde{x}]$ such as the quadratic interpolation method; see, e.g., Bazaraa et al. (1993). The procedure can then be repeated for all other primary regions on (p, q) and then for all the links on the network to find the overall MPArC. Of course, in the absence of the results on the unimodality of (11.14) within a primary region, only a locally optimal solution can be guaranteed.

We next turn our attention to the probabilistic median problem. As above, we consider a primary region $[\hat{x}, \tilde{x}]$ within a link $(p, q) \in L$. It can be shown that the optimum solution within the primary region must occur either at $\hat{x}$, or $\tilde{x}$, or at the following internal point y, whenever the expressions below evaluate to $y \in [\hat{x}, \tilde{x}]$.

$$y = \frac{bc + af}{bf + ae}, \tag{11.15}$$

and

$$a = R - \sum_{i \in L'} d(i,p)\mu_i - \sum_{i \in R'} [d(i,q) + \ell_{pq}]\mu_i,$$

$$b = \sum_{i \in R'} \mu_i - \sum_{i \in L'} \mu_i,$$

$$c = \sum_{i \in L'} \sum_{j \in L'} d(i,p)d(j,p)\sigma_{ij} + \sum_{i \in R'} \sum_{j \in R'} [d(i,q) + \ell_{pq}][d(j,q) + \ell_{pq}]\sigma_{ij}$$
$$+ 2 \sum_{i \in L'} \sum_{j \in R'} d(i,p)[d(j,q) + \ell_{pq}]\sigma_{ij},$$

$$e = \sum_{i \in R'} \sum_{j \in L'} \sigma_{ij} + \sum_{i \in R'} \sum_{j \in R'} \sigma_{ij} - 2 \sum_{j \in L'} \sum_{i \in R'} \sigma_{ij}, \text{ and}$$

$$f = te.$$

Observe that the expression (11.15) does not require computation of multivariate Normal densities and thus can be evaluated without difficulty. The optimal MPARM solution on a network is thus surprisingly easy to find even in the case of correlated node weights.

The final model addressed by Frank (1967) is the minimum variance median. Once again, we consider a primary region $[\hat{x}, \tilde{x}]$ on a link $(p, q) \in L$. The MVAM solution within this primary region is a point x defined by

$$x = \begin{cases} \hat{x}, & \text{if } t \leq \hat{x} \\ t, & \text{if } \hat{x} < t < \tilde{x} \\ \tilde{x}, & \text{if } t \geq \tilde{x} \end{cases},$$

$$\text{where } t = \left\{ \sum_{i \in R'} \sum_{j \in R'} [d(i,q) + \ell_{pq}]\sigma_{ij} + \sum_{i \in R'} \sum_{j \in L'} d(j,p)\,\sigma_{ij} \right\} \Big/ \left[\sum_{i \in L'} \sum_{j \in L'} \sigma_{ij} + \sum_{i \in R'} \sum_{j \in R'} \sigma_{ij} - 2 \sum_{j \in L'} \sum_{i \in R'} \sigma_{ij} \right]$$
$$- \left\{ \sum_{i \in L'} \sum_{j \in L'} d(i,p)\,\sigma_{ij} + \sum_{i \in L'} \sum_{j \in R'} [d(j,q) + \ell_{pq}]\sigma_{ij} \right\} \Big/ \left[\sum_{i \in L'} \sum_{j \in L'} \sigma_{ij} + \sum_{i \in R'} \sum_{j \in R'} \sigma_{ij} - 2 \sum_{j \in L'} \sum_{i \in R'} \sigma_{ij} \right].$$

To close this section, note that the question of whether the node weights are correlated or not is quite important; assuming independence of node weights where such

an assumption is not warranted may lead to wide departures from optimality. We illustrate this point with the following example.

Example 4: Consider again the single-link network of Example 1 and assume that the random weights associated with the two nodes A and B follow a joint Normal distribution with mean vector $\boldsymbol{\mu} = (15, 12)'$ and variance-covariance matrix

$$\mathbf{V} = \begin{bmatrix} 1.33 & 2.33 \\ 2.33 & 8.33 \end{bmatrix}$$

with a correlation coefficient of 0.7. Let $R = 160$.

Since the link (A, B) is clearly an isthmus and thus (as discussed earlier) consists of a single primary region, the optimal MPARM solution must be one of the endpoints or the internal point given by (11.15). We evaluate the latter expression, computing $a = 40$, $b = -3$, $c = 833.33$, $e = 5.00$, $t = 12.00$, $f = 60.00$, and $y = -4.99$. Since this y value does not result in an internal point of (A, B), the solution must be at one of the endpoints. The z-value is 1.38 at node A ($x = 0$) and 0.86 at node B ($x = 10$). Therefore, node A is the MPARM in this case with the objective value of 0.0829.

If the two random weights are independent, their covariance is 0, then (11.15) yields $a = 40$, $b = -3$, $c = 833.33$, $e = 9.67$, $t = 8.62$, $f = 83.33$, and $y = 6.1$. It is easy to check that y has lower objective value than either of the endpoints, and is the MPARM when the weights are independent. The objective function value at y is 0.0514, nearly 40% lower than in the previous case. The large gap between the optimal locations and objective function values in the two cases indicates the importance of the independence assumption.

We can also use the same example to illustrate the impact of different distributional assumptions. Assume that the weights of the two nodes are independent uniform random variables over intervals [13, 17] and [7, 17], respectively. It is easy to verify that $E(h_A) = 15$, $\sigma^2(h_A) = 1.33$, $E(h_B) = 12$, and $\sigma^2(h_B) = 8.33$, yielding the same values as for the Normally distributed weights considered earlier. As discussed above, for the case of independent uniform weights the objective function of MPARM can be computed in closed form, see Berman and Wang (2006) for details. Using the quadratic interpolation method for the resulting expression, we obtain the MPARM solution at $x = 5.0$, quite far from the optimal location under the independent Normal probability distributions, even though the random weights have the same mean and variance under the two distributions. The large gap between the two solutions is not unexpected as the network is quite small, and thus the Normal approximation is not expected to work well in this case.

11.3.5 Summary and Open Problems

The results presented in this (and the previous) section indicate that while much progress has been achieved in extending Frank's original results and providing ef-

ficient computational approaches, many open problems remain and deserve further investigation.

In the single-facility case, the minimum variance absolute median problem has not been analyzed as thoroughly as the other models. This is likely due to the obvious flaws in this model: it appears odd to look for a location optimizing the second moment (the variance), without imposing any restrictions on the first moment (the expected value). This has, no doubt, limited the interest in this model. Nevertheless, the idea of limiting risk through minimizing variance is sound and has been used in a variety of fields from statistical quality control to finance. However, a more reasonable model must incorporate both the expected value and the variance either by attaching a weight to each of these components in the objective function (leading to a mean-variance trade-off model) or constraining the value of one component (e.g., the expected value) while optimizing the other (e.g., the variance). Further investigation of such models could lead to interesting insights.

The difference between the maximum probability absolute median or center and their absolute expected counterparts also deserves further investigation. In particular, when an aspirational objective is used, can bounds be provided on the departures from optimality in the expected sense? Can the acceptable levels of such departures be specified as bounds? Can combined objectives (such as a weighted combination of maximum probability center and absolute expected center) be handled?

Relatively few results have been obtained for the multiple-facility extensions of the models discussed above, possibly because the analysis is substantially more difficult than for the single-facility case. To our knowledge, the only available results are by Berman and Wang (2008b, 2010), where a maximum probability absolute R m-median (MPARmM) and a maximum probability absolute r m-center (MPArmC), (with $m > 1$), are defined as the respective extensions of MPARM and MPArC models. For both models the authors identify solvable special cases similar to those described in Sect. 11.3.2. For the more general cases, as discussed in Sect. 11.3.3, the evaluation of the objective function of the MPARM problem at a given point is **NP**-hard; this conclusion easily extends to the multi-facility case as well. In view of this, the analysis by Berman and Wang (2010) limits the potential location sites for MPARmM to the set of nodes and assumes that node weights have independent discrete probability distributions. The MPARmM problem is formulated as an integer programming model, which can be solved only for small instances. Several heuristics and a Normal approximation method have been developed, and computational results suggest that the Normal approximation method is the only viable solution procedure, even for small problems.

Berman and Wang (2008b) consider the MPArmC problem with independent random node weights. Unlike the MPArC model, the MPArmC problem is **NP**-hard. It is also shown that at least one optimal solution consists of optimal points for the MPArC problem on subnetworks and nodal points, namely the dominant points. This implies that an optimal solution can be found by examining all subsets of cardinality m of the set of dominant points J'. If the probability distributions of node weights are independent and continuous, it is generally difficult to construct the set of dominant points *a priori*. A suggested approach is to approximate node weight

distributions with discrete probability distributions and then identify the dominant points by treating the node weights as discrete random variables. Sufficient conditions are derived for the case when the set of dominant points consists of all the nodal points, the antipodes with respect to any node, and the break points defined in Sect. 11.3.4, so no approximation is required. An exact solution procedure and some heuristics are proposed to search for the best solution in the set J'. However, it remains clear that further study of multi-facility extensions of problems discussed in the previous sections is required.

Finally, we note that while Frank's original work and all of the results described above have dealt with the network topology, the concepts of probabilistic centers, anti-centers, medians, anti-medians, and maximal cover extend easily to the planar topology as well. The only work to investigate the related issues on the plane appears to be Berman et al. (2003a), who study the maximum probability absolute center problem, which they refer to it as the probabilistic minimax problem, on the plane when random weights associated with demand points are uniformly distributed. They show that an optimal solution exists in the convex hull of the demand points. They also prove that the problem is equivalent to minimizing a convex function in a convex region and thus can be solved using a steepest descent approach when certain conditions on the weight distributions are satisfied. In case the conditions do not hold, an alternative branch and bound procedure is suggested. The authors report that problems with 100 demand points took negligible time to solve in Excel Solver. The investigation of the remaining objectives on the plane (as well as MPArC objective under more general distributional assumptions), remains to be done.

11.4 Extending Frank's Work

Arguably, the main contribution of Frank's work was not in the analysis of the specific location models discussed earlier in this chapter, but in the general motivation his work provided for the analysis of stochastic elements in facility location problems. This field came into its own in mid-1980s and continues to be quite active, having produced several hundred publications over twenty five years.

Frank's work focused rather narrowly on one aspect of potential uncertainty in location models: the node weights. However, most classical location objectives, including median and center objectives, involve three components: node weights (representing demand for service), travel time between customers' and facility locations, and the assignment rule of customers to facilities. Thus, a natural continuation of Frank's work is the analysis of stochasticity in travel times (or distances) of links on a network; this work is reviewed in Sect. 11.4.1.

Moreover, interpreting the "travel time" component more generally as the customer-facility interaction, several other sources of stochasticity arise: (1) the customer may be delayed in obtaining service due to congestion at the facility in cases where customers travel to facilities to obtain service, known as the "immobile

server" case, (2) the customer may be delayed in getting service because mobile servers are occupied elsewhere, in the "mobile server" case where servers travel to customers to provide service, and (3) the closest facility may not be the one providing service since, due to congestion delays or server availability, it may be faster to obtain service from a more distant but less busy facility, which may occur in both mobile and immobile server cases. The location models that take congestion into account are reviewed in Sect. 11.4.2 and 11.4.3. We note that these models, in addition to determining facility location, also seek to determine the required service capacity of the facilities, which is an important issue facing decision-makers. This has accounted for the continuing interest and large number of publications regarding models of this type.

Finally, a facility may fail to provide service due to a temporary or permanent break-down, causing customers to seek service elsewhere, thus creating stochasticity in both the travel times and the customer-to-facility assignments. The models seeking to analyze systems where facility reliability may be less than perfect are relatively recent and are reviewed in Sect. 11.4.4.

11.4.1 Probabilistic Links

The uncertainty about the lengths of the links of a network often arises in practical applications, particularly when "lengths" are measured in units of travel time rather than geographical distances. This uncertainty arises due to factors such as changes in traffic patterns during the day, car accidents, changes in weather conditions, and other unforeseen occurrences. Mirchandani and Odoni (1979) studied the p-median problem when travel times on the links are discrete random variables. They assume that the network can be in a finite number of states, where each state is a snapshot of the network and the probability of each state is either given or can be calculated from the probability distributions of the lengths of links, which are assumed to be known. Under the link homogeneity assumption—the time required to travel a fraction q of any link $(i, j) \in L$ for any state r is equal to q times the length of link (i, j) in state r—an optimal set of locations is proved to be nodal. Subsequent work (see Berman et al. 1990 and references therein) studied the p-median problem with probabilistic lengths when states of the network change according to a Markov chain and facilities (servers) can be moved at a cost in response to the change of states. The objective function is to minimize a weighted function of demand travel times and relocation costs.

The median problem with links that are continuous random variables was studied by Handler and Mirchandani (1979). They formulated the problem for locating p facilities and provided an algorithm for the 1-median problem. For more details and references the readers can refer to Chap. 1 of the book by Mirchandani and Francis (1990) for the p-median problem with probabilistic discrete lengths, and to Chap. 12 of the same book for the problem with continuous probabilistic links.

11.4.2 Location Models with Facility Congestions

As noted earlier, in this section we describe models which account for possible service delays due to congestion. We classify location models with congested facilities into two types: facilities with mobile servers that travel to customers to provide service, (typically emergency service systems such as fire, ambulance, and police); and facilities with immobile servers where customers travel to the facilities to obtain service (such as retail stores, hospitals, and banks). From a methodological point of view, these two types of models are quite different, with the former being significantly more complex since they include several travel time components: outbound travel to customer's location and travel back to the facility (some models allow for direct customer-to-customer travel) in addition to the on-scene service time. Moreover, dispatching rules must be specified, i.e., rules that determine which server is assigned to respond to a particular customer request for service. The literature can be further subdivided into two streams: models with median-type objective that represents some overall system-wide service time, and models with covering objectives that seek to provide an acceptable level of service to all customers with the minimal amount of resources, including servers and facilities. Both streams are reviewed in this section, while models with immobile servers are reviewed in Sect. 11.4.3.

11.4.2.1 Congested Facilities with Mobile Servers: The Median Objective

The following discussion is based on the material covered in the review papers by Berman et al. (1990) and Berman and Krass (2002), and the references therein.

Berman et al. (1985) studied the problem of locating a single-server facility, operating either as a $M/G/1/\infty$ (Poisson arrival of customers, general service time, one facility and all customer calls are queued) or as an $M/G/1/0$ (customer calls that arrive while the server is busy are lost). An important assumption of both models is that the server returns to the facility upon completion of each service and consequently the service times are *iid*. The objective function is the expected response time for the $M/G/1/\infty$, calculated by adding the expected travel time and the expected waiting time. For the $M/G/1/0$ the expected waiting time is replaced with the expected cost of dispatching a special reserve unit. For this model Berman et al. (1985) proved that the optimal solution is identical to that of the standard 1-median problem; for the $M/G/1/\infty$, they proved that the objective function is strictly convex on the portion of the link between two consecutive antipodes that they call a primary region. Therefore, the search for the optimal solution, called the stochastic queuing median (SQM) of a network is reduced to a finite set of locations, which includes nodes plus local minima within the primary regions. Many extensions of the problem with one facility and a single server are included in Berman et al. (1990).

In subsequent research, the $M/G/1/\infty$ was generalized to include $K > 1$ servers in a facility using an approximation for the expected waiting time, and the results of the $M/G/1/0$ model were generalized to include K servers stationed at a single facility.

For the problem of locating K congested facilities we distinguish between two cases: no-cooperation, where customers must always be served from the same facility, and co-operation, where customer calls may be served by any facility and thus a farther-away facility with a free server may provide service when the closest facility has no available servers. Berman and Mandowsky (1986) studied the problem of locating K facilities, each containing a single mobile server, for the no-cooperation case. The problem is solved by using a location/allocation scheme where the allocation part (given location of the facilities, how to allocate the customers to the various facilities) was based on an allocation model that finds the optimal territories for two facilities that do not cooperate. The same problem, but with facility cooperation allowed, was investigated by Berman et al. (1987), who developed a heuristic that uses the Hypercube Model of Larson (1974) to approximate the system as an $M/M/K$ queue with distinguishable servers. The approach to solve the problem is based on the following idea: given K locations, the Hypercube model provides K dispatching zones, such that for each zone, an optimal location of a single server can be found by using the standard 1-median or the *SQM*.

When relaxing the assumption that service units always return to the home location following the completion of service, the problem is very difficult since the service times are no longer *iid*. Berman and Vasudeva (2005) studied the problem when service units return to the home locations only if no calls are waiting, otherwise they travel from call to call to provide the service. They use queuing approximations for their model.

11.4.2.2 The Covering Objective

In models with the covering objective, two service level constraints are typically defined: in order to be covered, a customer must have a facility within the coverage radius, and customer calls must find an available server at least α% of the time, where $\alpha \in [0, 100]$ is an externally specified parameter (alternative specifications of this constraint may include setting limits on the expected waiting time). See Berman and Krass (2002) for further discussion of models of this type.

The study of covering models with congestion originated with Daskin (1983), who introduced a model assuming that the busy fraction for servers who operate independently is an externally specified parameter. Batta et al. (1989) relax the assumption that servers operate independently by incorporating the approach adopted by Larson (1975) for the Hypercube Model. ReVelle and Hogan (1989a, b) used region-based estimates for p. Marianov and ReVelle (1994, 1996) used the $M/M/K/K$ loss system to estimate node availability. Note that the models in the papers by ReVelle and Hogan (1989a, b) and Marianov and ReVelle (1994, 1996) do not ensure system feasibility, see Baron et al. (2008) for further discussion.

Ball and Lin (1993) introduce a model that ensures system availability, but with unrealistic number of servers. They assume that service times are deterministic and derive lower bounds for server availability. Borras and Pastor (2002) use simulation to examine ex-post the availability level of several known models. Recently,

Baron et al. (2008) showed that earlier models often overestimate servers' availability and thus may result in infeasible solutions. By analyzing the underlying partially accessible queuing system, they develop lower bounds on system availabilities that are used in two new models for which feasibility is guaranteed.

11.4.3 Congested Facilities with Immobile Servers

In these models, customers travel to facilities to obtain services and it is assumed that service delays may occur at the facilities due to congestion. Most of work on this class of models has used the covering objective.

Marianov and Serra (1998) introduced the problem of locating K facilities using the maximal weight cover form. The problem is to maximize the total demand covered by the facilities subject to the service level constraint, which ensures that for any facility the probability that the waiting time is less than or equal to a predetermined level is bounded from below. They considered two versions of the problem: in one, they assume that each facility behaves as a $M/M/1$ queuing system; in the other one they assume that each facility behaves as a $M/M/k$ queuing system. For both problems they show that the service level constraint can be linearized, resulting in a problem that can be formulated as a linear integer program. Marianov and Rios (2000) apply this methodology to find the locations of ATM machines where their number is also a decision variable.

Wang et al. (2002) studied the problem of locating facilities operating as $M/M/1$ queuing systems with the median-type objective of minimizing the total cost of the system, which is the sum of the expected waiting and travel costs. They incorporated service level constraints on the expected waiting time in all facilities. Berman et al. (2006) considered a similar model with the coverage-type objective function of minimizing the number of facilities, subject to an additional constraint on the demand that is lost due to congestion and insufficient coverage. Berman and Drezner (2007) generalized the queuing system of Wang et al. (2002) by modeling the facilities as an $M/M/K$ queuing system. In addition to finding the optimal location and number of facilities, they also showed how to allocate the servers among the facilities. Aboolian et al. (2008a) examined the same problem but with an objective that minimizes the maximum sum of travel and waiting time costs. Aboolian et al. (2008b) generalized the results of Berman and Drezner (2007) by including in the objective function in addition to the expected travel and waiting cost, also the fixed cost of opening facilities and the variable cost of the servers.

Castillo et al. (2009) studied a problem similar to that of Aboolian et al. (2008b). In their model, there is a centralized authority that determines the assignment of customers to the facilities (whereas Aboolian et al. (2008b) assign customers to the closest facility). Also, Castillo et al. (2009) use approximations for the expected waiting time. Finally, Baron et al. (2008) analyzed the problem of determining the number, location and capacity of congested facilities under general assumptions

such as spatially continuous demand, general server arrival and service distributions, and nonlinear costs.

11.4.4 Unreliable Facilities

In this class of models facilities may fail to provide service due to breakdowns, forcing customers to seek service elsewhere. A crucial assumption is whether customers have the knowledge about the state of the facility before they start their trip. If such an advanced knowledge is available, the customers are assumed to travel directly to the closest operating facility. When prior knowledge about the state of the facilities is not available, the customers must search for an operating facility.

The problem with advanced knowledge about the state of the facilities was analyzed by Berman et al. (2007b). The paper includes structural results, analysis of the model behavior and several exact and heuristic approaches. Drezner (1987) addressed a similar problem and suggested a heuristic. Lee (2001) proposed a different heuristic to Drezner's problem for the problem in the plane. Synder and Daskin (2006) developed several models that are similar to the work by Berman et al. (2007b) focussing on the development of Lagrangian relaxation algorithm.

The problem without the advanced information on the status of the facilities was introduced by Berman et al. (2009). They focused on studying the effects of reliability and information on the properties of the optimal solution. When the failure of facilities to receive customers for service is due to congestion at the facilities, Berman et al. (2007a) considered the problem studied by the earlier Berman et al. (2009) using queuing approximations. We note that both of these models assume that customers' search strategy consists of always traveling to the closest unexamined facility, which is not necessarily an optimal strategy. The problem of optimally locating the facilities when customers search in an optimal manner remains open.

11.5 Conclusion

In this chapter we described the pioneering work of H. Frank who played a key role in introducing stochastic models to location analysis. While Frank's interest was limited to just one aspect of uncertainty in location models—namely the uncertainty related to node weights—his research served as a springboard to broader research into various aspects of stochasticity in location models, as described in Sect. 11.4. Many of the ideas introduced by Frank, including maximizing the probability that a certain constraint is satisfied rather than some deterministic objective, have been applied in many other contexts within the field of location analysis and elsewhere. Stochastic location modeling is an active and exciting field where many important problems remain open. This ranges from questions related to the extensions of

Frank's original models, many of which were described in Sect. 11.3.5, to constructing a tractable combined model that would represent uncertainties related to node weight (demand), service, and travel times simultaneously.

References

Aboolian R, Berman O, Drezner Z (2008a) The multiple server center location problem. Ann Oper Res, 16:337–352

Aboolian R, Berman O, Drezner Z (2008b) Location and allocation of service units on a congested network. IIE Trans 40:422–433

Ball M, Lin F (1993) A reliability model applied to emergency service vehicle location. Oper Res 41:18–36

Baron O, Berman O, Krass D (2008) Facility location with stochastic demand and constraints on waiting time. Manuf Serv Oper Manage 10:484–505

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

Bazaraa MS, Sherali HD, Shetty CM (1993) Nonlinear programming: Theory and algorithms (2nd ed). Wiley, New York

Berman O, Drezner Z (2007) The multiple server location problem. J Oper Res Soc 58:91–99

Berman O, Krass D (2002) Facility location problems with stochastic demands and congestion. In: Drezner Z, Hamacher HW (eds) Location analysis: Applications and theory, pp. 329–371

Berman O, Mandowsky RR (1986) Location-allocation on congested networks. Eur J Oper Res 26:238–250

Berman O, Vasudeva S (2005)Approximate performance measures for public services. IEEE Trans Syst Man Cybern A Syst Hum 35:583–591

Berman O, Wang J (2004) Probabilistic location problems with discrete demand weights. Networks 44:47–57

Berman O, Wang J (2006) The 1-median and 1-antimedian problems with continuous probabilistic demand weights. INFOR 44:267–283

Berman O, Wang J (2007) The 1-minimax and 1-maximin problems with demand weights of general probability distributions. Networks 50:127–135

Berman O, Wang J (2008a) The probabilistic 1-maximal covering problem on a network with discrete demand weights. J Oper Res Soc 59:1398–1405

Berman O, Wang J (2008b) The probabilistic *p*-minimax problem with random demand weights. Working paper, Rotman School of Management, University of Toronto

Berman O, Wang J (2010) The network *p*-median problem with discrete probabilities demand weights. Comput Oper Res 37:1455–1463

Berman O, Larson R, Parkan C (1987) The stochastic queue *p*-median problem. Transp Sci 21:207–216

Berman O, Chiu SS, Larson RC, Odoni AR, Batta R (1990) Location of mobile units in a stochastic environment. In: Mirchandani PB, Francis RL (eds) Discrete location theory. Wiley Interscience Series in Discrete Mathematics and Optimization, pp. 503–549

Berman O, Drezner Z, Wesolowsky GO, Wang J (2003a) Probabilistic minimax location problem on the plane. Annuals Oper Res 122:59–70

Berman O, Wang J, Drezner Z, Wesolowsky G (2003b) The minimax and maximin location problems with uniform distributed weights. IIE Trans 35:1017–1025

Berman O, Krass D, Wang J (2006) Locating service facilities to reduce lost demand. IIE Trans 38:933–946

Berman O, Huang R, Kim S, Menezes MBC (2007a) Locating capacitated facilities to maximize captured demand. IIE Trans 39:105–1029

Berman O, Krass D, Menezes MBC (2007b) Reliability issues, strategic co-location and centralization in m-median problems. Oper Res 55:332–350
Berman O, Krass D, Menezes M (2009) Optimal location in the problem of disruptions and incomplete information. Decis Sci 40:845–868
Berman O, Larson RC, Chiu SS (1985) Optimal server location on a network operating as an M/G/1 queue. Oper Res 33:746–771
Borras F, Pastor J (2002) The ex-post evaluation of the minimum local reliability level: An enhanced probabilistic location set covering model. Ann Oper Res 111:51–74
Castillo I, Ingolfsson A, Sim T (2009) Socially optimal location of facilities with fixed servers, stochastic demand, and congestion. Prod Oper Manag 18:721–736
Church RL, Meadows ME (1978) Location modeling utilizing maximum service distance criteria. Geogr Anal 11:358–373
Daskin MS(1983) A maximum expected covering location model: formulation, properties and heuristic solution. Transp Sci 17:48–70
Drezner Z (1987) Heuristic solution methods for two location problems with unreliable facilities. J Oper Res Soc 38:509–514
Erkut E, Neuman S (1989) Analytical models for locating undesirable facilities. Eur J Oper Res 40:275–291
Frank H (1966) Optimum locations on a graph with probabilistic demands. Oper Res 14:409–421
Frank H (1967) Optimum locations on graphs with correlated normal demands. Oper Res 15:552–557
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) Optimum distribution of switching centers in a communication network and some related graph theoretic problems. Oper Res 13:462–475
Handler GY, Mirchandani PB (1979) Location on networks: theory and algorithm. The MIT Press, Cambridge, MA
Larson RC (1974) A hypercube queuing model for facility location and redistricting in urban emergency services. Comput Oper Res 1:67–95
Larson RC (1975) Approximating the performance of urban emergency service systems. Oper Res 23:845–868
Lee SD (2001) On solving unreliable planar location problem. Comput Oper Res 28:329–344
Marianov V, ReVelle C (1994) The queueing probabilistic location set covering problem and some extension. Socioecon Plann Sci 28:167–178
Marianov V, ReVelle C (1996) The queueing maximal availability location problem: a model for the siting of emergency vehicles. Eur J Oper Res 93:110–120
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 locating-allocation for congested systems. J Reg Sci 38:401–424
Mirchandani PB, Francis RL (eds) (1990) Discrete location theory. Wiley, New York
Mirchandani PB, Odoni AR (1979) Location of medians on stochastic networks. Transp Sci 13:85–97
ReVelle C, Hogan K (1989a) The maximum reliability location problem and α-reliable *p*-center problem: derivatives of the probabilistic location set covering problem. Ann Oper Res 18:155–174
ReVelle C, Hogan K (1989b) The maximum availability location problem. Transp Sci 23:192–200
Simon HA (1957) Models of man. Wiley, New York
Synder LV, Daskin MS (2006) Reliability models for facility location: the expected failure cost case. Transp Sci 39:400–416
Wang Q, Batta R, Rump CM (2002) Algorithms for a facility location problem with stochastic customer demand and immobile servers. In: Berman O, Krass D (eds) Recent developments in the theory and applications of location models Part II. Ann Oper Res 111, Kluwer Academic Publishers, pp. 17–34

Chapter 12
Hub Location Problems: The Location of Interacting Facilities

Bahar Y. Kara and Mehmet R. Taner

12.1 Introduction

O'Kelly's (1986) classical paper started a new research stream by identifying a connection between spatial interaction models and location theory. The traditional spatial interaction theory applies models of travel behavior to investigate demand patterns between fixed locations. Location theory, on the other hand, takes demand as given, assumes a simple view of travel behavior, and focuses on finding the best location for facilities.

Spatial interaction theory focuses on the problem of locating centers of special interest, and observes that the selected locations have an effect on the evolution of the associated network. O'Kelly's self-identified contribution in this context relates to the interaction effects between facility locations and spatial flows. He makes a distinction between endogenous and exogenous effects. In particular, he considers the given problem parameters as exogenous data, which are endogenously affected by the location of the hubs as well as the allocations. Hubs are special facilities acting as consolidation and dissemination points for the flows. Flows from the same origin with different destinations are consolidated enroute at a hub node where they are combined with flows from different origins with a common destination. The main idea is to keep the flow interactions in perspective at the design stage of the hub network. That is, the hubs need to be strategically located in view of their effects on the intensity and cost of the flow data. In general, the hub location problems are defined as analogous counterparts of the classical location problems with the addition of allocation decisions.

This chapter reviews and outlines the research on hub location problems that emerged as a new research stream led by O'Kelly's (1986) seminal paper. Section 12.2 discusses the geographical applications leading the way to the ideas proposed by O'Kelly. Section 12.3 summarizes the major findings presented in

B. Y. Kara (✉) · M. R. Taner
Department of Industrial Engineering, Bilkent University, Ankara, Turkey
e-mail: bkara@bilkent.edu.tr

M. R. Taner
e-mail: mrtaner@bilkent.edu.tr

H. A. Eiselt, V. Marianov (eds.), *Foundations of Location Analysis,*
International Series in Operations Research & Management Science 155,
DOI 10.1007/978-1-4419-7572-0_12,

O'Kelly's original paper. Prominent theoretical developments that emerged from these findings are discussed in Sect. 12.4. Section 12.5 reviews some related application oriented studies. Finally, Sect. 12.6 concludes the chapter with highlights of the current and future trends for research in the area.

12.2 Before Hub Location

The identification of the importance of consolidation and dissemination points as well as their endogenous effects was well known in spatial interaction theory before O'Kelly's work. For instance, the classical paper by Taaffe et al. (1963) discusses the issue in the context of formation of transportation infrastructure in third world countries. They observe that consolidation-dissemination points are located in administrative centers, political and military control centers, mineral exploitation areas, and areas of agricultural export production. Lines of penetration emerge between these points of demand concentration. Figure 12.1 illustrates a line of penetration between two fictitious centers of critical importance (centers A and B), resulting in indirect connections between the points previously connected to either one of these centers (i.e., points $a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_m$). Once such penetration lines are formed, they have an impact on both the surrounding area along these lines and the initial centers in terms of local development. These local developments are analogous to the endogenous attraction proposed by O'Kelly, and they in turn manifest themselves as a factor that further supports the structure of the penetration lines. Once the development in the centers and along the penetration lines stabilizes, the formation of the backbone is completed.

In classical spatial theory, there are also examples of active strategic development of the transportation backbones. For instance, Miehle (1958) constructs a mechanical model to simulate alternative backbone structures enforcing the passage of flows through certain designated locations functioning as hubs. Goodchild (1978) mathematically considers the role of endogenous attraction. He assumes fixed locations and solves only the allocation problem, where attraction to a facility is modeled as a function of both distance and usage. Distance is an exogenously given

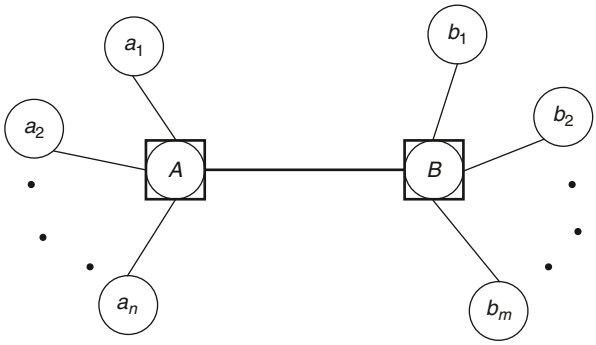


Fig. 12.1 Lines of penetration between centers A and B

factor, whereas attraction due to usage is endogenously or voluntarily determined based on the actual level of service that a facility provides.

Ducca and Wilson (1976) consider a similar problem in the context of the planned positioning of shopping centers. In their model, demand intensity is expressed as a simultaneous function of existing retail density, employment, and distances traveled. Allen and Sanglier (1979) develop a model of dynamically interacting urban centers. Each center has an associated attraction parameter, and once a center is located, the parameters of the original problem in terms of the demands generated at different locations are affected through immigration and emigration. In particular, there is a positive feedback loop due to the employment opportunities generated by a located center. In a subsequent paper, Allen and Sanglier (1981) improve their original model by also considering the negative feedback loop that reflects the crowding effect.

This brief discussion on inter-facility attraction shows that the fundamentals of the notion of endogenous attraction observed by O'Kelly date back to late 1950s. The formal definition of the problem in the context of location theory led to the development of a new field. The remainder of this chapter discusses this new field and ties its evolution back to O'Kelly's paper.

12.3 O'Kelly's Seminal Contribution

Genesis of location of interacting facilities as a new research area within location theory dates back to O'Kelly's paper, which was significantly impacted by the popular trend of simultaneous consideration of location and transportation decisions in spatial theory. In this paper, O'Kelly focuses on the interaction between hubs serving the United States inter-city air passenger streams, and studies the relevant data recorded in a Civil Aeronautics Board sample survey of 1970. He observes that, although airline companies in practice carefully consider the location of the hub facilities in view of their collective ability to efficiently connect the cities in their network, classical location research completely ignores these interactions. This perspective helps O'Kelly to identify a novel version of a location-allocation problem, in which the located facilities lie along the route between demand points.

O'Kelly studies both a single- and a two-hub version of this new problem in the 2-dimensional plane. The single-hub version is shown to be equivalent to the classical Weber least cost location problem. Regarding the economic advantages of building a single-hub network, O'Kelly points out that the only rational reason to justify such a system would be the potential savings in link costs due to the scale effects of routing the traffic through the hub. This issue in a problem with n demand points is mathematically expressed as

$$\sum_i \sum_j W_{ij} \left(C(p_i, Q) + C(Q, p_j) \right) + Kn < \sum_i \sum_j W_{ij} C(p_i, p_j) + {}^1\!/_2 n(n-1)K,$$

where the notation is defined as follows.

p_i: Demand point i, $i=1, 2, \ldots, n$
W_{ij}: Flow between demand points p_i and p_j, $i = 1, 2, ..., n,\ j=1, 2, \ldots, n$
Q: Hub to be located at (x, y)
$C(p_i, p_j)$: Cost per unit flow between points p_i and p_j measured in terms of the Euclidean distance
K: Cost of intercity linkage (which may include the cost of using the transportation mode and the operational expenses such as fuel cost, driver wages, etc.)

Observe that if function C satisfies the triangular inequality, the savings result from the fewer links to operate when the hub is utilized. The expression indicates that the total transfer cost is greater when the traffic is routed through the hub. However, this difference is compensated by the smaller cost of operating fewer flow links in the hub version, i.e., n vs. $\frac{1}{2}n(n-1)$ in the hub and non-hub versions, respectively.

O'Kelly acknowledges the need for using multiple hubs to accommodate a large area and discusses also multiple-hub problems. In such a network, the inter-hub linkages can be specially designed to efficiently handle bulk flow. In this way, the unit transportation costs between hubs can be significantly reduced. The reduced cost of these flows in turn appears as an endogenous function of the hub locations.

O'Kelly proposes a simple approximation, and discounts the inter-hub costs by a factor α, such that $0 \leq \alpha < 1$, $\alpha \in \mathbb{R}$. Because of the special structure of the cost function, the multiple-hub problems involve a two-fold decision in the sense that both the location of the hubs and the assignment of the demand points to the hubs must be decided upon.

The paper particularly focuses on solving the two-hub version of the problem, which is significantly easier than the more general p-hub version. Using decision variables

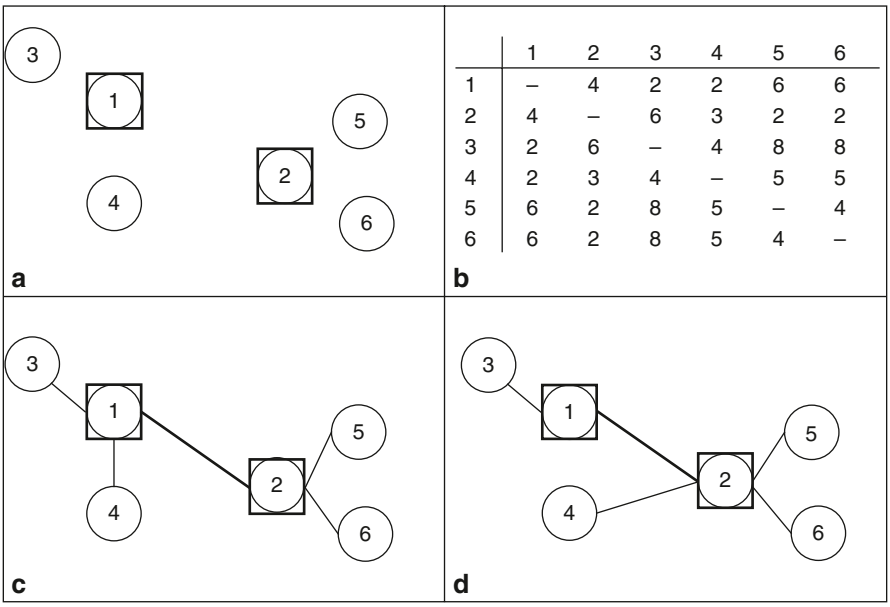
$$X_{ik} = \begin{cases} 1, & \text{if demand point } p_i \text{is assigned to hub } Q_k, k = 1, 2 \\ 0, & \text{otherwise} \end{cases}$$

the cost function to be minimized is characterized as

$$\underset{Q_1, Q_2}{\text{Min}} \sum_i \sum_j W_{ij} R_{ij},$$

where R_{ij} is the routing and transportation cost between points i and j conditional upon the corresponding hub location decision. This cost is mathematically expressed as follows.

$$\begin{aligned} R_{ij} = {} & X_{i1}X_{j1}(C(p_i, Q_1) + C(p_j, Q_1)) \\ & + X_{i2}X_{j2}(C(p_i, Q_2) + C(p_j, Q_2)) \\ & + X_{i1}X_{j2}(C(p_i, Q_1) + \alpha C(Q_1, Q_2) + C(p_j, Q_2)) \\ & + X_{i2}X_{j1}(C(p_i, Q_2) + \alpha C(Q_2, Q_1) + C(p_j, Q_1)) \end{aligned}$$



	1	2	3	4	5	6
1	–	4	2	2	6	6
2	4	–	6	3	2	2
3	2	6	–	4	8	8
4	2	3	4	–	5	5
5	6	2	8	5	–	4
6	6	2	8	5	4	–

Fig. 12.2 Two different allocation schemes on an example network. **a** Network configuration. **b** Distance matrix. **c** Allocation scheme 1. **d** Allocation scheme 2

Due to the binary nature of the decision variables, for each origin destination pair only one of the four possible components of the objective function will take on a positive value. The possibilities involved are the cost of flow from origin to destination via the same hub (either hub 1 or hub 2), and the transfer cost from origin to destination via both hubs (either from hub 1 to hub 2 or the reverse). Obviously, when both hubs are used, the cost of inter-hub transfer is discounted by factor α.

O'Kelly observes that due to the quadratic term and the discounting effect, assignment to the nearest hub may turn out to be suboptimal. We develop an example to illustrate this phenomenon in Fig. 12.2. The network configuration and corresponding distance matrix are shown in Fig. 12.2a and b, respectively. The magnitude of symmetric flows between point 4 and points 5 and 6 are equal to 10. Flow densities between all other pairs have a much smaller value of 1. The discount factor α is set equal to 0.60. Figure 12.2c shows allocation scheme 1, in which all points are allocated to their nearest hub. This scheme results in a total cost value of 353.60. On the other hand, allocation scheme 2, shown in Fig. 12.2d, assigns point 4 to the more distant hub, and gives a smaller total cost value of 308.40.

The proposed approach to solving the two-hub problem is to minimize the discounted cost function by simply taking the first order derivatives with respect to the location coordinates and setting them equal to zero. In this problem, however, the cost function is minimized for different partitions of demand points corresponding to the hubs. A partition refers to the set of demand points assigned to a given hub. The partitions whose convex hulls are non-overlapping are defined as non-overlap-

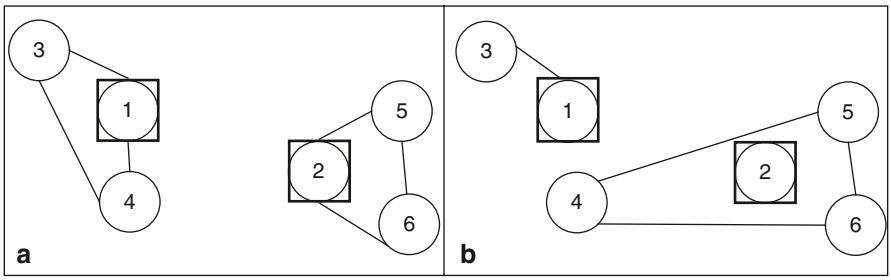


Fig. 12.3 Two different non-overlapping partitions of the example network of Fig. 12.2. **a** Partition 1. **b** Partition 2

ping partitions. Figure 12.3 shows two non-overlapping partitions on the previous example network. Partitions 1 and 2 correspond to allocation schemes 1 and 2, respectively. Motivated by the fact that consideration of only the non-overlapping partitions yields the optimum solution for the two-center location-allocation problem (Ostresh 1975), O'Kelly relies on the simplifying assumption that the assignment of demand points can be considered only for non-overlapping partitions, though he acknowledges that this approach may not necessarily yield the true optimum solution in the current problem.

Another of O'Kelly's observations relates to the effect of the hub network structure on the intensity of flow between demand points. He proposes the following function that updates the revised flow.

$$W_{ij} = \frac{O_i D_j \exp(-\beta R_{ij})}{\sum_{k=1}^{n} D_k \exp(-\beta R_{ik})}$$

Recall that the R_{ij} values are the routing and transportation costs considering all hub assignment possibilities. This function revises the flow density between each pair as a decreasing function of the relevant transportation costs. The sensitivity of the flow volume to the cost is governed by coefficient $\beta \geq 0$, where a larger value of this coefficient leads to a more significant effect. O'Kelly presents some computational analysis on the Civil Aeronautics Board data in which the effects of using different parameters (α and β) for modeling endogenous attraction are investigated.

The most significant contribution of this classical paper remains the identification of the hub location problem as a version of the p-median location-allocation problem involving interactions. The solution techniques for the multiple-hub problems are later improved by various researchers including O'Kelly himself (e.g. 1987, 1992). In addition, a multitude of studies focusing on hub-location counterparts of different classical location problems emerged, and the next section presents an overview of these studies within the framework of a new proposed taxonomy.

12.4 Theoretical Developments in the Hub Location Literature

This section starts out by identifying the connections between hub location problems with their counterparts in the classical location literature. After observing the factors that result in different types of hub location problems, the authors propose a new taxonomy that serves for a convenient classification of the relevant developments. These developments are presented for the cost minimization and minmax type of objectives in Sects. 12.4.3 and 12.4.4, respectively.

12.4.1 Analogies with Location Theory

Having identified hub-location problems involving interacting facilities, O'Kelly (1987) formulates a general version of the problem where flow between demand points is to be transferred via p hubs to be cited at a subset of the nodes corresponding to origins and/or destinations. The following additional notation is needed.

N: Set of nodes
C_{ij}: Transportation cost for a unit flow between nodes i and j

Note that the transportation cost is redefined to highlight its correspondence to the network distance of the shortest path linking the two nodes. Although this cost is defined here as related in some way only to the distances involved, it is important to acknowledge that there may be a multitude of different factors affecting the magnitude of the cost and the discount factor. The only decision variable X_{ik} is now redefined for $i, k=1, 2,\ldots, n$. Note that if $i=k$ and $X_{ii}=1$, node i is a hub.

The proposed formulation is as follows.

$$\text{Min } z = \sum_i \sum_j W_{ij} \left(\sum_k C_{ik} X_{ik} + \alpha \sum_k \sum_m C_{km} X_{ik} X_{jm} + \sum_m C_{jm} X_{jm} \right) \tag{12.1}$$

$$\text{s.t. } (n - p + 1) X_{jj} - \sum_i X_{ij} \geq 0 \ \forall j \tag{12.2}$$

$$\sum_j X_{ij} = 1 \ \forall i \tag{12.3}$$

$$\sum_j X_{jj} = p \tag{12.4}$$

$$x_{ij} \in \{0,1\} \ \forall i, j \tag{12.5}$$

The objective function, shown in (12.1), minimizes the total cost comprising the origin-to-hub, discounted inter-hub transportation, and hub-to-destination cost components. The inter-hub transportation cost in this basic formulation includes a quadratic term to account for the origin-destination pairs connected through their designated hubs. Note that the discount factor works on the transportation costs, not the distances. Constraint set (12.2) ensures that no switching is allowed through a non-hub node. Constraint sets (12.3) and (12.5) enforce allocation of each node to exactly one hub. Finally constraint (12.4) sets the number of hubs equal to p. Note that this initial formulation is subsequently considered as the "basic formulation" in the hub location literature.

O'Kelly (1987) reiterates that interaction is the factor differentiating this new problem from the p-median and multi-facility Weber problems, both of which are widely considered in the classical location literature. The novelty is that the locations of the hubs have a direct effect on the magnitude of the inter-hub flows and the associated linkage costs.

An immediate consequence of this endogenous effect is in the allocation of nodes to hubs. In classical location theory with uncapacitated facilities, once the locations are given, the allocation subproblem can be optimally solved by assigning a node to its nearest facility. In hub location problems, on the other hand, the assignment of a node to a facility is impacted also by that facility's ability to service the interaction pattern. Therefore, proximity of the hub to a node ceases to be the sole factor dictating the allocation.

The logical connections between hub location problems and the location theory literature were outlined by Campbell (1994a). Campbell defines location analogous versions of the hub-location problem, namely the p-hub median, hub location with fixed costs, p-hub center, and hub covering problems. The basic problem defined in O'Kelly (1987) is a p-hub median problem. A detailed discussion of the other problems defined by Campbell (1994a) will follow below. Campbell's principle contribution to the expansion of the hub location literature relates to the consideration of alternative criteria for objectives. On the constraint side, researchers identified the following three major factors to produce alternative versions of this basic problem:

1. Single- vs. multi-allocation (Campbell 1990),
2. Full vs. partial hub network (Chou 1990), and
3. Presence/absence of direct connectivity between non-hub nodes (Aykin 1995).

Recall that the basic model assumes each node is served by exactly one hub, all hubs are connected to each other, and any transfer between two non-hub nodes must be via at least one hub. O'Kelly and Miller (1994) suggest that different combinations of these three factors result in eight alternative versions of the basic problem.

Based on the alternative objective functions and possible variations in the constraint set, the present authors propose a taxonomy in the next section to facilitate a convenient and systematic discussion of the emerging literature.

12.4.2 A Taxonomy of Hub Location Problems

We observe that the factors that determine the nature of the problem can be considered in four categories. To also accommodate other problem-specific restrictions,

we propose a five-fold taxonomy in the following form $\varepsilon/\varphi/\kappa/\lambda/\omega$. The fields in this short-hand notation correspond to the following specific factors.

ε: Objective criterion
φ: Allocation structure
κ: Capacity
λ: Inter-hub connectivity
ω: Other restrictions.

The alternative objective criteria corresponding to the *p*-hub median, hub location with fixed cost, *p*-hub center and hub covering problems will be denoted shortly as *pH-median*, *fixH-cost*, *pH-center* and *H-cover*, respectively. The allocation structure refers to the degree of flexibility in terms of the number of hubs to which a node can be assigned. The corresponding parameter φ thus in turn may be either single or multi. Various types of capacities may be imposed upon the flow handled by the hubs and the transportation lines. The uncapacitated version of the problem is denoted by *U*, whereas the presence of node and arc capacities is indicated by *node* and *arc*, respectively. Finally, the underlying network topology appearing in the λ field may range from full to different partial structures such as path, tree, ring, and star. Since the other restrictions are expected to vary depending on the circumstances of a specific problem, the notation to be used in the ω field is left to the discretion of other authors. Note that the basic problem can be denoted as *pH-median*/*single*/*U*/*full*.

We remark here that Campbell et al. (2002) also provide a taxonomy to help classify the hub location problems. The alternative proposed herein is based on the review and synthesis provided in O'Kelly and Miller (1994) as well as the objective criteria discussed in Campbell (1994a).

After the basic problem was identified by O'Kelly (1986), for almost a decade researchers worked on mathematical formulations that would efficiently solve it. The initial formulation provided in O'Kelly (1987) was quadratic. Linear formulations were given in Aykin (1990), Campbell (1996), and Skorin-Kapov et al. (1996), among others. In these formulations, single- and/or multiple-allocation versions of the problem were considered under the cost objective. For the *pH-median* problems, the objective is the minimization of the total transportation cost. Conversely, in the *fixH-cost* problem a fixed cost associated with opening a new hub was considered alongside the transportation cost. We first discuss several important studies on the *pH-median* and *fixH-cost* problems. Then we proceed with the *pH-center* and *H-cover* versions of the problem investigated in the more recent literature.

12.4.3 Minisum Objectives

In the early 1990s, due to the quadratic nature of the formulation, researchers attempted to solve the single allocation version of the *pH-median* problem with heuristic approaches. Three important examples of such attempts can be found in Klincewicz (1991 and 1992), as well as Skorin-Kapov and Skorin-Kapov (1994).

The first solvable exact formulation dates to back to Campbell (1996), who studies *pH-median*/*single*/*U*/*full* and *pH-median*/*multi*/*U*/*full*. Campbell gives linear mathematical programming formulations for the two problems by defining a four-indexed binary variable X_{ijkm}, which takes on a value of 1 only if the flow between nodes i and j is routed via hubs k and m. He observes that the integrality of these variables can be conveniently relaxed when solving *pH-median*/*multi*/*U*/*full* without forgoing optimality. He also remarks that solution to the multiple allocation version of the problem constitutes a lower bound for the single allocation version.

The single allocation version of the problem with the Civil Aeronautics Board data was optimally solved for the first time in Skorin-Kapov et al. (1996) by using a branch-and-bound algorithm utilizing a tight lower bound obtained from the linear programming relaxation of their original formulation. Ernst and Krishnamoorthy (1996) provide an efficient network flow formulation to solve the same problem. This formulation relies on modeling flows generated by each node as a different commodity which results $O(n^3)$ binary integer variables as opposed to $O(n^4)$ in the previous formulations. Ernst and Krishnamoorthy (1998a) embed this notion in a branch-and-bound algorithm, which to the best knowledge of the present authors is the most efficient solution algorithm for this problem to date. Ernst and Krishnamoorthy (1998b) apply the network flow notion also to the multi-allocation version of the problem and obtain optimum solutions to large instances. In the same paper, they observe that the problem can be solved polynomially by an all-pairs shortest path algorithm when the hub locations are fixed.

Following O'Kelly (1986), all research until O'Kelly (1992) considered solely the transportation costs in the objective function. O'Kelly (1992) incorporates this fixed cost into the problem and addresses the capacitated version of the problem, *fixH-cost*/*single*/*node*/*full*. He uses a modified version of his basic formulation with the addition of the total hub cost, $\sum_j F_j X_{jj}$, in the quadratic objective function. The multi-allocation version of this problem with additional arc costs, *fixH-cost*/*multi*/*node*/*full*/{*direct*, *arc-costs*} was studied by Aykin (1994). Recall that the term "direct," in this context, implies that a non-stop connection between non-hub nodes is permissible. He proposed a branch and bound algorithm utilizing a Lagrangian-based lower bound. This problem, with the only difference of not allowing direct connections, was studied by Ernst and Krishnamoorthy (1999), who proposed efficient integer programming formulations. Recent exact solution approaches exploit the polyhedral structure of the hub location problems. Labbé and Yaman (2004) derive facet-defining inequalities for *fixH-cost*/*single*/*U*/*full*. For the multiple allocation version of the problem, Hamacher et al. (2004) propose valid inequalities by modifying the facet defining inequalities for the uncapacitated facility location problem. Similarly, Marin (2005) exploits the polyhedral structure of the set packing problem to develop valid inequalities for *fixH-cost*/*multi*/*U*/*full* with Euclidian distances.

12.4.4 Minmax Objectives

An inherently different class of hub location problems is of the minmax type. The 1-hub center problem was originally defined by O'Kelly and Miller (1991) to minimize the maximum cost incurred by any origin-destination pair. This problem was motivated by a desire to achieve equity between user nodes in terms of the transportation costs incurred. Following the remarks made in O'Kelly and Miller's (1991) conclusion, the minmax objectives in later studies focus on the service time concerns rather than the cost issues. In the *p*-hub center problem, the objective is to minimize the worst service time between any origin destination pair. Alternatively, the objective of the hub cover problem is to serve all node pairs with the minimum possible number of hubs while keeping the travel times below a predetermined threshold level. These problems received attention in the literature partly due to their practical applications in such systems as perishable goods transfer and overnight delivery.

The first paper which fully defines and classifies different versions of these problems is Campbell (1994a). In addition to providing integer programming formulations for *pH-center*/*single*/*U*/*full, pH-center*/*multi*/*U*/*full, H-cover*/*single*/*U*/*full, H-cover*/*multi*/*U*/*full*, Campbell identifies different types of service time restrictions. In particular, he additionally defines separate service times for the segments constituting a path between origin-destination pairs. He also proposes integer programming formulations for the *pH-center* and *H-cover* problems based on these new service time definitions. These alternative versions of the two problems are still open areas that require further investigation.

After being defined by Campbell (1994a), the *pH-center* and *H-cover* problems were not studied until Kara and Tansel (2000). They provide a proof of **NP**-hardness for *pH-center*/*single*/*U*/*full* and develop an efficient integer programming formulation with n^2 binary variables. Ernst et al. (2002) give a more efficient formulation for the same problem by using auxiliary variables. They also show that the multiple allocation version of the problem is **NP**-hard, and propose a modification of their original formulation for its solution. Baumgartner (2003) analyzes these two formulations, develops facet defining valid inequalities, and proposes a branch-and-cut algorithm based on these inequalities.

For the covering version of the problem with single allocation, Kara and Tansel (2003) provide an **NP**-hardness proof along with an efficient integer programming formulation. Ernst et al. (2005) present formulations for both the single and multiple allocation versions of this problem. Their formulation for the single allocation case outperforms that of Kara and Tansel (2003). Polyhedral properties of these *H-cover* problems are studied by Hamacher and Meyer (2006).

A variant of this problem is motivated by real life applications based on the observation that trucks are synchronized at the hub nodes by occasionally delaying their departures. Kara and Tansel (2001) call this variant the latest arrival hub location problem, defining the *pH-median*, *pH-center* and *H-cover* versions of the problem. They propose a formulation for *pH-center*/*single*/*U*/*full*/*latest-arrival* which

can efficiently solve all Civil Aeronautics Board instances. The *H-cover* version of this problem was investigated in a similar way by Tan and Kara (2007) who test the performance of their formulation based on a new data set of Turkish highway travel times.

Special cases of the *pH-center* problems with fixed hubs are investigated by Iyer and Ratliff (1990) and Campbell et al. (2007). Iyer and Ratliff (1990) consider a "guaranteed time distribution" problem, which is in fact equivalent to the uncapacitated *p*-hub center problem with a tree type network structure. They propose a polynomial time exact algorithm to solve this problem. Their algorithm was later modified by Campbell et al. (2007) to solve the 2*H-center*/*single*/*U*/*path*, *pH-center*/*single*/*U*/*tree*, 2*H-center*/*single*/*U*/*full*, *pH-center*/*multi*/*U*/*full* problems. Campbell et al. prove additionally that problems *pH-center*/*single*/arc/*full* and *pH-center*/*single*/*node*/*full* are **NP**-hard.

12.5 Application-Oriented Studies

In addition to the theoretical investigations discussed in the previous section, the hub location problem identified by O'Kelly (1986) has been widely studied in the past two decades regarding other practical applications than airline passenger streams. These different practical applications, which occasionally lead to alternative versions of the problem, can be broadly classified as telecommunication networks and cargo delivery practices. This section discusses major findings in these two areas.

In the context of telecommunication networks, data packets are transferred between user nodes through concentrators (servers, switches, multiplexers, etc.) which function as hubs. The user nodes are connected to the concentrators via access networks, whereas the concentrators are connected to each other and/or to a central root node through a backbone network. Different topologies of backbone/access networks such as clique, star, tree, path, ring, and their hybrids, are possible. Objectives considered in the design of telecommunication networks include equipment installation and routing cost as well as reliability (survivability), capacity, and expandability concerns. Klincewicz (1998) provides an extensive review of the literature in this area. More recent works on telecommunication network design are discussed in Gourdin et al. (2002) and Labbé et al. (2005). Motivated by ongoing technological developments, there has been extensive research in this area in the past few years, and this trend is expected to continue for the foreseeable future.

As discussed in the theoretical aspects presented in Sect. 12.4.2, in cargo delivery practice time issues overshadow the cost concerns, resulting in minimax objectives. Cargo delivery networks are designed and managed mostly either with a constraint on the delivery times or with the objective of minimizing the delivery times. Hall (1989) identifies the issues of critical concern in the design of cargo networks as the number of hub terminals, the routing strategies of the transportation modes serving these terminals, and the synchronization of the inflow and outflow at a terminal.

An application to the postal delivery systems was described by Ernst and Krishnamoorthy (1996) based on a data set obtained from the Australian Post. Due to the

possibility of having different modes of transfer in the collection and distribution segments, the cost structure in postal delivery services is different from that in the airline data. To model these differences, Ernst and Krishnamoorthy (1996) propose the use of two additional parameters apart from α. In particular, parameters χ and $\delta \geq \alpha$ correspond to differences in transportation costs in collection and distribution processes, respectively. Use of different factor coefficients allows for the consideration of possible differences in the collection, transportation, and distribution costs that may result due to the use of different transportation modes. Note that this problem is equivalent to the basic problem when $\chi = \delta = 1$.

Nickel et al. (2001) relax the assumption of all hubs being interconnected, and study a public transportation problem that can be denoted as *fixH-cost*/*multi*/*U*/*incomplete*/*hub-arc*. The authors are the first to address the incomplete hub network version of the problem. They propose four-indexed mixed-integer programming formulations for the single and multiple hub versions. Campbell et al. (2005a, b) exploit this same idea to address the *pH-median*/*single*/*U*/*incomplete*/*hub-arc* problem. They introduce a new perspective for the solution of this problem. In particular, instead of locating hubs, they locate discounted hub arcs. They develop mixed-integer programming models and two exact algorithms for four different versions of the problem accommodating different objective criteria. They give exact solutions for the Civil Aeronautics Board data.

Motivated by a Federal Express application, Kuby and Gray (1993) model the practical case, in which feeder links consolidate local flows at a convenient node. In their problem, transportation media serving the regional hub are allowed to make multiple stops along their way. This problem can be denoted as *1H-median*/*single*/*arc*/*full*/*stopover-feeder*. Kuby and Gray considered a single, fixed hub air network problem, and developed a path-based mixed-integer programming formulation to explore the savings provided by the consideration of stopovers and feeders. Later, Yaman et al. (2007) provide integer programming models for *H-cover*/*single*/*U*/*full*/{*latest-arrival, stopovers*}. The authors propose a different mixed-integer programming formulation, which is strengthened by valid inequalities and lifting. They test the performance of the model on the Turkish highway travel time data. Wasner and Zapfel (2004) suggest that the stopovers can be modeled in the form of a vehicle routing problem.

The modeling complications necessitated by these practical observations suggest that the basic problem proposed by O'Kelly (1986) has implications in a variety of real life applications. The specific needs of these applications provide many ideas that continuously support the evolution of research in this area.

12.6 Conclusion

O'Kelly's classical 1986 paper led to the emergence of a new research area by identifying a connection between location theory and spatial interaction theory. This connection mainly manifests itself in the form of an endogenous interaction that has an impact on both the intensity and cost of flow to be routed through the facili-

ties that are selected as hubs. The problem has been widely studied in the past two decades both from a theoretical and a practical perspective. Theoretical papers in this new area investigated various objectives including *pH-median*, *fixH-cost*, *pH-center* and *H-cover* problems as well as network topologies with fully and partially connected structures. On the practical side, many researchers modeled and solved various real life applications observed in airflow streams, telecommunication networks, cargo delivery systems, and urban transit.

This chapter discussed the most prominent research relevant to both the theoretical and practical aspects of the problem within the framework of a proposed new taxonomy. The interested reader is referred to the excellent review papers written by Campbell (1994b), Klincewicz (1998), Bryan and O'Kelly (1999), Campbell et al. (2002), and Alumur and Kara (2008) for more in-depth coverage of the area.

The authors would like to note that hub location is still a very active research area with many potentially fruitful extensions. One of these extensions is identified by Marianov et al. (1999), who study a multi-allocation hub location problem in the presence of competitors. This interesting problem offers an avenue for further research, as it has not received much attention since. Another important extension is observed by O'Kelly and Bryan (1998) on the fundamental assumption that characterizes the endogenous attraction via the constant scaling factor α. They propose a nonlinear cost function to more accurately model this attraction. Although a few other researchers later improved or modified this function, further research is necessary in this regard. Recall that O'Kelly (1986) proposed two different types of endogenous attraction. In the first type, cited hubs affect the cost of flow, whereas in the second category, the affected parameter is the intensity of flow. The entire literature stemming from this idea focused on the former type and investigated the hub location problem in view of the cost advantages provided by the economies of scale. The latter aspect, which requires modeling of the impact on the intensity of flow, received no attention other than O'Kelly's original proposal.

The authors would like to conclude by emphasizing that these are just a few avenues for future research in this area led by O'Kelly's classical paper (1986). The relevance of the problem to a number of application areas and the wide interest received from many researchers are expected to trigger further developments in the future.

References

Allen PM, Sanglier M (1979) A dynamic model of growth in a central placed system. Geogr Anal 11:256–272

Allen PM, Sanglier M (1981) A dynamic model of growth in a central placed system II. Geogr Anal 13:149–164

Alumur S, Kara BY (2008) Network hub location problems: the state of the art. Eur J Oper Res 190:1–21

Aykin T (1990) On a quadratic integer program for the location of interacting hub facilities. Eur J Oper Res 46:409–411

Aykin T (1994) Lagrangean 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 application to the air transportation system. Transp Sci 29:201–221
Baumgartner S (2003) Polyhedral analysis of hub center problems. Diploma Thesis, Universität Kaiserslautern, Germany
Bryan DL, O'Kelly ME (1999) Hub-and-spoke networks in air transportation: an analytical review. J Reg Sci 39:275–295
Campbell JF (1990) Freight consolidation and routing with transportation economies of scale. Transp Res B 24:345–361
Campbell JF (1994a) Integer programming formulations of discrete hub location problems. Eur J Oper Res 72:387–405
Campbell JF (1994b) A survey of network hub location. Stud Locat Anal 6:31–49
Campbell JF (1996) Integer Hub location and *p*-hub median problem. Oper Res 44:1–13
Campbell JF, Ernst AT, Krishnamoorthy M (2002) Hub location problems. In: Drezner Z, Hamacher HW (eds) Facility locations applications and theory. Springer, Berlin, pp 373–407
Campbell JF, Ernst AT, Krishnamoorthy M (2005a) Hub arc location problems. Part I: Introduction and results. Manag Sci 51:1540–1555
Campbell JF, Ernst AT, Krishnamoorthy M (2005b) Hub arc location problems. Part II: Formulations and optimum algorithms. Manag Sci 51:1556–1571
Campbell AM, Timothy JL, Zhang L (2007) The *p*-hub center allocation problem. Eur J Oper Res 176:819–835
Chou Y-H (1990) The hierarchical-hub model for airline networks. Transp Plan Technol 14:243–258
Ducca FW, Wilson RH (1976) A model of shopping center location. Environ Plan A 8:613–623
Ernst AT, Krishnamoorthy M (1996) Efficient algorithms for the uncapacitated single allocation *p*-hub median problem. Locat Sci 4:139–154
Ernst AT, Krishnamoorthy M (1998a) An exact solution approach based on shortest paths for *p*-hub median problems. INFORMS J Comput 10:149–162
Ernst AT, Krishnamoorthy M (1998b) Exact and heuristic algorithms for the uncapacitated multiple allocation *p*-hub median problem. 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
Ernst AT, Hamacher HW, Jiang H, Krishnamoorthy M, Woeginger G (2002) Uncapacitated single and multiple allocation *p*-hub center problems. Unpublished report, CSIRO Mathematical and Information Sciences, Melbourne
Ernst AT, Jiang H, Krishnamoorthy M (2005) Reformulations and computational results for uncapacitated single and multiple allocation hub covering problems. Unpublished report, CSIRO Mathematical and Information Sciences, Melbourne
Goodchild MF (1978) Spatial choice in location-allocation problems: the role of endogenous attraction. Geogr Anal 10:65–72
Gourdin E, Labbe M, Yaman H (2002) Telecommunication and location. In: Drezner Z, Hamacher HW (eds) Facility locations, applications, and theory. Springer, Berlin, pp 275–305
Hall RW (1989) Configuration of an overnight package air network. Transp Res A 23:139–149
Hamacher HW, Meyer T (2006) Hub cover and hub center problems. Working Paper, Department of Mathematics, University of Kaiserslautern, Germany
Hamacher HW, Labbe M, Nickel S, Sonneborn T (2004) Adopting polyhedral properties from facility to hub location problems. Discrete Appl Math 145:104–116
Iyer AV, Ratliff HD (1990) Accumulation point location on tree networks for guaranteed time distribution. Manag Sci 36:958–969
Kara BY, Tansel BC (2000) On the single-assignment *p*-hub center problem. Eur J Oper Res 125:648–655
Kara BY, Tansel BC (2001) The latest arrival hub location problem. Manag Sci 47:1408–1420
Kara BY, Tansel BC (2003) The single assignment hub covering problem: models and linearizations. J Oper Res Soc 54:59–64

Klincewicz JG (1991) Heuristics for the *p*-hub location problem. Eur J Oper Res 53:25–37
Klincewicz JG (1992) Avoiding local optima in the *p*-hub location problem using tabu search and GRASP. Ann Oper Res 40:283–302
Klincewicz JG (1998) Hub location in backbone/tributary network design: a review. Locat Sci 6:307–335
Kuby MJ, Gray RG (1993) The hub network design problem with stopovers and feeders: the case of Federal Express. Transp Res A 27:1–12
Labbé M, Yaman H (2004) Projecting flow variables for hub location problems. Networks 44:84–93
Labbé M, Yaman H, Gourdin E (2005) A branch and cut algorithm for hub location problems with single assignment. Math Program 102:371–405
Marianov V, Serra D, ReVelle C (1999) Location of hubs in a competitive environment. Eur J Oper Res 114:363–371
Marin A (2005) Uncapacitated Euclidean hub location: strengthened formulation, new facets and a relax-and-cut algorithm. J Glob Optim 33:393–422
Miehle W (1958) Link-length minimization in networks. Oper Res 6:232–243
Nickel S, Schöbel A, Sonneborn T (2001) Hub location problem in urban traffic networks. In: Niittymaki J, Pursula M (eds) Mathematics methods and optimization in transportation systems. Kluwer, Dordrecht, pp 95–107
O'Kelly ME (1986) The location of interacting hub facilities. Transp Sci 20:92–106
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 71:293–306
O'Kelly ME, Bryan DL (1998) Hub location with flow economies of scale. Transp Res B 32(8):605–616
O'Kelly ME, Miller HJ (1991) Solution strategies for the single facility minimax hub location problem. Pap Reg Sci 70:367–380
O'Kelly ME, Miller HJ (1994) The hub network design problem: a review and synthesis. J Transp Geogr 2:31–40
Ostresh LM (1975) An efficient algorithm for solving the two center location-allocation problem. J Rec Sci 15(2):209–216
Skorin-Kapov D, Skorin-Kapov J (1994) On tabu search for the location of interacting hub facilities. Eur J Oper Res 73:502–509
Skorin-Kapov D, Skorin-Kapov J, O'Kelly M (1996) Tight linear programming relaxations of uncapacitated hub location problems. Eur J Oper Res 94:582–593
Taaffe EJ, Morrill RL, Gould PR (1963) Transport expansion in underdeveloped countries: a comparative analysis. Geogr Rev 53:523–559
Tan PZ, Kara BY (2007) A hub covering model for cargo delivery systems. Networks 49:28–39
Wasner M, Zapfel G (2004) An integrated multi-depot hub-location vehicle routing model for network planning of parcel service. Int J Prod Econ 90:403–419
Yaman H, Kara BY, Tansel BC (2007) The latest arrival hub location problem for cargo delivery systems with stopovers. Transp Res B 41:906–919

Part VI
Solution Techniques

Chapter 13
Exact Solution of Two Location Problems via Branch-and-Bound

Timothy J. Lowe and Richard E. Wendell

13.1 Introduction

In 1960, Land and Doig published a paper that most scholars recognize as the first description of a now well-known technique for solving difficult optimization problems by solving a sequence of easier, restricted subproblems (Land and Doig 1960). Little et al. (1963) named this technique "Branch-and-Bound" (*B&B*), and used it to solve the traveling salesman problem. Although the method is described and used in several papers in the 1960s (see for example, Lawler and Wood 1966), the description below, provided by Hillier and Lieberman (1980), succinctly captures the idea.

> The basic idea of the branch-and-bound technique is the following: suppose (to be specific) that the objective function is to be *minimized*. Assume that an *upper bound* on the optimal value of the objective function is available. (This is usually the value of the objective function for the best feasible solution identified thus far.) The first step is to *partition* the set of all feasible solutions *into several subsets*, and for each one, a *lower bound* is obtained for the value of the objective function of the solutions within that subset. Those subsets whose lower bounds exceed the current upper bound on the objective value are then excluded from further consideration. (A subset that is excluded for this or other legitimate reasons is said to be *fathomed*.) One of the remaining subsets, say, the one with the smallest lower bound, is then partitioned further into several subsets. Their lower bounds are obtained in turn and used as before to exclude some of these subsets from further consideration. From *all* the *remaining* subsets, another one is selected for further partitioning and so on. This process is repeated again and again until a feasible solution is found such that the corresponding value of the objective function is no greater than the lower bound for any subset. Such a feasible solution must be optimal since none of the subsets can contain a better solution.

The method of partitioning the set of feasible solutions into subsets (branching) is relatively straightforward for integer variables, particularly when these can take on only one of two values, zero or one. Thus, a partition is created when one of these

T. J. Lowe (✉)
Tippie College of Business, University of Iowa, Iowa City, IA 52242-1994, USA
e-mail: timothy-lowe@uiowa.edu

R. E. Wendell
Joseph M. Katz Graduate School of Business, University of Pittsburgh, Pittsburgh, PA 15260, USA
e-mail: wendell@katz.pitt.edu

H. A. Eiselt, V. Marianov (eds.), *Foundations of Location Analysis,*
International Series in Operations Research & Management Science 155,
DOI 10.1007/978-1-4419-7572-0_13, © Springer Science+Business Media, LLC 2011

variables is set to zero in one subset, and set to one in the other subset. An easy way to envision the partitioning process is through what is called the *branch-and-bound tree*. The top node of the tree represents the original problem. Two branches can be created from this node by selecting one of the variables, and setting it equal to zero in one branch and to one in the other branch. Each of the resulting nodes can be further partitioned via the selection of another variable and a repetition of the above process.

To create bounds at each node, early implementers of the branch-and-bound method solved "relaxed" problems by treating integer variables as continuous. By doing this, the resulting relaxed problem was often a linear program that could be solved by existing codes. The key to branch-and-bound efficiency is to reduce the number of subsets that must be visited and to be able to create "strong" bounds. Early adopters realized this and those same issues are faced today by current users of the method. We will have more to say on this issue later in the chapter.

In spite of the fact that branch-and-bound can be painfully slow as a solution method for discrete optimization problems, it is still often applied as the technique of choice for these problems. Discrete optimization problems generally have locally optimal solutions and so sensible search methods are necessary to explore the solution space for a globally optimal solution. Branch-and-bound is such a method since it provides a means of exploring various subregions of the feasible set of solutions in an organized manner.

In this chapter, we give an overview of the use of branch-and-bound to solve two prototypical location problems: the *quadratic assignment problem QAP* and the *uncapacitated facility location problem UFLP*. Our focus will be on the early applications of branch-and-bound to these problems via a critical review of two papers from the 1960s. In providing these reviews we attempt to replicate the authors' thought process in the development of the reported solution method.

The remainder of this chapter is organized as follows. Section 13.2 discusses the work of Gavett and Plyter (1966) on the quadratic assignment problem, following which we discuss advancements in the application of branch-and-bound to the problem as well as special cases of the problem solvable in polynomial time. Section 13.3 is dedicated to the uncapacitated facility location problems where we first review the work of Efroymson and Ray (1966) on this problem. We then follow this review with a discussion of further work on the problem, and special cases solvable in polynomial time. Branching strategies for branch-and-bound methods are discussed in Sect. 13.4, and concluding remarks are offered in Sect. 13.5.

13.2 Gavett and Plyter (1966): The Quadratic Assignment Problem

The Quadratic Assignment Problem was formulated by Koopmans and Beckman (1957) over 50 years ago. The motivating and most popular application of the quadratic assignment problem is the facility layout problem of assigning n facilities to n locations where one and only one facility can be assigned to each location. Thus, there are $n!$ possible assignments. The cost of an assignment depends on both the

distance between each pair of locations, and the traffic intensity between facilities assigned to those locations. The objective is to find a minimal-cost solution among the $n!$ possible assignments. One of the earliest exact methods for solving it was the branch–and-bound approach given in the paper of Gavett and Plyter (1966). Herein we review their approach.

13.2.1 Solving the Quadratic Assignment Problem via Branch-and-Bound

To formally pose the quadratic assignment problem as an optimization problem, let $\mathbf{A} = [a_{j\ell}]$ denote a matrix of distances between locations j and ℓ for $j, \ell = 1, \ldots, n$. Also, let $\mathbf{B} = [b_{ik}]$ denote a matrix of rates at which material is transferred (traffic intensity) between facilities i and k where $i, k = 1, \ldots, n$. Letting $p = (p_1, p_2, \ldots, p_n)$ denote a permutation of $1, 2, \ldots, n$ and letting P_n denote the set of all permutations on $\{1, 2, \ldots, n\}$, we can state the quadratic assignment problem as follows:

$$QAP: \text{Min} \left\{ \sum a_{j\ell} b_{p_j p_\ell} : p \in P_n \quad \text{for } j, \ell = 1, 2, \ldots, n \right\} \tag{13.1}$$

In addition to facility location, there are many applications of the quadratic assignment problem in the literature. These include backboard wiring, economic problems, scheduling, the design of typewriter keyboards and control panels, archeology, statistical analysis, and reaction chemistry. For a further discussion see, for example, Loiola et al. (2007).

Consider the following simple example from Gavett and Plyter of 4 facilities to be assigned to 4 locations:

$$\mathbf{A} = \begin{bmatrix} 0 & 6 & 7 & 2 \\ 6 & 0 & 5 & 6 \\ 7 & 5 & 0 & 1 \\ 2 & 6 & 1 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 0 & 10 & 20 & 5 \\ 18 & 0 & 9 & 4 \\ 5 & 6 & 0 & 8 \\ 8 & 0 & 15 & 0 \end{bmatrix}.$$

Assigning the facilities 3, 1, 2 and 4 to the respective locations 1, 2, 3, 4 results in a total cost of 523. A better solution is assigning the facilities 2, 4, 3, 1 to the respective locations 1, 2, 3, 4 for a total cost of 403. With only 4 facilities/locations in this example, there are only 4! = 24 possible assignments and hence the problem is readily solvable (indeed, it is easy to verify that 403 is the minimal cost). However, with even modest values of n (e.g., $n = 25$) the number of permutations quickly becomes disturbingly large. The challenge is how to efficiently find an optimal solution to such problems. Following up on the work of Little et al. (1963) on applying branch-and-bound to the traveling salesman problem, Gavett and Plyter showed how branch-and-bound could be used to solve this problem.

The authors assume that the matrix $\mathbf{A}$ is symmetric (so that the distance from location j to ℓ equals the distance from ℓ to j). In this case (13.1) can be simplified as follows:

$$\text{Min}\Big\{ \sum a_{j\ell}(b_{p_j p_\ell} + b_{p_\ell p_j}) : p \in P_n \quad \text{for} j = 1, 2, \ldots, n \text{ and } l = j + 1, \ldots, n \Big\} \tag{13.2}$$

For a given permutation, observe that the objective in (13.2) is simply the sum of the products of distances $a_{j\ell}$ between pairs of locations and the total traffic intensity $(b_{p_j p_\ell} + b_{p_\ell p_j})$ between facilities assigned to them. Clearly, an ideal solution would match high intensities with small distances and low intensities with high distances. Doing this among permutations P_n can be difficult. However, as Gavett and Plyter noticed, the problem is easy to solve by expanding the permutations to match pairs of locations with pairs of intensities. Observe that the number of pairs of n locations is simply the combination of n locations taken 2 at a time, namely $N = n(n-1)/2$.

Let $\{\alpha_1, \alpha_2, \ldots, \alpha_N\}$ be the set of distances between pairs of locations. Thus $\alpha_r = a_{j\ell}$ for some pair (j, ℓ) of locations. Also let $\{\beta_1, \beta_2, \ldots, \beta_N\}$ be the set of traffic intensities between pairs of facilities so that $\beta_t = b_{ik} + b_{ki}$ for some pair (i, k) of facilities. Finally, let P_N be the set of permutations on $\{1, 2, \ldots, N\}$. With this notation, Gavett and Plyter's relaxed problem is

$$\text{Min}\left\{ \sum_{r=1}^{N} \alpha_r \beta_{p(r)} : p \in P_N \right\} \tag{13.3}$$

The matching in (13.3) of location pairs with facility pairs is sometimes referred to as "pair-assignment;" see, e.g., Pierce and Crowston (1971).

As suggested by Conway and Maxwell (1961) and independently established by Gilmore (1962), Gavett and Plyter prove that given two vectors of the same size, if the objective is to sort entries of the vectors so that the dot product is minimized, the solution is found by sorting one vector in nonincreasing order and the other in nondecreasing order. Thus, a permutation minimizes (13.3) when it corresponds to matching sorted elements of $\{\alpha_1, \alpha_2, \ldots, \alpha_N\}$ with reversely sorted elements of $\{\beta_1, \beta_2, \ldots, \beta_N\}$. In the example this corresponds to matching location pairs (1,3), (1,2), (2,4), (2,3), (1,4), and (3,5) respectively with traffic intensities of facility pairs (2,4), (1,4), (2,3), (3,4), (1,3), and (1,2) for an optimal value of 389 in (13.3). Note that this is not a feasible solution to (13.2) since matching the location pairs (1,3) and (1,2) respectively with the facility pairs (2,4) and (1,4) means that location 1 must correspond to facility 4. However, the location pair (1,4) matching with the facility pair (1,3) is inconsistent with location 1 corresponding to facility 4. This is not surprising since P_N is generally much larger than P_n. In the example, the number of permutations in P_n is 24 while the number in P_N is 720. While each permutation in P_n corresponds to one in P_N, the converse is not true. Thus, as we have seen, an optimal solution to (13.3) may not be admissible in that it may not correspond to a feasible solution in (13.2).

Given an efficient way to solve the relaxed problem (13.3), Gavett and Plyter turn their attention to using the branch-and-bound approach from Little et al. (1963) to solve (13.2). To relate the problem to this approach, they first define an

Table 13.1 The cost matrix **C**

Sorted intensities		4	13	15	23	25	28
Facility pairs		2 to 4	1 to 4	2 to 3	3 to 4	1 to 3	1 to 2
Sorted distance	Location pairs						
7	1 to 3	28	91	105	161	175	196
6	1 to 2	24	78	90	138	150	168
6	2 to 4	24	78	90	138	150	168
5	2 to 3	20	65	75	115	125	140
2	1 to 4	8	26	30	46	50	56
1	3 to 4	4	13	15	23	25	28

$[N \times N]$-dimensional cost matrix **C** of elements $\alpha_r\beta_t$ whose rows correspond to location pairs $\{\alpha_1, \alpha_2, \ldots, \alpha_N\}$ sorted by decreasing distances and whose columns correspond to facility pairs $\{\beta_1, \beta_2, \ldots, \beta_N\}$ sorted by increasing intensities. Thus, in the example, the $[6 \times 6]$-dimensional matrix **C** is shown in Table 13.1.

By construction, observe that the elements of **C** are nondecreasing across each row and are nonincreasing down each column. Also, observe that the optimal solution to (13.3) above corresponds to the diagonal of **C** with the sum of the diagonal elements being the optimal value of (13.3).

A primary difference between the matrix **C** in the quadratic assignment problem vs. the matrix considered by Little et al. is its interpretation. In Little et al., the element c_{ij} denotes the cost from i to j, whereas in Gavett and Plyter the element is the cost of assigning a location pair with a traffic intensity pair. In both cases, the relaxed problem is an assignment problem where each row will be assigned to exactly one column and where each column will be assigned to exactly one row. Little et al. point out that one could solve the assignment problem for the original cost matrix **C** and reduce the matrix by the cost of the optimal assignment. However, rather than doing this, they present a simple reduction technique to give a nonoptimal bound: reduce **C** by subtracting the smallest element in each row from the elements in the row, and then subtract the smallest element in each column from the elements in the column in the resulting matrix. All elements of the reduced matrix will be nonnegative. Thus, the sum of the reducing constants is a lower bound, since the cost of any permutation in **C** will differ from the cost under the reduced **C** by the sum and since the reduced matrix is nonnegative. Applying this technique to the quadratic assignment problem, the reducing constant (minimal element) of each row is simply its first element (which respectively are 28, 24, 24, 20, 8, 4). After subtracting these from their respective rows, we obtain the matrix

$$\begin{bmatrix} 0 & 63 & 77 & 133 & 147 & 168 \\ 0 & 54 & 66 & 114 & 126 & 144 \\ 0 & 54 & 66 & 114 & 126 & 144 \\ 0 & 45 & 55 & 95 & 105 & 120 \\ 0 & 18 & 22 & 38 & 42 & 48 \\ 0 & 9 & 11 & 19 & 21 & 24 \end{bmatrix}$$

Now, the reducing constants for each column are simply the minimal element in the column (which respectively are 0, 9, 11, 19, 21, 24). After subtracting these from their respective columns, we obtain the reduced matrix

$$\begin{bmatrix} 0 & 54 & 66 & 114 & 126 & 144 \\ 0 & 45 & 55 & 95 & 105 & 120 \\ 0 & 45 & 55 & 95 & 105 & 120 \\ 0 & 36 & 44 & 76 & 84 & 96 \\ 0 & 9 & 11 & 19 & 21 & 24 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Since by construction all elements of this reduced matrix are nonnegative, the sum of the reduced constants (namely, 192) is a lower bound. Of course, this lower bound is not nearly as good as the lower bound of 389, which as noted previously is the optimal cost value for the assignment problem. In the case of the traveling salesman problem Little et al. observed that the advantage of finding an optimal solution to the assignment problem in comparison to their simple reduction technique was mixed in terms of their computational results. The challenge tackled by Gavett and Plyter was to find for the case of the relaxed quadratic assignment problem an efficient method of obtaining an optimal basic feasible solution to the assignment problem (namely having zeroes along the diagonal of the reduced matrix and having nonnegative elements everywhere else). Their approach for doing this, which they called successive reduction, can be viewed as the primary technical contribution of their paper.

The successive reduction technique works as follows. Starting with the matrix **C**, the diagonal element in each column is subtracted from all other elements in its respective column. Then the smallest element in each row is subtracted from other elements in its row. After at most N repetitions of these two reductions, the desired reduced matrix is obtained.

We illustrate with the example. In the first iteration, the column reducing constants (namely 28, 78, 90, 115, 50, 28) are the diagonal elements of the original matrix **C**. Subtracting these from its respective column yields the matrix

$$\begin{bmatrix} 0 & 13 & 15 & 46 & 125 & 168 \\ -4 & 0 & 0 & 23 & 100 & 140 \\ -4 & 0 & 0 & 23 & 100 & 140 \\ -8 & -13 & -15 & 0 & 75 & 112 \\ -20 & -52 & -60 & -69 & 0 & 28 \\ -24 & -65 & -75 & -92 & -25 & 0 \end{bmatrix}$$

The row reducing constants are the respectively minimal elements in the rows (namely 0, –4, –4, –15, –69, –92). Subtracting these from the corresponding rows yields the matrix

$$\begin{bmatrix} 0 & 13 & 15 & 46 & 125 & 168 \\ 0 & 4 & 4 & 27 & 104 & 144 \\ 0 & 4 & 4 & 27 & 104 & 144 \\ 7 & 2 & 0 & 15 & 90 & 127 \\ 49 & 17 & 9 & 0 & 69 & 97 \\ 68 & 27 & 17 & 0 & 67 & 92 \end{bmatrix}$$

Observe that the matrix above is nonnegative and the sum of the reducing constants give a lower bound of 205.

In the second iteration, the column reducing constants are (0, 4, 4, 15, 69, 92), yielding the matrix

$$\begin{bmatrix} 0 & 9 & 11 & 31 & 56 & 76 \\ 0 & 0 & 0 & 12 & 35 & 52 \\ 0 & 0 & 0 & 12 & 35 & 52 \\ 7 & -2 & -4 & 0 & 21 & 35 \\ 49 & 13 & 5 & -15 & 0 & 5 \\ 68 & 23 & 13 & -15 & -2 & 0 \end{bmatrix}$$

and the row reducing constants are (0, 0, 0, –4, –15, –15), yielding the matrix

$$\begin{bmatrix} 0 & 9 & 11 & 31 & 56 & 76 \\ 0 & 0 & 0 & 12 & 35 & 52 \\ 0 & 0 & 0 & 12 & 35 & 52 \\ 11 & 2 & 0 & 4 & 25 & 39 \\ 64 & 28 & 20 & 0 & 15 & 20 \\ 83 & 38 & 28 & 0 & 13 & 15 \end{bmatrix}$$

The sum of all reducing constants from the first and second iterations yields a new lower bound of 355. Subsequent iterations proceed similarly with a nonnegative matrix at each iteration (as given in Gavett and Plyter) and with a lower bound given by the sum of the new and previous reducing constants. The reducing constants and lower bounds are given in Table 13.2.

Why does successive reduction work? At a technical level, the method iteratively reduces **C** in a way that adds zeroes in the diagonal and sub-diagonal elements,

Table 13.2 Gavett and Plyter's reduction constants and lower bounds for each iteration

Iteration	Column-reducing constant	Row reducing constant	Lower bound
1	28, 78, 90, 115, 50, 28	0, –4, –4, –15, –69, –92	205
2	0, 4, 4, 15, 69, 92	0, 0, 0, –4, –15, –15	355
3	0, 0, 0, 4, 15, 15	0, 0, 0, 0, –4, –4	381
4	0, 0, 0, 0, 4, 4	0, 0, 0, 0, 0, –2	387
5	0, 0, 0, 0, 0, 2	0, 0, 0, 0, 0, 0	389

increases the sum of the reducing constants, and ends each iteration with a nonnegative matrix (so that the sum of the reducing constants is a lower bound).

At a broader level, a key to understanding successive reduction (as well as the simple reduction proposed by Little et al.) is the dual to the assignment problem. To illustrate, we now consider the dual to the example assignment problem; see, e.g., Bazaraa and Jarvis (1977):

$$\text{Max } u_1 + u_2 + u_3 + u_4 + u_5 + u_6 + v_1 + v_2 + v_3 + v_4 + v_5 + v_6$$

s.t.

$u_1 + v_1 \le 28$	$u_1 + v_2 \le 91$	$u_1 + v_3 \le 105$	$u_1 + v_4 \le 161$	$u_1 + v_5 \le 175$	$u_1 + v_6 \le 196$
$u_2 + v_1 \le 24$	$u_2 + v_2 \le 78$	$u_2 + v_3 \le 90$	$u_2 + v_4 \le 138$	$u_2 + v_5 \le 150$	$u_2 + v_6 \le 168$
$u_3 + v_1 \le 24$	$u_3 + v_2 \le 78$	$u_3 + v_3 \le 90$	$u_3 + v_4 \le 138$	$u_3 + v_5 \le 150$	$u_3 + v_6 \le 168$
$u_4 + v_1 \le 20$	$u_4 + v_2 \le 65$	$u_4 + v_3 \le 75$	$u_4 + v_4 \le 115$	$u_4 + v_5 \le 125$	$u_4 + v_6 \le 140$
$u_5 + v_1 \le 8$	$u_5 + v_2 \le 26$	$u_5 + v_3 \le 30$	$u_5 + v_4 \le 46$	$u_5 + v_5 \le 50$	$u_5 + v_6 \le 56$
$u_6 + v_1 \le 4$	$u_6 + v_2 \le 13$	$u_6 + v_3 \le 15$	$u_6 + v_4 \le 23$	$u_6 + v_5 \le 25$	$u_6 + v_6 \le 28$

Observe that Little et al.'s simple reduction technique solves the first column and the last row of inequalities above as equations with $v_1 = 0$. This gives the dual solution $\mathbf{u} = (28, 24, 24, 20, 8, 4)$ and $\mathbf{v} = (0, 9, 11, 19, 21, 24)$. Using the fact that the coefficients of $\mathbf{C}$ result from products of increasing intensities across the columns and decreasing distances across the rows, it is straightforward to see that the solution is dual feasible and therefore by duality gives a lower bound (namely, 192). Thus, Little et al.'s technique is simply one way to find a dual feasible solution, and hence a lower bound.

Gavett and Plyter's method essentially generates a sequence of at most N dual feasible solutions where each component of a solution is the sum of the corresponding reducing constants at its iteration. In particular, for our example it yields the dual solutions given in Table 13.3.

Hence, Gavett and Plyter's technique is a way of optimally solving the dual problem in at most N iterations. It is just one of a number of possible ways of solving the dual problem of the relaxation of the quadratic assignment problem. Indeed,

Table 13.3 Dual feasible solutions and lower bounds for each iteration

Iteration	$(v_1, v_2, v_3, v_4, v_5, v_6)$	$(u_1, u_2, u_3, u_4, u_5, u_6)$	Lower bound
1	28, 78, 90, 115, 50, 28	0, –4, –4, –15, –69, –92	205
2	28, 82, 94, 130, 119, 120	0, –4, –4, –19, –84, –107	355
3	28, 82, 94, 134, 134, 135	0, –4, –4, –19, –88, –111	381
4	28, 82, 94, 134, 138, 139	0, –4, –4, –19, –88, –113	387
5	28, 82, 94, 134, 138, 141	0, –4, –4, –19, –88, –113	389

a simpler method than the successive reduction technique is to solve the inequalities along the diagonal and sub-diagonal (denoted above with a border) as equations and get the dual solution (with $u_1 = 0$) in just one iteration. Using the monotonic properties of **C**, it is straightforward to show that the solution obtained by doing this is dual optimal.

As for branching, Gavett and Plyter use the same framework (and even the same notation) as Little et al. with some minor modifications. Like Little et al., at each step "certain assignments are eliminated corresponding to pattern restrictions on the cost matrix. In the traveling salesman problem, this restriction involves eliminating subtours conflicting with already selected cities. In the facility-location problem *QAP*, this restriction means applying the labels associated with a selected element to eliminate other elements in the **C** matrix that would produce an unacceptable assignment at a future branch."

13.2.2 Alternative Branch-and-Bound Approaches to the Quadratic Assignment Problem

As previously noted, Gavett and Plyter (1966) used a pair-assignment formulation together with a row and column reduction technique to compute lower bounds at nodes of the branch-and-bound tree. A similar approach using only a column-reduced matrix was proposed independently by Land (1963); see, e.g., Pierce and Crowston (1971) for further discussion. While Gavett and Plyter's reduction technique gave an easy-to-compute optimal solution to the pair-assignment problem, this solution is often infeasible to the original quadratic assignment problem, resulting in a relatively weak bound; see, e.g., Christofides and Gerrard (1981). In their branching strategy, Gavett and Plyter implemented restrictions on branching variables in order to prevent multiple assignments, etc. These restrictions were computationally advantageous since the number of nodes in the branch-and-bound tree could be reduced through their use. Nevertheless, according to Burkard and Cela (1998), numerical results show that pair-assignment algorithms are outperformed by single-assignment algorithms.

Single-assignment strategies relate facilities directly to locations. The earliest strategies of this type were introduced by Gilmore (1962) and Lawler (1963). In his paper, Gilmore outlines an enumeration algorithm to solve the quadratic assignment problem, making use of lower bounds on the objective function. Also, he suggests two methods for computing lower bounds on partial permutations. As previously noted, one method uses the fact that a lower bound on the product of two given vectors of the same size can easily be determined by sorting them in opposite orders of magnitude and then taking the product of these sorted vectors. The other suggested method involves solving a linear assignment problem *LAP*. Lawler, on the other hand, used an integer linear

program to compute lower bounds with $n^4 + n^2$ variables $\{y_{ijkl}\}$ and $\{x_{ij}\}$ and $2n + n^4 + 1$ constraints:

$$MILP\text{: Min} \sum_{i,j,k,l} c_{ijk\ell} y_{ijk\ell}$$

$$\text{s.t.} \sum_{j} x_{ij} = 1, i = 1, \ldots, n \quad x_{ij} = 1, i = 1, \ldots, n$$

$$\sum_{i} x_{ij} = 1,\ j = 1, \ldots, n$$

$$\sum_{i,j,k,\ell} y_{ijk\ell} = n^2$$

$$x_{ij} + x_{k1} - 2y_{ijk1} \geq 0; \quad i, j, k, l = 1, 2, \ldots, n$$

$$x_{ij} = 0 \text{ or } 1, \quad i, j = 1, \ldots, n$$

$$y_{ijkl} = 0 \text{ or } 1, \quad i, j, k, l = 1, \ldots, n,$$

where c_{ijkl} is the joint cost of assigning entity i to location j and entity k to location ℓ. Lower bounds are created at partial assignment nodes by solving $O(n^2)$ linear assignment problems and then using the resulting objective function values as coefficients in a master *LAP*. Lawler acknowledges that his bounding technique is similar to that of Gilmore. The resulting bounds created are often cited as benchmarks in other research efforts regarding the quadratic assignment problem. As stated by Loiola et al. (2007), "the *QAP* lower bound presented by Gilmore and Lawler is one of the best known. Its importance is due to its simplicity and its low computational cost."

However, researchers have realized that the simplicity of computing the Gilmore and Lawler bound comes at a cost, as the bound is often not very tight for large instances of the quadratic assignment problem. Since the publication of the Gilmore-Lawler bound, research efforts have been directed toward finding improved bounds.

An obvious approach for obtaining lower bounds is to make use of the linear programming relaxation of the mixed integer linear program and its dual linear program (see for example, Assad and Xu (1985), Adams and Johnson (1994), Ramachandran and Pekny (1998), and Karisch et al. (1999). Using ideas from Drezner (1995), Resende et al. (1995) implemented an interior point algorithm to solve a relaxation of the mixed integer program.

A different formulation of the quadratic assignment problem has led to the generation of other bounding methods. Often, as in Gavett and Plyter, the coefficient c_{ijkl} is the product of b_{ik} (the flow or traffic between entity i, and entity k) and a_{jl} (the distance between locations j and l). With $\mathbf{B}$ the $[n \times n]$-dimensional flow matrix and $\mathbf{A}$ the $[n \times n]$-dimensional matrix, a *trace formulation* of the problem is

$$TF: \min tr(\mathbf{BXAX}^t),$$

$$\text{s.t. } \mathbf{X} \in S_n,$$

where $tr(\mathbf{M})$ is the trace of matrix $\mathbf{M}$ and S_n is the set of permutation matrices. Letting O_n represent the set of orthogonal matrices, it follows (since every permutation matrix is an orthogonal matrix) that a relaxation of the problem *TF* is

$$TFR: \min tr(\mathbf{BXAX}^t),$$

$$\text{s.t. } \mathbf{X} \in O_n.$$

The solution to the relaxation *TFR* is found by computing the eigenvalues of both matrices **B** and **A**, sorting one vector in nondecreasing order, the other in nonincreasing order, and then taking the product of the two resulting vectors. Unfortunately, the resulting *eigenvalue bound* has proven to be somewhat weak, but has been improved by enforcing additional constraints. For example, Hadley et al. (1992) enforce constraints on row and column sums resulting in a *projected eigenvalue bound*. Sometimes their bound was better than that by Gilmore and Lawler, and sometimes not.

Anstreicher and Brixius (2001) take a different approach by convexifying the quadratic objective function while making use of the derivation of the projected eigenvalue bound. Their formulation also makes use of optimal solutions of a semdefinite programming (*SDP*) problem related to the eigenvalue bound. They show that their bound is at least as good as the projected eigenvalue bound. Also, they have found that the value of their bound appears to increase much faster in comparison as branching occurs. This latter attribute is obviously very important in a branch-and-bound framework.

Use of this bound led to the first solution of several large benchmark problems, including the notorious "Nug 30" problem from Nugent et al. (1986). A nice summary of advances in quadratic assignment problem research as of the early 2000s can be found in Anstreicher (2003).

Recently, reformulation-linearization (*RL*) has been applied to the quadratic assignment problem to compute lower bounds. This technique involves multiplying equality constraints and nonnegativity constraints by product factors of the variables (reformulation). Then, each nonlinear term is replaced by a single variable, resulting in a mixed zero-one linear integer program (linearization). Reformulation creates redundant constraints, and different formulations are possible depending upon the product factors chosen in this step. As described by Adams et al. (2007) a level-1 reformulation (*RLT*-1) of the quadratic assignment problem is developed by multiplying each equality constraint and each nonnegativity constraint by each of the n^2 variables. For a level-2 reformulation (*RLT*-2), each constraint is multiplied by the product of two variables again creating redundant constraints. As before, reformulation is followed by linearization through substitution. Even higher levels of reformulation and linearization are possible through the use of higher level product forms, resulting in improved bounds, but at the cost of even larger zero-one linear programs. The resulting optimization problems can be quite large, but have been shown to provide relatively tight bounds. Adams et al. (2007) used Lagrangean

relaxation and dual ascent in a branch-and-bound framework to solve problems up to size $n = 30$ from Nugent et al. (1986) Although they found that their method required lower bound calculations at fewer nodes than competitive methods, computing each bound required a large amount of *RAM*. They cite a future research challenge as one of finding ways to reduce the *RAM* requirement.

Also recently, additional attention has focused on a semidefinite programming relaxation of the quadratic assignment problem, see Zhao et al. (1998) and Rendl and Sotirov (2007), as well as a reformulation-linearization semidefinite programming relaxation (also called a lift-and-project relaxation), see Burer and Vandenbussche (2006) and Lovasz and Schrijver (1991) for details. Interestingly, the equivalence between these two relaxations for the quadratic assignment problem was recently shown by Povh and Rendl (2009). Using a bundle method to solve the resulting problem, Rendl and Sotirov in 2003 obtained the tightest lower bounds at that time for a large number of test problems. More recently, Burer and Vandenbussche (2006) used an augmented Lagrangian method and derived even tighter bounds on a number of test problems. Exploiting a special structure in the data matrices of certain quadratic assignment problems, de Klerk and Sotirov (2008) have found even tighter lower bounds than Burer and Vandenbussche on some problems.

Loiola et al. (2007) provide a recent survey on the quadratic assignment problem, including a discussion on different approaches used to solve the problem. In particular, the paper includes data on lower bound values found and run times of several competing methods, including those mentioned above, applied to classical test problems.

13.2.3 Special Cases of the Quadratic Assignment Problem that are Solvable in Polynomial Time

We now briefly review some of the work that considers special cases of the quadratic assignment problem with particular emphasis on cases that can be solved in polynomial time. Burkard et al. (1997) considered the special case, in which c_{ijkl} is the product of the flow between facilities i and k, and the distance between locations j and l. They showed that if $2n$ numbers b_i^r, b_i^c, $i = 1, \ldots, n$ exist and can be associated with the rows and columns of the flow matrix such that $b_{ik} = b_i^r + b_k^c$ for all i and k, then the problem is reducible to the linear assignment problem and therefore is solvable in polynomial time. The result is also true if the distance matrix can be decomposed in a similar manner. Ergodan (2006) shows that this result can be generalized to a broader class of quadratic assignment problems that are "additively decomposed."

Ergodan also considers "multiplicative decomposition" and has the following result. Suppose there exists $\{v_{ij}: i, j = 1, \ldots, n\}$ where $c_{ijk\ell} = v_{ij}\, v_{k\ell}$, for all i, j, k, ℓ. Then if the optimal objective function value of the linear assignment problem with coefficients $\{v_{ij}\}$ is nonnegative, then the linear assignment problem solves the corresponding quadratic assignment problem.

Ergodan and Tansel (2006) consider the case where the n-node flow graph has a path structure (it has no cycles and every node has a degree of 0, 1, or 2) and the n by n distance matrix is induced by a grid graph in the following sense. With $rc = n$ for

two positive integers r and c, let G_{rc} be the undirected grid graph with rc nodes, where the nodes are arranged in r rows and c columns, and where the arc set consists of arcs connecting adjacent nodes in the same row, or adjacent nodes in the same column. Define D_{ab} as the $[n \times n]$-dimensional matrix of shortest path distances in G_{ab}. Then if the distance matrix **A** of the quadratic assignment problem is identical to hD_{ab} for some positive h, D is said to be induced by a grid graph. For this special structure, Erdogan and Tansel show that the quadratic assignment problem is solvable in $O(n)$ time.

For information on other special structures that lead to polynomial-time solvability, see Erdogan (2006).

13.3 Efroymson and Ray (1966): The Uncapacitated Facility Location Problem

Also in the early 1960s there was considerable research interest in another problem known today as the Uncapacitated Facility Location Problem (*UFLP*). Our purpose here is to report on perhaps the earliest published use of the branch-and-bound technique to solve the problem exactly. We will explain how branch-and-bound was used in the paper by Efroymson and Ray (1966). The problem setting involves several "demand points" (customers) requiring service from one or more potential "plant sites." There is a given supply cost between a given demand point and potential plant site that will be incurred if the plant is opened and the demand is serviced from the plant. In addition, there is a fixed cost to open each plant.

13.3.1 Solving the Uncapacitated Facility Location Problem via Branch-and-Bound

To formally pose the uncapacitated facility location problem as an optimization problem, suppose there are n customer locations $j = 1, \ldots, n$ and m potential plants $i = 1, \ldots, m$. The following mixed integer program is a prototypical formulation of the problem:

$$UFLP: \text{Min } Z = \sum_{i,j} c_{ij} x_{ij} + \sum_{i} f_i y_i \tag{13.4}$$

$$\text{s.t.} \sum_{i} x_{ij} = 1, j = 1, \ldots, n \tag{13.5}$$

$$x_{ij} \leq y_i \ \forall i, j \tag{13.6}$$

$$x_{ij} \geq 0 \ \forall i, j \tag{13.7}$$

$$y_i = 0 \text{ or } 1 \ \forall i \tag{13.8}$$

where we define the parameters

c_{ij}: the cost to service all of customer j's demand from plant i
f_i: the nonnegative cost of opening plant i

and the variables

x_{ij}: the fraction of customer j's demand satisfied by plant i, and
y_i = 1 if plant i is open, and 0 otherwise.

Thus, the decision problem is to decide which plants to open (which y_i values to set to one) and which open plant(s) will service each customer. The overall objective is to minimize total cost. Note that the allocation variables x_{ij} are continuous and take on values between zero and one. This is why c_{ij} represents the cost of servicing *all* demand and so $c_{ij}\, x_{ij}$ denotes proportional costing. In many applications, c_{ij} is determined by a transportation cost per unit multiplied by total demand of customer j. Finally, constraint (13.6) forces plant i to be open whenever $x_{ij} > 0$ for some j. There are many applications of this classical location problem and we outline two of these in what follows.

Krarup and Bilde (1977) describe an application in manufacturing called the dynamic economic lot size problem. A manufacturer of a single product needs to develop a production plan for the next n months in order to satisfy demand for the product in each of these months. Producing the product in month i incurs a fixed setup cost f_i as well as a per-unit manufacturing cost p_i. Demand for the product in month j is denoted as d_j, and d_j can be satisfied by production in month j and/or some earlier month. However, units produced earlier than needed incur a holding cost, where r_i is the per unit cost of holding one unit from month i to month $i + 1$. Define c_{ij} as the cost of manufacturing and (if necessary) holding all of month j's demand when production occurs in month $i \leq j$. Thus,

$$c_{ij} = \begin{cases} d_j(p_i + \sum_{t=i}^{j} r_t), & \text{for } i < j \\ d_j p_i & \text{for } i = j \\ \infty & \text{for } i > j \end{cases}$$

Note that if $i = j$, then no holding cost is incurred. Also units produced in month i cannot be used to satisfy demand in some earlier month. However, if it is possible to backorder demand, then c_{ij} for $j < i$ could be finite, but most likely would involve a per-unit (and per-period) backorder cost. Letting $y_i = 1$ if and only if production occurs in month i, and x_{ij} as the fraction of month j's demand produced in month i, the uncapacitated facility location problem is solved to minimize total setup, manufacturing, and holding cost over the n-month planning horizon.

In the days before electronic funds transfer, the time to clear a check often depended on which bank the check was drawn on, and the location of the recipient of the check. After all, checks were often delivered by the postal service. Thus, a company might want to maximize the total funds that are in transit. However, maintaining an account at a given bank is not costless. With c_{ij} as the "dollar days" (float)

in transit from bank i to customer j and f_i the cost to maintain an account at bank i, the firm is faced with the problem of maximizing $\sum_{i,j} c_{ij}x_{ij} - \sum_i f_i y_i$ subject to the constraints (13.5)–(13.8). Note that we are maximizing a modified version of (13.4), but structurally the problems are the same. Cornuejols et al. (1990) call this problem the Bank Account Location Problem. A mirror image of this problem is called the Lock Box Problem, where a firm collecting funds wishes to minimize "float." For more on the above problem see also Cornuejols et al. (1977).

Efroymson and Ray recognized that practical instances of the uncapacitated facility location problem might have several thousand rows and columns and that contemporary integer programming techniques could not hope to solve such large problems in a reasonable amount of time. They therefore sought methods to solve the overall problem via a sequence of smaller subproblems.

Note that for *fixed* values of the y_i variables $\{y_i', i = 1, \ldots, n\}$, where at least one $y_i' = 1$, an optimal **x**-vector can be found easily by setting, for each value of $j, x_{ij} = 1$ if $c_{ij} = \min\{c_{ij} : y_i' = 1\}$. In other words, for each j, find the smallest c_{ij} over those plants i for which the corresponding y variable is set to one. An efficient solution method is to find a means of computing good **y**-vectors that will eventually lead to an optimal **y**-vector. Combining the above observation with the fact that the solution to a linear programming relaxation of an mixed integer program creates a lower bound to it (given a minimization objective), Efroymson and Ray made extensive use of the linear program *LPR* defined below.

Let N_k be the set of indices of those plants that can supply customer k and P_i be the set of indices of those customers that can be supplied from plant i, where n_i is the number of elements in P_i. Note that N_k might be all plants and P_i might be all customers, but practical considerations often prohibit some links. With these definitions, consider the following linear program:

$$LPR: \text{Min}\, Z_L = \sum_{i,j} c_{ij}x_{ij} + \sum_i f_i y_i \tag{13.9}$$

$$\text{s.t.} \sum_{i \in N_j} x_{ij} = 1, \quad j = 1, \ldots, n \tag{13.10}$$

$$\sum_{j \in P_i} x_{ij} \leq n_i y_i, \quad i = 1, \ldots, m \tag{13.11}$$

$$x_{ij} \geq 0 \ \forall i, j \tag{13.12}$$

$$y_i = 0 \text{ or } 1 \ \forall i = 1, \ldots, m \tag{13.13}$$

Efroymson and Ray made use of the *LPR* formulation in their branch-and-bound method. Since *UFLP* has both continuous and integer variables, it is natural to

branch on the zero-one variables y. Thus, at some node in the branch-and-bound tree, some of the y variables may be set to zero (their indices collected in a set labeled K_0), some may be set to one (their indices are included in a set labeled K_1), while the status of some of the remaining variables y has not yet been decided. We denote this latter set of indices as K_2.

A simple procedure can be used to solve the problem *LPR* without the use of an linear programming solver. The authors observed that the optimal allocation variables $\{x_{ij}^*\}$ and corresponding allocation costs $\{AC_j^*\}$ could be constructed as shown in *Algorithm* 1.

Algorithm 1: Solution Algorithm for the LPR Problem

Step 1: Find $AC_j^* \equiv \min\{\min\{c_{ij}: i \in K_1\}, \min\{c_{ij} + f_i/n_i: i \in K_2\}$ for $j = 1, \ldots, n$.

Step 2: Set $x_{ij}^* = 1$ for that value of i that attains AC_j^* in Step 1, and $x_{ij}^* = 0$ otherwise.

The optimal y variables for those plants with indices in K_2 are then computed as $y_i* = (1/n_i) \sum_{j \in P_i} x_{ij}^*$. The optimal objective function value at the node, accounting for those plants $i \in K_1$ that are fixed open is then $Z_L^* = \sum_{i \in K_1} f_i + \sum_{j=1}^{m} AC_j^*$.

The above procedure solves *LPR* because relation (13.11) will hold as an equation at an optimal solution. Thus, those y_i variables $i \in K_2$ can be removed from (13.9) by substitution. Using the above ideas, *LPR* can be solved by finding the minimum entry in each column of a $[(|K_1| + |K_2|) \times m]$-dimensional matrix. Note that the value Z_L^* can often be a fairly weak lower bound on *UFLP* at the current node. This is especially true when the number of customers actually served by plant i, $i \in K_2$, is considerably smaller than n_i. When this occurs, only a fraction of the full cost f_i of opening the plant is accounted for. Realizing this fact, Efroymson and Ray developed "simplification rules," i.e., conditions that can be used to either optimally fix the values of some members of K_2 in all solutions that emanate from the current node, or to reduce n_i.

The first rule is to set $y_i = 1$, $i \in K_2$ if it is known that the net savings in allocation costs with this plant open is at least as large as the fixed cost to open the plant. For any j, if plant i, $i \in K_2$ is not open, then $c_{\sim j} \equiv \min\{c_{kj}: k \in K_1 \cup K_2, k \neq i\}$ is the minimum possible cost to serve demand j by either a plant k, $k \in K_1$ that is fixed open, or some other plant $k \in K_2$ that might be opened. But then if $c_{\sim j} - c_{ij} > 0$, opening plant i would certainly provide an allocation cost savings to serve demand j. If the sum of these savings over all demands is at least as large as f_i, it is optimal to open plant i. More formally, let

$$\Delta_{ij}^o \equiv \max\{(\min\{c_{kj} : k \in K_1 \cup K_2, k \neq i\} - c_{ij}), 0\} \tag{13.14}$$

Rule 1: If $\sum_j \Delta_{ij}^o > f_i$, set $y_i = 1$.

On the other hand, if the net savings in allocation costs with plant i open is known to be no more that the cost to open the plant, then set $y_i = 0$. To implement this rule, restrict k to be in K_1 in (13.14) and define Δ_{ij}^c to be the computed value. Then,

Rule 2: If $\sum_j \Delta_{ij}^c \leq f_i$, set $y_i = 0$.

The final simplification provided by Efroymson and Ray involves the reduction of n_i. Note that reducing n_i can provide a stronger lower bound at the node. Suppose that j is currently in the set P_i. If for some *open* plant k we find that $c_{kj} \leq c_{ij}$, then demand j will be no worse off by eliminating plant i as a potential server of j's demand, i.e., we can safely eliminate index j from the set P_i, thereby reducing $|P_i|$ by 1. More formally,

Rule 3: Let $J(i) \equiv \{j \in P_i: \min \{c_{kj} : k \in K_1\} - c_{ij} \leq 0\}$. Eliminate $J(i)$ from P_i and reduce n_i by $|J(i)|$.

13.3.2 Alternative Branch-and-Bound Approaches to the Uncapacitated Facility Location Problem

Perhaps the best-known contribution to solution methods for the uncapacitated facility location problem is by Erlenkotter (1978). His approach involves working with the dual problem, solving a reduced nonlinear form of the dual heuristically through ascent and adjustment of the dual variables. The result of this method is the *DUALOC* algorithm that is frequently cited in the literature. Bilde and Krarup's (1977) method is similar to Erlenkotter's and was developed at approximately the same time. The ascent/adjustment method often produces an optimal dual solution that can possibly be used to construct an optimal primal solution. If not, the dual objective function value can be effectively used in a branch-and-bound algorithm to solve the uncapacitated facility location problem.

Another approach is to strengthen the lower bounds created by the linear programming relaxation of *UFLP*. One way to do this is to find inequalities to add as constraints to the linear program which cut off portions of the linear programming polyhedron that are known to not contain an optimal solution to the problem. These added constraints are often called valid inequalities and have been studied by many researchers. In particular, it is of value to eliminate extreme points that correspond to fractional solutions, since such solutions are infeasible to the uncapacitated facility location problem.

Cho et al. (1983a) study the issue of generating so-called facet inequalities that describe the integer polyhedron of *UFLP*. Such an approach has great value since the integer polyhedron is contained in the linear programming polyhedron. The authors state:

> This approach deserves attention since facets are the "strongest cutting planes." One can thus reasonably expect to improve computational results for any solution method which is based on linear programming even if one can identify only a subset of these facets.

They make use of a node-packing reformulation of the uncapacitated facility location problem, and are able to characterize all facets for the case of three plants ($m = 3$) and several destinations. In a companion paper, Cho et al. (1983b) identify all facets for the case of three customers ($n = 3$) and several plants.

Goldengorin et al. (2003) use a pseudo-Boolean polynomial-based representation of *UFLP* to solve the problem. Their algorithm, called branch-and-peg, uses rules to determine (before branching) whether a plant will (or will not) be located at certain sites in the current subproblem under examination. This "pegging" operation is applied to each subproblem and reduces its size. The authors report that on a number of problems solved, branch-and-peg took on average less than 10% of the execution time of branch-and-bound when the transportation matrix was dense.

Beltran-Royo et al. (2007) apply a concept called *Semi-Lagrangean Relaxation* to *UFLP*. The idea is to dualize the equality constraints (13.5) to form the dual function, but then add the constraints $\sum_i x_{ij} \leq 1,\ j = 1, \ldots, n$ to the dual problem. Adding the constraints increases the lower bound when the subproblem is solved to optimality. Unfortunately, the resulting subproblem is **NP**-hard, but the authors found that often the subproblems are smaller in dimension that the original primal problem. In those instances, they used *CPLEX* to solve the dual problem.

Algorithm 2: Variable Neighborhood Search: A Generic Algorithm

Step 1: Identify a (perturbed) solution in the k-th neighborhood of an incumbent. This step is frequently referred to as "shaking").
Step 2: Perform a local search from the perturbed solution.
Step 3: Move to an improved solution.

In a recent paper, Hansen et al. (2007) use a three-phase approach to solve large instances of *UFLP*. A key feature of their method is the use of *variable neighborhood search* (*VNS*). The idea of variable neighborhood search is to explore the neighborhood of a current solution. Once a neighborhood structure is defined, a distance function must be developed that describes the dissimilarity of two solutions. Then, for a given solution, points in the k-th neighborhood can be identified. Variable neighborhood search consists of the repetitive sequence of three basic steps that are shown in *Algorithm* 2.

There are three phases to their overall approach to solving unconstrained facility location problems. These phases integrate variable neighborhood search as a key ingredient. The procedure can be described as shown in *Algorithm* 3.

Algorithm 3: Solving UFLP with Variable Neighborhood Search

Phase 1: Apply variable neighborhood search directly to *UFLP* to find a good primal solution. This step provides an upper bound of the problem.

Phase 2: Find an exact solution to the dual of the linear programming relaxation of *UFLP*. Variable neighborhood search is also used in this phase of this approach. The dual solution provides a lower bound.

Phase 3: A branch-and-bound procedure is then implemented making use of the upper and lower bounds from Phases 1 and 2. With their method, the authors reported success in solving very large problem instances.

13.3.3 Special Cases of the Uncapacitated Facility Location Problem that are Solvable in Polynomial Time

In addition to research efforts to improve bounds for the uncapacitated facility location problem, another research focus on the problem has been to identify special cases that can be solved to optimality in polynomial time. Kolen (1982) observed that *UFLP* could be transformed to an equivalent covering problem. Then, if the covering matrix of the resulting problem is *totally balanced,* it can be transformed through row and column operations into *standard greedy* form. (A totally balanced zero-one matrix contains no square submatrix with row and column sums equal to two, and such a matrix is in standard greedy form if it does not contain a submatrix of the form $\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$. Hoffman et al. (1985) give a polynomial time algorithm for this transformation.) When this can be done, Kolen shows that this covering problem can be solved in polynomial time, see also Kolen and Tamir (1990).

Jones et al. (1995) identified another class of uncapacitated facility location problems, where not every instance fits the Kolen framework but that can still be solved to optimality in polynomial time. An instance is in this class if facility and demand point indices can be ordered so that the following holds:

(a) *Continuity*: If j, $\ell \in P_i$, then $k \in \mathrm{P}_i$ where $j < k << \ell$.
(b) *Cascading*: For all $i<t$, $\min\{j: j \in \mathrm{P}_i\} \leq \min\{j: j \in \mathrm{P}_t\}$, and $\max\{j: j \in \mathrm{P}_i\} \leq \max\{j: j \in \mathrm{P}_t\}$.
(c) *Monotonicity*: For all j, if $i, t \in \mathrm{N}_j$, and if $c_{ij} \leq c_{tj}$, then $c_{ik} \leq c_{tk}$ for all k where $i, t \in N_k$.

In addition to giving an $O(nm)$ algorithm for such instances, the authors identify several problems that satisfy the conditions (a), (b) and (c). These problems in-

clude the tool selection problem of Daskin et al. (1990), a substitutable inventory problem, a stochastic demand problem, the discrete lot sizing problem discussed by Wagner and Whitin (1958), and a facility location problem on the line.

13.4 Branching Strategies in Branch-and-Bound Procedures

For the most part, in this chapter we have focused on the "bound" part of branch-and-bound methods for the two location problems considered, because bounding techniques by their very nature need to be problem dependent. Nevertheless, our chapter would not be complete without at least a brief discussion of what seem to be some promising areas of research in the "branch" part of branch-and-bound procedures. These ideas can be applied to any mixed integer programming problem and thus are not restricted to location problems. Two key references for these ideas are Linderoth and Savelsberg (1999), and Achterberg et al. (2005).

As mentioned in the introduction to this chapter, the branch-and-bound process is most easily envisioned via a tree, where the top node of the tree is the original problem, and various branches are created through partitioning the set of feasible solutions to the problem. The "deeper" one is in the tree, the more options there are for selecting the next node for partitioning of the subset of solutions represented by that node. A significant amount of research has taken place regarding the node to be selected for partitioning, as well as how to perform the partition. In what follows, we will continue to assume that the original problem is one of minimization.

Regarding node selection, a popular method is to choose the node that has the smallest lower bound, where this bound is often found via linear programming relaxation. This method, when applied in its purest sense, is often called best-bound (or breadth-first) search. Another method, called depth-first search, is to continue searching down the tree until a feasible solution is found. Other methods include estimating the value of the best feasible integer solution obtainable from a given node in the tree, or combining depth first search early in the process and breadth-first search methods later in the process.

As described by Linderoth and Savelsberg, one way to partition the feasible region represented by a given node is to select a single variable that does not take on an integer value in the linear programming relaxation solution, but must be integer in an over-all optimal solution; then create two subregions by constraining this variable with an upper bound and a lower bound (they call this *variable dichotomy*). Below we discuss some methods for determining the variable to be "dichotomized." Another method is applicable when certain generalized upper bounding constraints are present in the original problem. The generalized upper bounding dichotomy is a means of partitioning by bounding the sum of different subsets of the variables to create different subregions.

Returning to variable dichotomy, there remains the issue of variable selection. Some authors have tested the use of "pseudocosts," i.e., estimates of changes in the objective function value when a variable is rounded up or rounded down. The estimates are usually created by using the objective function values of the corresponding linear programming relaxations. Average pseudocosts for a given variable can also be determined by gathering "local" pseudocosts at several nodes and computing the mean of the set. However they are computed, these pseudocosts can be used to help select the partitioning variable.

Another promising approach is called "strong branching," which involves testing the set (or a subset of) the fractional variable candidates to find those that appear to give the best progress before actually branching on any of them. "Full strong branching" involves *all* fractional variables and thus it may be computationally prohibitive to solve all the corresponding linear programming problems to optimality. Thus, some authors have considered testing just a subset of these variables and not solving the linear programs to optimality, instead performing a limited number of dual simplex pivots. Hybridized versions of these techniques are also possible.

Both Linderoth and Savelsberg (1999), and Achterberg et al. (2005) provide results on computational testing of the above ideas applied to a number of mixed integer programming problems as well as references to the work of others.

13.5 Conclusions

Herein we reviewed the use of branch-and-bound in solving exactly two important location problems, the quadratic assignment problem and the uncapacitated facility location problem. Our focus was on the early application of branch-and-bound to these problems via a critical review of two classical papers from the 1960s, namely Gavett and Plyter (1966) on the quadratic assignment problem and Efroymson and Ray (1966) on the uncapacitated facility location problem. In providing these reviews we attempted to replicate the authors' thought processes in the development of the reported solution method and to discuss how these papers set the stage for subsequent research.

The quadratic assignment problem is generally recognized as one of the most difficult combinatorial optimization problems. After an initial lull of research activity in this problem (until the mid-1970s), research on this topic has exploded. In spite of this activity, however, an exact solution to the problem has remained elusive for modest and large size problems. Yet recently, significant results have been obtained; see, e.g., Adams et al. (2007), Anstreicher (2003), Burer and Vandenbussche (2006), De Klerk and Sotirov (2008, 2009), and Rendl and Sotirov (2007). The research activity and the results are nicely summarized in the comprehensive review paper of Loiola et al. (2007). The result of this research has been better lower bounds and an approximate doubling in the size of problems that can be solved exactly in the last 10 years (from about $n = 15$ to about 30). Unfortunately, $n = 30$ is still a relatively small problem. In practice, this means that heuristic and metaheuristic approaches

are needed to attempt to solve the problem. Again, see Loiola et al. (2007) for an excellent review.

In contrast, much progress has been made in solving the uncapacitated facility location problem. As noted herein, large instances of the *UFLP* can now be solved; see, e.g., Beltran-Royo et al. (2007) and Hansen et al. (2007).

Acknowledgements The authors would like to recognize and thank Kurt Anstreicher and Samuel Burer for their suggestions. The second author wishes to thank Renata Sotirov for many stimulating discussions on the quadratic assignment problem.

References

Achterberg T, Koch T, Martin A (2005) Branching rules revisited. Oper Res Lett 33:42–54

Adams WP, Johnson TA (1994) Improved linear programming-based lower bounds for the quadratic assignment problem. In: Pardalos PM, Wolkowicz H (eds) Quadratic assignment and related problems. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 16. American Mathematical Society, Rhode Island, pp 43–75

Adams WP, Guignard M, Hahn PM, Hightower WL (2007) A level-2 reformulation--linearization technique bound for the quadratic assignment problem. Eur J Oper Res 180:983–996

Anstreicher KM (2003) Recent advances in the solution of quadratic assignment problems. Math Program 97:27–42

Anstreicher KM, Brixius NW (2001) A new bound for the quadratic assignment problem based on convex quadratic programming. Math Program 89:341–357

Assad AA, Xu W (1985) On lower bounds for a class of quadratic 0,1 programs. Oper Res Lett 4:175–180

Bazaraa MS, Jarvis JJ (1977) Linear programming and network flows. Wiley, New York

Beltran-Royo C, Vial J-Ph, Alonso-Ayuso A (2007) Solving the uncapacitated facility location problem with semi-Lagrangian relaxation. Optimization On-line. http://www.optimization-online.org/DB_HTML/2007/02/1597.html. Accessed 22 Sept 2009

Bilde O, Krarup J (1977) Sharp lower bounds and efficient algorithms for the simple plant location problem. Ann Discrete Math 3:79–97

Burer S, Vandenbussche D (2006) Solving lift-and project relaxations of binary integer programs. SIAM J Optim 16:726–750

Burkard RE, Cela E (1998) The quadratic assignment problem. In: Du D-Z, Pardalos PM (eds) Handbook of combinatorial optimization, vol. 3. Kluwer, Dordrecht, pp 241–337

Burkard RE, Cela E, Demindenko VM, Metelski NN, Woeginger GJ (1997) Perspectives of easy and hard cases of the quadratic assignment problem. SFB Report No. 104, Institute of Mathematics, Technical University Graz, Austria

Christofides N, Gerrard M (1981) A graph theoretic analysis of bounds for the quadratic assignment problem. In: Hansen P (ed) Studies on graphs and discrete programming. North-Holland, New York, pp 61–68

Cho DC, Johnson EL, Padberg M, Rao MR (1983a) On the uncapacitated plant location problem I: valid inequalities and facets. Math Oper Res 8:579–589

Cho DC, Padberg M, Rao MR (1983b) On the uncapacitated plant location problem II: facets and lifting theorems. Math Oper Res 8:590–612

Conway RW, Maxwell WL (1961) A note on the assignment of facility location. J Ind Eng 12:34–36

Cornuejols G, Fisher ML, Nemhauser GL (1977) Location of bank accounts to optimize float: an analytical study of exact and approximate algorithms. Manag Sci 23:789–810

Cornuejols G, Nemhauser GL, Wolsey LA (1990) The uncapacitated facility location problem. In: Mirchandani P, Francis R (eds) Discrete location theory. Wiley, New York

Daskin M, Jones PC, Lowe TJ (1990) Rationalizing tool selection in a flexible manufacturing system for sheet metal products. Oper Res 38:1104–1115

De Klerk E, Sotirov R (2008) Exploiting group symmetry in semidefinite relaxations of the quadratic assignment problem. Math Program, series A. Published online http://www.springerlink.com/content/85302n245v250051/fulltext.pdf. Accessed 22 Sept 2009

De Klerk E, Sotirov R (2009) Improved semidefinite bounds for quadratic assignment problems with suitable symmetry. Technical report. Available at http://stuwww.uvt.nl/~sotirovr/B&B_QAP.pdf. Accessed 22 Sept 2009

Drezner Z (1995) Lower bounds based on linear programming for the quadratic assignment problem. Comput Optim Appl 4:159–165

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

Erlenkotter D (1978) A dual-based procedure for the uncapacitated facility location problem. Oper Research 26:992–1009

Erogodan G (2006) Quadratic assignment problem: linearizations and polynomial time solvable cases. PhD Thesis, Department of Industrial Engineering, Bilkent University, Ankara, Turkey

Ergodan G, Tansel B (2006) A Note on a polynomial time solvable case of the quadratic assignment problem. Discrete Optim 3:382–384

Gavett JW, Plyter NV (1966) The optimal assignment of the facilities to locations by branch-and-bound. Oper Res 14:210–232

Gilmore PC (1962) Optimal and suboptimal algorithms for the quadratic assignment problem. SIAM J Appl Math 10:305–313

Goldengorin B, Ghosh D, Sierksma G (2003) Branch and peg algorithms for the simple plant location problem. Comp Oper Res 30:967–981

Hadley SW, Rendl F, Wolkowicz H (1992) A new lower bound via projection for the quadratic assignment problem. Math Oper Res 17:727–739

Hansen P, Brimberg J, Urosevic D, Mladenovic N (2007) Primal-dual variable neighborhood search for the simple plant location problem. INFORMS J Comp 19:552–564

Hillier FS, Lieberman GJ (1980) Introduction to operations research, 3rd edn. Holder-Day, San Francisco

Hoffman A, Kolen A, Sakarovich M (1985) Totally balanced and greedy matrices. SIAM J Algebraic Discrete Methods 6:721–730

Jones PC, Lowe TJ, Muller G, Xu N, Ye Y, Zydiak JL (1995) Specially structured uncapacitated facility location problems. Oper Res 43:661–669

Karisch SE, Cela E, Clausen J, Espersen T (1999) A dual framework for lower bounds of the quadratic assignment problem based on linearization. Computing 63:351–403

Kolen A (1982) Location problems on trees and in the rectilinear plane. Stitchting Mathematisch Centrum, Amsterdam

Kolen A, Tamir A (1990) Covering problems. In: Mirchandani P, Francis R (eds) Discrete location theory. Wiley, New York, pp 263–304 (Chap. 6)

Koopmans TC, Beckmann M (1957) Assignment problems and the location of economic activities. Econometrica 25:53–76

Krarup J, Bilde O (1977) Plant location, set covering and economic lot size: an $O(mn)$ algorithm for structured problems. In: Collatz L, Wetterling W (eds) Numerische Methoden bei Optimierungsaufgaben. Optimierung in graphentheoritischen und ganzzahligen Problemen, vol 3. International Series of Numerical Mathematics 36. Birkhaeuser, Basel, pp 155–180

Land AH (1963) A problem of assignment with interrelated cost. Oper Res Quart 14:185–198

Land AH, Doig AG (1960) An automatic method for solving discrete programming problems. Econometrica 27:497–540

Lawler EL (1963) The quadratic assignment problem. Manag Sci 9:586–599

Lawler EL, Wood DE (1966) Branch and bound methods: a survey. Oper Res 14:699–719

Linderoth JT, Savelsbergh MWP (1999) A computational study of search strategies for mixed integer programming. INFORMS J Comput 11:173–187

Little JDC, Murty KG, Sweeney DW, Harel C (1963) An algorithm for the traveling salesman problem. Oper Res 11:972–989

Loiola E, de Abreu NMM, Boaventura-Netto PO, Hahn P, Querido T (2007) A survey for the quadratic assignment problem. Eur J Oper Res 176:657–690

Lovasz L, Schrijver A (1991) Cones of matrices and set-functions, and 0-1 optimization. SIAM J Optim 1:166–190

Nugent CE, Vollmann TE, Ruml J (1986) An experimental comparison of techniques for the assignment of facilities to locations. Oper Res 16:150–173

Pierce JF, Crowston WB (1971) Tree-search algorithms for the quadratic assignment problem. Nav Res Logist Quart 18:1–36

Povh J, Rendl F (2009) Copositive and semidefinite relaxations of the quadratic assignment problem. Discrete Optim 36:231–241

Ramachandran B, Pekny JF (1998) Lower bounds for nonlinear assignment problems using many body interactions. Eur J Oper Res 105:202–215

Rendl F, Sotirov R (2007) Bounds for the quadratic assignment problem using the bundle method. Math Program B 109:505–524

Resende MGC, Ramakrishnan KG, Drezner Z (1995) Computing lower bounds for the quadratic assignment with an interior point algorithm for linear programming. Oper Res 43:781–791

Wagner H M, Whitin TM (1958) Dynamic version of the economic lot size model. Manag Sci 5:89–96

Zhao Q, Karisch SE, Rendl F, Wolkowicz H (1998) Semidefinite programming relaxations for the quadratic assignment problem. J Comb Optim 2:71–109

Chapter 14
Exploiting Structure: Location Problems on Trees and Treelike Graphs

Rex K. Kincaid

14.1 Introduction

Posing location problems on graphs or networks has sharpened our understanding of what underlying structures can be exploited to prove theorems and to develop efficient algorithms. The construction of efficient algorithms for network location problems has been greatly aided by the work of computer scientists who have devised algorithms and data structures that allow efficient traversal and storage of graphs. In this chapter the contributions of three early location analysis papers are examined in detail. Key ideas are identified and their effects traced forward through the literature. In addition, natural extensions of these key ideas are included. However, no attempt is made to be encyclopedic when surveying the literature. The chapter concludes with comments on future research directions.

The first two early papers examined are less well known than the third, but for different reasons. The paper by Harary and Norman (1953), although published in a highly visible mathematics journal, is not well known to the network location community. Largely, this is due to the fact that the title and main results in the paper are not about location problems. The second paper was originally published in Chinese in 1961 by Hua and was translated into English by the American Mathematical Society and published in 1962 (Hua et al. 1961). The third paper studied is a seminal (Western) location paper. Goldman (1971) is a well known and highly cited paper in network location literature.

Each of these three papers addresses location problems on tree networks and hints at ways to extend these results for treelike graphs. Many authors have focused on location problems restricted to tree networks. The second part of Tansel, Francis and Lowe's 1983 survey paper is one such example and provides an extensive list of references and problem types for tree networks. At the close of Tansel et al. (1983), the authors ask two questions, the first of which is particularly interesting

R. K. Kincaid (✉)
Department of Mathematics, The College of William and Mary,
Williamsburg, VA 23187-8795, USA
e-mail: rrkinc@math.wm.edu

H. A. Eiselt, V. Marianov (eds.), *Foundations of Location Analysis,*
International Series in Operations Research & Management Science 155,
DOI 10.1007/978-1-4419-7572-0_14,

to this chapter: other than trees, what special network structure leads to efficient algorithms for network location problems?

14.2 Preliminaries

Before summarizing the contributions of the three papers, a few definitions provide a common language for the ensuing exposition. A *graph G* consists of a nonempty set of *vertices*, *V*, and a set of *edges*, *E*. In Fig. 14.1, *G* has nine vertices and nine edges. The number of edges incident to a vertex is the *degree* of the vertex. Vertex 1 has a degree of five while vertex 9 has a degree of one. A *path* is a sequence of consecutive edges in a graph (no repeated edges or vertices). Edges (4,1), (1, 2), and (2, 7) form a path from vertex 4 to vertex 7. A graph is *connected* if there is a path between every pair of vertices. A *cutvertex* of a connected graph is a vertex whose removal (along with all edges incident with it) disconnects the graph. The resulting disconnected pieces of the graph are called *subgraphs*. Throughout this chapter no distinction is made between the terms graph and network.

The notation $x \in G$ denotes x as any point on the graph G (along an edge or at a vertex). Associated with each edge $(v_i, v_j) \in E$ is a weight or length. Define $d(x, y)$ to be the shortest path distance between any pair of points x and y of G. That is, define $d(x, y)$ as the length of the shortest path, among all paths in G between x and y. The x and y components of the domain of d can be defined over different sets (typically either V or all of G). The *eccentricity* of a point $x \in G$ is the length of the longest shortest path from x in G. The *diameter* of a graph is the length of the longest shortest path. Any path of length equal to the diameter is called a *diametral path*. Associate a *weight function* f_i with each $v_i \in V$. In most cases, f_i is assumed to

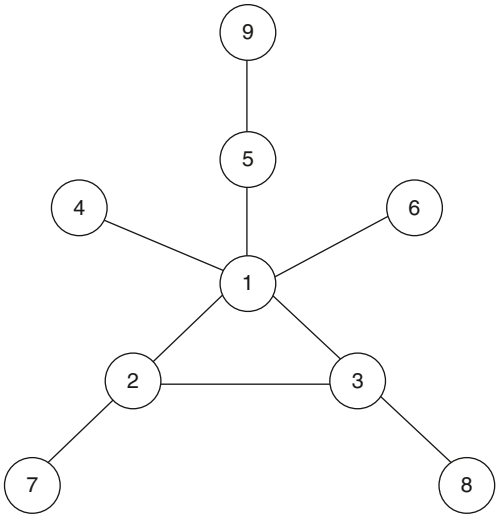


Fig. 14.1 A connected graph $G(V, E)$

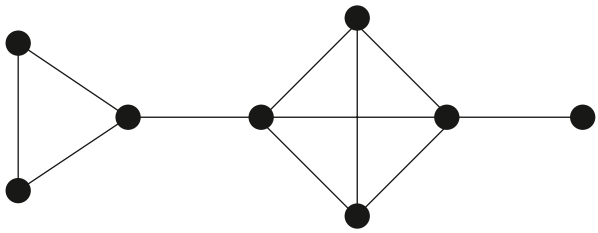


Fig. 14.2 A connected graph with four blocks

be linear, that is $f_i(d(v_i, x)) = w_i \cdot d(v_i, x)$. The cardinality of a set S is denoted by $|S|$. For optimization problems on graphs, a worst case complexity analysis is typically provided. For example, a (worst case) linear time algorithm is denoted $O(n)$ or $O(|V|)$. The typical convention assigns $n = |V|$.

In Fig. 14.2 the graph is connected and has eight vertices, eleven edges, four blocks and three cutvertices. A *block* of a graph G is a maximal connected subgraph of G that has no cutvertex. Each block in this figure is a complete subgraph. A complete graph (or subgraph) is denoted by K_n, where n is the number of vertices. The number of edges in K_n is $n(n-1)/2$. Figure 14.2 has two K_2 blocks, one K_3, and one K_4 block. Blocks need not be complete. Removing either the horizontal or the vertical edge (or both) in the K_4 block does not disrupt the block structure of the graph. Cycles are denoted by C_n, where $n \geq 3$ is the number of vertices in the cycle. For additional graph theory concepts and definitions, the interested reader is referred to Harary (1969) and West (2001).

A graph is said to be a *tree* if it is connected and contains no cycles. A graph that is a tree is designated by $T(V, E)$. Two important facts about trees: a tree contains at least two degree one vertices, and a path between any two vertices is unique and is, therefore, the shortest path.

A cactus graph (sometimes called a cactus tree) is a connected graph in which any two cycles have at most one vertex in common (a cutvertex). That is, each block of a cactus graph is either a K_2 or a C_n. Some authors use the term *m-cactus* to describe a cactus graph in which no cycle has more than m vertices. Cactus graphs were first studied under the name of Husimi trees in Harary and Norman (1953). The name *cactus* was coined in Harary and Norman (1953) for Husimi trees in which every cycle is a triangle. However, Husimi trees also came to refer to graphs in which every block is a K_n, and is now rarely used in the location literature. The term *treelike* is reserved for graphs in which each cyclic block may range from a cycle, C_n, to a complete graph, K_n. For example, the graph in Fig. 14.2 is treelike.

A common notation for distinguishing between location problem types lists the domain for the set from which locations are chosen, the set of points at which demand for service occurs, the number of points to be located, and the underlying network type. The standard choices for each of these components is shown in Table 14.1. Note that G is reserved for use as a network type in the fourth field. To avoid confusion with field four E replaces G in fields one and two, but still means a point can lie anywhere on the network. Later in this chapter we will consider *treelike* networks and denote them by *TL* in field four.

Table 14.1 Notation for location problems

Location set	Demand set	Number of location	Network type
V	V	1	T
E	E	P	G

Table 14.2 Classic network location problems

Name	Restrictions	Formulation
1-vertex median	$V/V/1/G$	$\min_{v_i \subseteq V} \sum_{i \neq j \in V} w_j d(v_i v_j)$
1-absolute median	$E/V/1/G$	$\min_{x \subseteq G} \sum_{j \in V} w_j d(x, v_j)$
1-vertex center	$V/V/1/G$	$\min_{v_i \in V} \max_{v_j \in V} w_j d(v_i, v_j)$
1-absolute center	$E/V/1/G$	$\min_{x \in G} \max_{v_j \in V} w_j d(x, v_j)$

The classical point location problems are median (Hakimi 1964) and center (Jordan 1869) problems. A median, x, is a point at which the average (or equivalently total) distance from x to all demand points is minimized. The notation $V/V/1/T$ states that the median is chosen amongst the vertices, demand occurs only at vertices, one vertex is to be selected, and the underlying network is a tree. Definitions for two 1-median and two 1-center problems are given in Table 14.2. When the first field is E, the solutions are called *absolute* centers or medians. The versions listed include linear vertex weights. Unweighted versions (or, equivalently, vertices of equal weight) are of interest as well. A key property of median problems, first noted by Hakimi (1964, 1965), is a vertex domination result; a solution to the $E/V/p/G$ median problem always resides on the vertex set V. Moreover, the result holds as long as the vertex weight functions f_j are concave (Levy 1967). Further domination results can be found in Hooker et al. (1991). The dominating sets described in that paper are not restricted to vertices but may include additional, well defined, points along edges. These authors survey the location literature and prove theorems that unify and generalize many scattered results. New finite dominating set theorems are proved as well.

14.3 Classical Contributions

In this section we summarize the contributions made by three early, and foundational, historical references—Harary and Norman (1953), Hua et al. (1961) and Goldman (1971).

14.3.1 Harary and Norman (1953): Husimi Trees

Most of the material in Harary and Norman (1953) is focused on developing mathematics for counting arguments on special classes of graphs and is of little

interest to location enthusiasts. However, embedded within the explanations for counting Husimi trees is an important lemma. The lemma establishes a fundamental result for center location problems on treelike graphs. To understand the statement of the lemma, the following definitions are needed. In this paper, as is typical in the mathematics literature, graphs have edges of unit length and there are no vertex weights.

Definition 1: *A graph is called a Husimi tree if no edge lies in more than one cycle.*

Definition 2: *A Husimi tree in which no cycle is greater than three is called a cactus.*

Definition 3: *A vertex is central if it lies on the midpoint of any diametral path.*

Definition 4: *The set of central points of G are those with minimum eccentricity and are denoted by C(G).*

Lemma 1: *If G is a connected graph, then there exists a block of G containing C(G).*

Lemma 1 is the first known extension of the notion of the center (vertex or absolute) of a tree to a treelike graph. The earliest reference defining the vertex center of a tree is generally attributed to Jordan (1869). The center of a tree, $C(T)$, is either a vertex or an edge. For Husimi trees (in the current literature almost always called cactus graphs), this lemma tells us that $C(G)$ is either a single vertex, a single edge, or a single cycle. Moreover, for treelike graphs (denoted *TL*), Lemma 1 guarantees that finding a vertex or absolute 1-center ($V/V/1/TL$ or $E/V/1/TL$) can be reduced to a problem on a single block. For example, in Fig. 14.1 let the edges of the triangle be of length two and all remaining edges be of length one; then, $C(TL)$ is the triangle block. Vertex 1 solves the $V/V/1/TL$ center problem and the midpoint of edge (1, 2) or edge (1, 3) solves the $E/V/1/TL$ center problem.

The use of the term *cactus* in Harary and Norman (1953) is a first as well. Here, cactus graphs are restricted to graphs in which all the cyclic blocks are triangles (K_3). It is not known why the name Husimi tree has dropped out of the location literature, or why the definition of a cactus graph (or *m*-cactus graph) has expanded to include blocks with cycles of any size (or of size at most *m*). Figure 14.1 fits Harary and Norman's definition of both, a Husimi tree and a cactus, since the only cyclic block is a triangle. Figure 14.2 is *not* a Husimi tree, since the K_4 blocks contains edges that are members of more than one cycle. The physics literature, however, continues to use the term Husimi trees—as defined by Harary and Norman (1953)—following a seminal reference by Essam and Fisher (1970), which provided definitions and applications of graph theory in physics.

Harary and Norman (1953) appear to be aware that Lemma 1 was a departure from the central theme of their paper and noteworthy on its own. In the introduction of their paper, the authors comment that Lemma 1 is "of some independent interest." Although Harary (1969) summarized a wide variety of graph structures associated with treelike graphs, it appears that he never returned to the topic of locating centers in treelike graphs.

14.3.2 Hua et al. (1961): 1-Median on a Tree

Hua et al. (1961) was assembled by the author, Lo-keng Hua, following an unidentified collaborative event of students and faculty from seven institutions (Peking Normal University, University of Science and Technology of China, Institute of Mathematics and the Institute of Mechanics of the Academia Sinica, Peking Agricultural Mechanization College, Peking Normal School–special curriculum, and the Peking Industrial and Agricultural Normal School) all under the "unified leadership of the Rural Work Section of the Peking Municipal Committee of the Chinese Communist Party." The summary recorded in Hua et al. (1961) was read at the *National Operations Research Shantung On-the-Spot Conference* in July of 1960.

A broad collection of topics in location theory are addressed. It is possible that some results were known by the authors prior to the conference reading in 1960. (Hua Lo-Keng published an article in 1959 in Chinese with an English translation in the *Notices of the American Mathematical Society* 6: 724–730 in the same year, whose topic was a summary of mathematical research in China between 1949 and 1959). The motivating example throughout is wheat harvesting in rural China. A nod is given to exogenous constraints that will *not* be considered—proximity of a site to water in order to fight fires, terrain topography for threshing sites, soil quality (do not build on sandy soil), and wind effects (lower and upper bounds on wind are needed to effectively winnow the wheat). Results for three separate classes of problems are summarized. First, the problem of how to select locations for threshing sites to serve a collection of small dispersed wheat fields is provided. Next, the problem of selecting a threshing site for a single large wheat field is discussed. Last, a method for estimating wheat production is given.

Only the first two classes of problems have a location flavor. The first gives rise to a set of location problems on a graph, while the second defines a location problem in the plane or some well defined polygonal shape. The first category of problems are the focus of this section: site selection for wheat threshing floors given a collection of small dispersed wheat fields. That is, the customers are the wheat fields, and the facilities to be located are the wheat threshing floors. It is assumed that routes exist connecting the fields to each other. In the graph setting, each wheat field is represented by a single vertex. Weights on the vertices are the (known) wheat production amounts. Edges are routes connecting the wheat fields. Additional non-wheat field vertices may exist representing route intersection points. These vertices have weight zero.

Five types of problems are examined. The author groups these problems under the name *string of grapes* problems. Although not explicitly stated, it seems clear that the grapes denote the dispersed tracts of wheat and the stems connecting the grapes represent the route network. Thus, a string-of-grapes problem is a facility location problem for which the underlying network is a tree (or can be simplified to a series of equivalent problems on a collection of trees). The vertices of the tree are the grapes plus the set of points representing where two or more stems

are joined together. The case descriptions below include algorithm names to link the cases, (which appear without names in Hua et al. 1961) to their modern-day equivalents.

Case 1: *The underlying graph is a tree and only one threshing site is to be selected.*

Algorithm 1 determines an optimal vertex for the threshing site selection. Note that no mention is made of a procedure for identifying degree one vertices.

Algorithm 1: 1-Median of a Tree

Step 1: Among the vertices of degree one find a vertex of the smallest weight.
Step 2: Add this weight to the weight of the adjacent vertex. Delete the degree one vertex and its appendant edge.
Step 3: Repeat until a single vertex remains.

The same algorithm is described in Goldman (1971). In addition, it appears that Hua et al. (1961) understand the vertex domination rule (Hakimi 1964, 1965) for the median objective. The authors mention that if the above algorithm reduces the graph to a single edge and both vertices have the same accumulated weight then any point on the edge (including the vertices) is optimal.

Case 2: *The underlying graph contains cycles.*

As in the first problem, only one threshing site is to be selected. It is not clear in what way cycles may be present. From the example and description given it appears that by "graphs with cycles," the authors mean graphs with cyclic blocks. In particular, each cyclic block seems to be restricted to a cycle, C_n. The algorithm proceeds by repeatedly removing edges from cycles.

Algorithm 2: 1-Median of a Cactus Graph

Step 1: Remove an edge from each cycle of the graph. The result is a tree.
Step 2: Apply the algorithm from Case 1.
Step 3: For the selected site (vertex), compute the sum of the weighted distance the wheat must travel to get to the site (median objective value).
Step 4: Repeat Steps 1–3 for all possible trees (edge deletion patterns). Select the site with smallest weighted distance sum.

The authors do not describe how to enumerate all possible trees. The example given has only one cycle and, as a result, the number of times Steps 1–3 are repeated is equal to the number of edges in the cycle.

A second example is given indicating that the authors understand that the search for a 1-median can be localized to a single block. As a result, blocks which can be ruled out in the search for a median are contracted to a single vertex, so edge deletions are not needed for that block and the total number of trees examined is reduced. For example, in Fig. 14.1 assume all vertices have a weight of 1. The weight of the nodes in the triangle block is 3 and is strictly less than 4 (half the total weight of 8). Thus, the triangle block can be contracted to a single vertex of weight 3. The vertex weight associated with a contracted block is the sum of all the weights in the block.

Case 3: *Extends Case 1 to the location of two or more sites.*

As in Case 1 the underlying graph is a tree. A byproduct of the algorithm is the allocation of wheat fields to threshing sites. The algorithm proceeds as follows.

Algorithm 3: 2-Median of a Tree

Step 1: Delete an edge of the tree. The result is two trees.
Step 2: Solve the site selection problem on each tree via the algorithm presented to solve Case 1.
Step 3: Solve the site selection problem on each tree via the algorithm presented to solve Case 1.
Step 4: Repeat Steps 1–3 for each possible edge deletion. Select the sites with the smallest weighted sums.

If $p>2$ sites are needed, the same algorithm is employed but $p-1$ edges are removed. The result is a set of p trees. Solve the site selection problem on each tree via the algorithm given in Case 1. For the selected sites (vertices), compute the sum of the weighted distance the wheat must travel to get to the site (median objective value). Repeat for all possible $p-1$ edge deletions and retain the best solution found.

Case 4: *Identical to the problem in Case 3 except that simple cycles are allowed.*

The suggested algorithm is to delete edges as described in Case 2. Then, for each tree, the process is repeated as described in Case 3. The process described is cumbersome and no attention is given to implementation issues.

Case 5: *The underlying graph is a tree. Existing sites for threshing floors exist.*

How are additional new sites to be located? The problem is dealt with in the same manner as the problem in Case 3. The explanation for the algorithm begins by analyzing the simplest instance, where one existing site and one additional site are to be located. A summary is found in *Algorithm* 4.

Algorithm 4: Conditional 1-Median of a Tree

Step 1: Delete an edge of the tree. Two trees result.
Step 2: Solve the site selection problem on the tree with no existing site as in Case 1.

Step 3: Repeat Steps 1 and 2 for each tree edge deletion.
Step 4: Select the best solution.

Case 5 is an example of a *conditional* location problem. A closely related problem is Hakimi's (1983) (*r*|*p*)-*medianoid*, in which *r* new facilities must be located to compete with *p* facilities already sited. Competition is not considered in Case 5, as all facilities are assumed to be a part of the same entity (presumably administered by the Chinese Communist Party).

The extension to more than one existing site and the siting of more than one additional threshing floor is mentioned. Details for an algorithm are not given. The authors describe a simplifying rule for the multifacility version of this problem for the conditions posed in cases 3, 4 and 5. Assume there is a fixed capacity, *U*, for any threshing site; if the weight of any subtree exceeds *U*, then discard that solution.

To unify the presentation of the algorithmic results a rhyme is given. The rhyme claims to be a mnemonic device for remembering the algorithm, but it is most likely more effective in the original language. It is included here (in English) for completeness.

String of Grapes Algorithm

When the routes have no loops,
Take all the ends into consideration,
The smallest advances one station.
When the routes do have loops,
A branch is dropped from each one,
Until there are no loops,
Then calculation as before is done.
There are many ways of dropping branches,
The calculation for each must be assessed,
After figuring all, we then compare,
And break the loop in the case which is best.

The paper continues with a second topic—selecting a threshing site for a large (contiguous) tract of wheat. Here the results are either obvious or a re-discovery of other earlier location problems in $\mathbb{R}^2$. For example, assuming the wheat is cut and collected into a finite set of sheaves, with known (x, y) coordinates, determine the location of a threshing site. This is the classic Weber problem originally described in Weber (1909). The Weber problem has a long and convoluted history which is nicely summarized by Drezner and Hamacher (2001).

The analysis of algorithms was still in its infancy when Hua's paper was published. However, some comments about the algorithms, as described in Hua et al. (1961), can be made. In Case 1, the algorithm to find an optimal vertex for the threshing site selection problem runs in linear time as long as vertices of degree one can be found efficiently. The class of graphs addressed in Case 2 are what are now called cactus graphs. The complexity of the proposed algorithm requires counting the number of trees needed to be examined. The number of trees grows rapidly unless one is able to determine blocks that may be contracted. The authors give no indication that a

technique to identify such blocks was known to them. In Case 3 it is easy to see that the algorithm requires ($n-1$) repetitions of the Case 1 algorithm when $p=2$. Thus, if the assessment of Case 1 is correct, the Case 3 algorithm, when $p=2$, is $O(n^2)$. When $p>2$, the general algorithm for Case 3 appears to have a complexity of $O(n^p)$. The procedure described to solve Case 4 requires applying the counting argument in Case 3 ($p>2$) to each possible tree generated from a cactus graph (the counting argument in Case 2). The worst case analysis could certainly be done. However, it is likely that the number of trees required to be examined would be quite large. The complexity analysis for the algorithm in Case 5 would be similar to Case 4.

The key contributions found in Hua et al. (1961) are the reduction algorithm for the 1-median problem on a tree (Case 1) and the hierarchical fashion in which additional problem feature complexities are included. The authors begin with a 1-median problem on a tree. Next, the tree assumption is relaxed and the 1-median problem on cactus graphs is considered (Case 2). Returning to trees, the number of sites to be selected is relaxed. First, $p=2$ and then $p>2$ are considered (Case 3). Case 4 relaxes both the tree and number of sites restrictions to solve the p-median problem on cactus graphs. Last, the authors extend the site selection problem to include the presence of existing facilities. The hierarchical approach reflects a typical scheme taken by current researchers in the location literature.

14.3.3 Goldman (1971): 1-Median on a Tree

Goldman published a series of four papers in *Transportation Science* (Goldman 1969, 1971, 1972 and Goldman and Witzgall 1970). The work in Goldman (1971) is described here but makes use of results in his earlier papers. The key assumptions for the underlying network are that the edges have positive length and the vertices have nonnegative weights, $w(i)$. An algorithm for the 1-median problem on a tree, or nearly a tree (one cycle is allowed), is given. In examining the complexity of the algorithm the author observes that a method for identifying degree one vertices is needed.

The beauty of the algorithm is its simplicity. Let $w(i)\geq 0$ denote the weight of vertex i. Let $W=\sum_{i\in V} w(i)$ denote the total vertex weight. The algorithm, when applied to a tree, reduces the tree to a single vertex—the 1-median of the tree. The algorithm is equivalent to the one given in Hua et al. (1961). The theoretical underpinnings of the algorithm were established in Goldman and Witzgall (1970).

Algorithm 5: Tree Reduction Algorithm

Step 1: If T is a single vertex v, stop v is a median.
Step 2: Find a degree one vertex v_i. If $w(i)>\frac{1}{2}W$, stop: v_i is the optimal solution.
Step 3: Otherwise, delete edge (v_i, v_j) (v_j is adjacent to v_i) and increment $w(j)$ by $w(i)$.
Step 4: Return to Step 1.

The paper also explains how to determine the median if it is located in a cycle. Consequently, the tree reduction procedure can be extended to graphs consisting of blocks that are single edges and/or simple cycles (a cactus graph). The paper describes a reduction algorithm for isolating the 1-median problem to a single block for any treelike graph. The modification is straightforward. In step 2 of the algorithm for trees, if no degree one vertex can be found we have reduced the treelike graph to a single block. If the block is a cycle, an algorithm is provided (described in the next paragraph) that finds the 1-median. If the block is not a cycle, then no procedure is given to identify the 1-median. Furthermore, no computational experience is given and the author points out that it may be non-trivial to determine the blocks of a treelike graph.

Why does the tree reduction algorithm fail if the median problem is reduced to a single cycle? The difficulty lies in the vertex weights. The reduction algorithm requires that the vertex weights be non-negative numbers. For cycles, the reduction process may result in negative weights on some of the vertices. To fully explain the algorithm on a single cycle requires additional terminology.

Label the vertices of the cycle C_n sequentially from 1 to n. Let (i,j) denote the length of edge (i,j). Let H denote half the total length of the cycle: $H = 1/2 \sum_{i=1}^{n} \ell(i,j)$. Two vertices i and j are *antipodal* if $d(i,j)=H$. Let $\hat{i}$ denote the antipode vertex for i. If i does not have an antipode, add a dummy vertex with zero weight so that it does. Consequently, the total number of vertices (original plus dummy) is even and $|V|=2k$ for some integer $k>0$. Let W denote the sum of the non-negative weights on the vertices. The value of the objective function of the weighted vertex median can then be re-written as

$$f(i) = \tfrac{1}{2}\left(\sum_{j\in V}[w(j)d(i,j) + w(\hat{j})d(i,j)]\right)$$

The ½ eliminates double counting since each vertex appears as both a j and a $\hat{j}$ in the summation. Noting that $d(i,\hat{j}) = H - d(i,j)$, we can further simplify the objective to

$$f(i) = \tfrac{1}{2}\left(HW + \sum_{j\neq i\in V}[w(j) - w(\hat{j})]d(i,j)\right)$$

In this form, it is easy to see that no change in $f(i)$ occurs if the same amount is added or subtracted from any pair of antipodal weights. That is, the value of $[w(j) - w(\hat{j})]$ does not change if some $\alpha>0$ is subtracted or added to both $w(j)$ and $w(\hat{j})$. As previously noted, the number of vertices in the cycle is even (each j has an antipode $\hat{j}$). The cycle algorithm splits the cycle in half (between the pairs of antipodal nodes) and solves a median problem separately on each half. Let V_1 denote the first k vertices of V (original plus dummy) and V_2 denote the remaining set of k vertices. To begin the search for a median on V_1 subtract $w(\hat{j})$ from $w(j)$ and $w(\hat{j})$

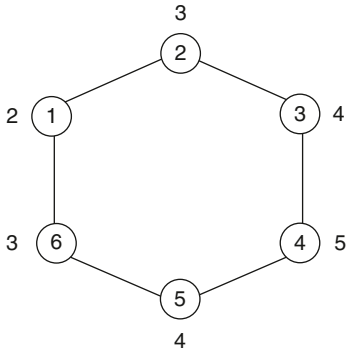


Fig. 14.3 Weighted cycle with unit edge lengths

for each $j \in V_1$. It follows that all vertices in V_2 have weight zero and possibly some of the vertex weights in V_1 are negative. In Fig. 14.3, the weights are displayed next to each vertex: $V_1 = 1, 2, 3$, and $V_2 = 4, 5, 6$. After subtraction, the adjusted antipodal pair weights are $w(4) = w(5) = w(6) = 0$, while $w(1) = -3$, $w(2) = -1$ and $w(3) = 1$. Thus, two of the adjusted weights in V_1 are negative and the nonnegative weight assumption for the tree reduction algorithm is violated.

Algorithm 6: Cycle Algorithm

Step 1: Initialization. Set $i := i^* = 1$. Set $f = F = U = 0$.
Step 2: Increment U by $w(i)$ and f by $(i, j)(2U - W)$.
Step 3: Set $i := i + 1$. If $i > m$, stop; i^* is a median (for V_1).
Step 4: If $f < F$, set $F := f$ and $i^* := i$ and return to Step 2.

Although a complexity analysis of the tree reduction algorithm was not formally done in Goldman (1971), if the identification of the degree one vertices is disregarded, the algorithm is $O(n)$. The key to the algorithm's low complexity is avoiding direct computation of the shortest path distance matrix, typically an $O(n^2)$ operation. Why can the edge weights be ignored? Let v_i denote any vertex of a tree. Deleting v_i partitions the tree into several (at least two) subtrees. Sum the weights of the vertices in each subtree. Let W_i be the weight of the maximum weight subtree. A *centroid* is a vertex that minimizes W_i for all $i \in V$. A median of a tree is the centroid. Zelinka (1968) established this result for trees with equal vertex weights. The result was extended to trees with weighted vertices by Kariv and Hakimi (1979a, b). An alternative approach in Goldman and Witzgall (1970) establishes sufficient conditions for a subgraph to contain a median. Goldman's (1971) algorithm makes direct use of these sufficient conditions.

The key contributions in Goldman (1971) are the linear time tree reduction algorithm for the 1-median problem on a tree and the observation that the same algorithm can be used to reduce the 1-median problem on a treelike graph to single block. Furthermore, Goldman (1971) notes the need for an algorithm to identify the blocks of a graph and provides an algorithm to solve the 1-median problem on a cycle, C_n.

14.4 Key Ideas in the Classical Contributions

Both Goldman (1971) and Hua et al. (1961) describe the same elegant algorithm directed at solving the $V|V|1|T$ median problem. It is clear that Goldman knew his approach also solved the $E|V|1|T$ median problem, and although not explicitly stated, there are indications that Hua was also aware of this. Both Goldman and Hua attempt to extend their results for trees to graphs containing cycles. Goldman (1971) provides a detailed algorithm to determine the location of a vertex median on a cycle, C_n. Beyond the algorithm for trees, Hua et al. (1961) provides no new structural insights to exploit when solving the vertex median problem on graphs with cycles. Instead he resorts to enumerating all possible spanning trees in the cyclic graph, solving a vertex median problem on each one, and selecting the observed best as a solution. For graphs with only a few cycles, Hua's enumerative procedures are reasonable. However, the number of enumerated trees can grow large. An upper bound on the number of possible spanning trees is given by n^{n-2}, Cayley's (1889) formula for $G=K_n$ (where $n=|V|$). To calculate the number of spanning trees for a given graph, Kirchhoff's matrix tree theorem (cf. Kirchhoff 1847 or West 2001) will do the job in $O(n^3)$ time.

The lemma in Harary and Norman (1953) extends the notion of absolute and vertex centers of a tree to treelike graphs. In particular the lemma reduces the search for a center on a treelike graph to a single block. Similarly Goldman (1971) localizes the search for a median to a single block for treelike graphs.

There are two common key ideas found in these three foundational references The first idea is to exploit structure in such a way that the results on trees can be applied. This idea has been carried forward by many authors. Graphs whose structure has been exploited in this way include cactus graphs, planar treelike graphs, block graphs, wheel graphs and bounded treewidth graphs. The second key idea is the reduction of a given location problem to an equivalent problem on a single block.

Both of these key ideas require an algorithm to identify the block structure of a graph. Tarjan (1972) provided the first efficient algorithm to do this with his introduction of depth first search algorithms for exploring graphs. A series of papers dedicated to this topic appeared in the early 1970s. In particular, Tarjan (1972) gives an $O(|V|+|E|)$ depth first search (*DFS*) algorithm to identify the blocks (biconnected components) of an undirected graph. However, it was more than ten years before these results were brought to bear on network location problems.

14.5 Key Ideas Carried Forward: Center Problems

Research that exploits graph structure for center problems is summarized in this section. Treelike graphs dominate the literature, with cactus graphs being the most prevalent. An algorithm whose complexity analysis relies upon how close a treelike graph is to a tree is highlighted. In addition, the influence of data structures that take advantage of the treelike structure is noted.

Gurevich et al. (1984) develop an algorithm for the *p*-center problem on treelike graphs. The algorithm makes explicit use of the depth first search approach in Tarjan (1972). The approach taken by Gurevich et al. (1984) is quite innovative. They exploit the fact that blocks must communicate through a cutvertex and that choices for locations in different blocks combine additively. In addition, the authors sought to find a way to interpolate between trees and general graphs when developing their algorithm and analyzing its complexity. The complexity analysis for the algorithm in Gurevich et al. (1984) depends on the number of edge deletions needed for a cyclic graph to become a tree. In particular, for each block, B_i of G, let $k(B_i)$ denote the number of edge deletions needed to reduce the block to a tree. Let $k = \max_i\{k(B_i\}$. Given a connected graph and a positive integer r determine the minimum number of centers so that every $x \in G$ is within r of some center. The complexity of the algorithm is $O(|E|(6r)^{\lceil k/2 \rceil}$. A second algorithm is given by Gurevich et al. (1982), whose complexity does not depend on r and solves the center problem when $k \leq 2$ in $O(n \log n)$.

The development of depth first search algorithms for graphs was a spring board for a wealth of activity in graph algorithms. One outgrowth of this activity, in the computer science community, was the development of efficient data structures for graphs. An important example with applications to location analysis is provided by Frederickson and Johnson (1983). They devised a data structure for storing information about trees (and treelike graphs) called sorted Cartesian matrices. These matrices store portions of the shortest path distance matrix efficiently. In particular, if $V_1 \subset V$ and $V_2 \subset V$, then a sorted Cartesian matrix is able to store the $|V_1||V_2|$ shortest path distances in only $|V_1|+|V_2|$ space. With these sorted Cartesian matrices, *p*-center problems on trees and cactus graphs can be solved efficiently.

Three *p*-center tree problems, $V/V/p/T$, $E/V/p/T$, and $V/E/p/T$, are solved in $O(n \log n)$ time while the implementation for $E/E/p/T$ has a complexity of $O(pn \log(2n/p))$ when $p < n = |V|$. Frederickson and Johnson (1983) are able to extend their algorithm for trees to $V/V/p/TL$, when TL is restricted to cactus graphs. The extension hinges on the authors' ability to adapt the sorted Cartesian matrix data structure to cactus graphs. Consequently, $V/V/p/TL$ can be solved in $O(n \log n)$ time on a cactus graph. No algorithms or discussion are provided for any of the remaining three *p*-center problems on cactus graphs, $V/E/p/TL$, $E/V/p/TL$, and $E/E/p/TL$. The success of Frederickson and Johnson's algorithm for center problems on trees and cactus graphs relies exclusively on their development of the sorted Cartesian matrix data structure.

A linear time algorithm for the $E/V/1/TL$ center problem on a subset of 3-cactus graphs is given in Kincaid and Lowe (1990). The restrictions required for 3-cactus graphs highlights the difficulty in extending the results in Frederickson and Johnson (1983) on trees for non-vertex restricted center location problems on cactus graphs. Two additional terms are needed to specify the restrictions. An *endblock* of a graph is any block with exactly one cutvertex (analogous to identifying degree one vertices in trees). The remaining blocks are called *interior blocks*. The algorithm in Kincaid and Lowe (1990) is restricted to 3-cactus graphs in which (1) the center cannot

lie in an endblock and (2) the eccentricity for any point x of any interior block must be achieved in at least two distinct endblocks.

Additional results for $E/V/1/TL$ when the center objective is replaced by a weighted obnoxious center objective, are found in Zmazek and Zerovnik (2004) and Cabello and Rote (2007). Zmazek and Zerovnik (2004) pose the $E/V/1/TL$ weighted obnoxious center problem on cactus graphs. An $O(c|V|)$ algorithm, where c denotes the number of different vertex weights, is given. For the unweighted version c can be dropped and a linear time algorithm results. Cabello and Rote (2007) give an $O(n\ log^3 n)$ algorithm that finds a weighted $E/V/1/G$ obnoxious center on any planar graph. The complexity simplifies to $O(n\ log\ n)$ for graphs with a bounded treewidth. The latter condition appears to hold some promise for defining treelike structure in graphs. Trees have a treewidth of one. Cactus graphs, series-parallel graphs and outerplanar graphs have treewidths of at most two. Treewidth measures the number of vertices mapped onto any tree vertex in an optimal tree decomposition. Bodlaender (1996) gives a linear time algorithm for constructing an optimal tree decomposition if the treewidth is bounded.

In another recent paper, Ben-Moshe et al. (2007) also developed an $O(n\ log^3 n)$ algorithm for the $E/V/1/TL$ weighted obnoxious center problem on cactus graphs. Algorithms on cactus graphs are also given for the weighted $E/V/1/TL$ and $V/V/1/TL$ center problems $O(n\ log\ n)$, the weighted $E/V/2/TL$ center problems, $O(n\ log^3 n)$, the weighted $V/V/p/TL$ center problem, $O(n\ log^2 n)$, the weighted $E/V/p/TL$ center problem, and the unweighted $E/E/p/TL$ problem, $O(n^2)$. We note that the algorithms presented in Ben-Moshe et al. (2007) make use of the nomenclature developed in Burkard and Krarup (1998) for the pos/neg median location problem.

14.6 Key Ideas Carried Forward: Median Problems

In this section, a number of approaches are summarized that construct a tree from a treelike graph. These approaches include a blocking graph, a skeleton graph, and a Y-Δ transformation. Graph planarity is seen to be a limiting feature of certain algorithms. The solution of 1-median problems on a single cycle are extended to a wheel graph. In addition, the difficulty with negative vertex weights, observed by Goldman (1971), is overcome in the algorithm for the pos/neg weighted median problem described in Burkard and Krarup (1998).

An early reference to extend and formalize the results of Goldman (1971) for treelike graphs is Chen et al. (1985). These authors provide the details of how to efficiently implement Goldman's linear time algorithm for 1-median problems on both trees and treelike graphs. As expected, the complexity relies heavily on the seminal work of Tarjan (1972). To extend the linear time algorithm from trees to treelike graphs, a *blocking graph* is constructed. The key property of a blocking graph is that it is a tree. (There are a wide variety of ways to create a tree from a treelike graph by utilizing its block structure, see, e.g., Harary 1969.) The vertex set of the blocking graph augments the original graph's vertex set by inserting verti-

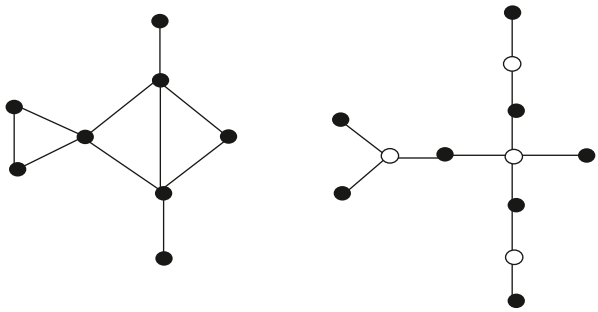


Fig. 14.4 Graph and blocking graph

ces to represent each block. Edges within the block are deleted and new edges are constructed between the new block vertex and the old vertices of the block. That is, each block is replaced by a tree. In Fig. 14.4 the graph on the left is the original graph and the graph on the right is the blocking graph. The vertices representing blocks are open circles.

When the demand and site location sets for center and median problems are restricted to the vertices, algorithms for trees are more easily extended to cactus graphs. One example, the $O(n \log n)$ algorithm in Frederickson and Johnson (1983) for the $V/V/p/TL$ center, was given in the previous section. In addition, linear time algorithms on 3-cactus graphs can be constructed for many location problems, including the vertex-restricted versions of the minimum weighted variance, the vertex restricted stochastic queue median problems, and a wide variety of other median type problems (Kincaid and Maimon 1989). The transformation of a K_3 into a tree is sometimes referred to as a Y-Δ transformation (see Fig. 14.5). The name arises from the shape of the block before, and after, an upside down Y, the transformation. The first use of this transformation is most likely in the electrical circuit transformation found in Kennelly (1899). In Fig. 14.5, the open circle vertex represents the block. Each block of a 3-cactus in which the edge lengths satisfy the triangle property are transformed as shown in Fig. 14.5. However, if a block's edge lengths do not satisfy the triangle property, a tree is formed by deleting the longest edge. The transformation preserves shortest path distances. A postorder traversal of the vertices of the resulting tree finds the optimal vertex for the location problem of interest.

The above approach provides a method for transforming a 3-cactus into a tree while preserving shortest path distances. Unfortunately, the transformation does not

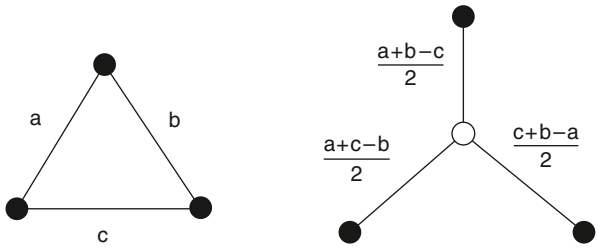


Fig. 14.5 Y-Δ transformation (blocks with triangle property)

generalize beyond 3-cactus graphs. It is possible, however, to exploit the structure of 4-cactus graphs and still solve the vertex median problem in linear time.

In Lan and Wang (2000) such an algorithm is constructed for the weighted version of the vertex median problem. The algorithm replaces each C_4 block with a K_4 and assigns edge weights to the K_4, so that all shortest paths through C_4 are represented. The linear time complexity analysis makes use of graph planarity results. A K_m block is planar for $m \leq 4$. The complexity of the algorithm for m-cactus graphs with $m \geq 5$ is $O(n^2)$.

In the same spirit as Goldman's cycle algorithm, Hatzl (2007) solves the weighted 1-median problem on wheel graphs in linear time. A wheel graph is defined as its name suggests: a wheel W_{n+1} has a cycle C_n that approximates the circular portion of a wheel, a hub vertex, and edges (spokes) connecting the hub to the vertices of C_n on the outside of the wheel. Hatzl (2007) also solves the weighted 2-median problem on cactus graphs in $O(n^2)$. The algorithm for the 2-median problem solves a collection of 1-median problems defined on pairs of subgraphs of the cactus graph. The subgraphs partition $V = V_1 \cup V_2$ with $V_1 \cap V_2 = \varnothing$. There are at most two edges joining the subgraphs induced by V_1 and V_2. (If there is only a single edge the complexity is reduced to $O(n \log n)$). The remainder of the algorithm relies on results developed for a parameterized version of the 1-median problem also found in Hatzl (2007).

An extension of the ideas associated with locating a median of a cycle in Goldman (1971) can be found in the pos/neg-weighted median problem. Goldman's cycle algorithm adjusts the vertex weights in the cycle so that some weights were likely to be negative. His cycle algorithm presents a way to incorporate negative weights and still solve the 1-median problem. Burkard and Krarup (1998) formulate the pos/neg-weighted median problem on a network as an extension of a complementary problem proposed by Courant and Robbins (1941). They provide a linear time algorithm that finds a vertex median of a cactus graph with vertex weights that may be positive or negative numbers. In addition, the edge lengths may be negative for edges not contained in cycles. That is, any K_2 block may have positive, zero, or negative length. The authors provide an example showing that vertex optimality is no longer guaranteed when vertex weights in a cycle are allowed to be negative.

An additional contribution of Burkard and Krarup (1998) is the framework they developed for their pos/neg location problem algorithm. Their nomenclature has been adopted by a wide variety of subsequent authors solving other location problems on cactus graphs, such as Ben-Moshe et al. (2007) for center problems. Burkard and Krarup (1998) partition the vertices of a cactus graph into three distinct classes. A *C-vertex* is a degree two vertex included in exactly one cycle. A *G-vertex* is a vertex not included in any cycle. The remaining vertices are called *H-vertices* or *hinges*. A *graft* is a maximal subtree in which no two hinges belong to the same cycle. These features are illustrated in the cactus graph below. The hinge vertices are drawn as solid squares. The grafts are the subtrees labeled G_1, G_2, and G_3. For example, G_1 is the subtree induced by three G-vertices and one hinge vertex. The cycles C_1 through C_6 are denoted by ellipses (no interior cycle vertices are drawn). Similar to the blocking graph in Chen et al. (1985), a tree, called a skeleton, is

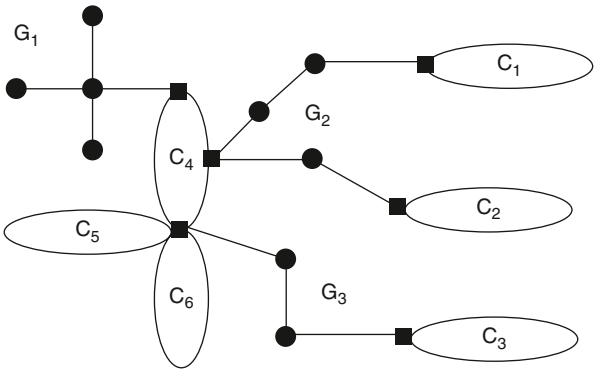


Fig. 14.6 A cactus with 6 cycles, 3 grafts and 6 hinges

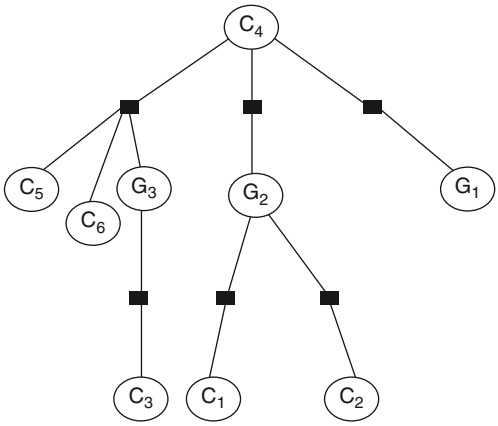


Fig. 14.7 A skeleton rooted at a block

constructed from any cactus graph. The skeleton for the cactus graph in Fig. 14.6 is drawn in Fig. 14.7 Cycles and grafts are collapsed to a single vertex. Hinges are cutvertices and remain unchanged.

14.7 Discussion and Future Research Directions

Two common key ideas were identified in the three foundational references. The first idea was to exploit network structures so that the results found on trees can be extended to treelike graphs. The question asked at the outset (taken from Tansel et al. 1983) follows directly from this idea; other than trees, what special network structure leads to efficient algorithms for network location problems? Researchers have embraced this question with vigor and have considered a wide variety of treelike graph structures, which include cactus graphs, planar treelike graphs, block graphs, wheel graphs, and bounded treewidth graphs. The advent of depth first search algorithms for graphs in 1972 was a major influence on the way algo-

rithms were constructed for treelike graphs. Nearly every algorithm mentioned in this chapter for cactus graphs makes use of depth-first search. Cactus graphs are planar, but as edges are added to the cyclic blocks it does not take long for planarity to fail. An interesting question to consider is for which location problems (or which class of algorithms) is graph planarity the limiting property?

The second idea gleaned from the three foundational references was that of reducing a given location problem to an equivalent problem on a single central component. This idea led Goldman (1971) to devise an algorithm for the median problem on a single cycle. For similar reasons, Hatzl (2007) also devised an algorithm for a single block, the wheel graph. In both cases the ability to reduce the original problem to an equivalent problem on a single block (cycle or wheel) is a major motivation for the single block algorithm development. Extensions to finding p locations on a graph often result in partitioning the graph into p subgraphs and solving single site location problems on each subgraph (as in Hua et al. 1961). Such partitioning procedures are often repeated and take advantage of efficient (mostly linear time) algorithms for single site location problems on each subgraph.

A different approach was taken in Gurevich et al. (1982, 1984) in the development of an algorithm for the p-center problem. The complexity of the algorithm includes an interpolating component associated with how near the graph is to a tree. The innovative complexity analysis allows an interpolation between trees and general graphs. It seems that no other research has attempted to extend this approach to other location problems, but their approach warrants further consideration. Finally, recent efforts associated with bounded treewidth graphs (see Bodlaender 1996) appear promising as a mechanism for measuring "treelikeness." Cactus graphs, series-parallel graphs, and outerplanar graphs all have treewidths of at most two. While it is **NP**-hard to determine the treewidth of a graph, many **NP**-hard combinatorial problems on graphs are solvable in polynomial time when restricted to graphs of bounded treewidth.

References

Ben-Moshe B, Bhattacharya B, Shi Q, Tamir A (2007) Efficient algorithms for center problems in cactus networks. Theor Comp Sci 378:237–252

Bodlaender HL (1996) A linear-time algorithm for finding tree-decompositions of small treewidth. SIAM J Comp 25(6):1305–1317

Burkard RE, Krarup J (1998) A linear algorithm for the pos/neg-weighted 1-median problem on a cactus. Computing 60:193–215

Cabello S, Rote G (2007) Obnoxious centers in graphs. Proceedings of the eighteenth annual ACM-SIAM Symposium on Discrete Algorithms New Orleans, Louisiana, pp 98–107

Cayley A (1889) A theorem on trees. Quart J Math 23;376–378

Chen ML, Francis RL, Lawrence JF, Lowe TJ, Tufecki S (1985) Block-vertex duality and the one-median problem. Networks 15;395–412

Courant R, Robbins H (1941) What is mathematics? Oxford University Press, Oxford

Drezner Z, Hamacher HW (eds) (2001) Facility location: applications and theory. Springer, Berlin

Essam JW, Fisher ME (1970) Some basic definitions in graph theory. Rev Mod Phys 42:272–288

Frederickson G, Johnson D (1983) Finding *k*-paths and *p*-centers by generating and searching good data structures. J Algorithms 4:61–80
Goldman AJ (1969) Optimum locations for centers in a network. Transp Sci 3:352–360
Goldman AJ (1971) Optimal center locations in simple networks. Transp Sci 5:212–221
Goldman AJ (1972) Minimax facility location of a facility in a network. Transp Sci 6:407–418
Goldman AJ, Witzgall CJ (1970) A localization theorem for optimal facility placement. Transp Sci 4:406–409
Gurevich Y, Stockmeyer L, Vishkin U (1982) Solving **NP**-hard problems on graphs that are almost trees and an application to facility location problems. Research Report RC 9348, IBM Watson Research Center, New York
Gurevich Y, Stockmeyer L, Vishkin U (1984) Solving **NP**-hard problems on graphs that are almost trees and an application to facility location problems. J ACM 31(3):459–473
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) Optimum distribution of switching centers in a communication network and some related graph theoretic problems. Oper Res 13;462–475
Hakimi SL (1983) On locating new facilities in a competitive environment. Eur J Oper Res 12;29–35
Harary F (1969) Graph theory. Addison-Wesley, Reading
Harary F, Norman RZ (1953) The dissimilarity characteristic of Husimi trees. Ann Math (Second Series) 58(1):131–141
Hatzl J (2007) Median problems on wheels and cactus graphs. Computing 80:377–393
Hooker JN, Chen CK, Garfinkel RS (1991) Finite dominating sets for network location problems. Oper Res 39:100–118
Hua Lo-Keng et al. (1961) Application of mathematical methods to wheat harvesting. Acta Math Sinica 11:77–91 (in Chinese). An English translation appeared in 1962 in Chinese Mathematics 2:77–91
Jordan C (1869) Sur les assemblages de lignes. J rein angew Math 70:185–190
Kariv O, Hakimi SL (1979a) An algorithmic approach to network location problems. Part I: The *p*-centers. SIAM J Appl Math 37:513–538
Kariv O, Hakimi SL (1979b) An algorithmic approach to network location problems. Part II: The *p*-medians. SIAM J Appl Math 37:539–560
Kennelly AE (1899) Equivalence of triangles and stars in conducting networks. Electr World Eng 34:413–414
Kincaid RK, Maimon O (1989) Locating a point of minimum variance on triangular graphs. Transp Sci 23;216–219
Kincaid RK, Lowe TJ (1990) Locating an absolute center on graphs that are almost trees. Eur J Oper Res 44:357–372
Kirchhoff G (1847) Über die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Verteilung galvanischer Ströme geführt wird. Ann Phys Chem 72:497–508
Levy J (1967) An extended theorem for location on a network. Oper Res Quart 18;433–442
Lan Y, Wang Y (2000) An optimal algorithm for solving the 1-median problem on weighted 4-cactus graphs. Eur J Oper Res 122;602–610
Tansel BC, Francis RL, Lowe TJ (1983) Location on networks: a survey. Part II: Exploiting tree network structure. Manag Sci 29;498–511
Tarjan R (1972) Depth first search and linear graph algorithms. SIAM J Comp 1:146–160
Weber A (1909) Über den Standort der Industrien, Tübingen, (1909). (English translation by Friedrich, C.G. (1929) Theory of Location of Industries, University of Chicago Press.)
West DB (2001) Introduction to graph theory, 2nd edn. Prentice Hall, Upper Saddle River
Zelinka B (1968) Medians and peripherians of trees. Arch Math (Brno) 4:87–95
Zmazek B, Zerovnik J (2004) The obnoxious center problem on weighted cactus graphs. Discrete Appl Math 136;377–386

Chapter 15
Heuristics for Location Models

Jack Brimberg and John M. Hodgson

15.1 Introduction

Location-allocation problems, due to their mathematical complexity, resist exact solutions for problems of more than moderate size. For this and other reasons, heuristic (approximative) approaches are widely used in solving them. This chapter considers the two seminal streams of heuristic solution procedures, both of which remain in use today often in a somewhat altered form. We begin by outlining the location-allocation problem that originally attracted the development of these approaches.

Consider the following basic scenario: a set of n points is given, either in the continuous plane or serving as nodes in a network, with each expressing a demand for some service. These demands can be met through travel to or from facilities also located at points. The objective is to determine a specified number of facility location points that provide the best possible service to these demand points. The studies considered here explicitly define this problem as determining that set of facility locations that minimizes a sum of demand-weighted distances between the demand points (customers) and their nearest facility. This problem is known as the multimedian or generalized median problem. Commonly, the number of facility locations sought is denoted by p, leading to the alternate and more popular title of p-median problem. Some researchers make a distinction between the continuous space *multisource Weber problem* (Brimberg et al. 2000) and the discrete or network space *p-median problem*; here, we do not.

Classic location problems including the location-allocation type occur in two-dimensional space, which may in turn be depicted in three ways: continuous, discrete, and network space. In the continuous case, referred to as site-generating models

J. Brimberg (✉)
Department of Mathematics and Computer Science, Royal Military College of Canada, Kingston, ON K7K 7B4, Canada
e-mail: jack.brimberg@rmc.ca

J. M. Hodgson
Department of Earth and Atmospheric Sciences, The University of Alberta, Calgary, AB T6G 2R3, Canada
e-mail: john.hodgson@ualberta.ca

H. A. Eiselt, V. Marianov (eds.), *Foundations of Location Analysis,* International Series in Operations Research & Management Science 155,
DOI 10.1007/978-1-4419-7572-0_15,

(see, e.g., Love et al. 1988), locations are given by Euclidean coordinates, distances are calculated endogenously from these coordinates, and facility locations are determined by *solving* for the p best pairs of (x, y) coordinates in the plane. In the discrete case, facility locations are *selected* from a set of potential sites (site selection models) and the distances, or more specifically, the shipping costs between demand points and potential sites are provided exogenously. In the network case, demands are expressed at the vertices of networks, facility locations (in the case of medians) are *selected* from network vertices, and distances are calculated over the shortest paths in the network. The classic papers of focus in this chapter were presented in continuous and network space; in the context of the p-median model, network and discrete space are identical in all practical terms.

Regardless of the space considered, the p-median problem is computationally difficult, having been shown to be **NP**-hard by Kariv and Hakimi (1979) for networks and Megiddo and Supowit (1984) for the continuous case. Thus, heuristic approaches are required to solve problems of reasonable size. This difficulty occurred to early researchers in location-allocation analysis, and seminal papers considered in this chapter presented the model formulations along with heuristic algorithms to solve them. They presented heuristic algorithms of two types. Cooper's early work with the continuous space problem gave rise to the *alternating locate/allocate* heuristic, which Maranzana adapted to the network space problem. Teitz and Bart dealt with the network problem using a *vertex substitution* heuristic in which vertices are systematically shifted in and out of a trial solution set.

The remainder of the chapter is organized as follows. Section 15.2 of this chapter reviews the pioneering papers of Cooper (1963, 1964) and Maranzana (1964) including their well-known alternating locate/allocate heuristics; we then proceed to the classical vertex substitution heuristic of Teitz and Bart (1968). Section 15.3 examines the impact of these seminal works on later developments. Section 15.4 provides a short discussion of the future direction of research on the p-median problem. Finally, Sect. 15.5 offers some concluding remarks.

15.2 The Classical Contributions

This section will follow the developments of ideas of four seminal papers in the field of location analysis. Each of these papers includes some heuristic methods that have had an impact on the way location problems are solved.

15.2.1 Cooper (1963, 1964)

Cooper (1963) posed the problem of locating a set of *sources* (facilities) in some optimal fashion in order to serve a set of *destinations* (customers) at fixed and known locations. The problem was described in the following general terms: given the

location of each destination, the requirements at each destination, and a set of shipping costs for the region of interest, determine the number of sources, the location of each source, and the capacity of each source.

Alluding to the theoretical difficulties of this problem, Cooper then added two important assumptions: there are no capacity restrictions on the facilities, and unit shipping costs are independent of facility output.

To put the new problem in context, Cooper (1963) reviewed the background literature on single facility problems, touching on the work of Cavalieri, Steiner, and others. Today, we know that this brief review overlooked two papers of major importance, those of Weiszfeld (1937) and Kuhn and Kuenne (1962), which provided a solution method for the single facility problem.

Cooper stated the problem formally, but we have changed the notation for overall consistency. The known customer locations are defined by their Cartesian coordinates

$$(a_i, b_i), \quad i = 1, 2, \ldots, n,$$

and the coordinates of the p facilities to be determined are

$$(x_j, y_j), \quad j = 1, 2, \ldots, p.$$

Note that although p is now given, we may repeat the analysis for various values of this parameter in order to ultimately determine the 'best' number of facilities.

Cooper also indicated that "in addition to not knowing the location of each of the p sources in the minimum cost solution, we also do not know which source is to serve which subset of destinations." We now term this *allocations*. He introduced a binary variable to deal with these allocations:

$$\alpha_{ij} = 1 \quad \text{if customer } i \text{ is served by facility } j, \ 0 \text{ if not.}$$

He further introduced a weighting factor w_{ij} relating to the multiplicity of supply trips or service calls, a measure of demand for the service.

Cooper framed the problem mathematically in general terms and then introduced the notion of optimal service being the minimization of a weighted sum of Euclidean distances between the customers and the facilities that serve them. This leads to the following formulation:

$$\text{Min}: \varphi = \sum_{j=1}^{p} \sum_{i=1}^{n} \alpha_{ij} w_{ij} \left[(a_i - x_j)^2 + (b_i - y_j)^2 \right]^{1/2}. \tag{15.1}$$

Setting the first-order partial derivatives of (15.1) to zero with respect to each x_j and y_j provides conditions for a minimum. Thus, after replacing the Euclidean distance terms in (15.1) with

$$D_{ij} = [(a_i - x_j)^2 + (b_i - y_j)^2]^{1/2}$$

and re-arranging, we obtain

$$x_j = \frac{\sum_{i=1}^{n} \frac{\alpha_{ij} w_{ij} a_i}{D_{ij}}}{\sum_{i=1}^{n} \frac{\alpha_{ij} w_{ij}}{D_{ij}}} \text{ and } y_j = \frac{\sum_{i=1}^{n} \frac{\alpha_{ij} w_{ij} b_i}{D_{ij}}}{\sum_{i=1}^{n} \frac{\alpha_{ij} w_{ij}}{D_{ij}}}, \quad j = 1, 2, \ldots, p. \tag{15.2}$$

Each pair of simultaneous equations provides the optimal solution for the single median problem defined for a known subset of customers who are allocated to a particular source. Where p facilities are sought, there will be p allocation groups; hence, p separate single facility problems must be solved for any given set of allocations (or partition of the customer set). Cooper presented an iterative scheme for solving (15.2) with fixed allocations, which is now commonly termed the Weiszfeld procedure, having rediscovered the same iterative scheme first proposed by Weiszfeld (1937) for the single median problem. The procedure simply updates the coordinates of each facility by substituting the values from the latest iteration in the right-hand side of (15.2), and continuing in this fashion until convergence is detected. (The interested reader is referred to Kuhn 1973, and Katz 1974, for convergence studies of the Weiszfeld procedure.) As a starting point for the iterations, Cooper recommended the weighted mean center of the customer points:

$$x_j^0 = \frac{\sum_{i=1}^{n} \alpha_{ij} w_{ij} a_i}{\sum_{i=1}^{n} \alpha_{ij} w_{ij}}, \text{ and } y_j^0 = \frac{\sum_{i=1}^{n} \alpha_{ij} w_{ij} b_i}{\sum_{i=1}^{n} \alpha_{ij} w_{ij}}, \quad j = 1, 2, \ldots, p. \tag{15.3}$$

(Cooper presents these values incorrectly in both the 1963 and 1964 papers, by neglecting the w_{ij} factor in the denominators. His case studies survive this error because all weights are assumed equal to unity.)

Cooper's initial solution approach is straightforward, based on his observation of the crux of the location-allocation relationship: "If, for a set of n destinations and p sources, the location of the sources is known, the determination of the optimal allocations is trivial. It is merely the set of weighted distances […] that is a minimum. Conversely, if the allocation is fixed, the determination of the optimum location of the sources is merely the exact calculation, with *known* α_{ij} that has been previously described" (Cooper 1964). The problem can thus be solved exactly by examining all possible allocation sets, $\{\alpha_{ij}\}$, and choosing the solution that minimizes (15.1). He presented a test problem, but acknowledged that this approach would not be computationally attractive for what in the 1963 computing environment was quaintly considered to be a large set of customers (>10). Cooper determined that the number of such allocations is the Stirling number of the second kind $S(n, p)$, a number that remains "formidably large" for problems that are considered of modest size today.

Cooper observed that a method for generating a "reasonable" number of facility location sets was required. He suggested considering the n customer sites as potential facility locations, thus reducing the problem to the discrete space form.

The problem is thereby reduced to size ${}_nC_p$, but he acknowledged that the method remained "inadequate for many problems of industrial importance because of the excessive amount of calculation involved". Moreover, he recognized that limiting facilities to a base set of customer sites might not yield correct allocations.

Having introduced the continuous location-allocation problem and discussed the issues in solving this problem in his 1963 paper, Cooper (1964) turned to the development of heuristics to solve it effectively. He surmised that the problem had no sharp minimum, but rather many alternative or close optima, which "makes feasible the use of heuristic algorithms with a reasonably high probability that a well constructed heuristic will find one of these near optimal solutions." (Here, we see an interesting distinction between classical and modern viewpoints: today, the existence of multiple local minima is considered to be regrettable, and the goal of heuristics is to come as close as possible to the optimal solution rather than to identify a good one. This goal is facilitated, of course, by the enormous increase in computing power available today.) The paper revisits the definition of the problem, the iterative procedure for the single facility problem (which he terms the "exact" procedure), and the suggestion that a direct solution to the multi-source problem is to minimize φ for all possible sets of α_{ij}. In order to treat the problem of very large customer sets ($n \leq 500$) and situations of nonlinear costs (an idea he did not pursue further), Cooper developed several heuristic algorithms. He began by proposing lower bounds for limited cases of the problem, which he used to rank the results of the heuristics, and also determined an obvious upper bound for the problem.

Cooper (1964) presented four basic heuristics that are summarized below. Three of the heuristics assume that the "destination set is a very favored set," and thus, use subsets of the customer set for locating facilities. The first two of these, upon termination, use the exact (Weiszfeld) procedure to determine the (optimal) continuous space origins with respect to the selected allocation.

A: The Destination Subset Algorithm. This considers all possible subsets of p customers as "sites" at which to locate facilities. This is the method of the 1963 paper and basically involves complete enumeration of all ${}_nC_p$ discrete space solutions to the problem, a reduction from the $S(n, p)$ possible continuous space solutions. Upon termination, a continuous adjustment is applied by using the exact procedure on the best discrete solution. Cooper again states that this procedure does not guarantee an optimal solution (the correct allocation may not be found), and warns that computation time for large problems is prohibitive.

B: The Random Destination Algorithm. Here, p random customers are selected to be facility locations. The algorithm is repeated a number of times and the best solution is retained. Cooper suggests a statistical approach to determine when the algorithm might reasonably be terminated. The procedure provides a continuous adjustment on the final solution as before.

C: The Successive Approximation Algorithm. The destination subset algorithm is run for $p = 2$ facilities. The best location for a third facility among the remaining customer sites is determined and then locked into the solution. Additional facilities

are similarly added. This is, after the initial two facilities are located, a greedy constructive algorithm of customer sites. It is unclear why Cooper does not apply the exact algorithm after termination.

D: The Alternate Location and Allocation Algorithm. This elegant heuristic by Cooper continues to be popular to this day. It is based on the simple observation alluded to earlier that the two components of the problem, locate and allocate, are easy to solve in isolation. That is, given the locations of the p facilities, and the fact that there are no capacity restrictions on them, the customers are simply allocated in turn to the facility that provides the lowest cost service. For homogeneous facilities ($w_{ij}=w_i$, for all i, j), this translates to assigning each customer to the closest facility. If the allocations are given, the problem reduces to p independent single facility problems that may be readily solved with the Weiszfeld procedure. The heuristic simply alternates between the two phases until no further improvement is possible. The steps provided by Cooper are summarized in *Algorithm 1*.

Cooper identifies the algorithm as a monotonic-decreasing convergent process that may not converge to the globally optimal solution. In fact, the process only guarantees a local minimum.

Cooper then uses solutions to the destination subset algorithm to demonstrate the lack of a sharp minimum, and reiterates that "it is this relative insensitivity to source location *with correct or near-correct allocations* which makes the use of heuristics feasible in this problem."

Algorithm 1: Alternating Locate/Allocate (ALT-1)

Step 1: Divide the customer set into p subsets of approximately equal size.
Step 2: For each subset, apply the exact procedure to determine the optimal facility location.
Step 3: Allocate all customers to the closest facility.
Step 4: Continue alternating between steps 2 and 3 until there is no allocation change.

Cooper tabulated results for 10 problems of size, $n=30$, $p=3$. For the first time in his experience, the destination subset algorithm arrived at an incorrect allocation. He also used 400 iterations of the random destination algorithm. As expected, the destination subset algorithm generally found the best solutions, but was by far the most computationally demanding of the methods. The random destination approach was next in quality, and much less costly. The successive approximation approach was not satisfactory—it is unfortunate that Cooper did not apply the exact solution procedure to its results for a fairer comparison with approaches *A* and *B*. The alternate location and allocation algorithm (ALT-1) was deemed to be "satisfactory." Cooper neglected to note that this algorithm actually performed best in three of the trials, even though its statistics were troubled by three spectacular failures.

He further tested the heuristics with 100 problems of size $n=40$, $p=3$, and concluded that: "it is apparent that the best *practical* method of solving large location-

allocation problems is with the use of the random destination algorithm and subsequent improvement by a single calculation of the exact location method." Again, he does not credit the alternate location and allocation approach, the solution statistics for which were almost identical to the random destination approach.

One questions why Cooper downplayed the abilities of the alternating location and allocation algorithm, which has been passed down as the major contribution of his early work. One further questions why Cooper did not think to combine algorithms *B* and *D* to apply the alternating algorithm to the random solutions. Cooper started the alternating heuristic with a rather messy selection of an initial allocation. Scott (1971) modified the algorithm by starting with a random selection of trial facility locations. Later work discovered that the influence of a single poor solution could be overcome by using several such random starts. This has become the common procedure for using the alternating heuristic today. The random multi-start version of the alternating heuristic allows us to obtain several local minima in different regions of the solution space, and thus improves the chances of obtaining a "good" solution.

Cooper's 1963 and 1964 papers first identified the location-allocation problem in mainstream literature. Moreover, they identify the computational characteristics of the problem, while the second paper provides solution techniques for what is otherwise a very difficult class of problems. The alternate location and allocation algorithm is still often used today. For several years, this popular approach has been dubbed "The Cooper Algorithm" (Scott 1971).

15.2.2 Maranzana (1964)

Maranzana (1964) defined the location-allocation problem on a network space as follows: given, in a network, a set V of n points (referred to as "sinks") $v_1, \ldots, v_n$, with associated nonnegative weights $w_1, \ldots, w_n$, and a nonnegative, n-dimensional, symmetric distance matrix $[d_{ij}]$, find p sources $v_{x_1}, ..., v_{x_p}$ among the points in V, and a partition of V into p subsets of sinks $V_{x_1}, ., V_{x_p}$ served respectively by the p sources so that

$$\sum_{j=1}^{p} \sum_{v_i \in V_{x_j}} D_{i,x_j} w_i \tag{15.4}$$

is a minimum, where D_{ij} is the minimum path length from v_i to v_j. (Again we have changed notation for consistency.) The total transport cost is assumed to be proportional to the weighted sum of shortest-path distances given in (15.4).

Maranzana concluded that direct enumeration would be impractical for "the typical problem," and proposed instead an iterative procedure to solve the problem heuristically. The method alternates between location and allocation phases as in Cooper's algorithm, except that since we are dealing with a network, the shortest paths between all pairs of nodes must be determined first in a preprocessing step. Maranzana adapted a dynamic programming approach attributed to

Bellman (1958) to accomplish this. The shortest path between a given pair of nodes is determined recursively as the shortest path that uses at most j edges, for $j = 1, 2, \ldots, n-1$. By using a fixed sequence and constant updates of the shortest paths between all pairs of nodes, Maranzana actually improved the efficiency of Bellman's algorithm.

Maranzana also noted that a separate routine was required to find the "center of gravity" of any subset Q of nodes, which he defined as a vertex v_j of V that provides the minimum weighted sum of shortest path lengths between itself, acting as a facility, and the vertices of Q, acting as its customers. (This should not be confused with the median of set Q, since v_j is not restricted to Q.) To find the center of gravity, Maranzana simply evaluated each vertex and chose the best one. The steps of his heuristic may be outlined as shown in *Algorithm 2*.

Algorithm 2: Alternating Locate/Allocate (ALT-2)

Step 1: Select p trial facility sites arbitrarily from the n vertices of V to specify the current location set.

Step 2: Partition the n vertices by assigning each customer to its nearest facility in the current location set.

Step 3: Determine a "center of gravity" for each subset in the partition.

Step 4: If the center of gravity is the same as the current location of the facility for each subset, stop (the current location set with associated partition is the final solution); else update the current locations to the new centers of gravity and return to Step 2.

Maranzana proved that the sequence of solutions generated by the algorithm is monotone non-increasing by showing that the allocation and location phases (Steps 2 and 3) may only improve the current solution. He then provided a simple numerical example to show that the procedure can converge to a non-optimal (local) minimum. Using a second simple example, he demonstrated the difficulty that may arise when the center of gravity (Step 3) is non-unique; that is, different decision rules for breaking ties may lead to very different solutions. To circumvent the above difficulties, Maranzana, unlike Cooper, who treated the alternating heuristic as being suited to a single application, suggested that "with a computer it is feasible to carry out the procedure on a number of initial selections so that one can be assured of arriving at a good solution." Finally, the algorithm was applied to problems of two and three facilities in a case study of 158 Italian cities, each given a hypothetical weight that appears to have been related roughly to city size. It is interesting that computation time was mainly devoted to the calculation of shortest path distances on the network, a problem that would be magnified later on for practical applications on much larger networks. Finally, we note that Maranzana seems to have assumed in his procedure that the optimal facility sites are located at the vertices of the network, a result that would be proven coincidentally by Hakimi (1964).

15.2.3 Teitz and Bart (1968)

Teitz and Bart (1968) addressed the problem of solving what they called the "generalized vertex median of a weighted graph," which today we call the "uncapacitated p-median problem on a network." Specifically, they considered "the problem of choice of location of p sources of unconstrained capacity from among n destinations having fixed demands and located at nodes of a network." They acknowledged that their problem was essentially the same as that investigated by Hakimi (1964) and Maranzana (1964), stating that their concern was with alternative methods of solution.

The problem is thus defined on an edge and vertex weighted graph, G. Each vertex v_i is weighted by a weight w_i and each i–j edge by the shortest path distance D_{ij}. The distance matrix $\mathbf{D}$ of G is an $[n \times n]$ symmetric matrix of shortest path distances between all pairs of vertices v_i, v_j. The weighted distance matrix $\mathbf{R}$, asymmetric for differentially weighted vertices, is defined as

$$[r_{ij}] = [w_i D_{ij}]. \tag{15.5}$$

The single vertex median solves

$$r_k = \min\{r_1, r_2, \ldots, r_n\}, \tag{15.6}$$

where

$$r_j = \sum_{i=1}^{n} r_{ij}, \quad j = 1, \ldots, n. \tag{15.7}$$

The generalized vertex median problem can be developed as follows: let V_p be some subset containing exactly p vertices of G. In an n-vertex graph there will be $\begin{pmatrix} n \\ p \end{pmatrix}$ possible such subsets, indexed V_p^m; $m = 1, 2, ..., \begin{pmatrix} n \\ p \end{pmatrix}$. For each such subset, we may construct a submatrix $\mathbf{R}_p^m$ of $\mathbf{R}$ by adjoining all columns of $\mathbf{R}$ for which the corresponding column vertices are contained in V_p^m. If facilities are limited to vertices in V_p^m, each customer v_i will be served by that facility in V_p^m for which r_{ik} is a minimum. The total weighted distance r_m for the V_p^m set of facilities is

$$r_m = \sum_{i=1}^{n} r_{ik}, \tag{15.8}$$

where in each row of $\mathbf{R}_p^m$, $k(=k(i))$ is the facility for which r_{ij} is minimized. The general vertex p-median of G is defined as some V_p^{m*} such that

$$r_{m*} = \min\{r_1, r_2, ..., r_{\binom{n}{p}}\}, \tag{15.9}$$

so that $r_{m*} \leq r_m$, $m = 1, 2, \ldots, \binom{n}{p}$. Teitz and Bart acknowledge that the p-median is not necessarily unique. They then addressed the task of finding it.

Teitz and Bart begin by outlining the direct enumeration method and Maranzana's alternating algorithm, termed the *partition method* for obvious reasons. The former was deemed too computationally demanding and the latter of suspect robustness. This is followed by the main contribution, their vertex substitution method, which in their words "concentrates upon the formal definition of the generalized vertex median and its associated weighted distance matrix."

The method proceeds as follows: for each possible subset of facility sites V_p^m, we may construct a submatrix $\mathbf{R}_p^m$ by combining the relevant p columns as described above. Consider what happens when one vertex v_j in the facility subset is replaced by another vertex v_b outside this set; that is, the v_b column takes the place of the v_j column in $\mathbf{R}_p^m$. If r_{ij} is the i-th row minimum of $\mathbf{R}_p^m$, then its replacement by r_{ib} could have one of several outcomes:

If $r_{ib} \leq r_{ij}$, the increment to the i-th row contribution to sum r would be

$$_i\Delta_{bj} = r_{ib} - r_{ij} \leq 0. \tag{15.10}$$

If $r_{ij} \leq r_{ib} \leq r_{is}$ (where r_{is} is the second-smallest i-th row element in $\mathbf{R}_p^m$),

$$_i\Delta_{bj} = r_{ib} - r_{ij} \geq 0. \tag{15.11}$$

If $r_{ij} \leq r_{is} \leq r_{ib}$,

$$_i\Delta_{bj} = r_{is} - r_{ij} \geq 0. \tag{15.12}$$

In the Teitz and Bart paper, the differences in these expressions are incorrectly reversed. For example, $r_{ib} - r_{ij}$ is incorrectly written $r_{ij} - r_{ib}$. The authors also seem to make a fundamental error by concluding that "if r_{ij} were not the i-th row minimum of $\mathbf{R}_p^m$, then no change in the i-th row contribution to r would result." This is not generally true, as implied by the analysis above. There can be no increase in the objective value r, but v_b may still become the new closest facility to v_i, resulting in a reduction in r.

It is worth substituting vertex v_b for v_j only if the net effect of all increments

$$\Delta_{bj} = \sum_{i=1}^{n} {}_i\Delta_{bj} \tag{15.13}$$

is less than zero, i.e., if it reduces the total weighted distance. An iterative process of single vertex substitutions as suggested by Teitz and Bart may now be employed to obtain a monotone decreasing sequence of solutions that ends when a local minimum is reached.

Algorithm 3: Vertex Substitution (VS-1)

Step 1: Choose some initial facility subset V_1 containing p (randomly-selected) vertices.

Step 2: For each vertex $v_j \in V_1$, find its associated customer subset of vertices for which it is the closest facility (no rules are given for breaking ties as in Maranzana); compute the total weighted distance r_1 for the resulting solution.

Step 3: Select some vertex v_b not in the facility subset, i.e., $v_b \in V \setminus V_1$.

Step 4: Substitute v_b in turn for each vertex $v_j \in V_1$, and compute Δ_{bj} each time.

Step 5: Find that vertex $v_k \in V_1$ that, when replaced by v_b, most reduces the total weighted distance, that is,

$$\Delta_{bk} < 0 \text{ and } \Delta_{bk} = \min_j \{\Delta_{bj}\} \tag{15.14}$$

Step 6: If such a vertex v_k can be found, substitute v_b for v_k in the facility subset; label the new subset V_2 and compute $r_2 (= r_1 + \Delta_{bk})$. If no vertex v_k satisfies relation (15.14), simply retain the facility subset V_1.

Step 7: Select another vertex, not previously tried, in the complement of V_1, and repeat Steps 4 through 6.

Step 8: When all vertices in the complement of V_1 have been tried, define the resulting facility subset V_t as a new V_1 and repeat Steps 2 through 7. Call each such complete repetition a cycle.

Step 9: When one complete cycle results in no reduction in r, terminate the procedure. The output is the last solution obtained.

Note: It appears to be unclear in Step 7 whether the authors intended that the new subset V_2 replace the original subset V_1 in the repetition of Steps 4–6. However, this would only affect the type of improvement strategy utilized, and not the gist of the procedure. We interpret the authors' intention as using the original V_1 in each such repetition of Steps 4–6, giving rise to what is termed today a "best" improvement strategy; i.e., look at all solutions in the one-interchange neighborhood of V_1 and select the best one. It is also unclear how they intend us to perform Step 6 in successive iterations; their use of V_t in Step 8 suggests that they would label further subsets V_3, V_4, …, V_t. This is not necessary, since we need only maintain a current "best" substitution at this step, which we could label V_t throughout. Moreover, the total weighted distance need not be calculated each time.

Teitz and Bart acknowledge that a situation could arise in which a single vertex substitution produces no further improvement, whereas pairwise or higher substitutions would further reduce the total weighted distance. However, they do not characterize this case or give examples. In their experiments on random graphs with 25 vertices, they observe that the procedure always terminated (i.e., reached a

local minimum) within four cycles. Most important—they note a very significant improvement in solution quality using vertex substitution compared to the partition method. The partition method furthermore exhibits considerable variation in performance. They conclude that vertex substitution is the preferable heuristic.

15.3 Impact of the Early Heuristics

This section investigates what further developments have been made on the basis of the heuristic methods described in the previous section. We first examine work that considers generalization of distance measurements, followed by a variety of location—allocation models and modern heuristics in continuous and discrete spaces.

15.3.1 Generalization of Distance Measurements

The Maranzana and Teitz and Bart approaches were defined in network space, but it is not necessary to have a network structure to define the *p*-median or to solve it using their procedures. Many examples in the literature do not rely on an underlying network structure. Both discrete and network spaces are defined with reference to a matrix of shortest distances, and operate through consideration of these internodal distances. This is possible because the Hakimi (1964) finding ensures that optimal locations can be limited to the vertices—hence the internodal distance matrix is all the information required. It follows that the relevant distance matrix among pairs of "vertices" can be defined other than within a network or if, within a network, without specifying the network structure. We can apply the partition and vertex substitution methods to any system where a matrix $[D_{ij}]$ is provided. Given a set of nodes in space, these distances might be specified for example as Euclidean distances, airline travel times, psychologically perceived travel costs, or in many other different ways.

The Teitz and Bart and Maranzana papers work on the assumption that the customer set represents the potential facility locations from which trial sites can be selected. In modern practice it is recognized that this is often not realistic. Some customer sites may be unsuited to facilities; some ideal facility locations may not express demand. Thus, we recommend that a separate set of potential facility locations be maintained in working with network and discrete space models. The distance matrix would be constructed and used in a similar fashion as before. In some realistic problems, facilities may already exist in some locations, and the heuristics above are easily adapted to deal with this situation.

The Cooper papers (1963, 1964) assumed that distances are measured by the Euclidean norm. Since that time more general distance functions, such as the l_p norm, have been incorporated in location models to provide more accurate measures of

travel distance. Given any two points, $X_1=(x_1, y_1)$, $X_2=(x_2, y_2)$ in the plane, the ℓ_p distance between them is given by:

$$\ell_p\,(X_1, X_2) = [|x_1 - x_2|^p + |y_1 - y_2|^p]^{1/p}$$

where the parameter $p \geq 1$. When $p=1$, we have the well-known rectangular (or Manhattan) norm; the Euclidean norm occurs with $p=2$. The Cooper algorithm is readily extended to the median problem with ℓ_p distances after modifying the Weiszfeld formulas appropriately; see, e.g., Love et al. (1988). The use of "block norms" allows the location step to be solved by linear programming techniques. In fact the problem may now be reduced from continuous space to a finite set of intersection points, thus allowing a vertex substitution heuristic to be used as well.

The distance function may be raised to some power in order to model more effectively the transportation costs or times; an example is the fire engine travel time study in Kolesar et al. (1975). Geodesic distances are typically used for location on a sphere as in the case of air travel. In *cluster analysis*, which has important applications, for example, in data mining, the fixed points (vertices) are located in higher-dimensional space according to the number of attributes involved. The well-known *k*-means model from the data mining literature is simply a version of the *p*-median model ($p=k$) with squared Euclidean distances.

The partition and vertex substitution methods have been readily adapted to such generalizations of the original continuous and discrete (network) problems.

15.3.2 Other Location-Allocation Models

The general principles of partition and vertex substitution may be extended to other forms of the location-allocation problem as introduced elsewhere in this book. With the vertex substitution method, we simply revise the procedure for calculating the incremental change in objective function associated with each swap move of vertex entering and vertex leaving the solution. The generation of a monotone sequence and convergence to a 'local' optimum are guaranteed. With the partition method, the location step is adjusted according to the type of objective function under consideration. For problems that are *separable* into location and allocation phases, we may show again that the sequence generated is monotone, which is essential for convergence of the heuristic. However, the partition method may not converge in more general cases. Consider, as an example, a form of the *covering* problem where the goal is to locate sensors on a grid in order to maximize the mean probability of detection measured at the grid points. The location step is no longer separable into *p* independent single facility problems due to additional interactions that exist with facilities (sensors) other than the closest one, and thus the partition method breaks down. A similar situation may occur in location-allocation models involving *noxious* facilities. It therefore appears that the partition method is not as universally

useful as the vertex substitution method for problems occurring on networks or in discrete space.

15.3.3 Modern-Day Heuristics

The partition and vertex substitution methods both fall in the category of *local search*; that is, the procedure finds a better solution in a local neighborhood of the current solution and iterates in this fashion until a local optimum is reached. It is interesting to note that in network (or discrete) space, any local optimum obtained by the vertex substitution method must also be a local optimum in the partition method, but the converse is not necessarily true. This is due to the fact that the location step in the partition method is equivalent to a "restricted" set of vertex swap moves. Thus, starting from some local optimum, a better partition of the customer set may be found by examining all possible swap moves as in the vertex substitution method. This may explain the superiority observed by Teitz and Bart of their heuristic, as well as the higher variability of results obtained by the partition method. It is also interesting to note that comparative studies of the two methods appear to be limited to the experiment of Teitz and Bart on a few small random instances, and some further testing by Rosing et al. (1979). Yet the vertex substitution method is widely used to this day, while Maranzana's work in comparison has been largely ignored. There may be computational advantages, for example, in using a two-stage approach where the fast partition method is applied first on a random initial solution, followed by vertex substitution with the solution from the first stage as the starting point.

The Cooper algorithm is still widely used on problems posed in continuous space. A few variants that seem to work better have been suggested including, as noted, Scott (1971) who starts with an initial random set of facility locations instead of an initial allocation. Care must be taken, since the Cooper algorithm may lead to a degenerate solution (Brimberg and Mladenovic 1999) where some facilities end up having no customers assigned to them. The shortcoming is easily remedied by inserting such facilities at unoccupied vertex locations (those customers that do not have coinciding facilities) whenever the situation arises within the solution process.

As problem size defined by n and p increases, an exponential growth in the number of local optima is observed. Thus, local search methods become inefficient. We will see next that the partition and vertex substitution methods still play an important role in the more advanced techniques used today.

15.3.3.1 The Continuous *p*-Median Problem

The random multi-start version of Cooper's algorithm remained the state-of-the-art for many years despite a number of other competing heuristics. Notable among these is a heuristic developed by Love and Juel (1982) that is the first method to

impose a set of neighborhood structures on the problem. A given neighborhood of a solution is defined here as the set of points around that solution that are obtained by exchanging a specified number of assignments of customers from their current facilities to new ones. The authors consider up to two exchanges, and show that the two-exchange neighborhood may be used (at a computational cost, of course) to 'jump out' of a local optimum trap in the one-exchange neighborhood. In their procedure the facilities are always optimally located with respect to any given allocation of the customers. Other heuristics include gradient-based methods (Murtagh and Niwattisyawong 1982, and Chen 1983) and a projection method by Bongartz et al. (1994). For further details, see, for example, the survey paper by Brimberg et al. (2008a).

Recall that one of Cooper's initial ideas was to solve a discrete version of the problem where the facility locations are restricted to the set of fixed points given by the customers and the shortest-path distance is simply the Euclidean distance between each pair of vertices. Hansen et al. (1998) revisit this idea several years later while taking advantage of an efficient code by Hanjoul and Peeters (1985) to solve the discrete problem exactly. A second stage involves a continuous improvement where p single facility problems resulting from the partition of the customer set by the discrete solution are solved. Excellent results are reported, but computation times become excessive. Brimberg et al. (2000) propose a new neighborhood structure based on the vertex substitution idea of Teitz and Bart (1968); that is, facilities are relocated one at a time to an unoccupied fixed point (a customer that does not have a coincident facility). The one-interchange neighborhood contains all such possible single moves. A local search using Cooper's algorithm is then conducted from all or selected points in this neighborhood. The authors investigate various "drop and add" strategies in the selection process, which allow a reduction in the size of the neighborhood from $O(np)$ to $O(n+p)$, and as a result, a much faster local search. When the full one-interchange neighborhood is verified, an efficient updating procedure by Whitaker (1983) is used. The relocation heuristics are able to obtain better results than the multi-start Cooper algorithm in a fraction of the time.

The recent application of *metaheuristics* to the continuous p-median problem has resulted in a significant advance in the state-of-the-art. Unlike local search that examines a narrow region of the solution space and terminates at a local optimum, metaheuristics are general frameworks that allow the search to expand to different regions of the solution space, and thus escape the "local optimum trap." A comparative study (Brimberg et al. 2000) shows that as problem size increases (and the number of local minima explodes), the performance of the multi-start Cooper algorithm deteriorates significantly relative to new heuristics based on Tabu search, variable neighborhood search, and the genetic algorithm. It is interesting to note, however, that these newer methods usually have Cooper's algorithm embedded within them. For example, the various versions of variable neighborhood search in the above comparative study use Cooper's algorithm in the local search step. The initial population in the genetic algorithm of Houck et al. (1996) is obtained by repeating Cooper's algorithm from random starting points until an adequate number of local minima is found, and after the crossover

operation, the new solution is improved (mutation operation) using the Cooper algorithm. For a further update on metaheuristic-based methods for solving the continuous *p*-median problem, see Brimberg et al. (2008a).

15.3.3.2 The Discrete *p*-Median Problem

As noted in Mladenovic et al. (2007), the vertex substitution method by Teitz and Bart, which they refer to as the *Interchange* procedure, is still "commonly used as a standard to compare with other methods." Both Maranzana's partition method, aptly named the *Alternate* heuristic, and the Interchange procedure have been used in composite type heuristics. For example, Captivo (1991) adds facilities one at a time in a greedy fashion that reduces total cost as much as possible, and then uses the Alternate procedure in each step to further improve the solution. Another composite method first constructs a greedy solution and then applies the Interchange procedure to that solution; it is often used for comparison with other new methods (see Voss 1996, and Hansen and Mladenovic 1997). Lagrangian-based procedures that "alternate" between solving for the primal variables and adjusting the Lagrange multipliers typically use a local search as above to improve the obtained solution (as in Beasley 1993).

The concept of neighborhood structure is intimately related to the vertex substitution method. We may view this method as a local search in the *1-interchange* neighborhood. Generalizations are now possible. For example, Kochetov et al. (2005) propose a new neighborhood structure, termed *LK* (Lin-Kernigham), which employs a depth parameter k that counts the number of interchange moves within one step of local search. The $LK(k)$ neighborhood may be described as follows: (*i*) find two vertices v_{add} and v_{drop} that give the best solution in the 1-interchange neighborhood; (*ii*) exchange these two vertices to get a new solution; (*iii*) repeat the above steps k times, not allowing any facility that has been dropped to re-enter the solution. The process is repeated until a local minimum in the *LK* neighborhood is reached. This type of local search has been used within Lagrangian relaxation, random rounding (after linear relaxation), and ant colony optimization (Dorigo and Di Caro 1999). The 1-interchange neighborhood can be modified in a straightforward way to handle the related *simple plant location problem.* In this case a fixed cost f_i is charged to open a facility at vertex v_i, and the number of facilities to open is unknown. Brimberg et al. (2008b) examine an extended version of the simple plant location problem with nonlinear objective function representing the return on investment. They use an expanded local search neighborhood that allows all single moves where either a vertex is opened (v_{add}), a vertex is closed (v_{drop}), or an interchange is made (v_{add} and v_{drop}).

Mladenovic et al. (2007) note that "the Interchange method is one of the most often used classical heuristics either alone or as a subroutine of other more complex methods or within metaheuristics." In large scale applications it is therefore critical that the procedure be implemented in an efficient manner. The popular CLARANS (Clustering Large Applications based on RANdomized Search) method in data min-

ing (Ng and Han 2002) conducts a local search using a small random sample of points in the 1-interchange neighborhood. Efficient implementations that can evaluate in reasonable time the entire neighborhood for very large instances, including the fast interchange of Whitaker (1983) mentioned previously, are summarized in Mladenovic et al. (2007).

Several modern heuristics that derive from metaheuristic rules use the vertex substitution method in some form. In the *Tabu search* procedure of Mladenovic et al. (1996), the 1-interchange move is extended to what they term the 1-chain-substitution move. In Rolland et al. (1996), the 1-interchange move is divided into add and drop moves that do not necessarily follow each other in an approach within Tabu search known as *strategic oscillation.* Note that this procedure allows the trajectory that is generated to oscillate between feasible and infeasible solutions. Kochetov (2001) proposes a simple probabilistic Tabu search in which a restricted (random) 1-interchange neighborhood is used. The *simulated annealing* heuristic of Chiyoshi and Galvao (2000) combines the 1-interchange neighborhood with the general methodology of simulated annealing. The *scatter search* method of Garcia-Lopez et al. (2003) uses the Interchange procedure in a final step to improve the combined solutions that are obtained.

The *Variable Neighborhood Search* methodology imposes a set of neighborhood structures on the solution space in order to conduct a systematic search at different distances from the current solution. The movement to different neighborhoods is accomplished by a 'shaking' operation. In the standard approach for the *p*-median problem (e.g., Hansen and Mladenovic 1997), the neighborhood structures are defined by moving 1, 2, …, k_{max} facilities from their currently occupied vertices to new unoccupied ones. The shaking operator thus selects a random point in the *k*-neighborhood by randomly moving *k* facilities in this manner. A local search from this point is conducted using the Interchange procedure.

Heuristic concentration (Rosing and ReVelle 1997) is a metaheuristic of special interest to this chapter as it was developed specifically for the *p*-median problem and is based straightforwardly on the Teitz and Bart algorithm. The local minima arising in repeated runs of the Teitz and Bart algorithm are identified as two-, three-, and so on "traps" by Rosing and Hodgson (2002); these identify clusters of nodes that cannot be avoided by single interchanges. The main idea behind heuristic concentration is to then create a concentration set of desirable facility sites (open facility sites that most often appeared in the solutions from the first stage), and use this concentration set as the set of potential facility locations, thus reducing the solution space of the problem. Nodes that occur in all solutions may be assumed to be in the optimal solution if so desired. The much smaller problem defined by the demand nodes and the concentration set may be solved optimally or approximately. Heuristic concentration has been shown to provide very good, usually optimal, solutions to problems of several hundred nodes, although these are considered relatively small instances by today's standards. Since the number of customers remains at its original size, computation times may become unmanageable for larger problems. This shortcoming may be addressed in the future by applying neighborhood approaches to the traps identified in the concentration set.

Thus we see that several of the new methods have at their hearts the fundamental interchange notion of Teitz and Bart (1968). Mladenovic et al. (2007) conclude that the more recent heuristics outperform Teitz and Bart, but due to the number of different data structures and implementations, they are unable to conclude "what metaheuristic dominates others." Another useful source to note is Reese (2006).

15.4 Future Research

We believe that interest in heuristics will continue to grow in the coming years in studies of combinatorial and global optimization problems including, of course, the *p*-median problem. Here are some reasons why.

Networks are getting increasingly larger in real applications. For example, Brimberg et al. (2000) motivate their work by citing two actual case studies, a transshipment center location problem and a districting problem with $(p, n) = (20, 1{,}700)$ and $(170, 1{,}400)$, respectively. A very large-scale study dealing with spare parts logistics for a Japanese manufacturing company with 6,000 customers and 380,000 potential warehouse sites is cited in Brimberg et al. (2008a). With such trends as globalization of business, we can expect the size and complexity of location and distribution problems to only increase. Given the limiting assumptions inherent in mathematical models, finding the "optimal" solution, even if it were possible, may be of questionable importance in practice. It seems a more sensible approach would be to use heuristics to find a set of alternative "good" solutions in a way that strikes a proper balance between quality of solution and computing time.

New important applications of the *p*-median model are materializing that are outside the original scope of locating physical assets such as warehouses. We mention as an example the importance of the *p*-median model and other related models in the field of *data mining*. One objective in data mining is to detect useful patterns within databases by using models such as the *p*-median that are able to partition the dataset into meaningful clusters. These databases generally contain several thousand entries, and thus the use of heuristics becomes a practical necessity.

New developments within the field of heuristics are extending the usefulness of these methods. For example, using decomposition to solve a series of smaller (decomposed) problems is proving to be a highly effective and efficient approach to handle large problem instances. In Hansen et al. (2001), a decomposition variant of variable neighborhood search, referred to as variable neighborhood decomposition search, obtains notably better results than basic variable neighborhood search in less computing time. In fact, the method finds much better results than fast-interchange in the same time fast-interchange takes for a single descent. Another example is the recent development of primal-dual heuristics that are able to obtain tight lower bounds on the optimal solution of the *p*-median and related simple plant location problems by solving exactly or approximately a relaxed version of the dual, see Hansen et al. (2007). Thus, a guaranteed bound on the quality of the solution obtained by the primal heuristic is now provided. Heuristics are also used in con-

junction with exact solution methods. Since these exact methods are generally very sensitive to the starting point, it is important to use a good heuristic initially. Brimberg et al. (2000) note that the improved solution quality from the newer heuristics available has enabled the exact solution of much larger instances of the continuous *p*-median problem than before.

We have seen a tremendous growth in the field of heuristics in recent years. Empirical studies have shown consistently that the new metaheuristics at our disposal work better than the older methods, but aside from this we do not understand much of what transpires. Research into the theoretical underpinnings of metaheuristics is still in its infancy, but judging from the recent *Seventh Metaheuristics International Conference* (June 2007), this is becoming a very hot area indeed. It seems a safe bet to predict that the field of metaheuristics will be subject to rigorous theoretical analysis in the years to come. We believe that statistical studies of the *landscapes* derived from various local search operators will play a useful role here. The ultimate goal, aside from designing better heuristics, will be a deeper understanding of the fundamental nature of combinatorial and global optimization problems.

15.5 Conclusions

We have reviewed the classical heuristics introduced by Cooper (1963, 1964), Maranzana (1964), and Teitz and Bart (1968). Maranzana's paper, although important also for its formulation of the network model and some fundamental results, may be considered less significant as its partition method was quickly superseded by the vertex substitution method of Teitz and Bart. The contributions of these original papers were timely and important, as they introduced a wide audience to some fundamental location problems and showed how they could be solved. They stand out as important way posts in location science and are truly deserving of their celebrity status.

Rather than review derivative work in detail, we have guided readers to the detailed reviews by Brimberg et al. (2000), Reese (2006), Mladenovic et al. (2007), and Brimberg et al. (2008a). These reviews indicate that the performance of the classical heuristics suffers in the face of the explosive number of local minima that arise in large problems. None of these classic approaches is ready for retirement, however.

The Cooper algorithm, accompanied by graphical illustration, is an excellent tool for teaching a fundamental lesson of location modeling and optimization in the classroom, as aptly demonstrated in Scott (1971).

The vertex substitution method of Teitz and Bart lies at the heart of all the interchange-based heuristics developed to solve the *p*-median problem more effectively. Although much progress has been made in recent years in the development of metaheuristics, much work still remains to understand the underlying theory, and in consequence, to design heuristics in a more intelligent fashion.

References

Beasley JE (1993) Lagrangian heuristics for location problems. Eur J Oper Res 65:383–399
Bellman R (1958) On a routing problem. Q Appl Math 16:87–90
Bongartz I, Calamai PH, Conn AR (1994) A projection method for ℓ_p norm location-allocation problems. Math Program 66:283–312
Brimberg J, Mladenovic N (1999) Degeneracy in the multi-source Weber problem. Math Program 85:213–220
Brimberg J, Hansen P, Mladenovic N, Taillard ED (2000) Improvements and comparison of heuristics for solving the uncapacitated multisource Weber problem. Oper Res 48:444–460
Brimberg J, Hansen P, Mladenovic N, Salhi S (2008a) A survey of solution methods for the continuous location-allocation problem. Int J Oper Res 5:1–12
Brimberg J, Hansen P, Laporte G, Mladenovic N, Urosevic D (2008b) The maximum return-on-investment plant location problem with market share. J Oper Res Soc 59:399–406
Captivo EM (1991) Fast primal and dual heuristics for the *p*-median location problem. Eur J Oper Res 52:65–74
Chen R (1983) Solution of minisum and minimax location-allocation problems with Euclidean distances. Nav Res Logist Q 30:449–459
Chiyoshi F, Galvao RD (2000) A statistical analysis of simulated annealing applied to the *p*-median problem. Ann Oper Res 96:61–74
Cooper L (1963) Location-allocation problems. Oper Res 11:331–343
Cooper L (1964) Heuristic methods for location-allocation problems. SIAM Rev 6:37–53
Dorigo M, Di Caro G (1999) The ant colony optimization meta-heuristic. In: Corne G, Dorigo M, Glover F (eds) New ideas in optimization, McGraw-Hill, New York
Garcia-Lopez F, Melian Batista B, Moreno Perez JA, Moreno Vega JM (2003) Parallelization of the scatter search for the *p*-median problem. Parallel Comput 29:575–589
Hakimi SL (1964) Optimum locations of switching centers and the absolute centers and medians of a graph. Oper Res 12:450–459
Hanjoul P, Peeters D (1985) A comparison of two dual-based procedures for solving the *p*-median problem. Eur J Oper Res 20:387–396
Hansen P, Mladenovic N (1997) Variable neighborhood search for the *p*-median. Locat Sci 5:207–226
Hansen P, Mladenovic N, Taillard E (1998) Heuristic solution of the multisource Weber problem as a *p*-median problem. Oper Res Lett 22:55–62
Hansen P, Mladenovic N, Perez-Brito D (2001) Variable neighborhood decomposition search. J Heuristics 7:335–350
Hansen P, Brimberg J, Urosevic D, Mladenovic N (2007) Primal-dual variable neighborhood search for the simple plant-location problem. INFORMS J Comput 19:552–564
Houck CR, Joines JA, Kay MG (1996) Comparison of genetic algorithms, random restart and two-opt switching for solving large location-allocation problems. Comput Oper Res 23:587–596
Kariv O, Hakimi SL (1979) An algorithmic approach to network location problems: The *p* medians. SIAM J Appl Math 37:539–560
Katz IN (1974) Local convergence in Fermat's problem. Math Program 6:89–104
Kochetov Y (2001) Probabilistic local search algorithms for the discrete optimization problems. Discrete Mathematics and Applications. Moscow, MSU, 84–117 (in Russian)
Kochetov Y, Alekseeva E, Levanova T, Loresh N (2005) Large neighborhood search for the *p*-median problem. Yugosl J Oper Res 15:53–63
Kolesar P, Walker WE, Hausner J (1975) Determining the relation between fire engine travel times and travel distances in New York City. Oper Res 23:614–627
Kuhn HW (1973) A note on Fermat's problem. Math Program 4:98–107
Kuhn HW, Kuenne RE (1962) An efficient algorithm for the numerical solution of the generalized Weber problem in spatial economics. J Reg Sci 4:21–34

Love RF, Juel H (1982) Properties and solution methods for large location-allocation problems. J Oper Res Soc 33:443–452
Love RF, Morris JG, Wesolowsky GO (1988) Facilities location: models and methods. North Holland, New York
Maranzana FE (1964) On the location of supply points to minimize transport costs. Oper Res Q 15:261–270
Megiddo M, Supowit KJ (1984) On the complexity of some common geometric location problems. SIAM J Comput 13:182–196
Mladenovic N, Moreno-Perez JA, Moreno-Vega JM (1996) A chain-interchange heuristic method. Yugosl J Oper Res 6:41–54
Mladenovic N, Brimberg J, Hansen P, Moreno-Perez JA (2007) The *p*-median problem: A survey of metaheuristic approaches. Eur J Oper Res 179:927–939
Murtagh BA, Niwattisyawong SR (1982) An efficient method for the multi-depot location-allocation problem. J Oper Res Soc 33:629–634
Ng R, Han J (2002) CLARANS: a method for clustering objects for spatial data mining. IEEE Trans Knowl Data Eng 14:1003–1016
Reese J (2006) Solution methods for the *p*-median problem: an annotated bibliography. Networks 48:125–142
Rolland E, Schilling DA, Current JA (1996) An efficient tabu search procedure for the *p*-median problem. Eur J Oper Res 96:329–342
Rosing KE, Hodgson MJ (2002) Heuristic concentration for the *p*-median: An example demonstrating how and why it works. Comput Oper Res 29:1317–1330
Rosing KE, ReVelle CS (1997) Heuristic concentration: Two stage solution construction. Eur J Oper Res 96:329–342
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
Scott AJ (1971) Combinatorial Programming, spatial analysis and planning. Methuen, London
Teitz MB, Bart P (1968) Heuristic methods for estimating the generalized vertex median of a weighted graph. Oper Res 16:955–961
Voss S (1996) A reverse elimination approach for the *p*-median problem. Stud Locat Anal 8:49–58
Weiszfeld E (1937) Sur le point pour lequel la somme des distances de *n* points donnés est minimum. Tohoku Math J 43:355–386
Whitaker R (1983) A fast algorithm for the greedy-interchange for large-scale clustering and median location problems. Inf Syst Oper Res 21:95–108

Chapter 16
The Weiszfeld Algorithm: Proof, Amendments, and Extensions

Frank Plastria

16.1 Introduction

16.1.1 A Brief History

Some time in the early seventeenth century, the following geometrical optimization problem was posed:

> …given three points in a plane, determine a point at which the sum of the distances to the three given points is smallest.

Opinions differ about the originator of this question, usually attributed to Pierre de Fermat (1601–1665), to Battista Cavalieri (1598–1647), or to Evangelista Torricelli (1608–1647); see the historical part of the overview papers by Kupitz and Martini (1997) and Drezner et al. (2002). Optimality conditions, purely geometric solutions, and corresponding classical "ruler and compass" constructions were soon found, in particular by Cavalieri and Torricelli. This involved considering two separate cases. In case the given points form a triangle with all angles less than 120°, one sought the point inside the triangle from which each side is seen under an angle of 120°. This angle is obtained by constructing an equilateral triangle and its circumscribed circle on each side; these three circles will meet in the desired point. Whenever there exists a vertex with an angle of at least 120°, we simply take that triangle's vertex as the sought point. These two solutions are shown in Figs. 16.1a and b.

The question extended to four given points in the plane turned out to be extremely easy, as shown a good century later by Fagnano (1775): if the four given points form a convex quadrilateral, take the intersection of the diagonals; otherwise take the given point inside the triangle formed by the three others. This solution is shown in Fig. 16.2.

F. Plastria (✉)
Department of Mathematics, Operational Research, Statistics and Information Systems for Management, MOSI, Vrije Universiteit Brussel, Pleinlaan 2, B 1050 Brussels, Belgium
e-mail: frank.plastria@vub.ac.be

H. A. Eiselt, V. Marianov (eds.), *Foundations of Location Analysis,*
International Series in Operations Research & Management Science 155,
DOI 10.1007/978-1-4419-7572-0_16,

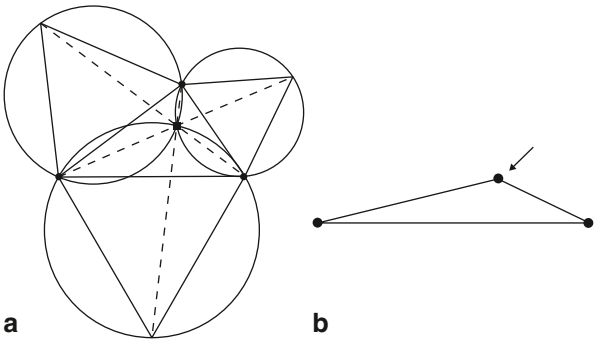


Fig. 16.1 Solution to the three-point problem

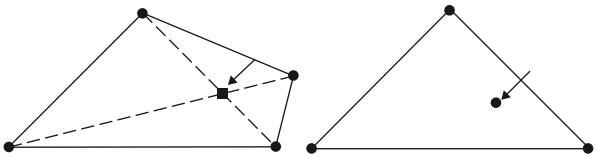


Fig. 16.2 Solution to the four-point problem

The move to more than four given points in the plane or in space was made by many scholars, including some quite well known mathematicians, such as Gauss, Bertrand, Simpson, Riesz, Steiner, and others. Optimality conditions were extended (see the next section), but geometric construction attempts seem to have remained largely undocumented likely because the problem has turned out to be unsolvable in general by simple geometric means when there are at least five given points. This fact, however, was only shown two centuries later by Cockayne and Melzak (1969) and Bajaj (1988) using advanced Galois theory. Nongeometric solution methods were also developed based on the principle of equilibrium of forces in physics, leading to a mechanical device called a Varignon frame, which was already described by Lamé and Clapeyron (1829). A very thorough treatment with extensive historical notes is given in Kupitz and Martini (1997).

One particular complete proof of the optimality conditions was obtained in the years just before the Second World War by Endré Weiszfeld, a young mathematician from Budapest, Hungary. Having fled the rising antisemitic movement in his country, he wrote its proof (together with two other proofs) in French during his stay in Paris, and sent it, together with several other manuscripts, to be published in the prestigious Tohoku Mathematical Journal in Japan, where it was published in the issue of 1936–37. This proof was based on the construction of an infinite sequence of points shown to converge to the optimal solution. At the time, this theoretical device was virtually useless for the numerical calculation of an optimal solution, since it involved far too many calculations to be of practical use. But with the advent of programmable numerical computing devices around the middle of the century, this numerical process became perfectly feasible and useful. During the 1950s and 1960s, virtually the same calculation method was rediscovered independently by several authors, in particular Miehle (1958), Cooper (1963), and Kuhn and Kuenne (1962), mostly without formal proof of convergence and in a more general setting (see Sect. 16.7 in

this Chapter). Weiszfeld's more complete work remained unknown in the west until former colleagues of Weiszfeld pointed to it during a seminar Kuhn gave in Budapest around 1963, as recounted in Kuhn (1973). Endre Weiszfeld changed his name to Andrew Vaszonyi when he emigrated to the United States after the war, where he has been active in applying mathematics to all kinds of business decisions and teaching "applicable math" until his death in 2003. He was one of the founders of *TIMS*, The Institute of Management Science, in 1954, and its first past-president, see Gass (2004). He candidly told his life's story in his autobiography Vazsonyi (2002a). But he never was really concerned with solving location problems (see Vazsonyi 2002b).

A slightly more complex variation of the geometrical problem considers a positive weight for each given point, and asks to

> find the point(s) at which the sum of *weighted* distances to the given points is minimum.

In (mining) engineering and mathematics this question emerged in 1829 in the previously cited memoir of Lamé and Clapeyron, under the name "general problem of the theory of least distances." Even more than the Weiszfeld paper, this paper has remained hidden and unknown until its recent rediscovery and translation by Franksen and Grattan-Guinness (1989). The following application is described by Lamé and Clapeyron as "the most simple case" (quoting the translation given in Franksen and Grattan-Guinness 1989):

> Let us suppose, for example, that one wants to establish a plant intended for the treatment of metallic minerals, obtained from different mines, and which must be mixed together in known proportions with the smelting flux, and with a fuel, extracted or bought at given locations. Let us suppose also that the metallic products obtained, for which the weights will be in a ratio known in advance with the initial materials used, should be distributed according to a certain law among different markets at known positions. Suppose finally that the only condition that it is important to satisfy in the choice of the position of this plant, is that of the best possible economy in terms of the price of transportation. Under these special circumstances, the plant ought to be placed, so that its distances from the different locations furnishing the original materials, or receiving the produced metals, multiplied respectively by the weights which must travel [these distances], and for which the ratios are known, give products the sum of which will be a minimum relative to all other positions.

The mechanical device they propose is as follows (quoting again Franksen and Grattan-Guinness 1989):

> One fastens horizontally a well planed board, on which will be traced the topographical map of the land. One will suspend vertical pulleys beneath this board, turning about vertical axes. These axes must cross the board at those points which on the map depict the mines and the above-mentioned markets. One will wind around each of these pulleys a wire, on which one suspends a weight; this weight must be proportional to that which should be transported on the path between the place, represented by that pulley, and the desired plant. Finally one will attach all the wires to one very small and mobile ring. The point at which this ring will stop, projected onto the map, will determine the location of the plant.

The device for points A, B, C, and D with weights w_A, w_B, w_C, and w_D is shown in Fig. 16.3.

In economics this question has often been said to appear first (with only three given points) in Weber's (1909) book about the location of industries, and in particular in the appendix written by Georg Pick (who describes the Varignon frame): where

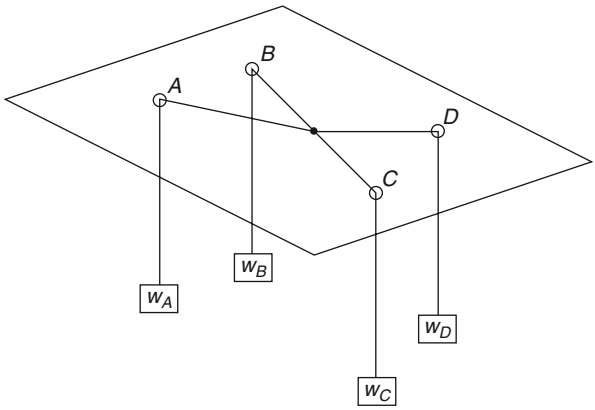


Fig. 16.3 Mechanical device for the weighted four-point problem

should one locate a factory that receives its raw materials from a single source but has to deliver its finished products to two different markets. Probably following Launhardt (1885), Weber follows the same idea that this location should be chosen such that the transportation costs of raw materials and products are the lowest possible, all other factors remaining constant. Consider these costs as proportional to (Euclidean) distance and therefore at different rates for the raw materials and for the products, thereby obtaining a simple instance of the question above. It has been popularized among economists and geographers as the search for the *point of minimum aggregate travel*, and has acquired the status of fundamental problem in spatial economics under various names, like *the Weber problem* (the name we will use here), the Steiner problem, and, more recently, as the single facility *Euclidean Minisum Location Problem* (Wesolowsky 1993). Most frequently, the problem is referred to as the *Euclidean median problem*. For a detailed history of this problem, see Drezner et al. (2002).

16.1.2 Choices in Setting and Notation

The next sections will develop the details of the Weber problem in step-by-step fashion, including Weiszfeld's method and proof of convergence. At the time it was written, it only referred to two and three-dimensional space, and the possibility of considering weighted distances was only shortly evoked as a final remark. Since all arguments directly extend to the weighted version in any dimension higher than 1, we will adapt his argument to this slightly more general case.

The notations used by Weiszfeld are now outdated, and new notations have been standardized when working in vector spaces. We will therefore also adopt these new standards. Readers interested in the original notations may refer to the original contribution in Weiszfeld (1937), or its recent annotated translation into English in Weiszfeld and Plastria (2009).

Weiszfeld (1937) described three proofs of the optimality conditions. The first is the convergent sequence argument, which forms the basis of the "Weiszfeld algorithm" and is the main subject of this chapter. The second proof is based on more

geometrical arguments but is, unfortunately, only valid in the plane (dimension 2) or for at most four points in three-dimensional space. Finally, the third proof is also quite remarkable since it uses, *avant la lettre*, arguments of convex analysis that were to be developed in general much later in the 1960s by Moreau and many others; see Rockafellar (1970), or, for a more recent exposition, Hiriart-Urruty and Lemaréchal (2001).

The next section starts by formulating the Weber problem, thereby introducing our notation together with the main concepts we will use, and stating the main theorem, which asserts existence and uniqueness of an optimal solution together with complete optimality conditions. This is followed by a short section about the standard error involving the center of gravity. Then we devote a section to following in detail the lines of Weiszfeld's first proof.

Weiszfeld's geometric second proof will not be discussed here, not only because it does not seem to lead to interesting generalizations (e.g. it does not readily work for the weighted problem), but mainly because it seems to contain some flaws that remain unresolved to the best of this author's knowledge. Instead, we devote a much shorter section to some of the basic ideas of convex analysis in order to derive the more modern view, initiated by Kuhn (1967), which finds its first traces in Weiszfeld's third proof, as annotated in Weiszfeld and Plastria (2007). This is followed by a discussion of several identified difficulties together with an overview of the ensuing research work and results.

The final section will have a look at the many ways in which Weiszfeld's algorithm has been and still is being extended to many variants of the Weber problem.

16.2 The Weber Problem and its Optimality Conditions

Let n different and not aligned points $\mathbf{a}_1,\ldots,\ \mathbf{a}_n$ be given in d-dimensional space $\mathbb{R}^d (d > 1)$, together with positive real weights $w_i > 0$, $i \in N = \{1, \ldots, n\}$.

The Weber problem seeks a point $\mathbf{m} \in \mathbb{R}^d$, where the weighted sum of Euclidean distances to these given points is minimal. In other words we want to minimize the following unconstrained function

$$f : \mathbb{R}^d \to \mathbb{R} : \mathbf{x} \to \sum_{i \in N} w_i \, \|\mathbf{a}_i - \mathbf{x}\|.$$

Here, $\|\mathbf{z}\|$ denotes the Euclidean length (or *norm*) of the vector $\mathbf{z} = (z_k)_{k=1,\cdots d} \in \mathbb{R}^d$, defined as

$$\|\mathbf{z}\| = \sqrt{\sum_{k=1}^{d} z_k^2},$$

and $\mathbf{a}_i - \mathbf{x}$ is the vector going from $\mathbf{x}$ to $\mathbf{a}_i$, so $\|\mathbf{a}_i - \mathbf{x}\|$ denotes the Euclidean distance between $\mathbf{x}$ and $\mathbf{a}_i$. Note that this norm is always strictly positive, except for the zero vector $\|\mathbf{0}\| = 0$. A unit vector is a vector of length 1.

We will also need the scalar product of two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$,

$$\langle \mathbf{u}; \mathbf{v} \rangle = \sum_{k=1}^{d} u_k v_k,$$

which is symmetric, i.e., $\langle \mathbf{u}; \mathbf{v} \rangle = \langle \mathbf{v}; \mathbf{u} \rangle$, and bilinear, i.e., $\langle \lambda \mathbf{u} + \mu \mathbf{u}'; \mathbf{v} \rangle = \lambda \langle \mathbf{u}; \mathbf{v} \rangle + \mu \langle \mathbf{u}'; \mathbf{v} \rangle$ holds for any $\mathbf{u}, \mathbf{u}', \mathbf{v} \in \mathbb{R}^d$ and any $\lambda, \mu \in \mathbb{R}$. We see that $\|\mathbf{z}\|^2 = \langle \mathbf{z}; \mathbf{z} \rangle$, and one may prove the well-known Cauchy-Schwartz-Buniakowski inequality

$$|\langle \mathbf{u}; \mathbf{v} \rangle| \leq \|\mathbf{u}\| \, \|\mathbf{v}\|,$$

where equality holds if and only if $\mathbf{v}$ is a multiple of $\mathbf{u}$. In fact, the geometric interpretation of this scalar product for nonzero vectors is $\langle \mathbf{u}; \mathbf{v} \rangle = \|\mathbf{u}\| \, \|\mathbf{v}\| \cos \alpha_{\mathbf{uv}}$, where $\alpha_{\mathbf{uv}}$ denotes the angle between vectors $\mathbf{u}$ and $\mathbf{v}$. Thus, in case $\mathbf{u}$ is a unit vector, $\langle \mathbf{u}; \mathbf{v} \rangle$ represents the length of the orthogonal projection of $\mathbf{v}$ on the line defined by $\mathbf{u}$ (connecting the origin to point $\mathbf{u}$). For any nonzero vector $\mathbf{z} \neq 0$ we may construct its corresponding unit vector, having the same direction as $\mathbf{z}$, by $u(\mathbf{z}) = \mathbf{z}/\|\mathbf{z}\|$. Note that $\langle \mathbf{u}, \mathbf{v} \rangle = 0$ either means that $\mathbf{u}$ or $\mathbf{v}$ is the zero vector, or that $\mathbf{u}$ and $\mathbf{v}$ are orthogonal.

We may now state the main theorem, the unweighted version of which is attributed by Weiszfeld to Sturm (1884). A proof will be developed in Sect. 16.4.

Theorem 1 (Existence and optimality conditions): *There exists a unique point m minimizing f. In other words, there exists an* $\mathbf{m} \in \mathbb{R}^d$, *such that* $f(\mathbf{m}) < f(\mathbf{x})$ *for any point* $\mathbf{x} \in \mathbb{R}^d$ *different from* $\mathbf{m}$ (*in which case it is evidently unique*).

This minimum point is characterized by the following optimality conditions

1. *If there exists a point* $\boldsymbol{m}$, *different from all* $\boldsymbol{a}_i$, $i \in N$, *for which*

$$\sum_{i \in N} w_i u(\mathbf{a}_i - \mathbf{m}) = 0, \tag{16.1}$$

then this $\boldsymbol{m}$ *is the minimum point.*

2. *If for some* $k \in N$ *we have*

$$\left\| \sum_{i \in N \setminus \{k\}} w_i u(\mathbf{a}_i - \mathbf{a}_k) \right\| \leq w_k, \tag{16.2}$$

then this $\boldsymbol{a}_k$ *is the minimum point.* (In this context, Weiszfeld uses the somewhat misleading symbol Σ' to indicate that the k-th term is left out of the summation. We prefer to state explicitly over which index-set the sum is made).

Note 1: It should be observed that this theorem also remains valid when the points $\mathbf{a}_i$ are aligned, except that uniqueness of the optimal solution is not guaranteed anymore.

Proof: When all $\mathbf{a}_i$ lie on a line of direction the unit vector $\mathbf{v}$, say, then all $\mathbf{a}_i$ are of the form $\mathbf{p}+t_i\mathbf{v}$, with $t_i \in \mathbb{R}^d$ and $\mathbf{p}$ an arbitrary fixed point of the line. Now in case $\mathbf{m}$ lies outside this line, let $\mathbf{m}' \neq \mathbf{m}$ be its orthogonal projection on this line. Then for all $\mathbf{a}_i$ we have $\mathbf{a}_i-\mathbf{m}=\mathbf{a}_i-\mathbf{m}'+\mathbf{m}'-\mathbf{m}$ and, since, by construction, $\mathbf{a}_i-\mathbf{m}'$ is orthogonal to $\mathbf{m}'-\mathbf{m}$, we have $\|\mathbf{a}_i-\mathbf{m}\|^2=\|\mathbf{a}_i-\mathbf{m}'\|^2+\|\mathbf{m}'-\mathbf{m}\|^2>\|\mathbf{a}_i-\mathbf{m}'\|^2$, so that $\mathbf{m}'$ lies strictly closer to all $\mathbf{a}_i$ than $\mathbf{m}$, and hence $f(\mathbf{m}')<f(\mathbf{m})$, showing that $\mathbf{m}$ cannot be optimal.

Choosing a basis of origin $\mathbf{m}'$ with the projection direction $\mathbf{m}-\mathbf{m}'$ as the first basevector, $\mathbf{v}$ as the second basevector, and arbitrary further basevectors when in dimension greater than 2, one also sees that all vectors $u(\mathbf{a}_i-\mathbf{m})$ have a strictly positive first component, showing that condition (16.1) does not hold. In case $\mathbf{m}$ lies on the line, all vectors $u(\mathbf{a}_i-\mathbf{m})$ either equal $\mathbf{v}$ or $-\mathbf{v}$, depending on whether $\mathbf{a}_i$ lies before or after $\mathbf{m}$ on the line oriented by $\mathbf{v}$.

The two conditions (16.1) and (16.2) then express exactly that the sum of weights before $\mathbf{m}$ equals the sum of weights after $\mathbf{m}$, possibly up to the weight present at $\mathbf{m}$ itself. In other words, $\mathbf{m}$ is a median of the points $\mathbf{a}_i$ with weights w_i, well-known as the optimal solution to a Weber problem on a line; see, e.g., Francis et al. (1992, p. 194). □

16.3 Intermezzo: About the Gravity Center

But let us first have a quick look at the slightly different question

> find the point(s) at which the sum of weighted *squared* distances to the given points is minimum.

Mathematically, given the points a_i ($i \in N$) and positive weights μ_i, we want to find a point where the following function is minimized

$$f^2 : \mathbb{R}^d \to \mathbb{R} : \mathbf{x} \to \sum_{i \in N} \mu_i \|\mathbf{a}_i - \mathbf{x}\|^2.$$

Below, we find the easy solution to this problem in two ways.

16.3.1 The Classical Argument

Written extensively using coordinates, with $\mathbf{x} = (x_k)_{k=1,\ldots,d}$ and $\mathbf{a}_i = (a_{ik})_{k=1,\ldots,d} \in \mathbb{R}^d$, the classical argument gives the somewhat simpler expression

$$f^2(\mathbf{x}) = \sum_{i \in N} \mu_i \left(\sum_{k=1}^{d} (a_{ik} - x_k)^2 \right),$$

which may be rewritten as

$$f^2(\mathbf{x}) = \sum_{k=1}^{d} \left[\mu x_k^2 - 2 \left(\sum_{i \in N} \mu_i a_{ik} \right) x_k + \sum_{i \in N} \mu_i a_{ik}^2 \right],$$

where $\mu = \sum_{i \in N} \mu_i > 0$. This is a simple sum of convex quadratic functions, each of a separate variable x_k, which may be minimized by minimizing them separately for each k (e.g., set the derivative to zero). This results in

$$x_k = \frac{1}{\mu} \sum_{i \in N} \mu_i a_{ik}, \text{or, in vector notation,}$$

$$\mathbf{x} = \frac{1}{\mu} \sum_{i \in N} \mu_i \mathbf{a}_i$$

which is the *center-of-gravity* or *centroid* of the μ_i-weighted points $\mathbf{a}_i$.

16.3.2 The Vectorial Argument

Define the center-of-gravity of the μ_i-weighted points $\mathbf{a}_i$ by

$$\mathbf{g} = \frac{\sum_{i \in N} \mu_i \mathbf{a}_i}{\sum_{i \in N} \mu_i}.$$

First, observe that this means that $\sum_{i \in N} \mu_i (\mathbf{a}_i - \mathbf{g}) = \mathbf{0}$.

Developing now the equality $\|\mathbf{z}_i\|^2 = \langle \mathbf{z}_i; \mathbf{z}_i \rangle$ for $\mathbf{z}_i = \mathbf{a}_i - \mathbf{x} = (\mathbf{a}_i - \mathbf{g}) + (\mathbf{g} - \mathbf{x})$, we obtain

$$\begin{aligned}
f^2(\mathbf{x}) &= \sum_{i \in N} \mu_i \|\mathbf{a}_i - \mathbf{x}\|^2 \\
&= \sum_{i \in N} \mu_i \langle (\mathbf{a}_i - \mathbf{g}) + (\mathbf{g} - \mathbf{x}); (\mathbf{a}_i - \mathbf{g}) + (\mathbf{g} - \mathbf{x}) \rangle \\
&= \sum_{i \in N} \mu_i \left[\langle \mathbf{a}_i - \mathbf{g}; \mathbf{a}_i - \mathbf{g} \rangle + \langle \mathbf{g} - \mathbf{x}; \mathbf{g} - \mathbf{x} \rangle + 2 \langle \mathbf{a}_i - \mathbf{g}; \mathbf{g} - \mathbf{x} \rangle \right] \\
&= \sum_{i \in N} \mu_i \|\mathbf{a}_i - \mathbf{g}\|^2 + \sum_{i \in N} \mu_i \|\mathbf{g} - \mathbf{x}\|^2 + 2 \left\langle \sum_{i \in N} \mu_i (\mathbf{a}_i - \mathbf{g}); \mathbf{g} - \mathbf{x} \right\rangle \\
&= f^2(\mathbf{g}) + \sum_{i \in N} \mu_i \|\mathbf{g} - \mathbf{x}\|^2,
\end{aligned}$$

and therefore $f^2(\mathbf{x}) > f^2(\mathbf{g})$ for all $\mathbf{x} \neq \mathbf{g}$, showing that $\mathbf{g}$ is the only minimizer of f^2.

16.3.3 The Standard Error

It has often been advocated to solve the Weber problem by taking the gravity center. But, as shown above, the center-of-gravity is the optimal solution to a quite different problem that does not consider Euclidean distances, but *squared* Euclidean distances. This was already clearly shown by Schärlig (1973) and by Francis et al. (1992), among others. For more details on this controversy and the history of the Weber problem, see Drezner et al. (2002). Unfortunately this error seems quite hard to eradicate: it is still found in most textbooks on Operations Management—ironically even in Weida et al. (2001), a recent introductory book with coauthor Vaszonyi.

It is also possible to view Weiszfeld's method as an iterated application of the center-of-gravity construction, but with varying weights, as we show next. When being at a tentative point $\mathbf{p}$, which easy weights μ_i should be chosen for the function f^2 in order to obtain the same objective function value as f at $\mathbf{p}$? The equality

$$f(\mathbf{p}) = \sum_{i \in N} w_i \|\mathbf{a}_i - \mathbf{p}\| = f^2(\mathbf{p}) = \sum_{i \in N} \mu_i \|\mathbf{a}_i - \mathbf{p}\|^2$$

has the easiest solution:

$$\mu_i = \frac{w_i}{\|\mathbf{a}_i - \mathbf{p}\|}.$$

It may therefore seem natural to have a look at the optimal solution for this function f^2, which is exactly what Weiszfeld proposes to do.

16.4 Weiszfeld's First Proof: the "Weiszfeld Algorithm"

Consider a point $\mathbf{p} \in \mathbb{R}^d$ different from all $\mathbf{a}_i$. We construct a new point $T(\mathbf{p})$ by taking the center-of-gravity of the points $\mathbf{a}_i$ with respective weights $w_i / \|\mathbf{a}_i - \mathbf{p}\|$, i.e;

$$T(\mathbf{p}) \stackrel{def}{=} \frac{\sum_{i \in N} \frac{w_i}{\|\mathbf{a}_i - \mathbf{p}\|} \mathbf{a}_i}{\sum_{i \in N} \frac{w_i}{\|\mathbf{a}_i - \mathbf{p}\|}}. \tag{16.3}$$

The Weiszfeld algorithm now consists of applying this construction iteratively. Start with any point $\mathbf{p}_1$, construct $\mathbf{p}_2 = T(\mathbf{p}_1)$, then $\mathbf{p}_3 = T(\mathbf{p}_2)$, etc. This will yield an infinite sequence (P) $\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \ldots$.

Note 2: It should be noted that an infinite sequence is only obtained when the points $\mathbf{p}_k$ are always different from all $\mathbf{a}_i$, since otherwise the new point $T(\mathbf{p}_k)$ is undefined. In the original paper by Weiszfeld this assumption is never mentioned, but remains implicit throughout. In the remainder of this section we will do likewise. This difficulty will be discussed in more detail in Sect. 16.6.1.

Theorem 2 (Convergence): *This sequence* (P) *is convergent, and its limit is the minimum point of the points* $\mathbf{a}_i$, $i \in N$, *independent of the starting point* $\mathbf{p}_1$.

The proof of this convergence theorem is obtained by way of a series of technical lemmas (called auxiliary theorems by Weiszfeld). Section 16.5 then discusses how the Convergence Theorem implies the main Theorem 1.

Lemma 3: *Given a finite number of points* $\mathbf{b}_j \in \mathbb{R}^d (j \in J)$ *with corresponding weights* $w_j > 0$ *and a line L not passing through all* $\mathbf{b}_j$, *let* $\mathbf{x}_0$ *be an arbitrary point of L, and* $\mathbf{v}$ *its direction. Then the points of L may be parametrically written as* $\mathbf{x}_t = \mathbf{x}_0 + t\mathbf{v}$, $t \in \mathbb{R}$. *The expression*

$$\left\langle \sum_{j \in J} w_j u(\mathbf{b}_j - \mathbf{x}_t); \mathbf{v} \right\rangle$$

is strictly decreasing with t.

Proof: Considering a fixed $\mathbf{b}_j \notin L$, $\langle u(\mathbf{b}_j - \mathbf{x}_t); \mathbf{v}\rangle$ is the cosine of the angle α_t between $u(\mathbf{b}_j - \mathbf{x}_t)$ and $\mathbf{v}$. This angle strictly increases with t from 0 to π, so its cosine strictly decreases with t. Therefore, for any $t < t'$ we have $\langle u(\mathbf{b}_j - \mathbf{x}_t); \mathbf{v}\rangle > \langle u(\mathbf{b}_j - \mathbf{x}_{t'}); \mathbf{v}\rangle$. In case $\mathbf{b}_j \in L$, we rather have for any $\mathbf{x}_t \neq \mathbf{b}_j$ either $u(\mathbf{b}_j - \mathbf{x}_t) = -\mathbf{v}$ or $u(\mathbf{b}_j - \mathbf{x}_t) = \mathbf{v}$ according to whether $\mathbf{x}_t$ lies before or after $\mathbf{b}_j$ on L, respectively yielding $\langle u(\mathbf{b}_j - \mathbf{x}_t); \mathbf{v}\rangle = -1$ or 1. So in this case for any $t < t'$, the relation $\langle u(\mathbf{b}_j - \mathbf{x}_t); \mathbf{v}\rangle \geq \langle u(\mathbf{b}_j - \mathbf{x}_{t'}); \mathbf{v}\rangle$ holds.

Multiplying each inequality (at least one of which is strict) with its corresponding $w_j > 0$ and summing yields for any $t < t'$

$$\left\langle \sum_{j \in J} w_j u(\mathbf{b}_j - \mathbf{x}_t); \mathbf{v} \right\rangle = \sum_{j \in J} w_j \langle u(\mathbf{b}_j - \mathbf{x}_t); \mathbf{v}\rangle > \sum_{j \in J} w_j \langle u(\mathbf{b}_j - \mathbf{x}_{t'}); \mathbf{v}\rangle$$
$$= \left\langle \sum_{j \in J} w_j u(\mathbf{b}_j - \mathbf{x}_{t'}); \mathbf{v} \right\rangle,$$

which proves the lemma. □

Lemma 4: *There can exist at most one point* $\mathbf{m}$ *satisfying Condition* (16.1).

Proof: We show that the existence of two different points $\mathbf{m}$ and $\mathbf{m}'$ satisfying condition (1) leads to a contradiction. The assumption means one would have

$$\sum_{i \in N} w_i u(\mathbf{a}_i - \mathbf{m}) = \sum_{i \in N} w_i u(\mathbf{a}_i - \mathbf{m}') = 0.$$

In case $\mathbf{m} \neq \mathbf{m}'$, these points would define a line L with direction $\mathbf{v}$, say. The previous equality would then yield

$$\left\langle \sum_{i \in N} w_i u(\mathbf{a}_i - \mathbf{m}); \mathbf{v} \right\rangle = \left\langle \sum_{i \in N} w_i u(\mathbf{a}_i - \mathbf{m}'); \mathbf{v} \right\rangle (= 0),$$

whereas by virtue of Lemma 3 with $J = N$ and $\mathbf{b}_i = \mathbf{a}_i$ applied to the points $\mathbf{m}$ and $\mathbf{m}'$ on L, one should have a strict inequality. □

Lemma 5: *In case there exists some point* $\mathbf{m}$ *that satisfies Condition* (16.1), *none of the points* $\mathbf{a}_i$ *satisfy Condition* (16.2).

Proof: Here, we provide the proof for the point $\mathbf{a}_1$; the reasoning for other points $\mathbf{a}_i$ being fully similar. Thus assume for some $\mathbf{m} \neq \mathbf{a}_1$ we have

$$\sum_{i \in N} w_i u(\mathbf{a}_i - \mathbf{m}) = 0.$$

Consider the line L passing through $\mathbf{a}_1$ towards $\mathbf{m}$. This line has the direction $\mathbf{v} = u(\mathbf{m} - \mathbf{a}_1)$. On one hand, the previous equality implies that $\left\langle \sum_{i \in N} w_i u(\mathbf{a}_i - \mathbf{m}); \mathbf{v} \right\rangle = 0$, whereas by virtue of the choice of $\mathbf{v}$, we have $\langle u(\mathbf{a}_1 - \mathbf{m}); \mathbf{v} \rangle = -1$, from which we conclude that $\left\langle \sum_{i \in N \setminus \{1\}} w_i u(\mathbf{a}_i - \mathbf{m}); \mathbf{v} \right\rangle = w_1$. However, from Lemma 3 with $J = N \setminus \{1\}$ and $\mathbf{b}_i = \mathbf{a}_i$, applied to the points $\mathbf{a}_1$ and $\mathbf{m}$ on L, and noting that since $\mathbf{a}_1$ lies on L, not all these points may lie on L, we obtain

$$\left\langle \sum_{i \in N \setminus \{1\}} w_i u(\mathbf{a}_i - \mathbf{a}_1); \mathbf{v} \right\rangle > \left\langle \sum_{i \in N \setminus \{1\}} w_i u(\mathbf{a}_i - \mathbf{m}); \mathbf{v} \right\rangle = w_1.$$

By virtue of the Cauchy-Schwartz-Buniakowski inequality, and since $\|\mathbf{v}\| = 1$, we then have

$$\left\| \sum_{i \in N \setminus \{1\}} w_i u(\mathbf{a}_i - \mathbf{a}_1) \right\| \geq \left\langle \sum_{i \in N \setminus \{1\}} w_i u(\mathbf{a}_i - \mathbf{a}_1); \mathbf{v} \right\rangle > w_1,$$

showing that $\mathbf{a}_1$ does not satisfy Condition (16.2). □

Lemma 6: *At most one of the points* $\mathbf{a}_i$ *may satisfy Condition* (16.2).

Proof: We will only prove that in case $\mathbf{a}_1$ satisfies condition (16.2), then $\mathbf{a}_2$ does not. All other cases are completely similar. Consider then the line L moving through $\mathbf{a}_1$ towards $\mathbf{a}_2$, which has direction $\mathbf{v} = u(\mathbf{a}_2 - \mathbf{a}_1)$. From Lemma 3 with $J = N \setminus \{1, 2\}$ and $\mathbf{b}_i = \mathbf{a}_i$ (which do not all lie on L), applied to the points $\mathbf{a}_1$ and $\mathbf{a}_2$ on L, we obtain

$$\left\langle \sum_{i \in N \setminus \{1,2\}} w_i u(\mathbf{a}_i - \mathbf{a}_1); \mathbf{v} \right\rangle > \left\langle \sum_{i \in N \setminus \{1,2\}} w_i u(\mathbf{a}_i - \mathbf{a}_2); \mathbf{v} \right\rangle.$$

Since by assumption $\mathbf{a}_1$ satisfies Condition (16.2), i.e. $\left\| \sum_{i \in N \setminus \{1\}} w_i u(\mathbf{a}_i - \mathbf{a}_1) \right\| \leq w_1$, we also have, using $\|\mathbf{v}\| = 1$,

$$\left\langle \sum_{i\in N\setminus\{1\}} w_i u(\mathbf{a}_i - \mathbf{a}_1); \mathbf{v} \right\rangle \le \left\| \sum_{i\in N\setminus\{1\}} w_i u(\mathbf{a}_i - \mathbf{a}_1) \right\| \le w_1.$$

By the choice of $\mathbf{v}$ we have $\langle u(\mathbf{a}_2 - \mathbf{a}_1); \mathbf{v}\rangle = 1$, and thus the equality

$$\left\langle \sum_{i\in N\setminus\{1\}} w_i u(\mathbf{a}_i - \mathbf{a}_1); \mathbf{v} \right\rangle = w_2 + \left\langle \sum_{i\in N\setminus\{1,2\}} w_i u(\mathbf{a}_i - \mathbf{a}_1); \mathbf{v} \right\rangle,$$

which, together with the previous inequalities, yields

$$\left\langle \sum_{i\in N\setminus\{1,2\}} w_i u(\mathbf{a}_i - \mathbf{a}_2); \mathbf{v} \right\rangle < w_1 - w_2.$$

It now follows that

$$\begin{aligned}
&\left\langle \sum_{i\in N\setminus\{2\}} w_i u(\mathbf{a}_i - \mathbf{a}_2); \mathbf{v} \right\rangle \\
&\quad = w_1 \langle u(\mathbf{a}_1 - \mathbf{a}_2); \mathbf{v}\rangle + \left\langle \sum_{i\in N\setminus\{1,2\}} w_i u(\mathbf{a}_i - \mathbf{a}_2); \mathbf{v} \right\rangle \\
&\quad = -w_1 + \left\langle \sum_{i\in N\setminus\{1,2\}} w_i u(\mathbf{a}_i - \mathbf{a}_2); \mathbf{v} \right\rangle < -w_1 + w_1 - w_2 = -w_2,
\end{aligned}$$

and due to the Cauchy-Schwartz-Buniakowski inequality we obtain

$$\left\| \sum_{i\in N\setminus\{2\}} w_i u(\mathbf{a}_i - \mathbf{a}_2) \right\| \ge \left| \left\langle \sum_{i\in N\setminus\{2\}} w_i u(\mathbf{a}_i - \mathbf{a}_2); \mathbf{v} \right\rangle \right| > w_2.$$

This indicates that $\mathbf{a}_2$ does not satisfy Condition (16.2). □

Lemma 7: *For any point* $\mathbf{p} \neq \mathbf{a}_i \; \forall$ *i with* $T(\mathbf{p}) \neq \mathbf{p}$, *we have* $f(T(\mathbf{p})) < f(\mathbf{p})$.

Proof: Applying the optimality property of the center of gravity discussed in Sect. 16.3 for the choice of weights $\mu_i = \frac{w_i}{\|\mathbf{a}_i - \mathbf{p}\|}$, for which the center-of-gravity is $\mathbf{g} = T(\mathbf{p})$, we obtain, choosing $\mathbf{x} = \mathbf{p} \neq T(\mathbf{p})$,

$$\sum_{i\in N} \frac{w_i}{\|\mathbf{a}_i - \mathbf{p}\|} \|\mathbf{a}_i - T(\mathbf{p})\|^2 < \sum_{i\in N} \frac{w_i}{\|\mathbf{a}_i - \mathbf{p}\|} \|\mathbf{a}_i - \mathbf{p}\|^2 = \sum_{i\in N} w_i \|\mathbf{a}_i - \mathbf{p}\|$$

But we also have

$$\begin{aligned}
&\sum_{i\in N}\frac{w_i}{\|\mathbf{a}_i-\mathbf{p}\|}\|\mathbf{a}_i-T(\mathbf{p})\|^2\\
&=\sum_{i\in N}\frac{w_i}{\|\mathbf{a}_i-\mathbf{p}\|}\big[\|\mathbf{a}_i-\mathbf{p}\|+\big(\|\mathbf{a}_i-T(\mathbf{p})\|-\|\mathbf{a}_i-\mathbf{p}\|\big)\big]^2\\
&=\sum_{i\in N}w_i\|\mathbf{a}_i-\mathbf{p}\|+2\sum_{i\in N}w_i\|\mathbf{a}_i-T(\mathbf{p})\|-2\sum_{i\in N}w_i\|\mathbf{a}_i-\mathbf{p}\|\\
&\quad+\sum_{i\in N}\frac{w_i}{\|\mathbf{a}_i-\mathbf{p}\|}\big(\|\mathbf{a}_i-T(\mathbf{p})\|-\|\mathbf{a}_i-\mathbf{p}\|\big)^2
\end{aligned}$$

which, combined with previous inequality, yields

$$\begin{aligned}
&2\sum_{i\in N}w_i\|\mathbf{a}_i-T(\mathbf{p})\|-2\sum_{i\in N}w_i\|\mathbf{a}_i-\mathbf{p}\|\\
&<-\sum_{i\in N}\frac{w_i}{\|\mathbf{a}_i-\mathbf{p}\|}\big(\|\mathbf{a}_i-T(\mathbf{p})\|-\|\mathbf{a}_i-\mathbf{p}\|\big)^2<0,
\end{aligned}$$

from which it follows that

$$f(T(\mathbf{p}))=\sum_{i\in N}w_i\|\mathbf{a}_i-T(\mathbf{p})\|<\sum_{i\in N}w_i\|\mathbf{a}_i-\mathbf{p}\|=f(\mathbf{p}).\qquad\square$$

Lemma 8: *The sequence* (P) *and any of its subsequences has an accumulation point*.

Proof: Each point $\mathbf{p}_k$ of (P) with $k>1$ equals $T(\mathbf{p}_{k-1})$, and therefore, by its definition, is a convex combination of the points $\mathbf{a}_i$. Therefore, all these points lie in the convex hull of the points $\mathbf{a}_i$, which is a bounded and closed set. The well-known theorem of Bolzano-Weierstrass then asserts the existence of an accumulation point of the sequence (P). The same argument also applies to any subsequence of (P). □

The following lemma is not present in Weiszfeld's paper, but is implicitly used.

Lemma 9 (Additional): *The function f is continuous everywhere, and the map T is continuous at any point* $\mathbf{p}\neq\mathbf{a}_i$, $\forall$ *i.*

Proof All distances $\|\mathbf{a}_i-\mathbf{p}\|$ are continuous functions of $\mathbf{p}$, from which one directly obtains the continuity of f. As long as $\mathbf{p}$ is different from any $\mathbf{a}_i$, these distances are also strictly positive. Therefore, the weights $\frac{w_i}{\|\mathbf{a}_i-\mathbf{p}\|}$, used to obtain T as a gravity center, are all continuous functions of $\mathbf{p}$, implying the continuity of T at $\mathbf{p}$. □

Lemma 10: *If* $\mathbf{m}$ *is an accumulation point of* (P) *which is different from any* $\mathbf{a}_i$, *then* $T(\mathbf{m})=\mathbf{m}$.

Proof: Applying Lemma 7 to each $\mathbf{p}_{k+1}=T(\mathbf{p}_k)$, we have $0\le f(\mathbf{p}_{k+1})\le f(\mathbf{p}_k)$ for each k, so the sequence of real numbers $f(\mathbf{p}_k)$ is decreasing and bounded below,

hence converges to some value μ. Therefore, by the continuity of f, for any accumulation point $\mathbf{m}$ of (P) we will have $f(\mathbf{m}) = \mu$. By construction of (P), $T(\mathbf{m})$ would then also be an accumulation point of (P). If then $T(\mathbf{m})$ were different from $\mathbf{m}$, Lemma 7 would imply $f(T(\mathbf{m})) < f(\mathbf{m})$, contradicting that $f(T(\mathbf{m})) = \mu$. Therefore $T(\mathbf{m}) = \mathbf{m}$. □

Lemma 11: *Let* $\mathbf{m}$ *be an accumulation point of* (P) *that is different from any* $\mathbf{a}_i$, *then* $\mathbf{m}$ *satisfies condition* (16.1).

Proof: By virtue of Lemma 10, we have $T(\mathbf{m}) = \mathbf{m}$. Using the definition of T and bringing the denominator to the right-hand side, we obtain

$$\sum_{i \in N} \frac{w_i}{\|\mathbf{a}_i - \mathbf{m}\|} \mathbf{a}_i = \sum_{i \in N} \frac{w_i}{\|\mathbf{a}_i - \mathbf{m}\|} \mathbf{m},$$

which is equivalent to

$$0 = \sum_{i \in N} \frac{w_i}{\|\mathbf{a}_i - \mathbf{m}\|} (\mathbf{a}_i - \mathbf{m}) = \sum_{i \in N} w_i u(\mathbf{a}_i - \mathbf{m}),$$

which is exactly Condition (16.1). □

Lemma 12: *At most one accumulation point of* (P) *is different from any* $\mathbf{a}_i$.

Proof Combining Lemmas 11 and 4, the proof follows immediately. □

Lemma 13: *In case* $\mathbf{a}_k$ *is an accumulation point of* (P), *it is the only accumulation point of* (P).

Proof: Without loss of generality let $\mathbf{a}_1$ be an accumulation point of (P), and let us assume there exists at least one other accumulation point of this sequence. This will lead us to a contradiction.

Such a point must then either be one of the $\mathbf{a}_i$ $(i \in N \backslash \{1\})$ or, if it exists, some by previous lemma unique $\mathbf{m}$ different from any $\mathbf{a}_i$. Since these possibilities form a finite set, one may construct a closed ball B of center $\mathbf{a}_1$ and a sufficiently small radius $\varepsilon > 0$ that does not contain any other accumulation point than $\mathbf{a}_1$. Under our assumption, (P) contains some subsequence converging to $\mathbf{a}_1$, and some subsequence converging to some point outside the ball B. This allows us to choose points $\mathbf{p}_{k_j} (j \in \mathbb{N})$ of (P) all lying in B, such that the corresponding sequence of next points $\mathbf{p}_{k_j+1} (j \in \mathbb{N})$ all lie outside B.

Any accumulation point of $\mathbf{p}_{k_j} (j \in \mathbb{N})$ then lies in B, and hence, as an accumulation point of (P), must equal $\mathbf{a}_1$. This means that

$$\lim_{j \to \infty} \left\| \mathbf{a}_1 - \mathbf{p}_{k_j} \right\| = 0 \tag{16.4}$$

and for all $i \in N \backslash \{1\}$

$$\lim_{j \to \infty} \left\| \mathbf{a}_i - \mathbf{p}_{k_j} \right\| = \|\mathbf{a}_i - \mathbf{a}_1\| \tag{16.5}$$

Since all $\mathbf{p}_{k_j+1} \notin B$, we also have

$$\frac{\left\|\mathbf{a}_1 - \mathbf{p}_{k_j+1}\right\|}{\left\|\mathbf{a}_1 - \mathbf{p}_{k_j}\right\|} > \frac{\varepsilon}{\left\|\mathbf{a}_1 - \mathbf{p}_{k_j}\right\|}$$

and by taking limits, using equation (16.4), we obtain

$$\lim_{j\to\infty} \frac{\left\|\mathbf{a}_1 - \mathbf{p}_{k_j+1}\right\|}{\left\|\mathbf{a}_1 - \mathbf{p}_{k_j}\right\|} = +\infty. \tag{16.6}$$

On the other hand, we have $\mathbf{p}_{k_j+1} = T\left(\mathbf{p}_{k_j}\right) = \dfrac{\sum\limits_{i\in N} \frac{w_i}{\left\|\mathbf{a}_i - \mathbf{p}_{k_j}\right\|}\mathbf{a}_i}{\sum\limits_{i\in N} \frac{w_i}{\left\|\mathbf{a}_i - \mathbf{p}_{k_j}\right\|}}$, which may be rewritten as

$$\mathbf{a}_1 - \mathbf{p}_{k_j+1} = \frac{\sum\limits_{i\in N\setminus\{1\}} \frac{w_i}{\left\|\mathbf{a}_i - \mathbf{p}_{k_j}\right\|}(\mathbf{a}_1 - \mathbf{a}_i)}{\sum\limits_{i\in N} \frac{w_i}{\left\|\mathbf{a}_i - \mathbf{p}_{k_j}\right\|}}.$$

After division by $\left\|\mathbf{a}_1 - \mathbf{p}_{k_j}\right\|$ and taking the norm, we obtain

$$\frac{\left\|\mathbf{a}_1 - \mathbf{p}_{k_j+1}\right\|}{\left\|\mathbf{a}_1 - \mathbf{p}_{k_j}\right\|} = \frac{\left\|\sum\limits_{i\in N\setminus\{1\}} \frac{w_i}{\left\|\mathbf{a}_i - \mathbf{p}_{k_j}\right\|}(\mathbf{a}_1 - \mathbf{a}_i)\right\|}{w_1 + \left\|\mathbf{a}_1 - \mathbf{p}_{k_j}\right\| \sum\limits_{i\in N\setminus\{1\}} \frac{w_i}{\left\|\mathbf{a}_i - \mathbf{p}_{k_j}\right\|}}.$$

By (16.4) and (16.5) the limit of the right-hand side denominator equals w_1, so that

$$\lim_{j\to\infty} \frac{\left\|\mathbf{a}_1 - \mathbf{p}_{k_j+1}\right\|}{\left\|\mathbf{a}_1 - \mathbf{p}_{k_j}\right\|} = \frac{1}{w_1}\left\|\sum_{i\in N\setminus\{1\}} \frac{w_i}{\|(\mathbf{a}_i - \mathbf{a}_1)\|}(\mathbf{a}_1 - \mathbf{a}_i)\right\|, \tag{16.7}$$

which contradicts (16.6). □

Lemma 14: *If* $\mathbf{a}_k$ *is an accumulation point of* (P), *it satisfies condition* (16.2).

Proof: Without loss of generality, suppose that $\mathbf{a}_1$ is an accumulation point of (P). Therefore, we know by Lemma 13 that it is the only accumulation point of the sequence, meaning that (P) converges to $\mathbf{a}_1$. The reasoning of the previous lemma may then be repeated now for the whole sequence (P) yielding, similarly to (16.7),

$$\begin{aligned}\lim_{k\to\infty} \frac{\|\mathbf{a}_1 - \mathbf{p}_{k+1}\|}{\|\mathbf{a}_1 - \mathbf{p}_k\|} &= \frac{1}{w_1}\left\|\sum_{i\in N\setminus\{1\}} \frac{w_i}{\|\mathbf{a}_i - \mathbf{a}_1\|}(\mathbf{a}_1 - \mathbf{a}_i)\right\| \\ &= \frac{1}{w_1}\left\|\sum_{i\in N\setminus\{1\}} w_i u(\mathbf{a}_1 - \mathbf{a}_i)\right\| \end{aligned} \tag{16.8}$$

But since (P) converges to $\mathbf{a}_1$, we also have

$$\lim_{k\to\infty} \frac{\|\mathbf{a}_1 - \mathbf{p}_{k+1}\|}{\|\mathbf{a}_1 - \mathbf{p}_k\|} \leq 1,$$

and combined with (16.8) we obtain

$$\left\| \sum_{i\in N\setminus\{1\}} w_i u(\mathbf{a}_1 - \mathbf{a}_i) \right\| \leq w_1,$$

which is exactly condition (16.2). □

Lemma 15: *The sequence* (P) *is convergent to a limit point satisfying either condition* (16.1) *or condition* (16.2). *Moreover, this limit point is independent of the starting point* $\mathbf{p}_1$.

Proof: We know that (P) has an accumulation point (Lemma 8). In case one of the given $\mathbf{a}_i$ is an accumulation point, it is the only accumulation point (Lemma 13), and it satisfies condition (16.2) (Lemma 14). Otherwise there exists an accumulation point different from all $\mathbf{a}_i$, and then it satisfies condition (16.1) (Lemma 11), and is the only accumulation point of (P) (Lemma 12). It follows that (P) has a unique accumulation point $\mathbf{m}$, and thus, since it remains bounded, converges to $\mathbf{m}$, and this limit point satisfies either condition (16.1) or condition (16.2).

Consider then another sequence (P') constructed in the same way as (P) but starting from another point $\mathbf{p}'_1 \neq \mathbf{p}_1$. Then (P') also converges to some limit point $\mathbf{m}'$ satisfying either condition (16.1) or condition (16.2). If now $\mathbf{m}'$ would differ from $\mathbf{m}$, this would contradict one of the Lemmas 4, 5 or 6, showing that we must have $\mathbf{m}' = \mathbf{m}$. □

Lemma 16: *The limit point* $\mathbf{m}$ *of* (P) *is the sought minimum point*.

Proof: We will not reproduce Weiszfeld's reasoning here since it is incorrect; it makes use of sequences (P) with any starting point and applies previous results on it, thereby forgetting the implicit assumption that (P) never reaches any of the given points $\mathbf{a}_i$. However, as will be discussed in a direct way without any use of the results in this section, this does not always hold. In Sect. 16.5 below we show in a direct way that Conditions (16.1) and (16.2) imply optimality. Combined with previous Lemma 15, the current lemma will follow. □

16.5 Some Glimpses of the Modern View

A function $g : \mathbb{R}^d \to \mathbb{R}$ is called *convex* if and only if

$$g((1-t)\mathbf{x} + t\mathbf{y}) \leq (1-t)g(\mathbf{x}) + tg(\mathbf{y}) \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \quad \forall\, t \in [0,1]$$

and it is called *strictly convex* when this inequality is strict as soon as $\mathbf{x} \neq \mathbf{y}$ and $0 < t < 1$.

The point $\mathbf{z} = (1-t)\mathbf{x} + t\mathbf{y}$ for $t \in [0,1]$ runs over the line segment between $\mathbf{x}$ and $\mathbf{y}$, and the expression $(1-t)g(\mathbf{x}) + tg(\mathbf{y})$ is the linear interpolation of g at $\mathbf{z}$ from $\mathbf{x}$ and $\mathbf{y}$. Hence g being (strictly) convex means that linear interpolation always (strictly) overestimates g, except at the points from which one interpolates.

From this definition, it is easy to see that multiplying a convex function by some positive constant or summing several convex functions always yields another convex function. This is also true of any "translate" defined by $g_b(\mathbf{x}) = g(\mathbf{b} - \mathbf{x})$ for some fixed $\mathbf{b}$.

We also have for any $\mathbf{x}$, $\mathbf{y}$ and $0 \leq t \leq 1$, using the Cauchy-Schwartz-Buniakowsky inequality,

$$\begin{aligned}
\|(1-t)\mathbf{x} + t\mathbf{y}\|^2 &= \langle (1-t)\mathbf{x} + t\mathbf{y};\ (1-t)\mathbf{x} + t\mathbf{y} \rangle \\
&= (1-t)^2 \langle \mathbf{x}; \mathbf{x} \rangle + t^2 \langle \mathbf{y}; \mathbf{y} \rangle + 2(1-t)t \langle \mathbf{x}; \mathbf{y} \rangle \\
&\leq (1-t)^2 \|\mathbf{x}\|^2 + t^2 \|\mathbf{y}\|^2 + 2(1-t)t \|\mathbf{x}\| \|\mathbf{y}\| \\
&= \left[(1-t) \|\mathbf{x}\| + t \|\mathbf{y}\| \right]^2,
\end{aligned}$$

which yields, taking square roots, $\|(1-t)\mathbf{x} + t\mathbf{y}\| \leq (1-t)\|\mathbf{x}\| + t\|\mathbf{y}\|$, proving that the "norm" $\|\bullet\|$ is a convex function.

Therefore the Weber-function $f : \mathbb{R}^d \to \mathbb{R} : \mathbf{x} \to \sum_{i \in N} w_i \|\mathrm{a}_i - \mathbf{x}\|$ is a *convex* function.

The norm-inequality above is an equality only when the Cauchy-Schwartz-Buniakowsky inequality is an equality, that is, when $t\mathbf{y}$ is a positive multiple of $(1-t)\mathbf{x}$. Whenever 0, $\mathbf{x}$ and $\mathbf{y}$ are not aligned and $t \neq 0, 1$, we always have a strict inequality.

From this last observation one may derive that the Weber-function is strictly convex as soon as the given points $\mathbf{a}_i$ are not colinear, an assumption we will continue to make following Weiszfeld.

Convex functions are fundamental objects in nonlinear and nondifferentiable optimization. They have been studied in depth and turn out to have many extremely important properties too numerous to discuss here. For reference, see Hirriart-Urruty and Lemaréchal (2001). Some of these properties of particular interest are as follows, discussed here without all the necessary details and proofs (except if very short).

Property 1: First of all, convex functions are always continuous, but not necessarily differentiable. The Euclidean norm function, in particular, is differentiable everywhere, except at the origin. This implies that the Weber function is nondifferentiable only in the given points $\mathbf{a}_i$.

Property 2: However, convex functions have directional derivatives at all points and in any direction. The directional derivative of g at a point $\mathbf{z} \in \mathbb{R}^d$ in direction $d \neq 0$ is defined as:

$$g'(\mathbf{z}, d) \stackrel{def}{=} \lim_{t \downarrow 0} \frac{g(\mathbf{z} + td) - g(\mathbf{z})}{t}.$$

If the function g is differentiable at point $\mathbf{z}$, this means that

$$g'(\mathbf{z}, d) = \langle \nabla g(\mathbf{z}); d \rangle$$

is a linear function of the direction d, defined by the gradient $\nabla g(\mathbf{z})$ of g at $\mathbf{z}$.

The Euclidean distance function $g_{\mathbf{a}}(\mathbf{x}) = \|\mathbf{a} - \mathbf{x}\|$ to point $\mathbf{a}$ is differentiable at any point $\mathbf{z} \neq \mathbf{a}$, with gradient $\nabla g_{\mathbf{a}}(z) = \dfrac{\mathbf{z} - \mathbf{a}}{\|\mathbf{a} - \mathbf{z}\|}$, i.e., the unit vector $u(\mathbf{z}-\mathbf{a})$ at $\mathbf{a}$ towards $\mathbf{z}$. At the point of nondifferentiability $\mathbf{z} = \mathbf{a}$, however, we find

$$g'_{\mathbf{a}}(\mathbf{a}, d) = \lim_{t \downarrow 0} \frac{g_{\mathbf{a}}(\mathbf{a} + td) - g_{\mathbf{a}}(\mathbf{a})}{t} \tag{16.9}$$

$$= \lim_{t \downarrow 0} \frac{\|td\| - 0}{t} \tag{16.10}$$

$$= \|d\| . \tag{16.11}$$

Directional derivatives behave somewhat similarly to usual derivatives: in particular, at a fixed point $\mathbf{z}$ and for fixed direction d, they sum nicely and scale positively. For convex functions g_1 and g_2 and $h = w_1 g_1 + w_2 g_2$ (w, $w' > 0$) we always have $h'(\mathbf{z}, d) = w_1 g'_1(\mathbf{z}, d) + w_2 g'_2(\mathbf{z}, d)$.

Therefore, for the Weber function we obtain the following directional derivatives:

If $\mathbf{z}$ is different from all $\mathbf{a}_i$, then

$$f'(\mathbf{z}, d) = \left\langle \sum_{i \in N} w_i u(\mathbf{a}_i - \mathbf{z}); d \right\rangle \tag{16.12}$$

If $\mathbf{z} = \mathbf{a}_k$, then

$$f'(\mathbf{z}, d) = \left\langle \sum_{i \in N \setminus \{k\}} w_i u(\mathbf{a}_i - \mathbf{a}_k); d \right\rangle + w_k \|d\| \tag{16.13}$$

Property 3: For a convex function g a point $\mathbf{z}$ is a local minimum of g if and only if its directional derivative at $\mathbf{z}$ in any direction d is nonnegative. In other words, $g'(\mathbf{z}, d) \geq 0$ for all $d \neq 0$.

In order to apply this property to the Weber function f, we must again consider the two cases:

Case 1: If $\mathbf{z}$ is different from all $\mathbf{a}_i$ this local optimality condition becomes $\left\langle \sum_{i \in N} w_i u(\mathbf{a}_i - \mathbf{z}); d \right\rangle \geq 0 \; \forall \; d \neq 0$. However, in case $\left\langle \sum_{i \in N} w_i u(\mathbf{a}_i - \mathbf{z}); d \right\rangle > 0$, for the opposite direction $-d$ we would have $\left\langle \sum_{i \in N} w_i u(\mathbf{a}_i - \mathbf{z}); -d \right\rangle$ $= -\left\langle \sum_{i \in N} w_i u(\mathbf{a}_i - \mathbf{z}); d \right\rangle < 0$, contradicting the local optimality condition. There-

fore, it follows that $\left\langle \sum_{i\in N} w_i u(\mathbf{a}_i - \mathbf{z}); d \right\rangle = 0$ for all $d \neq 0$, which can only happen if $\sum_{i\in N} w_i u(\mathbf{a}_i - \mathbf{z}) = 0$, i.e., when condition (1) is satisfied.

Case 2: If $\mathbf{z} = \mathbf{a}_k$, we must have for all $d \neq 0$ that $\langle D_f; d\rangle + w_k \|d\| \geq 0$ where $D_f = \sum_{i\in N\setminus\{k\}} w_i u(\mathbf{a}_i - \mathbf{a}_k)$. This may be rewritten as $w_k \geq \langle D_f; \frac{-d}{\|d\|}\rangle$, and the particular choice of $d = -D_f$, using $\|-d\| = \|d\|$, leads to $w_k \geq \left\langle D_f; \frac{D_f}{\|-D_f\|} \right\rangle = \|D_f\|$. In this case, for any other d, using one part of the Cauchy-Schwartz-Buniakowsky inequality allows us to write the expression $\langle D_f; d\rangle + w_k\|d\| \geq -\|D_f\|.\|d\| + \|D_f\|.\|d\| = 0$. Replacing D_f by its definition, we just proved that the local optimality condition at $\mathbf{z} = \mathbf{a}_k$ is given by

$$\left\| \sum_{i\in N\setminus\{k\}} w_i u(\mathbf{a}_i - \mathbf{a}_k) \right\| \leq w_k,$$

which is exactly Condition (16.2).

Property 4: Any local minimum of a convex function g is also a global minimum. To prove this, assume $\mathbf{x}$ is a local minimum of g, and consider any point $\mathbf{y}$ for which we have to prove that $g(\mathbf{y}) \geq g(\mathbf{x})$. We may choose some $\mathbf{z} = (1-t)\mathbf{x}+t\mathbf{y})$ with $t > 0$ on the line segment connecting $\mathbf{x}$ with $\mathbf{y}$, close enough to $\mathbf{x}$ such that $g(\mathbf{z}) \geq g(\mathbf{x})$. By convexity of g we then have $g(\mathbf{x}) \leq g(\mathbf{z}) \leq (1-t)\mathrm{g}(\mathbf{x})+t\mathrm{g}(\mathbf{y})$, implying that $g(\mathbf{y})-g(\mathbf{x}) \geq 0$. □

This last statement, together with previous point, proves Theorem 1.

Property 5: A strictly convex function admits at most one minimal solution.

Indeed, any solution on the open line segment connecting two different minimal solutions would, by strict convexity, yield a strictly lower objective value, which is a contradiction. Together with previous statement, this confirms the uniqueness results of Weiszfeld for Lemmas 4 and 6.

A final property is more technical, but very useful, as will be seen in next sections:

Property 6: If the function $g : \mathbb{R}^d \to \mathbb{R}$ is convex, it admits subgradients at any point.

A vector $\mathbf{q} \in \mathbb{R}^d$ is called a *subgradient* of g at the point $\mathbf{c} \in \mathbb{R}^d$ if and only if

$$\forall\, \mathbf{x} \in \mathbb{R}^d : g(\mathbf{c}) + \langle \mathbf{x} - \mathbf{c}; q\rangle \leq g(\mathbf{x}) \tag{16.14}$$

This generalizes the more classical property that if the convex function $g : \mathbb{R}^d \to \mathbb{R}$ is differentiable at the point $\mathbf{c} \in \mathbb{R}^d$, then we have

$$\forall\, \mathbf{x} \in \mathbb{R}^d : g(\mathbf{c}) + \langle \mathbf{x} - \mathbf{c}; \nabla g(\mathbf{c})\rangle \leq g(\mathbf{x}) \tag{16.15}$$

or, in other words, the gradient $\nabla g(\mathbf{c})$ is a subgradient of g at $\mathbf{c}$. In fact, it is the only subgradient of g at $\mathbf{c}$, and the fact that it is unique guarantees differentiability.

In order to apply this at any point of the convex Weber function f we need also to know a good subgradient of f at the points $\mathbf{c} = \mathbf{a}_k$ where it is not differentiable. It may be shown that the choice

$$q^k = \left(1 - \frac{w_k}{\|D_f\|}\right) D_f \tag{16.16}$$

is such a subgradient when $\mathbf{a}_k \in N$ is not an optimal solution as it does not satisfy Condition (16.2), which is equivalent to saying that the coefficient $1 - \frac{w_k}{\|D_f\|} \geq 0$. When $\mathbf{a}_k$ is the optimal solution, however, the choice $q = 0$ is a subgradient of f at $\mathbf{a}_k$.

16.6 Additions and Improvements

16.6.1 Fixed Point Iterate

The Weiszfeld algorithm has been rediscovered independently several times, e.g. by Cooper (1963) and Kuhn and Kuenne (1962). It was Kuhn (1973) who provided the first independent proof of convergence, and reported (in some notes added in proof) the rediscovery of Weiszfeld's early work. Kuhn did observe Weiszfeld's omission to consider the case where some iterate of the constructed sequence falls exactly on one of the given fixed points $\mathbf{a}_i$; but he did not solve this question. He simply considered it "a very unlikely event," and "corrected" Weiszfeld by stating "for all but a denumerable number of starting points, the sequence converges."

However, many years later, Chandrasekaran and Tamir (1989), and Cánovas et al. (2002) constructed examples in which the set of starting points $\mathbf{p}$, such that $T(\mathbf{p})$ equals one of the $\mathbf{a}_i$, is a continuum, contradicting the "denumerable number" of Kuhn. In all these examples, though, the given points $\mathbf{a}_i$ all lie in some lower dimensional subspace. Brimberg (2003) finally succeeded in proving the following conjecture made by Chandrasekaran and Tamir (1989): when the convex hull of the fixed points $\mathbf{a}_i$ has full dimension, the set of "bad" starting points, from which Weiszfeld's algorithm has a "fixed point iterate," is indeed always denumerable.

The occurrence of such a *fixed point iterate* may be considered quite exceptional. However, mathematicians usually do not like exceptions and want an answer for any possible situation, however unlikely. Similarly, a good programmer will try to make certain that no "bugs" will arise in programs, so should envisage all possible situations and offer a way out in each of them.

Suppose our sequence arrives exactly at (or starts from) some $\mathbf{a}_k$. We have learned how to check the optimality of this point: use the optimality Condition (16.2). What is needed is to specify some new iterate $\overline{T}(\mathbf{a}_k) \neq \mathbf{a}_k$ in case this

condition is not met. Preferably, this new iterate should be better than $\mathbf{a}_k$ in the sense that $f(\overline{T}(\mathbf{a}_k)) < f(\mathbf{a}_k)$. As we have seen in the previous section that the directional derivative of f in the direction $-D_f$ (the negative gradient at $\mathbf{a}_k$ of the differentiable part of f) is then negative, this means that points with lower value are to be found in this direction. Therefore for sufficiently small λ, points of the form $\mathbf{c}_\lambda = \mathbf{a}_k - \lambda D_f$ will satisfy $f(\mathbf{c}_\lambda) < f(\mathbf{a}_k)$. Indeed, all proposals in the literature do so, but differ in the choice of λ. Chronologically, Balas and Yu (1982) propose to find an adequate λ by a binary search, which might take a few steps. Vardi and Zhang (2001) show that a good choice is

$$\lambda = \frac{\left(1 - \dfrac{w_k}{\|D_f\|}\right)}{\displaystyle\sum_{i \in N \setminus \{k\}} \frac{w_i}{\|\mathbf{a}_k - \mathbf{a}_i\|}}.$$

Rautenbach et al. (2004) prove that another adequate choice is given by

$$\min \left\{ \min_{i \in N \setminus \{k\}} \tfrac{1}{2} \|\mathbf{a}_k - \mathbf{a}_i\|; \frac{\|D_f\| - w_k}{\displaystyle\sum_{\ell \in N \setminus \{k\}} \frac{4w_\ell}{\|\mathbf{a}_k - \mathbf{a}_\ell\|}} \right\}.$$

Yet another more involved stepsize that may be used for minimization of more general functions was already developed by Illgen (1979).

16.6.2 The Starting Point

A major algorithmic question is how to choose a good starting point. Two main ideas may guide this choice. First, one will probably want to make calculations short, starting from some point already close to the sought optimum point. Second, one might want to avoid the fixed point iterate situation discussed above.

The first concern is not easy to answer. The most common suggestion seems to be to start with the gravity center. This idea was probably prompted by an attempt to remedy the standard error discussed in Sect. 16.3.3. However, the only (unconvincing) reason for this choice might be that the center-of-gravity and the sought optimum point both lie "between" the given points or, more precisely, in their convex hull. This convex hull is the smallest convex set that contains them all, and consists of all convex combinations of the $\mathbf{a}_i$, i.e., points of the form

$$\sum_{i \in N} \lambda_i \mathbf{a}_i,$$

where the λ_i are nonnegative numbers with $\sum_{i\in N} \lambda_i = 1$. In other words, this means being a gravity center of the $\mathbf{a}_i$ for some choice of weights. For the choice $\lambda_i = \left(w_i / \sum_{i\in N} w_i\right)$ the convex combination yields the gravity center of the $\mathbf{a}_i$, weighted by w_i. For any $\mathbf{p} \in \mathbb{R}^d$ (where $p \notin N$), the choice

$$\lambda_i = \frac{\dfrac{w_i}{\|\mathbf{a}_i - \mathbf{p}\|}}{s(\mathbf{p})},$$

where $s(\mathbf{p}) = \sum_{j\in N} w_j / \|\mathbf{a}_j - \mathbf{p}\|$ gives as convex combination the Weiszfeld iterate $T(\mathbf{p})$. On the one hand, this shows that $T(\mathbf{p})$ always lies in the convex hull. On the other hand, if $\mathbf{p}$ is the sought optimum point $\mathbf{m}$, we have $\mathbf{p} = T(\mathbf{p})$ by the results of previous section, so this optimum also lies in the convex hull. Note that this also holds if the optimum point is a fixed point such as, $\mathbf{m} = \mathbf{a}_k$; in that case it clearly also lies in the convex hull, which may also be seen by the choice of $\lambda_i = 0$ ($i \in N\backslash\{k\}$, and $\lambda_k = 1$.

Since from any starting point $\mathbf{p}$ a single Weiszfeld iteration brings us immediately to $T(\mathbf{p})$ in the convex hull, the reasons to prefer starting at the gravity center are rather thin. Simple experiments show that usually the iterations converge quite quickly from any starting point. But it has been observed that in some cases, in particular when for some given point $\mathbf{a}_k$ the fixed point optimality inequality (16.2) is almost an equality (either way), the method may become extremely slow, see, e.g., the example by Drezner et al. (2002). This was confirmed by the theoretical analysis of the convergence rate by Katz (1974).

By all means it seems therefore a good idea to start by checking each of the given points $\mathbf{a}_i$, $i \in N$ for optimality. Thanks to the optimality Condition (16.2) this is easy to do. In case the condition holds at some $\mathbf{a}_k$, we are immediately done with the whole optimization process because $\mathbf{m} = \mathbf{a}_k$. It is only when unlucky at all $\mathbf{a}_i$, that we have to go through the cumbersome iterative procedure from some starting point. What has been gained, however, is the certainty that the optimal solution will not be one of the given points. In addition to taking care of calculating each value $f(\mathbf{a}_i)$, one also determines the lowest of these values, say $f(\mathbf{a}_k)$, and one knows for sure that the optimal value $f(\mathbf{m}) < f(\mathbf{a}_k)$. In that case, a good idea is to start the iterations from this point $\mathbf{a}_k$, using one of the proposals for $\overline{T}(\mathbf{a}_k)$ discussed in previous Sect. 16.6.1. Since we know that $f(\overline{T}(\mathbf{a}_k)) < f(\mathbf{a}_k) < f(\mathbf{a}_i)$, $i \in N\backslash\{k\}$, this has the advantage that none of the further iterations will be (or even come close to) a fixed point.

This therefore yields an answer to the second concern.

Note 3: It may be argued that checking all $\mathbf{a}_i$, $i \in N$ is a lot of work, particularly for large sets N. Since each evaluation of $f(\mathbf{c})$ or optimality check at some point $\mathbf{c}$ calls for calculation of $|N|$ terms, the total work will be of order $O(|N|^2)$. It is usually possible to avoid a good part of this work by using the subgradients mentioned in Sect. 16.5. Checking $\mathbf{a}_k$ involves calculating the corresponding vector D_f and its

norm, which, in case $\mathbf{a}_k$ turns out not to be optimal, also yields immediately the subgradient q^k defined by (16.16) in Sect. 16.5. We may therefore write

$$\forall\, \mathbf{x} \in \mathbb{R}^d\colon\ f(\mathbf{a}_k) + \left\langle \mathbf{x} - \mathbf{a}_k; q^k \right\rangle \leq f(\mathbf{x}).$$

It follows that any $\mathbf{a}_i \in \mathbb{R}^d$ for which $\langle \mathbf{a}_i - \mathbf{a}_k; q^k \rangle > 0$ will satisfy $f(\mathbf{a}_k) < f(\mathbf{a}_i)$ (substitute $\mathbf{x} = \mathbf{a}_i$ in the previous inequality), so does not need to be checked anymore for optimality.

Another even simpler algorithmic strategy was suggested by Ostresh (1978a): start from any point with the Weiszfeld algorithm, and at each step check for optimality the given point $\mathbf{a}_i$ closest to the current iterate unless already checked earlier. Almost no additional work is involved in determining this closest given point, because all distances to the $\mathbf{a}_i$ have to be calculated anyway for finding the next iterate.

16.6.3 The Stepsize

The standard optimality condition for having a minimum of the convex function f at a point $\mathbf{p}$ is $\nabla f(\mathbf{p}) = 0$. Now

$$\nabla f(\mathbf{p}) = \sum_{i \in N} w_i \frac{\mathbf{p} - \mathbf{a}_i}{\|\mathbf{p} - \mathbf{a}_i\|} = s(\mathbf{p})(\mathbf{p} - T(\mathbf{p}))$$

where, as before, $s(\mathbf{p}) = \sum_{j \in N} \frac{w_j}{\|\mathbf{a}_j - \mathbf{p}\|}$. On the one hand, this shows (again) that the optimal solution $\mathbf{m}$ must satisfy $T(\mathbf{m}) = \mathbf{m}$. This observation has been the motivation for Cooper's (1963) and Kuhn's (1967) (re-)discovery of the iterative scheme $\mathbf{p}_{n+1} = T(\mathbf{p}_n)$ for finding $\mathbf{m}$, which was found in a different way by Weiszfeld.

On the other hand, we also obtain the equality

$$T(\mathbf{p}) = \mathbf{p} - \frac{1}{s(\mathbf{p})} \nabla f(\mathbf{p}),$$

which shows that the Weiszfeld method belongs in fact to the general class of gradient descent methods: at each step $n = 0, 1, \ldots$, a move is made from the current point $\mathbf{p}_n$ in the direction of steepest descent (i.e., the negative gradient) of the objective function at that point; in formula $\mathbf{p}_{n+1} = \mathbf{p}_n - \sigma_k \nabla f(\mathbf{p}_n)$. In general optimization problems, one has to search for a stepsize σ_n that will make the objective value decrease such that $f(\mathbf{p}_{n+1}) < f(\mathbf{p}_n)$. What is remarkable in Weiszfeld's method is that such a stepsize may be directly calculated as $\sigma_n = \frac{1}{s(\mathbf{p}_n)}$.

Ostresh (1978a) has even shown that, in this gradient method, one may in fact choose independently at each step any stepsize

$$\frac{1}{s(\mathbf{p}_n)} \leq \sigma_n \leq \frac{2}{s(\mathbf{p}_n)}$$

and still obtain a descent method converging to the optimal solution. This property yields yet another way to avoid fixed point iterates, which was the main goal behind Ostresh's study: in case the Weiszfeld method (or a similar one with other stepsizes) would fall exactly on some non-optimal given point $\mathbf{a}_k$, simply repeat the previous step using a slightly modified stepsize within the allowed bounds, thereby avoiding $\mathbf{a}_k$.

16.6.4 Stopping Rules

The convergence property of an iterative algorithm says that the infinite generated sequence $(\mathbf{p}_n)_{n\in N}$ will have a limit point, and that this limit is the sought solution $\mathbf{m}$. A solution may therefore be reached that is arbitrarily close to $\mathbf{m}$, provided the calculations continue up to sufficiently large n. Unfortunately, the convergence property does not indicate how large the value of n must be in order to be "sufficiently large" to satisfy the stopping criterion.

There have been many different viewpoints regarding an adequate stopping rule. The simplest and therefore most popular rule is to fix in advance the number of steps. This is a totally blind rule and does not yield any clear indication of how good the attained approximate solution is, particularly in view of the widely different convergence rates that may occur, see Katz (1974).

In the beginning of the computer age, another concern was to limit the calculation time. This is in a sense even more blind than limiting the number of steps, since each step depends (linearly) on the size N of the data. With modern day computers this kind of rule has fortunately disappeared, except in the area of metaheuristics calling for very intensive calculations.

This means we can now focus on obtaining "rational" stopping rules, which yield some quality stamp on the accuracy of the solution obtained. But what is accuracy?

- A first type of accuracy hinted at above, is proximity to the optimal solution $\mathbf{m}$. This is, however, very difficult to measure directly, since $\mathbf{m}$ is unknown. The only attempts in this direction seem to have been to approximate the full level set of f at the same level $f(\mathbf{p}_n)$ that is reached at the current iterate $\mathbf{p}_n$, in other `words by the set $L_n = \{\mathbf{x} \in \mathbb{R}^d \,|\, f(\mathbf{x}) \leq f(p_n)\}$. These attempts occur not in the context of the Weiszfeld algorithm, but in Plastria (1987) using a cutting plane method, or in Plastria (1992a) using the "Big Square Small Square" global optimization method.
- A second type of accuracy, which is much easier to calculate, is related to the fact that the sought point $\mathbf{m}$ is defined by $\nabla f(\mathbf{m}) = 0$. One can test how close to this equality one is at the current iterate $\mathbf{p}_n$ by calculating $\|\nabla f(\mathbf{p}_n)\|$. One then

decides to stop whenever this is "sufficiently" small, meaning in practice when $\|\nabla f(\mathbf{p}_n)\| \leq \varepsilon$, for some small chosen $\varepsilon > 0$.

- A third accuracy measure is in terms of values: how closely does the current objective value $f(\mathbf{p}_n)$ approximate the optimal value $f(\mathbf{m})$? And although $f(\mathbf{m})$ is unknown, several procedures have been developed to calculate a good lower bound $\ell_n \leq f(\mathbf{m})$ at each step n, which converges increasingly to $f(\mathbf{m})$ from below when n increases. When such a lower bound ℓ_n is known, the accuracy in value of the current iterate may be overestimated, either in absolute terms by the difference, $f(\mathbf{p}_n) - \ell_n$, or in relative terms (or percentage wise) by the ratio $\frac{f(\mathbf{p}_n)-\ell_n}{f(\mathbf{p}_n)}$. The calculations are then stopped if these have become "sufficiently small." Clearly the effectiveness of this last stopping rule will depend in part on how quickly this bound comes close to $f(\mathbf{m})$, but also on the effort spent in calculating it.

Bounds have been developed by Love and Yeong (1981), Juel (1984), Drezner (1984), Wendell and Peterson (1984), and Love and Dowling (1989). Here we only derive Juel's bound, which is easy using the tools developed above.

The (sub)gradient inequality (16.15) at any iterate $c = \mathbf{p}_n \notin N$, applied at the point $\mathbf{x} = \mathbf{m} \in N$ results in

$$f(\mathbf{p}_n) - \langle \mathbf{p}_n; \nabla f(\mathbf{p}_n)\rangle + \langle \mathbf{m}; \nabla f(\mathbf{p}_n)\rangle = f(\mathbf{p}_n) + \langle \mathbf{m} - \mathbf{p}_n; \nabla f(\mathbf{p}_n)\rangle \leq f(\mathbf{m}).$$

We also know that $\mathbf{m}$ is a convex combination of the given points $\mathbf{a}_i$, from which one easily derives that for any vector q we have $\min_{i \in N} \langle \mathbf{a}_i; q\rangle \leq \langle \mathbf{m}; q\rangle$.

Plugging this property for $q = \nabla f(\mathbf{p}_n)$ into the inequality above we obtain Juel's lower bound

$$\ell_n^J = f(\mathbf{p}_n) - \langle \mathbf{p}_n; \nabla f(\mathbf{p}_n)\rangle + \min_{i \in N} \langle \mathbf{a}_i; \nabla f(\mathbf{p}_n)\rangle \leq f(\mathbf{m})$$

This bound was shown by Elzinga and Hearn (1983) to be always better than the original bound of Love and Yeong. The other bounds are more involved, and a comparison between them was conducted by Dowling and Love (1986, 1987).

16.6.5 Acceleration Attempts

Several attempts have been made to try to accelerate Weiszfeld's method, particularly in view of cases of slow convergence. In most cases, general types of acceleration techniques from nonlinear programming were proposed, like Steffensen's scheme by Katz (1974), the Armijo rule by Cooper and Katz (1981) or a Newton-Raphson modification by Ostresh (1978b); these are tested experimentally, but without formal proof of acceleration. Ostresh's (1978a) hope that his extended step-

sizes discussed in Sect. 16.6.3 would enable "proving the descent property of yet to be discovered, very fast, nongradient algorithms" has not yet been fulfilled. More recent work in this context is found in Drezner (1995), Li (1998), and Brimberg et al. (1998).

Weber's problem has been, and still is, considered a simple and educational example of a nonlinear and nondifferentiable optimization problem, and has therefore often been used as a kind of benchmark for testing new ideas in nonlinear optimization. One may safely state that virtually any proposed nonlinear optimization algorithm has been at some point tried out on the Weber problem. A list of these is given by Wesolowsky (1993), to which we may add a cutting plane method by Plastria (1987), a quadratically convergent method by Overton (1983), several interior point methods, such as that by Wu (1994), and a recent Newton bracketing method by Levin and Ben-Israel (2002).

16.7 Extensions to Other Problems

Weiszfeld's method has been adapted to an enormous number of variants of the classical Weber problem. Here we will list a number of these applications, without attempt at exhaustiveness. As far as possible we indicate also if proofs of convergence are available.

As many of these extended problems do not have a convex objective function, local, non-global, optima may occur. In such circumstances convergence of a method should be understood as convergence to some locally optimal solution.

The adaptation of Weiszfeld's method always follows the reasoning of Cooper and Kuhn, suggested in Sect. 16.6.3: consider the optimality condition $\nabla f(\mathbf{p}) = 0$, rewrite it in some way as a fixed point equation $\mathbf{p} = T(\mathbf{p})$ where the particular form of T will depend on the function f, and then try to "solve" this by an iterative scheme $\mathbf{p}_{n+1} = T(\mathbf{p}_n)$.

16.7.1 Modified Transport Cost Functions

The most common modification of the Weber problem is to consider a more general objective function

$$f(\mathbf{x}) = \sum_{i \in N} g_i(\mathbf{a}_i - \mathbf{x}),$$

where the g_i are nondecreasing functions defined on the positive real numbers. The traditional Weber problem is obtained by the choice of linear functions $g_i(t) = w_i t$.

A first modification of this type was made by Cooper (1968), choosing $g_i(t) = w_i t^K$ for some fixed $K > 0$. He proposed Weiszfeld's scheme, but without conver-

gence proof. It was Katz (1969) who introduced the general model above, and proved convergence of a Weiszfeld scheme under some technical conditions on the g_i.

When the functions g_i are convex, such as Cooper's proposal with $K \geq 1$, the resulting f is still convex and we may hope for convergence to a globally optimal solution. When the functions g_i are nonconvex this cannot be ensured; usually one will have many local optima, in particular often at the given points $\mathbf{a}_i$.

Several authors contributed to solving such models, among whom Drezner (2009) offers one of the most recent and encompassing results. For a much larger related class of models in the much more general setting of Banach spaces, Eckhardt (1980) proved convergence for a Weiszfeld scheme.

16.7.2 General ℓ_p Distances

Many authors have studied the Weber problem where for a given $\mathbf{x} = (\mathbf{x}_k)_{k=1,\ldots,d} \in \mathbb{R}^d$ the Euclidean norm $\|\mathbf{x}\|$ is replaced by the more general ℓ_p norm given by the formula

$$\ell_p(\mathbf{x}) = \left(\sum_{k=1}^{d} |\mathbf{x}_k|^p \right)^{\frac{1}{p}}$$

This also yields a convex function f of $\mathbf{x}$ for any $\mathbf{p} \geq 1$, which, for $\mathbf{p} > 1$, is differentiable everywhere, except at the points $\mathbf{a}_i$, $i \in N$. For $\mathbf{p} = 1$ one obtains the well-known "rectilinear," "rectangular," "Manhattan," or "taxi" distance, which leads to Weber problems that may be studied and solved directly, as developed for example in the excellent book by Francis et al. (1992). Therefore, in what follows we assume $\mathbf{p} > 1$.

When Weiszfeld's algorithm is applied directly to the ℓ_p distance Weber problem, the descent property and hence convergence is found only for $\mathbf{p} \leq 2$, see Brimberg and Love (1993). In order to avoid the non-differentiabilities of f, the ℓ_p distance is often replaced, following Morris and Verdini (1979) by its "hyperbolic approximation"

$$\ell_p^{\varepsilon}(\mathbf{x}) = \left(\sum_{k=1}^{d} \left(x_k^2 + \varepsilon^2 \right)^{\frac{p}{2}} \right)^{\frac{1}{p}}$$

The developments may be found in the book by Love et al. (1988), providing many pertinent references. The most general convergence result of a Weiszfeld scheme for a model combining hyperbolic approximation of ℓ_p distances with nonlinear cost functions extending those described in Sect. 16.7.1 has been proven by Frenk et al. (1994). Convergence is found, however, only when $1 \leq \mathbf{p} \leq 2$.

16.7.3 Other Single Facility Location Problems

The Weiszfeld scheme converges very well when adapted to several other distance measures. Examples include Chen's (1991) method for locations on an inclined plane and Cera and Ortega's (2002) model for locating a hunter of fleeing prey. Both examples involve the use of some asymmetric ellipsoidal distance. Distance measures of this nature are studied in general by Plastria (1993). Katz and Vogl (2010) redefine Weiszfeld's method for problems, in which distance to each given point is an individually rescaled version of Euclidean distance, and extend Katz's (1974) convergence proof and convergence rate analysis,

Other examples of good convergence of Weiszfeld-type techniques involve location models on the sphere that use geodesic distances; see, e.g., Drezner and Wesolowsky (1978) or Katz and Cooper (1980). Weiszfeld's scheme has also successfully been adapted to situations with area demand, in which the objective function involves an integral instead of a sum, see Drezner and Wesolowsky (1980) and Chen (2001).

Further variants include Weber problems with possible negative weights (Drezner and Wesolowsky 1991), Weber problems taking queuing into account (Drezner et al. 1990), Weber problems within buildings (Arriola et al. 2005), competitive location models (Drezner and Drezner 2004), models that include price decisions (Fernández et al. (2007), and single facility location-allocation problems (Plastria and Elosmani 2008). In most of these applications, only convergence to a local optimum may be expected.

16.7.4 Multifacility Location Problems

Simultaneous location of several facilities $f \in F$ leads to more difficult questions. When the exact facilities-demand points and inter-facilities interactions are given as weighted distances with respective weights w_{if} and w_{fg} for demand point i and facilities f and g, we obtain the so-called Multifacility location problems of type

$$\min \sum_{f\in F}\sum_{i\in N} w_{if}\left\|\mathbf{a}_i - \mathbf{x}_f\right\| + \sum_{f\in F}\sum_{g\in F} w_{fg}\left\|\mathbf{x}_g - \mathbf{x}_f\right\|,$$

for which Weiszfeld's method has been adapted by many authors. After several proposals, among them one of the pioneers of the method (Miehle 1958), the first proven convergent method was given by Rado (1988), and later improved by Rosen and Xue (1992). A full general description of the optimality conditions was obtained by Plastria (1992b), and is valid for any type of norm. The solution algorithm was extended to hyperbolic approximation of ℓ_p distances by Rosen and Xue (1993).

16.7.5 Location-Allocation Problems

When the model includes the allocation decision of which facility will serve which given point, we obtain multiple facility problems usually called location-allocation problems, or multi-Weber problems. For each fixed allocation, the corresponding location question may be split into several single facility Weber problems (one for each facility) that may be solved by a Weiszfeld scheme. But the major difficulty here is the search for an adequate allocation, and no tractable method is known which guarantees to find a globally optimal solution, except when the number of facilities to locate is as small as two (Ostresh 1975).

Eilon and Watson-Gandy (1971) proposed to use a pure Weiszfeld scheme, as though the allocation is fixed, but adapting this allocation at each step. However, to the best of our knowledge, no convergence proof of such a scheme has been published. Therefore, in general, heuristic solution methods are proposed, usually with built-in Weiszfeld schemes, for solving single facility subproblems. The most popular proposal, usually called *Alternate*, was made by Cooper (1963). It consists of two alternating steps, the *allocation step* that finds the best allocation for fixed locations of the facilities, and the *location step* that finds the best locations for the facilities for a fixed allocation. These two steps are performed in an alternating sequence, until no further changes occur. For an enhancement of this method, see Rosing and Harris (1992).

Nowadays, heuristic methods have evolved and are becoming a field of enquiry unto themselves, as in Hoos and Stützle (2005) and Gendreau and Potvin (2008). One may state that almost any metaheuristic idea has been applied to the location-allocation problems with various success. Usually these involve either Weiszfeld steps or the more elaborate *Alternate* procedure as "local search" substeps. An overview of these methods, together with a comparison of some of them, is provided by Brimberg et al. (2000). One of the most successful approaches in this particular setting is of variable neighborhood search type, and is described by Brimberg et al. (2006).

16.7.6 Outside the Facility Location Field

Weiszfeld's method has recently known applications to problems quite different from location questions. Examples are those by Shi et al. (2007) for an application in signal processing, and Valkonen (2006) for an application in functional analysis.

16.8 Outlook

It is quite extraordinary how a proof developed some eighty years ago for attacking a centuries old and allegedly purely theoretical question has grown, thanks to its simplicity, into a quite efficient computer solution method for a problem that is

considered nowadays as one of the quintessences of locational modeling coming in many guises to which the methodology adapts. Such phenomena are rather rare in science, and probably mainly limited to the field of mathematics and its applications, corroborating the doubts about the validity of some current quality evaluation methods for scientific research, e.g., impact factors that are based on very short term periods of time, often no longer than a year.

We hope that this exposition and overview of extensions will help interest in it to continue to grow with further developments and extensions to many variants of the basic Weber problem. Presumably several fields of application of the model beyond spatial economics remain unexplored and it may be expected that the first steps of the Weiszfeld method outside the facility location field as discussed in Sect. 16.7.6 will know further discovery.

References

Arriola R, Laporte G, Ortega FA (2005) The Weber problem in a multi-storey building. INFOR 43(3):157–169.

Bajaj C (1988) The algebraic degree of geometric optimisation problems. Discrete and Comput Geom 3:177–191

Balas E, Yu CS (1982) A note on the Weiszfeld-Kuhn algorithm for the general Fermat problem. Management Science Report 484:1–16, Graduate School of Industrial Administration, Carnegie Mellon University, Pittsburgh

Brimberg J (2003) Further notes on convergence of the Weiszfeld algorithm. Yugoslav J Oper Res 13:199–206

Brimberg J, Chen R, Chen D (1998) Accelerating convergence in the Fermat–Weber location problem. Oper Res Lett 22:151–157

Brimberg J, Hansen P, Mladenović N (2006) Decomposition strategies for large-scale continuous location–allocation problems IMA J Manag Math 17:307–316

Brimberg J, Hansen P, Mladenović N, Taillard ED (2000) Improvements and comparison of heuristics for solving the multisource Weber problem. Oper Res 48:444–460

Brimberg J, Love RF (1993) Global convergence of a generalized iterative procedure for the minisum location problem with ℓ_p distances. Oper Res 41:1153–1163

Cánovas L, Canavate R, Marín A (2002) On the convergence of the Weiszfeld algorithm. Math Program 93:327–330

Cera M, Ortega FA (2002) Locating the median hunter among n mobile prey on the plane. Inter J Ind Eng 9:6–15

Chandrasekaran R, Tamir A (1989) Open questions concerning Weiszfeld's algorithm for the Fermat-Weber location problem. Math Program 44:293–295

Chen R (1991) An improved method for the solution of the problem of location on a inclined plane. RAIRO Rech Oper—Oper Res 25:45–53

Chen R (2001) Optimal location of a single facility with circular demand areas. Comput Math Appl 41:1049–1061

Cockayne EJ, Melzak ZA (1969) Euclidean constructibility in graph-minimization problems. Math Mag 42:206–208

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

Cooper L (1968) An extension of the generalized Weber problem. J Reg Sci 8:181–197

Cooper L, Katz IN (1981) The Weber problem revisited. Comput Math Appl 7:225–234

Dowling PD, Love RF (1986) Bounding methods for facilities location algorithms. Nav Res Logist Q 33:775–787
Dowling PD, Love RF (1987) An evaluation of the dual as a lower bound in facilities location problems. Inst Ind Eng Trans 19:160–166
Drezner Z (1984) The planar two center and two median problems. Transp Sci 18:351–361
Drezner Z (1995) A note on accelerating the Weiszfeld procedure. Locat Sci 3:275–279
Drezner Z (2009) On the convergence of the generalized Weiszfeld algorithm. Ann Oper Res 167:327–336
Drezner T, Drezner Z (2004) Finding the optimal solution to the Huff based competitive location model. Comput Manag Sci 1:193–208
Drezner Z, Klamroth K, Schöbel A, Wesolowsky GO (2002) The Weber problem. In Drezner Z, Hamacher H (eds) Facility Location: Applications and Theory. Springer, Berlin, pp. 1–136
Drezner Z, Schaible S, Wesolowsky GO (1990) Queuing-location problems in the plane. Nav Res Logist Q 37:929–935
Drezner Z, Wesolowsky GO (1978) Facility location on a sphere. J Oper Res Soc 29:997–1004
Drezner Z, Wesolowsky GO (1980) Optimal location of a facility relative to area demands. Nav Res Logist Q 27:199–206
Drezner Z, Wesolowsky GO (1991) The Weber problem on the plane with some negative weights. INFOR 29:87–99
Eckhardt U (1980) Weber's problem and Weiszfeld's algorithm in general spaces. Math Program 18:186–196
Eilon S, Watson-Gandy CDT, Christofides N (1971) Distribution management: Mathematical modelling and practical analysis. Charles Griffin & Co., Ltd., London
Elzinga DJ, Hearn DW (1983) On stopping rules for facilities location algorithms. Inst of Ind Eng Trans 15:81–83
Fagnano GF (1775) Problemata quaedam ad methodum maximorum et minimorum spectantia, Nova Acta Eruditorum 1775 Mensis Iunii (published in 1779), 281–303
Fernández J, Pelegrín B, Plastria F, Tóth B (2007) Solving a Huff-like competitive location and design model for profit maximization in the plane. Eur J Oper Res 179:1274–1287
Francis RL, McGinnis LF, White JA (1992) Facility layout and location: an analytical approach (2nd edition). Prentice Hall, Englewood Cliffs
Franksen OI, Grattan-Guinness I (1989) The earliest contribution to location theory? Spatio-temporal equilibrium with Lamé and Clapeyron, 1829. Math Comput Simul 31:195–220
Frenk JBG, Melo MT, Zhang S (1994) A Weiszfeld method for a generalized ℓ_p distance minisum location model in continuous space. Locat Sci 2:111–127
Gass SA (2004) In Memoriam, Andrew (Andy) Vazsonyi: 1916–2003, OR/MS Today, February 2004, http://www.lionhrtpub.com/orms/orms-2-04/frmemoriam.html, see also Ann Oper Res 167:2–5 (2009)
Gendreau M, Potvin J-Y (2008) Metaheuristics: a Canadian Perspective. INFOR 46:71–80
Hoos HH, Stützle T (2005) Stochastic local search: foundations and applications. Elsevier, Amsterdam
Hiriart-Urruty, J-B, Lemaréchal C (2001) Fundamentals of convex analysis. Springer, Berlin
Illgen A (1979) Das Verhalten von Abstiegsverfahren an einer Singularität des Gradienten. Mathematik, Operationsforschung und Statistik, Ser Optim 10:39–55
Juel H (1984) On a rational stopping rule for facilities location algorithms. Nav Res Logist Q 31:9–11
Katz IN (1969) On the convergence of a numerical scheme for solving some locational equilibrium problems. SIAM J Appl Math 17:1224–1231
Katz IN (1974) Local convergence in Fermat's problem. Math Program 6:89–104
Katz IN, Cooper L (1980) Optimal location on a sphere. Comput Math Appl 6:175–196
Katz IN, Vogl SR (2010) A Weiszfeld algorithm for the solution of an asymmetric extension of the generalized Fermat location problem. Comput Math Appl 59:399–410
Kuhn H (1967) On a pair of dual nonlinear programs. In: Abadie J (ed) Methods of nonlinear programming. North-Holland, Amsterdam, pp. 38–54

Kuhn H (1973) A note on Fermat's problem. Math Program 4:98–107
Kuhn HW, Kuenne RE (1962) An efficient algorithm for the numerical solution of the generalized Weber problem in spatial economics. J Reg Sci 4:21–33
Kupitz YS, Martini H (1997) Geometric aspects of the generalized Fermat-Torricelli problem. In: Intuitive geometry, Bolyai Society, Math Stud 6:55–127
Lamé G, Clapeyron BPE (1829) Mémoire sur l'application de la statique à la solution des problèmes relatifs à la théorie des moindres distances. Journal des Voies de Communications 10: 26–49. (In French—Memoir on the application of statics to the solution of problems concerning the theory of least distances.) For a translation into English see Franksen & Grattan-Guinness (1989)
Launhardt W (1885) Mathematische Begründung der Volkswirtschaftslehre, Wilhelm Engelmann, Leipzig
Levin Y, Ben-Israel A. (2002) The Newton bracketing method for convex minimization. Comput Optim and Appl 21:213–229
Li Y (1998) A Newton acceleration of the Weiszfeld algorithm for minimizing the sum of Euclidean distances. Comput Optim Appl 10:219–242
Love RF, Dowling PD (1989) A new bounding method for single facility location models. Ann Oper Res 18:103–112
Love RF, Morris JG, Wesolowsky GO (1988) Facilities location: models & methods. North-Holland, New York
Love RF, Yeong WY (1981) A stopping rule for facilities location algorithms. Am Inst Ind Eng Trans 13:357–362
Miehle W (1958) Link-length minimization in networks. Oper Res 6:232–243
Morris JG, Verdini WA (1979) Minisum ℓ_p distance location problems solved via a perturbed problem and Weiszfeld's algorithm. Oper Res 27:1180–1188
Ostresh LM Jr (1975) An efficient algorithm for solving the two-center location-allocation problem. J Reg Sci 15:209–216
Ostresh LM Jr (1978a) Convergence and descent in the Fermat location problem. Transportation Science 12:153–164
Ostresh LM Jr (1978b) On the convergence of a class of iterative methods for solving the Weber location problem. Oper Res 26:597–609
Overton ML (1983) A quadratically convergent method for minimizing a sum of Euclidean norms. Math Program 27:34–63
Plastria F (1987) Solving general continuous single facility location problems by cutting planes. Eur J Oper Res 29:98–110
Plastria F (1992a) GBSSS, the generalized big square small square method for planar single facility location. Eur J Oper Res 62:163–174
Plastria F (1992b) When facilities coincide: exact optimality conditions in multifacility location. J Math Anal and Appl 169:476–498
Plastria F (1993) On destination optimality in asymmetric distance Fermat-Weber problems. Ann Oper Res 40:355–369
Plastria F, Elosmani M (2008) On the convergence of the Weiszfeld algorithm for continuous single facility location-allocation problems. TOP 16:388–406
Rado F (1988) The Euclidean multifacility location problem. Oper Res 36:485–492
Rautenbach D, Struzyna M, Szegedy C, Vygen J (2004) Weiszfeld's algorithm revisited once again. Report 04946-OR, Research Institute for Discrete Mathematics, University of Bonn, Germany
Rockafellar T (1970) Convex analysis. Princeton University Press, Princeton, NJ
Rosen JB, Xue GL (1992) On the convergence of Miehle's algorithm for the Euclidean multifacility location problem. Oper Res 40:188–191
Rosen JB, Xue GL (1993) On the convergence of a hyperboloid approximation procedure for the perturbed Euclidean multifacility location problem. Oper Res 41:1164–1171
Rosing K, Harris B (1992) Algorithmic and technical improvements: optimal solutions to the (generalized) multi-Weber problem. Pap in Reg Sci 71:331–352

Schärlig A (1973) About the confusion between the center of gravity and Weber's optimum. Reg Urban Econ 13:371–382
Sturm R (1884) Über den Punkt kleinster Entfernungssumme von gegebenen Punkten. Journal für die reine und angewandte Mathematik 97:49–61. (In German: On the point of smallest distance sum from given points)
Shi Y, Chang Q, Xu J (2007) Convergence of fixed point iteration for deblurring and denoising problem. Appl Math and Comput 189:1178–1185
Valkonen T (2006) Convergence of a SOR-Weiszfeld Type Algorithm for Incomplete Data Sets. Numer Funct Anal and Optim 27;931–952. See also Errata, same journal, (2008), volume 29:1201–1203
Vardi Y, Zhang C-H (2001) A modified Weiszfeld algorithm for the Fermat-Weber location problem. Math Program 90:559–566
Vazsonyi A (2002) Which Door has the Cadillac. Writers Club Press, New York
Vazsonyi A (2002) Pure mathematics and the Weiszfeld algorithm. Decision Line 33:12–13, http://www.decisionsciences.org/DecisionLine/Vol33/33_3/index.htm
Weber A (1909) Über den Standort der Industrien. Tübingen, Germany. (English translation: Friedrich CJ (translator) (1929), Theory of the location of industries. University of Chicago Press, Chicago)
Weida NC, Richardson R, Vazsonyi A (2001) Operations analysis using Excel. Duxbury, Pacific Grove
Weiszfeld E (1937) Sur le point pour lequel la somme des distances de *n* points données est minimum. Tôhoku Math J (first series) 43:355–386
Weiszfeld E, Plastria F (2009) On the point for which the sum of the distances to *n* given points is minimum. Ann Oper Res 167:7–41
Wendell RE, Peterson EL (1984) A dual approach for obtaining lower bounds to the Weber problem. J Reg Sci 24:219–228
Wesolowsky GO (1993) The Weber problem: history and perspectives. Locat Sci 1:5–23
Wu S (1994) A polynomial time algorithm for solving the Fermat-Weber location problem with mixed norms. Optimization 30:227–234

Chapter 17
Lagrangean Relaxation-Based Techniques for Solving Facility Location Problems

Roberto D. Galvão and Vladimir Marianov

17.1 Introduction

Though it is generally agreed that the term "Lagrangean relaxation" was first used by Geoffrion (1974), the use of this technique, either explicitly or implicitly (through special applications of Lagrangean relaxation ideas), precedes the work of Geoffrion by a number of years. For example, among others, Held and Karp (1970, 1971) used the concept in their successful algorithm for the traveling salesman problem.

In location theory, it appears that the first use of Lagrangean relaxation ideas should be credited to Bilde and Krarup (1967, 1977). Although the publication of their work in English dates from 1977, the corresponding lower bounding generating procedure for the *Simple Plant Location Problem* (*SPLP*) was originally developed by the authors in 1967. This procedure was published in the period 1967–1969 in a series of research reports written in Danish, which attracted limited (if any) attention outside Scandinavia. As the authors testify in the abstract of their 1977 paper, due to their simplicity and high standard of performance, their algorithms were still competitive ten years after they were first published. This encouraged them to write an English version of their original work.

Erlenkotter (1978) acknowledges the early contribution of Bilde and Krarup in the last paragraph of the introductory remarks to his seminal paper on Lagrangean relaxation-based techniques for optimally solving the simple plant location problem. Erlenkotter states (1978, p. 993):

> After this paper was submitted for publication, I learned of closely related work by Bilde and Krarup (1977). Their paper, although published recently, is essentially a translation into English of a report originally prepared in Danish in 1967. From a different perspective, Bilde and Krarup develop a procedure essentially the same as the ascent procedure given

V. Marianov (✉)
Department of Electrical Engineering, Pontificia Universidad Católica de Chile, Santiago, Chile
e-mail: marianov@ing.puc.cl

R. D. Galvão
COPPE, Federal University of Rio de Janeiro, Brazil, deceased

H. A. Eiselt, V. Marianov (eds.), *Foundations of Location Analysis,*
International Series in Operations Research & Management Science 155,
DOI 10.1007/978-1-4419-7572-0_17, © Springer Science+Business Media, LLC 2011

here and incorporate it into a branch-and-bound procedure. They do not, however, consider the adjustment approach and other improvements developed here or provide explicit computational comparisons with other approaches.

Diehr (1972) and Marsten (1972) were the first authors to use Lagrangean relaxation-based ideas to solve the *p*-median problem. Diehr (1972) developed a heuristic algorithm which provides upper and lower bounds for the *p*-median problem. His algorithm exploits the structure of the dual of the linear programming relaxation of *p*-median problems to determine generally very tight lower bounds for this problem. In his computational experiments, using networks of up to 100 vertices, the average difference between upper and lower bounds was less than 2%.

Marsten (1972) showed that all medians of a weighted network are the extreme points of the same polyhedron P_p. He defines P_p starting from an equivalent form of the linear programming relaxation of the *p*-median problem. He then dualizes one of its constraints in Lagrangean fashion. The dual of the corresponding Lagrangean problem is linear program whose maximization objective is parameterized, from which he defines the polyhedron P_p. Marsten then developed an algorithm which makes a path of P_p that passes through most of the integer extreme points of the polyhedron and through very few others. The tour may, however, include extreme points that correspond to fractional values of *p*; furthermore, it may also not encounter the optimal *p*-median for certain integer values of *p*.

Future developments of Lagrangean relaxation-based ideas for solving simple plant location problems and *p*-median problems followed different paths. While dual ascent procedures, of the type developed by Erlenkotter (1978), continue to be very effective for solving simple plant location problems to this date, as for example in Körkel (1989), and in the several dual-based procedures developed for hub location problems (further discussed in Sect. 17.3), the use of Lagrangean relaxation proper did not prosper much in this field, perhaps due to the very successful results obtained through the dual procedures. On the other hand, several Lagrangean relaxation algorithms have been successfully developed for *p*-median problems; for example, Narula et al. (1977), Christofides and Beasley (1982), Boffey and Karkazis (1984), Hanjoul and Peeters (1985), and Mirchandani et al. (1985).

Apart from the pioneering work of Diehr (1972), the only known attempts to solve *p*-median problems through dual ascent procedures appear to be the works of Mavrides (1979) and Galvão (1980). In the latter case it was found that the restriction on *p*, the number of facilities in *p*-median problems, complicates the solution of the dual. Consequently, the corresponding computational results, while strong at the time, are modest by today's standards.

The previous paragraphs focused on simple plant location problems and *p*-median problems, which are location problems related to classical (early) applications of Lagrangean relaxation-based ideas. A general model for static *uncapacitated facility location problems* (*UFLP*) was defined by Galvão and Raggi (1989). This model, which has both simple plant location problems and *p*-median problems as special cases, was solved by a three-phase method composed of (*i*) a primal-dual algorithm; (*ii*) a Lagrangean-based subgradient optimization procedure; and (*iii*) a branch-and-bound algorithm. A detailed review of the use of Lagrangean relax-

ation in the solution of uncapacitated facility location problems is given in Galvão (1993). Further applications of Lagrangean relaxation to different classes of location problems, such as capacitated and hierarchical problems, may be found elsewhere in the literature, and a brief review of major contributions for these classes of location problems is presented in Sect. 17.4 of this Chapter.

The remainder of this chapter is organized as follows. The main concepts of Lagrangean relaxation are briefly reviewed in Sect. 17.2. The next section surveys the classical contributions by Bilde and Karup, Diehr, and Marsten. Section 17.4 provides a brief survey of important works that followed the classical contributions. Finally, Sect. 17.5 provides an outlook and a conclusions of the chapter.

17.2 A Brief Review of Lagrangean Relaxation

This brief review is based on the works of Fisher (1981) and Galvão (1993). It contains only the main results available in the literature; no proofs are included. For detailed reviews of the topic, see the excellent surveys by Geoffrion (1974), Shapiro (1979), and Fisher (1981). For a discussion of subgradient optimization strategies see Sarin and Karwan (1987), Sherali and Myers (1988), and Baker and Sheasby (1999). The use of conditional subgradient optimization is analyzed in Larsson et al. (1996). The choice of the step size in subgradient optimization algorithms is addressed by Sherali and Myers and by Poljak (1967, 1969), Held et al. (1974) and Bazaraa and Sherali (1981).

Lagrangean relaxation may, in general, be applied to any combinatorial optimization problem formulated as an integer program. However, as applications reviewed in this chapter are modeled as zero-one integer problems, we consider the following formulation:

$$
\begin{aligned}
&P: v = \min \mathbf{cx} \\
&\text{s.t. } \mathbf{Ax} \leq \mathbf{b} \\
&\mathbf{Dx} \leq \mathbf{e} \\
&\mathbf{x} \in \{0,1\}^n,
\end{aligned}
$$

where $\mathbf{c}$, $\mathbf{x}$, $\mathbf{b}$, and $\mathbf{e}$ are $[1 \times n]$, $[n \times 1]$, $[m \times 1]$, and $[\ell \times 1]$-dimensional vectors, respectively, and $\mathbf{A}$ and $\mathbf{D}$ are matrices of appropriate dimensions. Suppose that $\mathbf{Ax} \leq \mathbf{b}$ is the set of complicating constraints, that is, the set of constraints that if relaxed (dualized) into the objective function will make the resulting Lagrangean problem "easier" to solve than problem P. These constraints are thus added to the objective function through a vector $\boldsymbol{\lambda}$ of Lagrangean multipliers. The corresponding Lagrangean relaxation of P is given by:

$$
\begin{aligned}
&LR_\lambda: v_\lambda = \min \mathbf{cx} + \boldsymbol{\lambda}\,(Ax - b) \\
&\text{s.t. } \mathbf{Dx} \leq \mathbf{e} \\
&\mathbf{x} \in \{0,1\}^n.
\end{aligned}
$$

Without loss of generality, assume that P has at least one feasible solution, and that the set of feasible solutions to $LR_{\boldsymbol{\lambda}}$, which does not depend on $\boldsymbol{\lambda}$, is finite. Note that since $\mathbf{Ax} \leq \mathbf{b}$, we must have $\boldsymbol{\lambda} \geq \mathbf{0}$ for v_λ to be a lower bound for problem P. If the complicating constraints were of the form $\mathbf{Ax} \leq \mathbf{b}$, we would have to have $\boldsymbol{\lambda} \leq \mathbf{0}$ for $v_\lambda \leq v$ to hold; correspondingly, $\boldsymbol{\lambda}$ would be unrestricted in sign in the case $\mathbf{Ax} = \mathbf{b}$. When we solve LR_λ for any vector $\boldsymbol{\lambda} \geq \mathbf{0}$, we obtain a lower bound for P; we are interested, however, in obtaining the best possible lower bound.

The best choice for $\boldsymbol{\lambda}$ is given by the optimal solution of the problem:

$$D\text{: } v_D = \max_{\lambda \geq 0} v_\lambda.$$

This problem is designated by D because it corresponds to the Lagrangean dual of P with respect to the constraints $\mathbf{Ax} \leq \mathbf{b}$, see Geoffrion (1974). It is a concave problem whose structural properties make it solvable. Since we assumed that the set of feasible solutions to LR_λ is finite, the set $X = \{\mathbf{x}\text{: } \mathbf{Dx} \leq \mathbf{e},\ \mathbf{x} \in \{0, 1\}^n\}$ may be expressed as $X = \{\mathbf{x}^t,\ t = 1, \ldots, T\}$, and problem D may be formulated as a linear program with many constraints as problem $\overline{D}$ as follows:

$$\begin{aligned}
&\overline{D} : v_D = \max w \\
&\text{s.t. } w \leq \mathbf{c}\mathbf{x}^t + \boldsymbol{\lambda}(\mathbf{A}\mathbf{x}^t - \mathbf{b}),\ t = 1, \ldots, T \\
&\boldsymbol{\lambda} \geq \mathbf{0}.
\end{aligned}$$

The linear programming dual of $\overline{D}$ is a linear program with many columns (see for example Fisher 1981):

$$\begin{aligned}
&\overline{P} : v_D = \min \sum_{t=1}^{T} \xi_t \mathbf{c}\mathbf{x}^t \\
&\text{s.t. } \sum_{t=1}^{T} \xi_t \mathbf{A}\mathbf{x}^t \leq \mathbf{b} \\
&\sum_{t=1}^{T} \xi_t = 1, \\
&\xi_t \geq 0, t = 1, \ldots, T.
\end{aligned}$$

Problem $\overline{P}$ with ξ_t integer is equivalent to P, although $\overline{P}$ and $\overline{LP}$, the linear programming relaxation of P, are generally not equivalent.

Problems $\overline{D}$ and $\overline{P}$ have important structural properties that allow the formulation of algorithms to find the optimal $\boldsymbol{\lambda}$ vector that solves D. It is possible to solve $\overline{D}$ through a subgradient optimization method, and $\overline{P}$ can be solved, for example, by generalized linear programming, such as the Dantzig-Wolfe decomposition. We will briefly review each of these two possibilities.

From $\overline{D}$ it follows that $w = v_\lambda$ is the lower envelope of a finite family of linear functions. Function v_λ has the desirable properties of continuity and concavity, but is not differentiable at all points. It is, however, subdifferentiable everywhere, and the subgradient optimization method, an adaptation of the gradient method in which gradients are replaced by subgradients, can be used to solve $\overline{D}$.

An m-vector γ is called a sugradient of v_λ at $\boldsymbol{\lambda} = \overline{\boldsymbol{\lambda}}$ if $v_\lambda \leq v_{\overline{\lambda}} + (\boldsymbol{\lambda} - \overline{\boldsymbol{\lambda}})\gamma$ for all $\boldsymbol{\lambda}$. It is clear that v_λ can have several subgradients at $\boldsymbol{\lambda} = \overline{\boldsymbol{\lambda}}$ and that $\gamma = (\mathbf{Ax} - \mathbf{b})$ is a subgradient at any $\boldsymbol{\lambda}$ for which $\mathbf{x}$ solves $LR_{\boldsymbol{\lambda}}$. Any convex combination of subgradients at $\boldsymbol{\lambda} = \overline{\boldsymbol{\lambda}}$ is also a subgradient at this point.

In the subgradient optimization method, given an initial vector λ^0, a sequence $\{\lambda^k\}$ is generated by the rule

$$\lambda_i^{k+1} = \max\{0, \lambda_i^k + \theta_k . \gamma_i^k\}, \quad i = 1, \ldots, m,$$

where γ^k is any subgradient selected at $\boldsymbol{\lambda} = \overline{\boldsymbol{\lambda}}$ and $\theta_k > 0$ is the step size. Poljak (1967) proved that if θ_k satisfies $\theta_k \to 0$ and $\sum_k \theta_k \to +\infty$, then $v_{\lambda^k} \to v_D$, where v_{λ^k} is the value of v_λ at iteration k.

The step size most commonly used in practical applications is

$$\theta_k = \alpha_k (v^* - v_{\lambda^k}) / \mid\mid \gamma^k \mid\mid^2,$$

where the step size parameter α_k is a scalar satisfying $0 < \alpha_k \leq 2$, v^* is an upper bound on v_D, usually obtained by applying a heuristic to solve P, and $\mid\mid \gamma^k \mid\mid^2$ is the norm of the subgradient vector at iteration k. Although the above rule does not satisfy the second Poljak sufficient condition for optimal convergence, it has performed well in practice.

As emphasized by Sherali and Myers (1988), one of the most influential factors in the convergence of the subgradient optimization algorithm is the choice of the step size, especially the updating of the step size parameter α_k throughout the procedure. Held et al. (1974) initialize α_k as 2 and halve it whenever its current value fails to improve the value of v_D after a predetermined number of iterations. They report satisfactory convergence results. Bazaraa and Sherali (1981) use a different strategy, which they claim to yield faster convergence rates.

We will now examine the solution of $\overline{P}$ by generalized linear programming. This procedure starts with a subset of the T columns, which form a master problem. A column generating technique is used to find a new column for the master problem and eventually an optimal solution for this problem will yield, as dual multipliers, the $\boldsymbol{\lambda}$ vector of Lagrangean multipliers. This approach has not performed consistently well and modifications of this technique, for example the Boxstep algorithm of Hogan et al. (1975), have performed better. A hybrid approach is to use subgradient optimization as a starting strategy and then switch to generalized linear programming when the convergence rate slows down.

The experience to date indicates that subgradient optimization is a more effective method to determine $\boldsymbol{\lambda}$, and for this reason it has been used in the majority of the applications. The subgradient algorithm is easy to implement; a version of it appears, for example, in Sherali and Myers (1988). This algorithm terminates when the value of v_D coincides with an upper bound calculated for P, in which case the optimal solution for P is available, or when $\gamma = 0$, in which case a duality gap may exist. Other stopping rules commonly used are to halt the procedure when the algorithm fails to converge after a predetermined number of iterations,

or when the step size θ_k becomes very small. In these two latter cases a duality gap is present.

We will omit discussion of Lagrangean decomposition, a solution technique closely related to Lagrangean relaxation. For an overview of Lagrangean decomposition, see, for example, Guignard and Kim (1987).

17.3 The Classical Contributions and Their Impact

This section surveys three of the original contributions that laid the foundations for the use of Lagrangean Relaxation techniques for the solution of location problems. Each subsection surveys one of the three contributions.

17.3.1 Bilde and Krarup (1967, 1977): Sharp Lower Bounds for the Simple Plant Location Problem

17.3.1.1 Statement of the Problem and some Fundamental Results

In order to follow the arguments of Bilde and Krarup, we first need to formulate the simple plant location problem. In order to do so, we need the following notation:

Sets

I: Set of sites available for the location of facilities: $i \in I = \{1, 2 \ldots m\}$, and
J: set of customers: $j \in J = \{1, 2 \ldots n\}$.

Parameters

m: Number of potential facility sites,
n: number of customers,
f_i: fixed cost associated with facility i,
b_j: demand (number of units) of customer j,
t_{ij}: unit transportation cost from facility i to customer j,
$c_{ij} = t_{ij} b_j$: total transportation cost incurred when demand j is totally supplied from facility i,
$\Delta = [\Delta_{ij}]$: an $[m \times n]$-dimensional matrix of real numbers, and
λ_j: level number of column j.

Variables

x_{ij}: Fraction of the demand of customer j supplied from facility i. If the single source property holds for a given problem, then we have

$$x_{ij} = \begin{cases} 1 & \text{if customer } j \text{ is supplied from facility } i; \\ 0, & \text{otherwise.} \end{cases}$$

$$y_i = \begin{cases} 1 & \text{if facility } i \text{ is open;} \\ 0 & \text{otherwise.} \end{cases}$$

r_i: slack (auxiliary) variable related to facility i.

Given these sets, parameters, and variables, we can now write the simple plant location problem as

$$SPLP: \; Z_{SPLP}(\min) = \sum_{i=1}^{m} f_i y_i + \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij} \tag{17.1}$$

$$\text{s.t.} \sum_{i=1}^{m} x_{ij} = 1 \; \forall \, j$$

$$y_i - x_{ij} \geq 0, \quad \forall i, j$$

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

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

In the formulation above the objective function minimizes fixed plus transportation costs. The first constraints guarantee that the demand of all customers is satisfied, the following constraints ensure that products are shipped only from open facilities, and the remaining constraints define the nature of the decision variables. Notice that since there are no capacity constraints or economies of scale, this formulation has the single source property, with each customer fully assigned to its closest facility. Also, without loss of generality, Bilde and Krarup assume that all fixed costs are nonnegative. Finally, if it is also assumed that the parameters c_{ij} are nonnegative, the demand satisfaction constraints may be replaced by $\sum_{i=1}^{m} x_{ij} \geq 1, \; \forall j$, since, given the nonnegative nature of the parameters f_i and c_{ij}, these constraints will be satisfied as equalities in the optimal solution. This latter form of expressing the demand satisfaction constraints is convenient for some of the developments that follow.

A direct way of generating a lower bound on $Z^0_{SPLP} = \min\{Z_{SPLP}\}$ could be to relax the integrality constraints to $x_{ij}, y_i \geq 0, \forall \, i,j$, and solve the resulting linear program problem. Instead, the authors developed a highly effective heuristic method to maximize the lower bound. Matrix $\mathbf{\Delta}$ is said to be *feasible* if it satisfies the following conditions:

$$\sum_{j=1}^{n} \Delta_{ij} \leq f_i, \; \forall i, \quad \text{and} \quad \Delta_{ij} \geq 0 \; \forall i, j. \tag{17.2}$$

The matrix $\mathbf{\Delta}$ is now introduced into the formulation (17.1) of the simple plant location problem by adding and subtracting the same term in the objective function, which results in the following formulation:

$$SPLP: Z_{SPLP}(\min) = \sum_{i=1}^{m} (f_i y_i - \sum_{j=1}^{n} \Delta_{ij} x_{ij}) + \sum_{i=1}^{m} \sum_{j=1}^{n} (c_{ij} + \Delta_{ij}) x_{ij} \tag{17.3}$$

$$\text{s.t.} \sum_{i=1}^{m} x_{ij} \geq 1 \ \forall j$$

$$y_i - x_{ij} \geq 0 \ \ \forall i, j$$

$$x_{ij} \in \{0,1\} \ \forall i, \ j$$

$$y_i \in \{0,1\} \ \forall i.$$

The optimal solution to problem (17.1) or, alternatively, problem (17.3) is denoted by $(\mathbf{x}^0, \mathbf{y}^0)$. It is not difficult to see that

$$f_i y_i - \sum_{j=1}^{n} \Delta_{ij} x_{ij} \geq 0, \ \forall i \tag{17.3a}$$

for any feasible $\mathbf{\Delta}$ and for any $(\mathbf{x}, \mathbf{y})$ representing a feasible solution to (17.3).

For any *fixed* set of feasible parameters Δ_{ij}, consider the linear programming problem

$$LBSPLP : Z_{LBSPLP}(\min) = \sum_{i=1}^{m} \sum_{j=1}^{n} (c_{ij} + \Delta_{ij}) x_{ij} \tag{17.4}$$

$$\text{s.t.} \sum_{i=1}^{m} x_{ij} \geq 1 \ \forall j$$

$$x_{ij} \geq 0 \ \ \forall i, j$$

with $\min \{Z_{LBSPLP}\} = Z^*_{LBSPLP}$.

For $(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^0, \mathbf{y}^0)$ we obtain, from (17.3a)

$$\begin{aligned} Z^0_{SPLP} &= \sum_{i=1}^{m} (f_i y_i^0 - \sum_{j=1}^{n} \Delta_{ij} x_{ij}^0) + \sum_{i=1}^{m} \sum_{j=1}^{n} (c_{ij} + \Delta_{ij}) x_{ij}^0 \\ &\geq \sum_{i=1}^{m} \sum_{j=1}^{n} (c_{ij} + \Delta_{ij}) x_{ij}^0 \geq Z^*_{LBSPLP} \end{aligned} \tag{17.5}$$

The problem *LBSPLP* may be solved by inspection. It is easy to see that

$$Z^*_{LBSPLP} = \sum_{j=1}^{n} \min_i \{c_{ij} + \Delta_{ij}\}, \tag{17.6}$$

i.e. Z^*_{LBSPLP} is the sum of the column minima of the $(\mathbf{C}+\mathbf{\Delta})$ matrix. This immediately implies

Theorem 1: *Given any set of nonnegative numbers* Δ_{ij} *satisfying* $\sum_{j=1}^{n} \Delta_{ij} \leq f_i \ \forall i$, *we have*

$$\sum_{j=1}^{n} \min_i \{c_{ij} + \Delta_{ij}\} \leq Z^0_{SPLP}.$$

This lower bound is evidently dependent on the way $\mathbf{\Delta}$ is determined. From Theorem 1 it follows that the *sharpest lower bound* is found as the optimal solution to the problem

$$W_{LBSPLP}(\max) = \sum_{j=1}^{n} \min_i \{c_{ij} + \Delta_{ij}\} \tag{17.7}$$

$$\text{s.t.} \sum_{j=1}^{n} \Delta_{ij} \leq f_i \ \forall i$$

$$\Delta_{ij} \geq 0 \ \forall i, j.$$

In fact, problem (17.7) could be reformulated and solved as a linear program, but the authors decided instead to seek a bounding procedure which—rather than attempting to find an optimal solution to the bounding problem—could obtain sharp lower bounds using limited computational effort. To this end, the authors developed a heuristic procedure, the principles of which are described in the following paragraphs.

This heuristic procedure is initiated with the given matrix $\mathbf{C}$ and a matrix $\mathbf{\Delta}$ that consists entirely of zeroes. A set r_i of auxiliary (slack) variables is then introduced as

$$r_i = f_i - \sum_{j=1}^{n} \Delta_{ij} \ \forall i \tag{17.8}$$

where the slack variables r_i initially equal the costs f_i. The $(n+1)$ numbers $(r_i, \ \Delta_{i_1} \ldots \ \Delta_{i_n})$ may be viewed, throughout the computations, as a *partitioning* of the corresponding f_i.

The idea of Bilde and Krarup is to find partitionings of the fixed costs so as to maximize the summed column minima of the $(\mathbf{C}+\mathbf{\Delta})$ matrix. All parameters c_{ij} preserve their original values and the elements of $\mathbf{\Delta}$ are increased iteratively. The procedure operates in the columns of the matrix $(\mathbf{C}+\mathbf{\Delta})$, one at a time. In each step a column is selected and an attempt is made to alter a subset of its elements by increasing the respective parameters Δ_{ij} in such a way that it has a maximum effect on the corresponding column minimum, with a minimum "consumption" of the slack variables r_i involved. (Note that the slack variables r_i cannot become negative).

Finally, in order to guide the search for the column to be the next candidate for further augmentation, the authors associate a *level-number* λ_j with each column j of the matrix $(\mathbf{C}+\mathbf{\Delta})$. This number is equal to the number of occurrences of the smallest element in that column. At any stage of the computation the next candidate for selection, j^*, is the column with the smallest level number, ties being broken by selecting the column with the smallest subscript. The selection rule can then be written as

$$j^* = \min\{j : \lambda_j = \min_{s \in J}(\lambda_s)\}. \tag{17.9}$$

For further details of the heuristic procedure, together with a small numerical example solved in detail to illustrate it, readers are referred to Bilde and Krarup (1977).

17.3.1.2 The Bounding Procedure and a Lagrangean Relaxation of Formulation (17.1)

Bilde and Krarup (1977) relate their bounding procedure to a Lagrangean relaxation of formulation (17.1), corresponding to the definition of $y_i - x_{ij} \geq 0 \ \forall i, j$, as the set of "complicating constraints." The parameters Δ_{ij} are the corresponding set of nonnegative Lagrangean multipliers. The Lagrangean problem then becomes

$$\begin{aligned} LR1\_SPLP\text{: } & Z_{LR1\_SPLP}(\min) \\ & = \sum_{i=1}^{m} f_i y_i + \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij} x_{ij} + \sum_{i=`1}^{m}\sum_{j=1}^{n} \Delta_{ij}(x_{ij} - y_i) \\ & \equiv \sum_{i=1}^{m} (f_i - \sum_{j=1}^{n} \Delta_{ij})\, y_i + \sum_{i=1}^{m}\sum_{j=1}^{n} (c_{ij} + \Delta_{ij})\, x_{ij} \end{aligned} \tag{17.10}$$

$$\text{s.t.} \sum_{i=1}^{m} x_{ij} \geq 1 \ \forall j$$

$$x_{ij} \in \{0,1\} \ \forall i, j$$

$$y_i \in \{0,1\} \ \forall i.$$

If we consider the multipliers Δ_{ij} satisfying (17.2), the variables y_i may be removed from *LR*1_*SPLP* because $\sum_j \Delta_{ij} \leq f_i \ \forall i$ (and, therefore, $y_i = 0 \ \forall i$). After the variables y_i are removed from (17.10), the problem *LR*1_*SPLP* coincides with the formulation (17.4) and may be thus solved by inspection. Due to relation (17.6), the minimum value of Z_{LR1_SPLP} is determined by

$$Z^0_{LR1\_SPLP} = \sum_{j=1}^{n} \min_i \{c_{ij} + \Delta_{ij}\}.$$

In terms of Lagrangean relaxation, Bilde and Krarup's approach may be viewed as a parameterized relaxation, where the bounding procedure is a rule for setting the parameters Δ_{ij} to obtain sharp lower bounds. Notice that the Lagrangean dual

$$\max_{\Delta_{ij} \geq 0} \{Z^0_{LR1\_SPLP}\},$$

and the lower bound maximization problem (17.7) are equivalent problems.

The remainder of the paper by Bilde and Krarup (1977) is dedicated to demonstrating that, if certain conditions are met, the optimal solution of simple plant location problems may be found by inspection for moderately sized problems as a follow-up of the bounding procedure. Otherwise, a branch-and-bound algorithm is used to find such optimal solutions, and some related computational experience is reported. These topics, however, are beyond the scope of this chapter.

17.3.1.3 The Impact of the Work on Future Developments

The very successful algorithm *DUALOC*, developed by Erlenkotter (1978) for the simple plant location problem cannot be seen as a development originating from the work of Bilde and Krarup, since the two procedures, in spite of their similarity, were developed independently, as explained in the introductory remarks of this chapter. The work by Körkel (1989), on the other hand, was directly built upon the work of Erlenkotter, since it specifically improves *DUALOC*, cutting computational times considerably. The use of dual ascent procedures for solving simple plant location problems and related problems, however, continue to be successfully used to this date, based on the same principles of these early developments.

Consider for example the *uncapacitated multiple allocation hub location problem* (*UMAHLP*), an important problem in the design of logistic networks. Its objective is to find minimum cost solutions to the problem of locating hubs and allocating terminals to them, in the presence of installation and transportation costs. According to Klincewicz (1996), the dual ascent method he developed for *UMAHLP*, comprising basically a dual ascent algorithm, a dual adjustment routine and a branch-and-bound algorithm, had its origins in the works of Bilde and Krarup (1977) and Erlenkotter (1978) for the simple plant location problem.

The first integer programming formulations of *UMAHLP* are given in Campbell (1994). Klincewicz (1996) identified a close relationship between the simple plant location problem and *UMAHLP* and proposed a solution method that may be seen as an extension of the method of Erlenkotter (1978) for the hub location problem. Sung and Jin (2001) and Mayer and Wagner (2002) presented alternative mathematical formulations for *UMAHLP* and used dual-based methods for their solution. Finally, Cánovas et al. (2006) presented a formulation that has the tightest linear

programming relaxation of those developed so far. They also developed a complex dual ascent procedure, used to solve the relaxed problem and produce tight bounds in each node of their branch-and-bound algorithm.

While the papers reviewed above are not intended to be an exhaustive survey of the use of dual-based methods to solve location problems, they do demonstrate their importance. The close relationship that exists between dual-based and Lagrangean relaxation methods, as demonstrated by Bilde and Krarup (1977), are a proof of the importance of the pioneering developments in these fields.

17.3.2 Diehr (1972): Upper and Lower Bounds for the p-Median Problem

As previously discussed, Diehr (1972) developed a heuristic algorithm that provides upper and lower bounds for the p-median problem. He considers a network $G=(V, E)$ with n vertices V joined with edges E. Associated with each vertex j is a nonnegative real valued weight $h_j, j=1, 2\ldots n$. Associated with each edge in E is a nonnegative distance or length. Let $V^p = \{i_1, i_2, \ldots, i_p\}$ be a set of any p vertices contained in V, i.e., $V^p \subset V$. Define $d(V^p, j) = \min\{d(i_1, j), d(i_2, j), \ldots, d(i_p, j)\}$, where $d(i_k, j)$ is the length of the shortest path from vertex i_k to vertex j. Finally, define an $[n\times n]$-dimensional matrix of weighted distances $\mathbf{D}=[D_{ij}]=[h_j d(i, j)]$.

Associated with any subset V^p is a value $S(V^p)$ given by $S(V^p) = \sum_{j=1}^{n} \min_{i\in V^p}\{D_{ij}\}$. Thus, for a given subset V^p the value of $S(V^p)$ is determined by summing the minimum D_{ij} values for each column (vertex of the network). The p-median is then given by the subset V_0^p, where $S(V_0^p) = \min_{V^p\subset V}\{S(V^p)\}$. Notice that the p-median was defined as a set of p vertices of the network. This is correct because Hakimi (1964, 1965) proved that there exists at least one subset $V^p \subset V$ containing exactly p vertices, such that $S(V^p) \leq S(Y^p)$ for any arbitrary set Y^p of p points on the links or vertices of the network $G=(V, E)$. The optimal p-median of a network can therefore be sought as a subset of p vertices of the network.

The solution algorithm of Diehr is composed of two phases: (i) a heuristic algorithm to find an approximate solution (upper bound) for the problem, and (ii) a heuristic search on a dual problem to determine a lower bound. The primal heuristic is a "greedy"-interchange heuristic that has an iterative (vertex substitution) phase similar to the algorithm of Teitz and Bart (1968); it will not be discussed here, as our interest lies in the solution of the dual problem.

17.3.2.1 The Dual Problem

Theorem 2: *Given any subset V^p of p vertices, a lower bound $\underline{S}(V^p)$ to the value of $S(V^p)$ is given by*

$$\underline{S}(V^p) = \sum_{j=1}^{n} [B_j - \sum_{i \in V^p} \max(0, B_j - D_{ij})] \leq S(V^p), \tag{17.11}$$

where $B_j, j = 1, 2 \ldots n$ are any real-valued variables.

Proof: Assume, without loss of generality that $D_{i_1 j} \leq \ldots . \leq D_{i_p j}$ for some j. Diehr (1972) proves that (17.11) provides a loss for each value of j which is less than the loss for each j in the value of $S(V^p)$. That is, he proves that the following relationship holds:

$$B_j - \sum_{i \in V^p} \max\left(0, B_j - D_{ij}\right) \leq \min_{i \in V^p} D_{ij} = D_{i_1 j}. \tag{17.12}$$

Consider the following two cases, one of which must be true:

(i) $B_j \leq D_{i_1 j}$. In this case the left hand side of (17.12) is strictly less than $D_{i_1 j}$ by assumption.

(ii) $B_j > D_{i_1 j}$ In this case the left-hand side of (17. 12) can be rewritten as

$$\begin{aligned} &B_j - (B_j - D_{i_1 j}) - \sum_{k=2}^{p} \max(0, B_j - D_{i_k j}) \\ &= D_{i_1 j} - K, K = \sum_{k=2}^{p} \max\left(0, B_j - D_{i_1 k}\right) \end{aligned}$$

which is clearly less than or equal to $D_{i_1 j}$ since K is nonnegative. Since the bound holds for each j it clearly holds for the sum over j, establishing (17.11). □

Diehr argues that since a lower bound to the optimal solution value $S(V_p^0)$ is desired, it is necessary to minimize $\underline{S}(V^p)$ over all V^p contained in V, i.e.,

$$\min_{V^p \subset V} \underline{S}(V^p) \leq S(V_p^0). \tag{17.13}$$

The subset V^p which minimizes the left-hand side of (17.13) is determined as follows. Define for each $i \in V$ a "gain" G_i given by $G_i = \sum_{j=1}^{n} \max(0, B_j - D_{ij})$. Then rank the gains such that $G_{i_1} \geq G_{i_2} \geq \ldots \geq G_{i_n}$. The subset V^p which minimizes the left-hand side of (17.13) is simply $V^p = (i_1, i_2, \ldots, i_p)$. The lower bound on the optimal solution is thus given by:

$$\underline{S}(V^p) = \sum_{j=1}^{n} B_j - \sum_{k=1}^{p} G_{i_k}. \tag{17.14}$$

The goal is now to find the maximum possible lower bound through an adequate selection of the variables B_j. The method used to determine starting values for the variables B_j and to seek for "good" B_j values is discussed in the following section that is dedicated to the dual heuristic.

It is worth mentioning that the corresponding dual problem is equivalent to the dual of the linear programming relaxation of the primal problem, as formulated by ReVelle et al. (1970). The variables B_j are the dual variables corresponding to the restrictions that each vertex in the network must be assigned to one of the vertices in V^p, i.e., $\sum_{i=1}^{n} x_{ij} \geq 1 \ \forall j$. Note that for p-median problems, we must also satisfy the restriction $\sum_{i=1}^{n} y_i = p$, where p is the number of facilities to be established in the network and the variables x_{ij} and y_i are as defined in Sect. 17.3.1.

17.3.2.2 The Dual Heuristic

This phase of the algorithm uses solution values from the primal phase to determine initial values for the variables B_j. The primal "greedy"-interchange heuristic, because of its "greedy" phase that sequentially selects vertices for the solution, provides approximate solution values $S(V^p)$ and $S(V^{p-1})$, for the p- and (p-1)-median problems, respectively. Diehr then argues that if it is supposed that (*i*) these values are optimal and equal to the maximum lower bounds, and (*ii*) the values of the dual variables are the same for the maximum lower bounds for the p- and (p-1)-median problems, then from (17.14) it is possible to write:

$$S(V^{p-1}) - S(V^p) = \left(\sum_{j=1}^{n} B_j - \sum_{k=1}^{p-1} G_{i_k} \right) - \left(\sum_{j=1}^{n} B_j - \sum_{k=1}^{p} G_{i_k} \right) = G_{i_p}.$$

Assuming that the p largest gains are equal, i.e., $G_{i_1} = G_{i_2} = \ldots = G_{i_p}$, we then have

$$S(V^p) = \sum_{j=1}^{n} B_j - pG_{i_p}, \text{ and}$$

$$\sum_{j=1}^{n} B_j = S(V^p) + p(S(V^{p-1}) - S(V^p)) \tag{17.15}$$

Finally, the initial values of the variables B_j are determined so that their sum satisfies (17.15), with individual variables B_j biased above or below the mean, depending whether the values D_{ij} in the corresponding columns are large or small when compared to the D_{ij} values in other columns.

The determination of a good lower bound involves a local search on the variables B_j. Starting with B_1, it is determined whether the lower bound can be increased by increasing or decreasing B_1 to the next larger or smaller value of D_{ij}. If an improvement in the lower bound is obtained, the change which results in the largest increase in the lower bound is made. The search is then performed on B_2, B_3,…, B_n. The algorithm cycles through the columns until it is no longer possible to improve the lower bound by changing the values of the variables B_j.

Diehr also found that, when such local optimum is achieved, further improvements are often possible if a randomly selected B_j is allowed to increase or decrease to the next higher or lower D_{ij} in its column; if an improvement is thus obtained, the algorithm reverts to the local search cycling through the columns. The algorithm terminates when a sequence of n random selections of the variables B_j is completed without further improvements in the value of the lower bound. The algorithm also terminates if the lower bound coincides, at any given stage of the search, with the value found by the primal heuristic.

Finally, Diehr conjectured that comparable results could be obtained if a modified search of his method were applied to the simple plant location problem. In that case the lower bound would be given by

$$\sum_{j=1}^{n} B_j - \sum_{i=1}^{m} \max\ (0, G_i - f_i). \tag{17.16}$$

That is, a vertex should be included in the lower bound computation whenever its gain G_i was greater than the corresponding fixed cost f_i. Though Diehr did not conduct any experimentation to support his conjecture, we can compare his conjecture to the Lagrangean relaxation of the simple plant location problem with respect to the constraints

$$\sum_{i=1}^{m} x_{ij} \geq 1\ \ \forall j.$$

Consider formulation (17.1) of the simple plant location problem with constraints $\sum_{i=1}^{m} x_{ij} = 1 \forall j$ replaced by $\sum_{i=1}^{m} x_{ij} \geq 1 \forall j$. As already noted, the two formulations are equivalent. If we dualize these constraints in usual Lagrangean fashion using the variables $B_j \geq 0\ \forall j$ as Lagrangean multipliers, we obtain the following formulation:

$$\begin{aligned} LR2\_SPLP : Z_{(LR2\_SPLP)_B} &= \min \left\{ \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij} + \sum_{i=1}^{m} f_i y_i + \sum_{j=1}^{n} B_j \left(1 - \sum_{i=1}^{m} x_{ij}\right) \right\} \\ &\equiv \min \left\{ \sum_{i=1}^{m}\sum_{j=1}^{n} \left(c_{ij} - B_j\right) x_{ij} + \sum_{i=1}^{m} f_i y_i + \sum_{j=1}^{n} B_j \right\} \end{aligned} \tag{17.17}$$

$$\text{s.t. } y_i - x_{ij} \geq 0\ \ \forall i,\ j$$

$$x_{ij} \in \{0,1\}\ \ \forall i,\ j$$

$$y_i \in \{0,1\}\ \ \forall i.$$

This Lagrangean problem is easily analyzed for fixed values of the variables B_j. From the form of the objective function and considering the constraints it follows that

$$x_{ij} = \begin{cases} y_i & \text{if } c_{ij} - B_j \leq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Now we define, for a fixed vector $\mathbf{B} = \{B_1, B_2 \ldots B_n\}$:

$$\Psi_i(\mathbf{B}) = \sum_{j=1}^{n} \min(0, c_{ij} - \mathbf{B}_j) + f_i.$$

The optimal values of the variables y_i are obtained by solving the reduced problem

$$LR2\_SPLPR : Z_{(LR2\_SPLPR)_B} = \min \sum_{i=1}^{m} \Psi_i(\mathbf{B}) y_i, \text{s.t. } y_i \in \{0, 1)\ \forall i. \tag{17.18}$$

It is easy to see that this problem can be solved by inspection by setting y_i=1 if $\Psi_i(\mathbf{B}) \leq 0$, and y_i=0, otherwise. The best choice for the vector $\mathbf{B}$ is obtained by solving the Lagrangean dual

$$LD : Z_{LD} = \max_{\mathbf{B} \geq 0} Z_{(LR2\_SPLPR)_{\mathbf{B}}} \tag{17.19}$$

for example through a subgradient optimization method.

Return now to (17.16), the expression proposed by Diehr (1972), where G_i was defined as $G_i = \sum_{j=1}^{n} \max(0, B_j - D_{ij})$. Replacing the parameters D_{ij} by c_{ij}, it is then possible to write

$$\begin{aligned} \sum_{j=1}^{n} B_j - \sum_{i=1}^{m} \max(0, G_i - f_i) &= \sum_{j=1}^{n} B_j + \sum_{i=1}^{m} \min(0, f_i - G_i) \\ &= \sum_{j=1}^{n} B_j + \sum_{i=1}^{m} \min\left(0, \left\{ f_i - \sum_{j=1}^{n} \max(0, B_j - c_{ij}) \right\}\right) \\ &= \sum_{j=1}^{n} B_j + \sum_{i=1}^{m} \min\left(0, \left\{ f_i + \sum_{j=1}^{n} \min(0, c_{ij} - B_j) \right\}\right) \\ &= \sum_{j=1}^{n} B_j + \sum_{i=1}^{m} \min\{0, \Psi_i(\mathbf{B})\}. \end{aligned} \tag{17.20}$$

For a given vector $\mathbf{B} = \{B_1, B_2 \ldots B_n\}$, the final expression of (17.20) is equivalent to the objective function of *LR2_SPLPR*, given in (17.18). The search for the vector $\mathbf{B}$

that maximizes the lower bound is therefore equivalent to solving the Lagrangean dual formulated in (17.19).

The development shown above not only proves the conjecture of Diehr (1972), but also indicates that his approach is related to a Lagrangean relaxation of the simple plant location problem. His proposed bound, however, would differ from the bound of Bilde and Krarup (1967, 1977) for the simple plant location problem (see Sect. 17.3), since in that case the dualized constraints were the constraints $y_i - x_{ij} \geq 0 \ \forall i, j$. To the best of our knowledge, however, no numerical experimentation has ever been carried out with either Lagrangean relaxation of the simple plant location problem.

Given that the conjecture of Diehr followed from an immediate extension of his detailed work for the *p*-median problem, it is easy to prove, following the same steps, that his dual heuristic for the *p*-median problem is also related to a Lagrangean relaxation of that problem.

17.3.2.3 Impact of the Work on Future Developments

Contrary to the simple plant location problem, several Lagrangean relaxation schemes have been successfully developed for the *p*-median problem (see Sect. 17.1). On the other hand, dual-based procedures for this problem did not receive much attention after the early paper by Diehr (1972), except perhaps for the work of Mavrides (1979) and Galvão (1980).

Galvão (1980) developed his dual-bounded algorithm to solve *p*-median problems having as a basis the paper by Diehr. It consists of a dual ascent algorithm that solves the dual of the linear programming relaxation of the *p*-median problem. This procedure produces sharp lower bounds for the *p*-median problem and was later embedded into a branch-and-bound algorithm. It used relations (17.15) to obtain initial values for the variables B_j, which have a prominent role in the procedure.

Consider the following formulation of the *p*-median problem:

$$p-MP : Z_{p-MP}(\min) = \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} x_{ij} \tag{17.21}$$

$$\text{s.t.} \sum_{i=1}^{n} x_{ij} \geq 1 \quad \forall j \qquad (B_j)$$

$$y_i - x_{ij} \geq 0 \quad \forall i, j \qquad (\Delta_{ij})$$

$$\sum_{i=1}^{n} y_i = p \qquad (\Omega)$$

$$x_{ij} \in \{0,1\} \ \forall i, j$$

$$y_i \in \{0,1\} \ \forall i$$

where the variables shown in parentheses are the dual variables corresponding to each set of constraints. The dual algorithm of Galvão may also be related to a Lagrangean relaxation of the p-median problem, if constraints $\sum_{i=1}^{n} x_{ij} \geq 1 \ \forall j$, and $\sum_{i=1}^{n} y_i = p$, are simultaneously dualized into the objective function using the indicated dual variables as Lagrangean multipliers.

The work of Galvão was later reviewed as a modification of the algorithm of Erlenkotter (1978) for the simple plant location problem. It may be indeed viewed as a specialization of Erlenkotter's algorithm, but was in fact developed independently, before Erlenkotter's paper was published, at a time when ideas about dual-based procedures for this type of problem were just beginning to be considered.

17.3.3 Marsten (1972): An Algorithm for Finding Almost all Medians of a Network

17.3.3.1 Essentials of Marsten's Contribution

In order to describe the procedure developed by Marsten (1972) it is necessary to rewrite the p-median problem using the formulation of ReVelle and Swain (1970) for location in a network. Using h_j as defined in Sect. 17.3.2, let

d_{ij} denote the length of the shortest path from node i to node j, denote

$$c'_{ij} = h_j d_{ij}, \text{ and define}$$

$$y_{ij} = \begin{cases} 1 & \text{if node } j \text{ is assigned to a median at node } i; \\ 0, & \text{otherwise.} \end{cases}$$

Note that $y_{ii}=1$ if and only if a median is placed at node i. The problem may then be formulated as:

$$RS: \text{Min} \sum_{i=1}^{n} \sum_{j=1}^{n} c'_{ij} y_{ij}$$

$$\text{s.t.} \sum_{i=1}^{n} y_{ij} = 1, \ j = 1, \ldots, n \qquad (17.22.1)$$

$$y_{ij} \leq y_{ii} \ \forall i \neq j \tag{17.22.2}$$

$$\sum_{i=1}^{n} y_{ii} = p \tag{17.22.3}$$

$$y_{ij} \in \{0,1\} \ \forall i, j. \tag{17.22.4}$$

Problem *RS* is an integer linear program. It can be relaxed to a linear program by replacing (17.22.4) with $y_{ij} \geq 0 \ \forall i, j$. Marsten calls this linear program $\overline{RS}$, it is a very large linear program with n^2 variables and (n^2+1) constraints.

As already noted in Sect. 17.1, Marsten (1972) showed that every p-median of the network is an extreme point of the same polyhedron P_p, and that it is possible to take a tour around P_p that passes through most of these special (integer) extreme points and through very few others. Marsten begins by defining polyhedron P_p. He eliminates the y_{ii} variables by means of eq. (17.22.1), which results in

$$y_{ii} = 1 - \sum_{k \neq i} y_{ik}, \ i = 1, \ldots, n$$

Carrying out this elimination gives an equivalent form of $\overline{RS}$, namely

$$\overline{RS} : \text{Min} \sum_{i} \sum_{k \neq i} c'_{ij} y_{ik} \tag{17.23.1}$$

$$\text{s.t. } y_{ij} + \sum_{k \neq i} y_{ik} \leq 1 \ \forall i \neq j \tag{17.23.2}$$

$$\sum_{i} \sum_{k \neq i} y_{ik} = n - p \tag{17.23.3}$$

$$y_{ij} \geq 0 \ \forall i \neq j. \tag{17.23.4}$$

The polyhedron P_p is now defined by Marsten as

$$P_p = \{\mathbf{y} \in E^{n(n-1)} : \mathbf{y} \geq \{0\}^{n(n-1)} \text{and } \mathbf{y} \text{ satisfies relation (17.23.2)}\}, \tag{17.24}$$

where $E^{n(n-1)}$ is the space of vectors with $n(n-1)$ zero-one elements. It is clear that any *integer* solution of $\overline{RS}$ is an extreme point of P_p and that this is true regardless of the value of p.

Marsten now dualizes $\overline{RS}$ with respect to the single constraint (17.23.3), using λ (scalar) as the dual variable, obtaining the objective

$$\underset{\mathbf{y}\in P_p}{\text{Min}} \sum_i \sum_{k\neq i} (c'_{ik} - \lambda) y_{ik} + (n-p)\lambda \tag{17.25}$$

He then defines

$$v(\lambda) = \min_{\mathbf{y}\in P_p} \sum_i \sum_{k\neq i} (c'_{ik} - \lambda) y_{ik}, \text{ and} \tag{17.26}$$

$$Z_p(\lambda) = (n-p)\lambda + v(\lambda) \tag{17.27}$$

The dual problem is then

$$\underset{\lambda}{\text{Max}}\, Z_p(\lambda) \tag{17.28}$$

If $m = n(n-1)$, and $\mathbf{e}_m$ is a vector of m ones, then (17.26) can be written more compactly as

$$v(\lambda) = \min_{\mathbf{y}\in P_p} (\mathbf{c} - \lambda \mathbf{e}_m)\mathbf{y} \tag{17.29}$$

Marsten (1972) notes that Eq. (17.29) leads to the following observations:

(a) $v(\lambda)$ does not depend on p
(b) $v(\lambda)$ is the optimal value of a linear program parameterized in its objective function. Noting that λ is a scalar, $v(\lambda)$ is a piecewise-linear concave function
(c) $Z_p(\lambda)$ is just the linear function $(n-p)\lambda$ plus $v(\lambda)$. Therefore $v(\lambda)$ gives the whole family of $Z_p(\lambda)$ functions for $p=1\ldots n$. Note that each $Z_p(\lambda)$ is also piecewise-linear and concave.

Let now $Y(\lambda)$ denote the set of optimal solutions for fixed λ, i.e.,

$$Y(\lambda) = \{\mathbf{y} \in P_p : v(\lambda) = (c - \lambda \mathbf{e}_m)\mathbf{y}\}. \tag{17.30}$$

Inspection of the objective function (17.26) reveals that $v(\lambda) \leq 0$ for all λ and that y_{ik} cannot participate in an optimal solution as long as $\lambda < c'_{ik}$. That is, $\lambda < c'_{ik}$. implies that $y_{ik}=0$ for all $\mathbf{y} \in Y(\lambda)$. In fact $\mathbf{y}=\mathbf{0}$ is an optimal solution as long as $\lambda \leq c^*, c^* = \min_{i\neq k} c'_{ik}$, i.e., as long as λ is less than or equal to the smallest number in the weighted distance matrix. So for $\lambda \leq c^*$ we have $v(\lambda) = 0$ and $Z_p(\lambda) = (n-p)\lambda$.

Marsten makes the tour of the polyhedron P_p in the course of constructing the $v(\lambda)$ function. The construction of $v(\lambda)$ is a straightforward application of parametric linear programming. Let $\mathbf{y}^0 \in Y(\lambda^0)$. Then

$$v(\lambda^0 + \delta) = [\mathbf{c} - (\lambda^0 + \delta)\mathbf{e}_m]\mathbf{y}^0 \tag{17.31}$$

as long as

$$\mathbf{y}^0 \in Y(\lambda^0 + \delta). \tag{17.32}$$

Equation (17.31) can be written as

$$v(\lambda^0 + \delta) = v(\lambda^0) - \delta \mathbf{e}_m \mathbf{y}^0, \tag{17.33}$$

which reveals that $v(\lambda)$ has slope $-\mathbf{e}_m \mathbf{y}^0$ as long as (17.32) holds.

Marsten observes that if $\mathbf{y}^0 \in P_p$ is a vector of zeroes and ones, then $\mathbf{e}_m \mathbf{y}^0$ is simply the number of ones, hence the number of assignments. When $\mathbf{y} = \{\mathbf{0}\}^m$ there are no assignments and therefore n medians are located. Each assignment reduces the number of medians by one. An integer solution $\mathbf{y}^0$ must therefore have $(n - \mathbf{e}_m \mathbf{y}^0)$ medians. Finally, taking $p = n - \mathbf{e}_m \mathbf{y}^0$ and using (17.27) and (17.33), Marsten obtains

$$\begin{aligned} Z_p\left(\lambda^0 + \delta\right) &= (n - p)\left(\lambda^0 + \delta\right) + v\left(\lambda^0 + \delta\right) \\ &= \left[n - \left(n - \mathbf{e}_m \mathbf{y}^0\right)\right]\left(\lambda^0 + \delta\right) + v\left(\lambda^0\right) - \delta \mathbf{e}_m \mathbf{y}^0 \\ &= v\left(\lambda^0\right) + \lambda^0 \mathbf{e}_m \mathbf{y}^0, \end{aligned} \tag{17.34}$$

as long as relation (17.32) holds. Therefore, subject to (17.32), $Z_p(\lambda)$ has a zero slope for $p = n - \mathbf{e}_m \mathbf{y}^0$. This means that λ^0 maximizes the dual objective function $Z_p(\lambda)$, and since $\mathbf{y}^0 \in Y(\lambda^0)$ it follows that $\mathbf{y}^0$ is an optimal solution of the primal problem $\overline{RS}$. Consequently, $\mathbf{y}^0$ is a p-median of the network for $p = n - \mathbf{e}_m \mathbf{y}^0$.

Marsten's argument proves that every integer extreme point $\mathbf{y}^*$ of P_p that belongs to $Y(\lambda^*)$ for some value of λ^* is a p-median of the network for $p = n - \mathbf{e}_m \mathbf{y}^*$. He notes, however, that it is possible that none of the p-medians of a network belong to $Y(\lambda)$ for any value of λ for certain values of p, for example $p = 9$. On the other hand, fractional extreme points of P_p that appear in some $Y(\lambda)$ may be thought of as "generalized" p-medians.

The procedure that Marsten gives for generating the entire $v(\lambda)$ function starts at $\mathbf{y} = 0$ and $\lambda = c^*$ and increases λ, moving from one extreme point of P_p to another, so that the procedure is always at a member of $Y(\lambda)$ for the current value of λ. Every integer extreme point of P_p is a p-median of the network. Marsten notes that these medians are encountered in decreasing order of p, from $p = n$ to $p = 1$, since $v(\lambda)$ is concave with slope $-\mathbf{e}_m \mathbf{y}^0$ for $\mathbf{y}^0 \in Y(\lambda)$. Thus $\mathbf{y} = \mathbf{0}$ belongs to $v(c^*)$ and the procedure is started with a $p = n - \mathbf{e}_m \mathbf{0} = n$ medians.

Marsten then sets to discuss in great detail the pivoting mechanism, where the special structure of P_p is exploited. This is a lengthy discussion, which takes several pages of his report; it is beyond the scope of the present work to reproduce it. In his discussion Marsten addresses: (*i*) the coefficient matrix for the constraints (17.23.2) that determine P_p; (*ii*) the determination of the entering basic variable; (*iii*) the determination of the exiting basic variable.

This detailed discussion is followed by a section on computational results. The author starts with a small test problem of a network having $n=10$ nodes. In this case, the tour passes through exactly 10 extreme points, including the origin. Each of these is an integer extreme point and hence a median of the network; the entire tour took less than 1 second on a CDC 6400 computer. Marsten tested next a 33-node network, with all the node weights being equal. The corresponding distance matrix was taken from Karg and Thompson (1964). Starting at the origin ($p=n$), integer extreme points corresponding to p-medians for $p=33$ down to $p=10$ were encountered. The next extreme point visited, however, was fractional, corresponding to a "median" for $p=9½$.

The tour for the 33-node network did not produce a 9-median or an 8-median. Marsten states that, in general, the series of solutions that *are* available should make it easy to find the missing ones by means of a branch-and-bound search, but no such experience is reported. He also notes that the computational burden increases as the number of medians decreases. This is apparently because the necessary changes in the configuration of the solution become more drastic as the value of p decreases; the number of basis changes preceding a breakthrough increases as the number of medians decreases.

It is unfortunate that, to the best of our knowledge, this work of Marsten was never published as a paper, which would have made it available to a much wider audience than in the form of an internal report. The report itself is very well documented, not only in the formal development of his proposed procedure, but also with detailed results related to Marsten's computational experiments.

17.3.3.2 Impact on Future Developments

We are not aware of direct impacts of this work of Marsten on future developments related to the p-median problem. A possible indirect impact of his work is the decomposition formulation of Garfinkel et al. (1974). These authors solved the linear programming relaxation of *RS* by decomposition, thus considerably reducing the size of the problem. In their decomposition formulation, the linear programming basis of the master problem contains only $(n+2)$ rows, and each of the n subproblems can be solved by inspection. Due to the very degenerate nature of the linear programming basis of the master problem, however, in many cases the algorithm fails to converge. This lack of convergence is a very serious problem, and prevents the decomposition formulation from effectively solving the problem; see, e.g., Galvão (1981).

The linear programming decomposition formulation represents only part of the Garfinkel et al. (1974) paper. In cases of noninteger termination of the linear program, integrality is achieved through group theoretic techniques and a dynamic programming recursion. Garfinkel et al. report some computational experience with their proposed procedures.

17.4 A Survey of Subsequent Work

There are numerous papers related to the use of Lagrangean relaxation in facility location problems. A search of the *Scopus* database, using "Lagrangean relaxation *and* location" as keywords, produces 74 references to papers published between 1978 and 2007. Twenty-seven of these papers are related to capacitated facility location, both with and without single-source constraints. These are followed by papers related to general facility location problems (10 papers), design of communication/computer networks (8 papers), and design of mobile/wireless networks (4 papers). Classical location problems such as the simple plant location problem, the *p*-median, the *Maximal Covering Location Problem* (MCLP) and hierarchical and dynamic location problems are also in the list. Finally, there are 8 papers that may be considered of theoretical nature.

In relation to classical location problems we may cite the Lagrangean dual ascent heuristic developed by Guignard (1988) for the simple plant location problem, a comparison of Lagrangean and surrogate relaxations for the maximal covering location problem by Galvão et al. (2000), the development of dual-based heuristics for a hierarchical covering location problem by Espejo et al. (2003), a maximal covering location model in the presence of partial coverage by Karasakal and Karasakal (2004), and a branch-and-price approach to *p*-median location problems by Senne et al. (2005).

It is clearly beyond the scope of the present chapter to make a detailed survey of all of these works. Nevertheless, since capacitated problems are very important, it is worthwhile to take a closer look at the corresponding applications of Lagrangean relaxation. We will concentrate on the variant of the capacitated facility location problem with single source constraints, i.e., problems for which the demand of each customer must be totally supplied from a single facility, a problem referred to here as *CFLP_SSC*.

Let q_j be the demand of customer $j \in J$ and Q_i the capacity of a facility located in $i \in I$. Using the notation defined for the simple plant location problem in Sect. 17.3, the problem may be formulated as

$$CFLP\_SSC\_1\text{: } Z_{CFLP\_1}(\min) = \sum_{i=1}^{m} f_i y_i + \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij}$$

$$\text{s.t. } \sum_{i=1}^{m} x_{ij} = 1 \ \forall j \tag{17.35.1}$$

$$\sum_{j=1}^{n} q_j x_{ij} \leq Q_i y_i \ \forall i \tag{17.35.2}$$

$$x_{ij} \in \{0,1\}\ \forall i,\ j$$

$$y_i \in \{0,1\}\ \forall i$$

If we choose to relax restrictions (17.35.1) using Lagragean multipliers $u_j \geq 0, j \in J$ (remembering that $\sum_{i=1}^{m} x_{ij} = 1$ may be replaced by $\sum_{i=1}^{m} x_{ij} \geq 1$ if we consider $f_i \geq 0\ \forall\ i$ and $c_{ij} \geq 0\ \forall\ i, j$), we obtain obtain the Lagrangean problem

$$LR1\_CFLP\_SSC: Z_{LR1\_CFLP}(\min) = \sum_{i=1}^{m} f_i y_i + \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij} + \sum_{j=1}^{n} u_j \left(1 - \sum_{i=1}^{m} x_{ij}\right)$$

$$\text{s.t.} \sum_{j=1}^{n} q_j x_{ij} \leq Q_i y_i\ \forall i \tag{17.36}$$

$$x_{ij} \in \{0,1\}\ \forall i,\ j$$

$$y_i \in \{0,1\}\ \forall i.$$

This problem decomposes into m subproblems of the form

$$Z^i_{LR1\_CFLP}(\min) = \sum_{j=1}^{n} (c_{ij} - u_j)x_{ij} + f_i y_i \tag{17.37}$$

$$\text{s.t.} \sum_{j=1}^{n} q_j x_{ij} \leq Q_i y_i$$

$$x_{ij} \in \{0,1\}\ \forall\ i, j$$

$$y_i \in \{0,1\}\ \forall\ i.$$

If $y_i = 0$, then $Z^i_{LR1\_MCLP} = 0$. If $y_i = 1$ then (17.37) is a zero-one knapsack problem with $Z^i_{LR1\_MCLP} = z_i^*$, where z_i^* is the optimal solution of knapsack problem i.

Now, if we were to add to *LR1_CFLP_SSC* the redundant constraints $\sum_{i=1}^{m} Q_i y_i \geq \sum_{j=1}^{n} q_j$, a stronger relaxation *LR2_CFLP_SSC* is obtained, which also decomposes into m zero-one knapsack problems in variables x_{ij}, one for each $i \in I$. Finally, an alternative model to the one given by (17.35) may be defined as

$$CFLP\_SSC\_2: Z_{CFLP\_2}(\min) = \sum_{i=1}^{m} f_i y_i + \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij}$$

$$\text{s.t.} \sum_{i=1}^{m} x_{ij} = 1 \ \forall j \tag{17.38.1}$$

$$\sum_{j=1}^{n} q_j x_{ij} \leq Q_i y_i \ \forall i \tag{17.38.2}$$

$$x_{ij} \leq y_i \ \forall i, j \tag{17.38.3}$$

$$x_{ij} \in \{0,1\} \ \forall i, j$$

$$y_i \in \{0,1\} \ \forall i.$$

Let now $v_i \geq 0, i \in I$ be the Lagrangean multipliers associated with constraints (17.38.2). The relaxation of this set of constraints results in the problem

$$LR3\_CFLP\_SSC: Z_{LR3\_CFLP}(\min)$$
$$= \sum_{i=1}^{m} f_i y_i + \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij} x_{ij} + \sum_{i=1}^{m} v_i \left(\sum_{j=1}^{n} q_j x_{ij} - Q_i y_i \right)$$

$$\equiv \sum_{i=1}^{m} (f_i - Q_i v_i) y_i + \sum_{i=1}^{m}\sum_{j=1}^{n} (c_{ij} + q_j v_i) x_{ij},$$

or, if we define $F_i = f_i - Q_i v_i$ and $C_{ij} = c_{ij} + q_j v_i$

$$Z_{LR3\_CFLP}(\min) \equiv \sum_{i=1}^{m} F_i y_i + \sum_{i=1}^{m}\sum_{j=1}^{n} C_{ij} x_{ij}$$

$$\text{s.t.} \sum_{i=1}^{m} x_{ij} = 1 \ \forall j$$

$$x_{ij} \leq y_i \ \forall i, j$$

$$x_{ij} \in \{0,1\} \ \forall i, j$$

$$y_i \in \{0,1\} \ \forall i,$$

which corresponds to the formulation of the simple plant location problem in (17.1).

Barceló and Casanovas (1984) proposed a Lagrangean heuristic to solve the problem *CFLP_SSC* in which the maximum number of facilities p is predefined and part of the model (using constraint $\sum_{i=1}^{m} y_i \leq p$). They present 3 Lagrang-

ean relaxations of *CFLP_SSC*: *LR*1_*CFLP_SSC*, *LR*3_*CFLP_SSC*, and a relaxation that dualizes constraints (17.35.1) and the restriction on the maximum number of facilities. They use *LR*1_*CFLP_SSC* and a two-phase heuristic to find approximate solutions to the problem.

Klincewicz and Luss (1986) developed a different Lagrangean heuristic to solve *CFLP_SSC*. They use the relaxation *LR*3_*CFLP_SSC*; the Lagrangean problems are simple plant location problems, which they solve by means of the *DUALOC* method of Erlenkotter (1978), but without using that method's branch-and-bound procedure. The initial solution is given by an *ADD* heuristic; another heuristic, based on differential costs, makes the necessary adjustments in an attempt to find better solutions through better allocations of clients to facilities. Klincewicz and Luss tested their method using 12 problems derived from the test problems created by Kuehn and Hamburger (1963), having 25 and 26 potential facility sites and 50 clients; the results obtained were of good quality.

A solution procedure based on relaxation *LR*2_*CFLP_SSC* was developed by Pirkul (1987) for the concentrator location problem. According to Beasley (1993), this solution procedure produced the best solutions available until then for constrained facility location problems with single source constraints. Pirkul presents two heuristic procedures to find approximate solutions to this problem: one involving two phases (in the first phase locations of the concentrators are determined, followed by the allocation of terminals to the concentrators in the second phase), the other attempting to find a primal feasible solution in each step of the subgradient optimization method that solves the Lagrangean dual.

Chen and Guignard (1998) study the polyhedral structure of two primal relaxations of a class of specially structured mixed integer programming problems. This class includes as special cases the generalized capacitated plant location problem and the production scheduling problem. The authors show that, for this class of problems, two polyhedra, constructed from the constraint sets in two different primal relaxations, are identical. These results have the surprising implication that the bounds from two a priori different primal relaxations of the capacitated plant location problem are actually equal. This means that a simple Lagrangean substitution yields exactly the same strong bound as the computationally more expensive Lagrangean decomposition method introduced by Guignard and Kim (1987).

Finally, Cortinhal and Captivo (2003) developed a method that uses Lagrangean relaxation, a Lagrangean heuristic and local and tabu searches to find lower and upper bounds to *CFLP_SSC*. They use relaxation *LR*2_*CFLP_SSC*; the Lagrangean problem decomposes into knapsack problems. They solve the knapsack problems using the *MT2* code of Martello and Toth (1990). The solutions of the Lagrangean problems are submitted to a two-phase heuristic. The first phase finds feasible solutions to the original problem. The second phase, consisting of a local search and a tabu search meta-heuristic, improves the solutions obtained in Phase I. The combined use of a Lagrangean heuristic and tabu and local searches produced very good results for several test problems, both created by the authors and available in the literature.

17.5 Conclusions

We have seen that Lagrangean relaxations, and dual-based related techniques, are very important solution strategies for facility location problems. The pioneering works of Bilde and Krarup (1967, 1977), Diehr (1972) and Marsten (1972) were analyzed in detail. The dual-based procedures of Bilde and Krarup and Diehr were proven equivalent to Lagrangean relaxations of the simple plant location problem and the *p*-median problem, respectively.

The method developed by Marsten (1972), in which every *p*-median of a network is shown to be an extreme point of the same polyhedron P_p, takes a grand tour of P_p that passes through most of the special (integer) extreme points and through very few others (fractional extreme points). This tour successively generates the *p*-medians of a network in descending order of *p*, although for some values of *p* the solution may be missed and never generated, or, conversely, extreme points of P_p may be generated which contain fractional values of y_{ii}. Thus, although Marsten's method is both theoretically and computationally attractive, it may fail to produce the *p*-median of a network for the specific value of *p* that may be required.

Important works that followed the classical contributions were then analyzed. We dedicated special attention to the single source capacitated location problem *CFLP_SSC*, which is a problem that has many important practical applications, such as the location of concentrators in a computer network, but is difficult to solve to optimality. Correspondingly, we found out that many of the Lagrangean relaxation-based solution methods for location problems are related to *CFLP_SSC*. We analyzed different relaxations of *CFLP_SSC* and discussed important papers dedicated to this problem.

It was not the purpose of this chapter to make a survey of Lagrangean relaxation applied to location problems. Therefore, although an effort was made to cite the main contributions in this area, many important papers may not have been included. A survey of this area is both important and presently unavailable, so we suggest that this task be undertaken by a researcher familiar with the topic.

Finally, we would like to point out potential future research directions. This is not an easy task, given the numerous possibilities that Lagrangean relaxation-based techniques present in the solution of facility location problems. However, an area that has been growing lately is design of mobile/wireless networks (see for example Wen et al. 2007). This topic is not yet fully developed and may become one of the main fields for future research work.

Acknowledgments R. Galvão would like to thank his Ph.D. student Ormeu Coelho da Silva Junior for helping him with a detailed revision of the original manuscript.

References

Baker BM, Sheasby J (1999) Accelerating the convergence of subgradient optimisation. Euro J Oper Res 117:136–144

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

Bazaraa M, Sherali HD (1981) On the choice of step sizes in subgradient optimization. Euro J Oper Res **7**:380–388

Beasley JE (1993) Lagrangean heuristics for location problems. Euro J Oper Res 65: 383–399

Bilde O, Krarup J (1967) Bestemmelse of optimal beliggenhed af produktionssteder. Research Report, IMSOR, Technical University of Denmark

Bilde O, Krarup J (1977) Sharp lower bounds and efficient algorithms for the simple plant location problem. Ann Discrete Math 1:79–97

Boffey TB, Karkazis J (1984) *p*-medians and multi-medians. J Oper Res Soc 35:57–64

Campbell JF (1994) Integer programming formulations of discrete hub location problems. Euro J Oper Res 72: 387–405

Cánovas L, García SE, Marín A (2006) Solving the uncapacitated multiple allocation hub location problem by means of a dual-ascent technique. Ann Oper Res 130: 163–178

Chen B, Guignard M (1998) Polyhedral analysis and decompositions for capacitated plant location-type problems. Discrete Appl Math 82:79–91

Christofides N, Beasley JE (1982) A tree search algorithm for the *p*-median problem. Euro J Oper Res 10:196–204

Cortinhal MJ, Captivo ME (2003) Upper and lower bounds for the single source capacitated location problem. Euro J Oper Res 151:333–351

Diehr G (1972) An algorithm for the p-median problem. Working Paper No. 191, Western Management Science Institute, University of California, Los Angeles

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

Fisher ML (1981) The Lagrangian relaxation method for solving integer programming problems. Manage Sci 27:1–18

Galvão RD (1980) A dual-bounded algorithm for the *p*-median problem. Oper Res 28: 1112–1121

Galvão RD (1981) A note on Garfinkel, Neebe and Rao's LP decomposition for the *p*-median problem. Transp Sci 15: 175–182

Galvão RD (1993) The use of Lagrangean relaxation in the solution of uncapacitated facility location problems. Location Sci 1:57–79

Galvão RD, Raggi LA (1989) A method for solving to optimality uncapacitated location problems. Ann Oper Res 18:225–244

Galvão RD, Espejo LGA, Boffey B (2000) A comparison of Lagrangean and surrogate relaxations for the maximal covering location problem. Eur J Oper Res 124: 377–389

Garfinkel RS, Neebe AW, Rao MR (1974) An algorithm for the *m*-median plant location problem. Transp Sci 8:217–236

Geoffrion AM (1974) Lagrangean relaxation for integer programming. Math Program Study 2: 82–114

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

Guignard M, Kim S (1987) Lagrangean decomposition for integer programming: theory and applications. RAIRO— Recherche Opérationnelle 21:307–323

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

Hanjoul P, Peeters D (1985) A comparison of two dual-based procedures for solving the *p*-median problem. Eur J Oper Res 20:287–396
Held M, Karp RM (1970) The traveling salesman problem and minimum spanning trees. Oper Res 18:1138–1162
Held M, Karp RM (1971) The traveling salesman problem and minimum spanning trees: part II. Math Program 1:6–25
Held M, Wolfe P, Crowder H (1974) Validation of subgradient optimization. Math Program 6: 62–88
Hogan WW, Marsten RE, Blakenship JW (1975) The boxstep method for large scale optimization. Oper Res 23:389–405
Karasakal O, Karasakal EK (2004) A maximal covering location model in the presence of partial coverage. Comput Oper Res 31:1515–1526
Karg RL, Thompson GL (1964) A heuristic approach to solving travelling salesman problems. Manage Sci 10: 225–248
Klincewicz JG (1996) A dual algorithm for the uncapacitated hub location problem. Location Sci 4: 173–184
Klincewicz JG, Luss H (1986) A Lagrangean relaxation heuristic for capacitated facility location with single-source constraints. J Oper Res Soc 37:495–500
Körkel M (1989) On the exact solution of large-scale simple plant location problems. Eur J Oper Res 39: 157–173
Kuehn AA, Hamburger MJ (1963) A heuristic program for locating warehouses. Manage Sci 9:643–666
Larsson T, Patriksson M, Strömberg A-B (1996) Conditional subgradient optimization: theory and applications. Eur J Oper Res 88: 382–403
Marsten RE (1972) An algorithm for finding almost all of the medians of a network. Discussion Paper No. 23, Northwestern University, Evanston, Illinois
Martello S, Toth P (1990) Knapsack problems: algorithms and computer implementations. Wiley, New York
Mavrides LP (1979) An indirect method for the generalized *k*-median problem applied to lock box location. Manage Sci 25: 990–996
Mayer G, Wagner B (2002) Hub locator: an exact solution method for the multiple allocation hub location problem. Comput Oper Res 29:715–739
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
Narula SC, Ogbu UI, Samuelsson HM (1977) An algorithm for the *p*-median problem. Oper Res 25:709–713
Pirkul H (1987) Efficient algorithms for the capacitated concentrator location problem. Comput Oper Res 14: 197–208
Poljak BT (1967) A general method for solving extremum problems. Sov Math 8: 593–597
Poljak BT (1969) Minimization of unsmooth functionals. U.S.S.R. Comput Math Math Phys 9: 1429
Revelle CS, Marks D, Liebman JC (1970) An analysis of private and public sector location models. Manage Sci 16: 692–707
ReVelle CS, Swain RW (1970) Central facilities location. Geogr Anal 2:30–42
Sarin S, Karwan MH (1987) A computational evaluation of two subgradient search methods. Comput Oper Res 14:241–247
Senne ELF, Lorena LAN, Pereira MA (2005) A branch-and-price approach to *p*-median location problems. Comput Oper Res 32: 1655–1664
Shapiro JF (1979) A survey of Lagrangean techniques for discrete optimization. Ann Discrete Math 5:113–138
Sherali HD, Myers DC (1988) Dual formulations and subgradient optimization strategies for linear programming relaxations of mixed-integer programs. Discrete Appl Math 20:51–68
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

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

Wen Y-F, Lin FY-S, Kuo W-C (2007) A tree-based energy-efficient algorithm for data-centric wireless sensor networks. Proceedings of the IEEE AINA, Ontario, Canada, May 21–23, 2007, Art. No. 4220895, 202–209

Part VII
Customer Choice and Location Patterns

Chapter 18
Gravity Modeling and its Impacts on Location Analysis

Lawrence Joseph and Michael Kuby

18.1 Introduction

Throughout the 20th century, geographers have developed a variety of models that assist public and private entities in locating facilities from factories to emergency services. In the area of retail location analysis, the early gravity models of geographers William J. Reilly and David L. Huff played a pioneering role in delineating retail trade areas and modeling many other kinds of spatial interaction. Their fundamental insight was that customers do not necessarily shop at the closest store, but patronize locations in proportion to the attractiveness of the retail centers and in inverse proportion to their distance. In this chapter, we elucidate the early history, structure, and significance of these models.

Gravity models are based on the laws of Newtonian physics (Young 1975). With respect to the social sciences, Berkeley (1713) and Carey (1858) are believed to have been the first to suggest an association between gravitation and the social implications of spatial interactions. In the early twentieth century, Reilly (1931) introduced his Law of Retail Gravitation following the results of an empirical survey to identify the breaking points of retail influence between two competing cities. Amidst a flurry of controversy over the accuracy and applicability of Reilly's model, Huff followed in 1963 with an alternative gravity model. Depending on the complexity of the chosen model, gravity models are designed to allow flexible criteria choices for application (Wagner 1974). The original gravity models were structured with two key elements: attraction and friction of distance. According to Huff and Jenks (1968), attraction can be measured by population, employment, retail or wholesale sales volumes, or other measures. The second component is to infer some sort of friction of distance, or distance decay. As distance from a center

L. Joseph (✉) · M. Kuby
School of Geographical Sciences and Urban Planning,
Arizona State University, Tempe, AZ 85287-0104, USA
e-mail: Lawrence.joseph@asu.edu

M. Kuby
e-mail: mikekuby@asu.edu

H. A. Eiselt, V. Marianov (eds.), *Foundations of Location Analysis,*
International Series in Operations Research & Management Science 155,
DOI 10.1007/978-1-4419-7572-0_18, © Springer Science+Business Media, LLC 2011

increases, the influence of that center decreases by some value. This may be measured by Euclidian distance, rectilinear distance, travel distance, travel time, travel cost, or other factors (Huff and Jenks 1968; Drezner and Drezner 2002).

The gravity model is obviously a simplification of the real world. Spatial interactions between people and places can be a complex process depending on a variety of factors. The shopping choices of the individual consumer are affected by factors including age, lifestyle, family, quality, price, or convenience (Bucklin 1971). Furthermore, the site and situation of the center can also be influential with respect to factors such as the type and age of center, proximity of competitors, accessibility, visibility, clientele, topography, crime, and socioeconomic factors. Thus, it is essential to account for market structure and the types of products or services for proper use of gravity models.

This chapter commences by presenting this field in chronological order, discussing first how the gravity model was used initially by Reilly to identify breaking points and how Huff transformed the model to predict shopping probabilities from demand points. We then discuss how the models diffused with popularity amidst the deliberations of proper application and calibration. Although gravity models have been popular, they have also been a topic of debate and controversy. The issue is not with the fundamental concept of a gravity model, but in the details of how the models should be structured, especially with respect to the distance-decay parameter in different location settings. There were several empirical studies that focused on how to select an appropriate parameter value to account for this phenomenon. Although our primary focus is to present the original models of Reilly and Huff, we lend some of our discussion to this issue as it affects the utilization and diffusion of the modeling.

18.2 The Classical Contributions

This section will survey the classical contributions by Reilly and Huff. We have kept our discussion as close to the original work as possible.

18.2.1 The Work of William J. Reilly

The Law of Retail Gravitation was a prominent early finding for the location analysis community (Thrall 2002). This law emerged from the efforts of William Reilly and his attempts to identify a reliable means of finding breaking points of retail influence between various sized cities at different distances. In this section, we review Reilly's landmark study from 1931, starting with an explanation of the law. We present the original formulation of his model with accompanying examples and his suggestions regarding its potential applications. In the end, Reilly's original book offers an interesting blend of theoretical and practical geography and marketing.

18.2.1.1 The Law of Retail Gravitation

Reilly's Law of Retail Gravitation relates two key location factors, city population and distance, to identifying the breaking points of retail influence between any two cities. Specifically, the breaking point is the location where the retail trading influence of two cities is identical. As you move closer to either city from the breaking point, that city asserts its retail trading dominance.

The model was based on two general assumptions. First, cities with larger populations attract more outside retail trade than smaller cities and towns. To that end, Reilly discovered that as the population of the cities increased, outside trade increased at a similar rate. For instance, a city with three times the population of a smaller city would attract three times more trade originating from outside the city than the smaller city. Second, cities attract more trade from closer cities and towns than from farther locations. Unlike the population factor, however, Reilly did not discover a one-to-one relationship between distance and retail trade influence; as distances increased, retail trade influence decreased at a faster rate. In fact, trade decreased at a rate nearly twice as fast as distance increased. It should be noted that the application and understanding of the importance of population and distance were not necessarily conceptually new to scholars and practitioners. Indeed, Reilly's major contribution was to identify the specific rates at which these factors affected retail trading influences between cities. The functional form of his law evolved from analyzing the results of a few key empirical studies.

18.2.1.2 Historical Backdrop to the Paradigm Shift in Trade Area Delineation

Somewhat lost in today's academic wrangling over the exact structure and composition of spatial interaction models is the motivation behind the original research. Reilly's contributions to delineating areas of retail trading influence were a response to the proliferation of the automobile among households across the United States. Unexpectedly, greater mobility from smaller towns and cities to larger markets greatly transformed the range and shape of retail trading areas in the early twentieth century. The improved accessibility of smaller towns to larger markets was accompanied by improved standards of living and the increased media outreach to these smaller communities. Populations in smaller cities and towns began to demand the same merchandise as populations in larger cities. While this national phenomenon was occurring, the major stakeholders, including retailers, manufacturers, distributors, and advertisers, were ill-prepared to respond to this dynamic spatial redistribution of consumers. Consequently, Reilly conducted an empirical investigation to analyze how smaller towns and larger cities interacted and the subsequent attraction of retail trade between them. This research eventually led to modeling the Law of Retail Gravitation. Before we elucidate the details of the model, the following discussions highlight the process that contributed to its development.

18.2.1.3 Pilot Study

While Reilly's best-known book was published in 1931, the major findings that led to developing the law were derived from a pilot study that began in 1927 (Reilly 1929). The study analyzed the relationships between retail trade and seven differently sized cities in Texas including Fort Worth, Dallas, Houston, Austin, San Antonio, El Paso, and Waco. More specifically, this survey was designed to demarcate the types of people and products that were involved with retail trade between smaller and larger cities, and the types of products these cities were capable of supporting. In addition, Reilly sought to identify a reliable method for measuring the extent of a city's retail trading influence.

At the time of his study, Reilly was not fortunate enough to commence his research with a complete list of customer location data. He realized, however, that there would be a need for some form of customer data in order to generate reliable results. Reilly started by examining transaction data obtained through associations of retail merchants. In fact, these associations provided customer lists from cooperating retail stores, including customers from outside the city where the store was located. Furthermore, he utilized data from charge accounts from the leading stores to map the spatial distribution of the stores and their respective account holders. To ensure accuracy, he followed up on this mapping with door-to-door field surveying. This allowed Reilly to calculate the proportion of charge accounts that were outside the seven studied Texas cities. Not surprisingly, Reilly found that some customers would often travel through smaller cities to patronize stores in larger cities. For instance, he found that there were numerous customers of stores who would pass through Fort Worth from the west to shop in Dallas. In an isolated city such as El Paso, however, he found that the retail influence would stretch for several hundreds of miles before another city attracted more trade. Finally, he compared newspaper circulation with charge account locations for the seven Texas cities and 1,204 surrounding cities and towns. The results matched fairly closely, but there were still enough differences to necessitate further investigation. In other words, newspaper circulation was not sufficient as the only means of measuring retail trading influence.

In this pilot study, Reilly came to the conclusion that retail influence could not simply be explained by a two-dimensional flat surface with assignment of consumers to their nearest retail outlets. Instead, it needed to be examined in terms of three dimensions including length, width, and height. The height is based on the population of the cities. It allows the analyst to visualize the retail influence of different sized cities at various distances. This would become an important component to his continuing research.

18.2.1.4 The National Study

Following his initial empirical findings, Reilly extended his study area to the United States as a whole. For this endeavor, he used data provided by the secretaries of the retail credit associations for various cities to determine the trade to and from these cities. These data provided the evidence for proper delineation of breaking points

of retail trade influence between the cities. Furthermore, Reilly and his colleagues drove on the connecting highways between the larger cities while using the credit data to identify the breaking points. At these locations, he conducted field checks to identify any potentially influential local factors. In particular, he conducted interviews in the communities with local merchants' associations as well as door-to-door interviews. Although it may seem that the charge account data would have been sufficient for determining the breaking points and that field investigation would be unnecessary, Reilly noted that these data were relative and not absolute, and therefore, he still sought to test it empirically.

In an effort to reduce the burden on his audience, Reilly provided a straightforward matrix for determining the breaking points between two cities. An abbreviated example from Reilly (1931) is provided in Table 18.1.

The first row and first column represent the urban-area population of Cities *A* and *B* respectively, including the city and its suburbs, in thousands. The body of the matrix contains a decimal value between 0 and 1 representing the location of the breaking point as a percentage of the distance from City *B* towards City *A*. We have shaded the cells where the breaking point is exactly half of the distance from *B* to *A*, and notice that the populations of these cities are equal. To use the matrix, line up the population values for any two studied cities and identify the corresponding cell of the matrix. Then, take that decimal value and multiply it by the travel distance between the cities to identify the breaking point from City *B*. This method is easily understood and applicable for any pairwise city combination. For example, consider an attempt to identify the breaking point between City *A* with a population of one million and City *B* with a population of two million. According to Table 18.1, the

Table 18.1 Reilly's breaking point factors. (Abbreviated example from Reilly 1931)

		POPULATION OF CITY *A* in (1000s)										
		700	**750**	**800**	**850**	**900**	**950**	**1000**	**1100**	**1200**	**1300**	**1400**
POPULATION OF CITY *B* in (1000s)	**1100**	.55	.55	.54	.53	.52	.52	.51	.50	.49	.48	.47
	1200	.57	.56	.55	.54	.54	.53	.52	.51	.50	.49	.48
	1300	.58	.57	.56	.55	.55	.54	.53	.52	.51	.50	.49
	1400	.58	.58	.57	.56	.56	.55	.54	.53	.52	.51	.50
	1500	.59	.58	.58	.57	.56	.56	.55	.54	.53	.52	.51
	1600	.60	.59	.58	.58	.57	.56	.56	.55	.54	.53	.52
	1700	.61	.60	.59	.59	.58	.57	.57	.56	.55	.54	.53
	1800	.62	.61	.60	.59	.58	.58	.57	.56	.55	.54	.53
	1900	.62	.61	.61	.60	.59	.58	.58	.57	.56	.55	.54
	2000	.63	.62	.61	.60	.60	.59	.58	.57	.56	.55	.54
	2200	.64	.63	.62	.62	.61	.60	.60	.58	.57	.56	.55
	2400	.65	.64	.63	.63	.62	.61	.61	.59	.58	.57	.57
	2600	.66	.65	.64	.64	.63	.62	.62	.60	.59	.58	.58
	2800	.67	.66	.65	.65	.64	.63	.63	.61	.60	.59	.58
	3000	.67	.67	.66	.65	.65	.64	.63	.62	.61	.60	.59
	3200	.68	.67	.67	.66	.65	.65	.64	.63	.62	.61	.60
	3400	.69	.68	.67	.67	.66	.65	.65	.64	.63	.62	.61
	3600	.69	.69	.68	.67	.67	.66	.66	.64	.63	.62	.62
	3800	.70	.70	.69	.68	.67	.67	.66	.65	.64	.63	.62
	4000	.70	.70	.69	.68	.68	.67	.67	.66	.65	.64	.63

Table 18.2 Breaking points (B.P.) between various cities. (From Reilly 1931)

Between[a]	Location of B.P. discovered in field study	Automobile highway distance (in miles) from each city to B.P. based upon	
		Field study	Application of law
Atlanta, GA… (270,367)	Collier, GA	64	66
Macon, GA… (53,866)		31	29
Atlanta, GA… (270,367)	Heflin, AL	87	89
Birmingham, AL… (257,657)		87	85
Austin, TX… (53,118)	Salada, TX	55	55
Waco, TX… (52,825)		55	55
Buffalo, NY… (620,007)	Westfield, NY	58	60
Erie, PA… (115,922)		29	27

[a] Population given immediately after each city–1930 Census

breaking point factor is 0.58. If the travel distance between these cities is 100 miles, the breaking point would be 58 miles away from City *B* and 42 miles away from City *A*.

Interestingly enough, Reilly did not hesitate to assert confidence in his breaking points. In fact, he provided several examples of how accurately his model chose the breaking points when compared to the field surveys. These comparative figures are given in Table 18.2. He notes that this methodology is applicable throughout the entire U.S. when comparing cities in the same region. However, according to Reilly, one of the limitations of the model is that is does not account for impedances. For instance, he provides an example of how a toll bridge crossing the Ohio River altered the breaking point between Indianapolis, IN and Louisville, KY. Thus, it appeared fewer Hoosiers were willing to pay the toll to shop at stores in Kentucky.

As previously stated, Reilly's model is based solely on two primary factors: distance and population. Although he does recognize the existence of many other secondary factors, he deemed them unnecessary to include in the model as the primary factors were powerful enough to identify the breaking points. Altogether, Reilly asserts his confidence that he is presenting a firm "law" and not merely a theoretical discussion. Of course, other researchers would eventually describe the limitations of his model while presenting their own modifications and extensions. We will discuss these later in this chapter. Still, Reilly's confidence was such that he openly challenged any reader to dispute his breaking points in the U.S., referring to any incorrect prediction as an "exceptional case."

18.2.1.5 The Original Model

Reilly presents a rather simple and straightforward quantitative model. Mathematically, to revisit the components, Reilly states "that the amount of outside trade enjoyed by a city increases directly in proportion to some power of the population" and "that the amount of outside trade which a city draws from a surrounding town varies

inversely in proportion to some power of the distance of the town from that city." Towards that end, the original model is presented with the following Eq. (18.1).

$$\frac{B_a}{B_b} = \left(\frac{P_a}{P_b}\right)^N \times \left(\frac{D_b}{D_a}\right)^n \tag{18.1}$$

where

B_a: the business which City A draws from any given intermediate town,
B_b: the business which City B draws from that intermediate town,
P_a: population of City A,
P_b: population of City B,
D_a: distance of City A from the intermediate town, and
D_b: distance of City B from the intermediate town.

Reilly presents Eq. (18.1) as a means to quantify his law. Much of his discussion is reserved for explaining N and n, as these are the exponents of population and distance respectively. N explains the rate at which outside trade increases as the population of the city increases, while n explains the rate at which outside trade decreases as distance from the city increases. From the field observations, Reilly found that as the size of a city's population increases, its retail trading influence increases at a similar rate (i.e., linearly), and thus allows for the use of the first power or $N=1$. To solve for n, however, Reilly uses the Eq. (18.2).

$$\left(\frac{D_b}{D_a}\right)^n = \frac{B_a}{B_b} \times \frac{P_b}{P_a} \tag{18.2}$$

$$n \log \frac{D_b}{D_a} = \log\left(\frac{B_a}{B_b} \times \frac{P_b}{P_a}\right)$$

$$n = \frac{\log\left(\frac{B_a}{B_b} \times \frac{P_b}{P_a}\right)}{\log\left(\frac{D_b}{D_a}\right)}$$

Reilly found that n was consistently between 1.5 and 2.5, and that these values were ubiquitous throughout the entire U.S. "with no exceptions." Accordingly, he uses the integer number 2, which indicates that a city's retail trading influence decreases faster than the distance from a city increases, or, more specifically, the influence decreases with the square of distance.

Reilly recognizes several individualistic secondary factors for each of the cities that may have affected the breadth of their retail trading influence. Some of the factors includes transportation, communication, type of consumer, population density, proximity to larger markets, business attractions, social amusement attractions, competition, topography, climate, and business leadership. Specifically, Reilly

highlights that communication and transportation were considered important influences over the spatial distribution of retail trade. Items such as a newspaper would not exist, however, without a sufficient population base. Furthermore, convenience is directly linked to distance and the time it takes to travel from origin to destination, which depends on the existing transportation network.

18.2.1.6 Applicable Uses and Intended Audience

The process of identifying breaking points can provide several opportunities for analysis. In fact, the model has been useful for items such as trade area delineation, newspaper circulation, and determining fiscal territories for manufacturing. The following discussions highlight some of these uses as put forth by Reilly.

Store location affects what segments of a community a retailer can serve. Strategies for advertising and product mix can then be designed to target those specific neighborhoods. Reilly presented how his model can assist the development of selective growth strategies for retailers. This includes choosing the right large cities for expansion. Along those lines, the gravity model is useful for determining the spatial distribution of shoppers by specific product and service categories. In general, consumers in smaller cities and towns travel to larger cities especially for style and specialty goods. Thus, the trade areas were larger for higher-end goods. Notwithstanding, if stores in smaller towns decided to offer such higher-end items, Reilly suggested (with an insight that pre-dated Christaller's threshold concept in central place theory, see Chap. 20 in this volume) that they would be unsuccessful because these small towns do not have enough population to support these types of goods. There are some markets, therefore, that could only offer common goods. This suggestion emerged in response to many small towns that were attempting to attract more higher-end businesses to their communities. Since these communities lacked the necessary population base to support the style and specialty merchandise, however, the stores were unable to offer deep lines of merchandise and diverse selections. Consequently, Reilly described how retailers may actually end up alienating their customers from shopping in the community altogether by a creating a negative shopping experience. This could also spread to avoiding stores even for common goods because shoppers may become frustrated when they are unable to find the products they want.

Given the era of Reilly's original studies, deciding in which newspapers to advertise was one of the more important marketing decisions for retailing businesses. Alternatively, newspapers were attempting to extend their circulations to the extent of the retail trading areas of the larger stores in their cities. Many times, the area served by the newspaper circulation may only be in the immediate area of a particular city, though the breadth of that city's retail influence may stretch much farther, or vice versa. The importance of this debate centered on whether it was worth the additional costs because of higher advertising rates for retail stores. By merging trade areas with newspaper circulation, retailers could eliminate waste in their advertising budget. If a newspaper circulated to an area where the city did not have

retail trading influence, then it might not be worth the expense to advertise in that paper. There were also implications concerning the extent of trading influence for particular categories. It may be worthwhile to advertise goods and services such as fur coats or theatrical plays in a newspaper with a large circulation area, but not a neighborhood bakery or barber shop.

Reilly also noted how his model could be useful for manufacturers. "Budget territories" could be based on consumer purchasing habits for various products and distribution channels. For instance, it is important to know which large cities attract the most stylish goods. He also explained how individuals from smaller towns often purchased their common goods in large cities for various reasons, such as the assumption that goods were less expensive in larger cities. Furthermore, customers might purchase some common goods while shopping for higher-end goods, a phenomenon we now refer to as "multi-purpose shopping" (O'Kelly 1983). Finally, many customers from outside the city patronized stores near their places of employment in the city.

Reilly provided his law as a means to "stimulate many minds to conceive possible uses of the law within the sphere of their own interests." While Reilly certainly accomplished this feat, he also stimulated many location researchers to alter his model and identify its limitations. To that end, David L. Huff's alternative model would further transform quantitative methods for trade area delineation of shopping centers.

18.2.2 The Work by David L. Huff

Three decades following the introduction of Reilly's Law of Retail Gravitation, David L. Huff developed an alternative model for trade area estimation. Although Huff credited the "gravitationalists" such as Reilly for their important contributions, he concluded that there was a dearth of models that could be utilized for empirical testing. Furthermore, he opined that the existing models at the time were ambiguous and debatable. Towards that end, Huff (1963, 1964) presented a conceptually superior means of predictive analysis in what would become a very influential study in retail research. First, he analyzed existing trade area delineation techniques. This was followed with the presentation of an alternative technique that would come to be referred to as the Huff Model. The model would eventually be adopted widely with significant implications for retail location decisions. Although Huff has been actively involved in both geography and marketing at the University of Texas, Austin, his analytical methods have been utilized by not only scores of businesses, but also a number of public and private organizations including the U.S. Census Bureau, U.S. Department of Transportation, cultural and health institutions, and many others (Huff 2003). In this section, we discuss the development of the Huff Model including the original quantitative presentation. We follow by extending the discussion to how the model has been influential for both academics and practitioners.

18.2.2.1 Existing Methods and Limitations

Huff provided a critical review of existing trade area delineation techniques. First, he discussed the process of surveying. This generally involved interviewing households or individuals at the point of sale to collect various data. These data may include what types of products or services were purchased, frequency of shopping trips, and home locations of customers. Following such a survey, trade areas could be drawn from the spatial distribution of the customers. The researchers reached the following consistent conclusions. First, distance to the shopping center was a key determinant of the proportion of customers that patronize that shopping center. Second, shopping centers with deeper and more diverse product lines were drawing sales from a larger geographic region. It should also be noted that different types of products and services had varying ranges of distances that customers were willing to travel. Lastly, competing opportunities were affecting the directional "pull" of shopping centers.

Huff was critical of basing location decisions on analogous conclusions from empirical studies. For instance, as noted from the surveying, customers may only be willing to travel a few miles for general food purchases but they might travel much farther for higher-end goods such as furniture or automobiles. Indeed, Huff objected to analysts relying on the same constant parameters for all types of shopping journeys because different products and services may have unique trading areas. Huff was also not as confident as Reilly with respect to excluding the secondary location factors from the analysis. In fact, Huff elucidated a number of factors that can affect the accuracy of modeling with fixed assumptions. Huff presented a critical quantitative review of Reilly's model, and reviewed the work of P.D. Converse.

18.2.2.2 The Influence of P.D. Converse

Huff credited Converse (1949) with modifying Reilly's formula making it easier to calculate the breaking points of retail influence between two competing cities. According to Huff, the advantage of Converse's formula is that it expedited the process of identifying the breaking points.

$$D_b = \frac{D_{ab}}{1 + \sqrt{\frac{P_a}{P_b}}} \tag{18.3}$$

where

D_b: the breaking point between City A and City B in miles from B,
D_{ab}: the distance separating City A from City B,
P_b: the population of City B, and
P_a: the population of City A.

Although Huff credited Reilly and Converse for developing a "systematic basis" for trade area delineation, he believed that there were limitations to these models both conceptually and operationally. One of his objections was that analysts were treating the model as all-or-nothing, where all the potential sales of one trading area were assigned to only one particular store or city. Moreover, there were no potential sales assigned to the store or city outside that trading area. Certainly this is unrealistic; Huff believed it was more accurate to use gradual declines of sales potential as distances increased to the cities or shopping centers. There was also the quandary of dealing with multiple trading areas in a given geographical area. In reality, trade areas of retail stores frequently overlap. Since the objective of Reilly's formula is to only identify the breaking points, however, it is inconsistent in those regions with overlapping trade areas. Consequently, Huff asserted the impracticalities of objectively appraising the total demand from a particular city or shopping center using the Reilly type gravity model.

18.2.2.3 The Huff Model

Huff addressed the aforementioned limitations by presenting an alternative probabilistic model. The Huff Model focuses on the origin or customer data as opposed to the destination or shopping center data. The model seeks to explain how customers make their patronage decisions among competing opportunities for products and services. This model (18.4) estimates the probability that a customer at an origin point i will shop at a retail center j.

$$P_{ij} = \frac{\dfrac{S_j}{T_{ij}^{\lambda}}}{\sum_{j=1}^{n} \dfrac{S_j}{T_{ij}^{\lambda}}} \tag{18.4}$$

where

P_{ij}: denotes the probability of a consumer at a given origin i traveling to a particular shopping center j,
S_j: is the size of a shopping center j (measured in terms of the square footage of selling area devoted to the sale of a particular class of goods),
T_{ij}: indicates the travel time involved in getting from a consumer's travel base i to a given shopping center j, and
λ: is a parameter which is to be estimated empirically reflecting the effect of travel time on various kinds of shopping trips.

Equation (18.4) can be further modified to estimate the total number of expected customers from i that would patronize retail center j. This simple extension is provided by (18.5), which multiplies P_{ij} by the number of customers at point i. This

modification can be particularly useful in estimating sales as well as predicting cannibalization of sister stores.

$$E_{ij} = P_{ij} \times C_i \tag{18.5}$$

where E_{ij} denotes the expected number of consumers at i that are likely to travel to shopper center j, and C_i is the number of consumers at i.

The Huff Model is not simply formulated from empirical data. Instead, it is based on an abstract theory of the geographical nature of customer behavior. The Huff Model also allows for the simultaneous estimation of probabilities of many retail centers at once. Furthermore, the model can be customized by assigning the parameter λ to any given power. This allows the analyst to control for the unique trading areas of various products and services. In a pilot study, Huff found that customers are willing to travel longer for furniture than for clothes. Larger values of λ represent a more rapid decay of patronage as travel time increases. Thus the parameter for shopping for clothing (3.191) was larger than for furniture (2.723).

Equations (18.4) and (18.5) are designed to permit graduated demand at different points, ranging from zero (no demand) to one (all demand). Thus, it is possible to have overlapping trade areas while at the same time identifying the breaking points of retail influence among competing shopping centers. The breaking points are where the probability values are equal for two or more shopping centers and half the consumers travel to each center, as illustrated in Fig. 18.1. Of course, this assumes that there are no other points farther away that capture greater demand. Thus, identifying the breaking points with this method would not account for additional pockets of demand outside the zone of continuous decline.

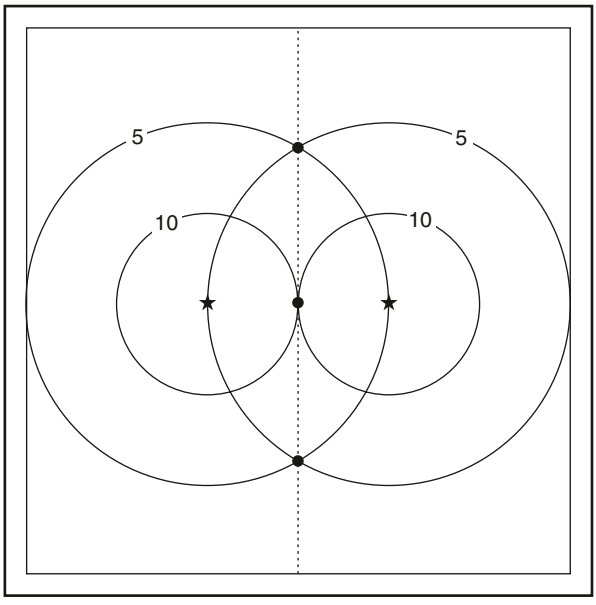


Fig. 18.1 Identifying breaking points using probability isolines. (From Huff and Jenks 1968)

Next, Huff developed Eq. (18.6) to sum up the trade a shopping center j would capture from all surrounding population centers i. This equation incorporated the spatial distribution of customers and their associated demand.

$$T_j = \sum_{i=1}^{n} (P_{ij} \times C_i) \tag{18.6}$$

where

T_j: is the trading area of a particular firm or agglomeration of firms j, that is, the total expected number of consumers within a given region who are likely to patronize j for a specific class of products or services (Huff referred to T_j as the "trading area" but it is really more of a market size as measured in customers),

P_{ij}: denotes the probability of an individual consumer residing within a given gradient i shopping at j, and

C_i: represents the number of consumers residing within a given gradient i.

The following points summarize Huff's definition of a retail trading area:

- the trade area represents a demand surface that consists of potential customers for particular products or services from a particular shopping center,
- a shopping center, also referred to as a distribution center, may consist of a single firm or several firms (agglomeration),
- the demand surface may consist of many subareas (demand gradients) with individual levels of demand (sales potential),
- the demand gradients are based on probabilities, ranging from zero to one,
- trade areas may overlap,
- the point where demand is equal for competing firms constitutes a "spatial competitive equilibrium," and
- the total market size for a shopping center is the sum of the expected consumers from the various surrounding cities and towns.

A few years later, Huff and Jenks (1968) displayed the advantages of illustrating the model using a three-dimensional surface instead of just two dimensions (Fig. 18.2).

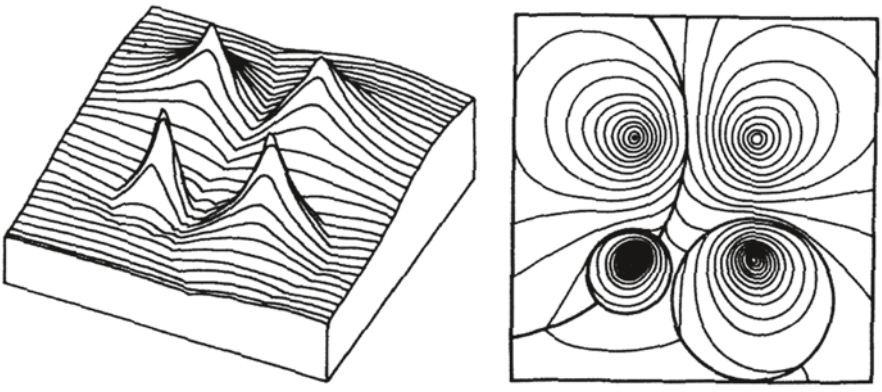


Fig. 18.2 Illustrating probabilities in three dimensions. (From Huff and Jenks 1968)

Greater heights signify higher probabilities. At each point of attraction, there is an attraction-distance ratio that provides the probability at that point. The three-dimensional graphic makes it easier to understand the spatial behavior of shoppers, especially for multiple center comparisons (Huff and Jenks 1968).

18.3 Implications of the Work by Reilly and Huff

The Huff Model is easy to use and can be applied to a variety of problems, even beyond delineating trade areas (Huff 2003). Still, the model was developed and is best known for its retail applications, and its popularity remains on the rise. Many fields and types of organizations are adopting the model. In the contemporary retail real estate market, for example, there may be more analysts using the Huff Model than ever before. This is partially a result of the development and proliferation of Geographic Information Systems (GIS), which offer built-in routines to apply the Huff Model (Huff 2003). In fact, David Huff joined Environmental Systems Research Institute, Inc. (ESRI) in 2003 as an advisor and technical contributor to the Business Analyst extension of ArcView (Huff 2003). With these tools, analysts can easily determine the proportion of customers in a particular neighborhood shopping at a particular store and the proportion of a particular store's customers that come from a particular neighborhood.

The Huff Model took its time to diffuse and has evolved over the years. The same could be said about Reilly's Law of Retail Gravitation. One of the key components of these models is the distance-decay parameter. Unfortunately, some early adopters of these models had issues dealing with appropriate spatial applications and the associated parameter estimation. We next describe the diffusion of calibration methods and applications.

The two most prominent debates about gravity models involved parameter estimation and acceptable spatial applications. Douglas (1949) found Reilly's model to be very precise compared to credit and banking data. Still, Reilly's law was designed to identify the breaking point between two differently sized cities surrounded by rural lands. Wagner (1974) listed how the model was being applied in much broader situations such as within market areas of one urban region. For instance, Reynolds (1953) and Converse (1953) engaged in series of disagreements over the success of the model in Iowa. Reynolds found that the model accurately predicted retail trade in southwestern Iowa, but not across the entire state. Converse, however, refuted the results, claiming the model should only be used for fashion shopping and, when used properly, it worked for the whole state.

The Huff Model was introduced as an alternative means to accurately appraise a center's influence in various location situations. Although this alternative model was garnering new attention from academics and practitioners, Reilly's original contribution was still at the forefront of trade area analysis in the 1960s and 1970s. Wagner (1974) empirically tested the accuracy of Reilly's model in central Ohio for thirteen goods and services from specialty to convenience for two markets: Spring-

field (population 83,723) and Columbus (475,316). While Wagner's 1974 study came out over a decade after the Huff model, it actually began eight years earlier. Over a succession of years (1964, 1967, and 1972), he interviewed 50 people in both the geographic center and at the breaking point calculated according to Reilly's law. While Columbus, the larger city, received more of the trade at the geographic midpoint as expected, Springfield received a disproportionate amount of trade at the theoretical breaking point, where the probabilities should have been equal. There was some variation, however, with respect to different products and services: customers would travel greater distances for music events and holiday shopping. The greatest lesson learned from Wagner's study was that the Reilly model is not as precise when comparing cities with greatly different populations. Instead of using population, Wagner (1974) suggested using square footage, sales, or advertising spending. By using population, the distance-decay value of two was not very precise.

Drezner and Drezner (2002) introduced a method to validate the gravity model by inferring attractiveness. This method was based on a set of independent variables including retail center attractiveness, buying income, and distance from customers to retailers, with market share as the dependent variable. In order to measure attractiveness, Drezner and Drezner (2002) obtained data for buying income, distance from customers to retailers, and market share, calculated from sales volumes obtained from secondary sources. Still, they affirm that sales volumes are not a sufficient sole measure to calculate attractiveness and that sales should be adjusted by key demographic factors such as buying income. In other words, the attractiveness of a retail center is affected by its trade area characteristics (Drezner and Drezner 2002). They tested their method for ten retail centers in Orange County, CA. Their results matched closely when compared with an independent survey.

Numerous academics, such as Bucklin (1971) and O'Kelly and Miller (1989), have concluded that using a distance-decay exponent of two may lead to inaccuracies. While Gautschi (1981) and Eppli and Shilling (1996) found that the distance parameter was overstated, Lee and Pace (2005) found the parameter was understated. Fotheringham (1981) suggested that the parameter is a function of spatial structure. Drezner and Drezner (2002) stated that the parameter is related to the amount of time spent shopping; as the time spent shopping increases, the parameter decreases.

Young (1975) suggested moving from a fixed parameter for all centers to a variable parameter to account for specific site characteristics. He analyzed how the size and type of a shopping center affects the distances consumers are willing to drive. He compared customer patterns at three large regional shopping centers and three smaller neighborhood shopping centers in Philadelphia. The large retail centers with hundreds of thousands of square feet and thousands of parking spots, structured around major department stores and many other off-retailers, drew trade from a mean travel time of over twenty minutes with less than a quarter of shoppers traveling less than ten minutes. The neighborhood centers, anchored by supermarkets and fewer off-retailers, only drew trade from about a ten minute drive, with the vast majority traveling less than ten minutes. The neighborhood centers had a decay exponent of 2.3 compared with 0.967 for the regional centers. Consequently, Young

(1975) declared that the distance decay value should be two for neighborhood centers and one for regional centers. The largest center, The King of Prussia Mall, did not have the smallest decay value. In fact, Neshaminy, which was only half the size in square feet and had less than one third of the parking spaces, had a lower distance decay exponent, meaning that trade fell off more gradually from the smaller center. Thus, generalization can be inaccurate even within empirical categorizations.

Retail gravity models may not have been fully comprehended by all analysts who use them. This has led to errors in analysis, especially by incorrectly estimating the distance-decay parameter (O'Kelly and Miller 1989). Certainly, most users agree that a parameter value of two is too restrictive. Sometimes retail influence may decrease more or less steeply than $1/d_{ij}^2$. Measuring the interactions between different places is a dynamic process and may not always translate into similar results. Besides, parameter estimation may vary in interurban and intraurban situations (Huff 1962; O'Kelly and Miller 1989). This led to the development of proper methods for calibrating spatial interaction models (Fotheringham and O'Kelly 1989; O'Kelly and Miller 1989). Nakanishi and Cooper (1974) developed a technique for estimating parameters by least squares. This necessitated transforming the model into linear form using logarithms to then estimate the parameters using linear regression techniques. Jain and Mahajan (1979) and Bell et al. (1998) also studied this problem. Although it is beyond the scope of this chapter to specifically describe this process, it is important to note that analysts should be cognizant of the models' limitations and the proper applications. For further reading, we suggest Cooper and Nakanishi (1988) as they provide a good review of model calibration.

18.4 New Approaches and Outlook

In the past two decades, there have been a growing number of papers in the literature that have included the principles of gravity modeling in other types of models. For instance, O'Kelly and Miller (1989) created a model to link Reilly's Law of Retail Gravitation and the Huff Model with Applebaum's customer spotting technique. According to Applebaum (1966), customer spotting involves mapping the spatial distribution of customers. The analyst can then use these data to define a trade area based on a chosen amount or percentage of total customers. By linking the two together, analysts can calculate market share in designated regions based on probabilities of patronage. This required the relaxation of a key constraint of the original Reilly model. Instead of only identifying breaking points, the model was modified to identify the proportional influence of two competing centers at any point, such as where a customer is three times more likely to shop *A* than *B*. Furthermore, Huff's probability isolines were useful for analyzing varying types of situations with dynamic distance-decay parameters (O'Kelly and Miller 1989).

One of the main limitations of the Reilly model was that it was designed to find the breaking point between only two competing centers. Thus, it was insufficient to

handle multiple choices because of the Luce Choice axiom (Luce 1959; Sheppard 1978). Calculating the breaking points between two competing centers may not be as meaningful if there are more than just two competitors. Consider the situation where the influence of a point *A* and point *B* is found to be equal at a particular location, but upon review of probabilities, we find the influence of some point *C* is higher. Along those lines, O'Kelly and Miller (1989) provided a model to estimate probabilities with multiple competing facilities. In addition to their contribution of linking Reilly and Applebaum, they also linked Huff's contour probability isoline mapping with Applebaum's "primary trade area" to identify the contour line that can be chosen to provide sufficient demand to allow a fixed proportion of store sales. In other words, comparing the necessary catchment area against the probability at its spatial extent can help analysts and senior management analyze risk. Choosing an isoline with higher probability values would imply less risk in generating the necessary store sales. Conversely, if the desired sales required attracting from low probability areas, then it would have greater risk.

We have already mentioned that gravity models have much broader applications for location analysis than retail. For instance, Drezner and Drezner (2001) created a model based on the Huff Model to identify the best hub locations for airlines. Bruno and Improta (2008) used a gravity model to estimate university selection by students in Italy and how new facilities would impact enrollment. Certainly, there are many other examples of contributions beyond the scope of a single chapter. We do, however, discuss the details of select papers from one industry in particular: the health care industry.

Bucklin (1971) investigated the gravity model in an intraurban context to determine whether patients were more willing to travel past competing hospitals if the distances were shorter. Of course, shoppers and patients do not always choose the closest facility. Bucklin suggested varying exponents for competing facilities at various distances from one another. Using data for patient locations of Alameda County hospitals in California, Bucklin found that closer pairs of competitors had lower exponents. For his eleven pairs of hospitals, 6–25 minutes apart, he found his exponents were as low as 0.033 and as high as 4.518. Conversely, Carrothers (1956) found that the exponents were higher in an intraurban context than intercity. In that hospital choice is more risky than ordinary purchases such as groceries, it is logical to not always choose the closest hospital, especially when there are other opportunities in close proximity. Bucklin asserted that this phenomenon may be best explained by a logistic curve with respect to distance to two competitors, where you can expect very small exponents when competitors are close and rapid increase of the exponent as distance increases to a tapering off at a maximum value, an S-shaped curve. Nevertheless, Bucklin recognized the flaw in the fixed-exponent gravity model.

McLafferty (1988) used a gravity model to evaluate the dynamic patterns of patients in accordance with the closure of the Sydenham Hospital in New York City. Using patient origin data obtained before and after the closure, she found that although the gravity model accurately predicted patient flows, the distance parameter was inconsistent. Following the closure, the parameter changed from 1.4 to

1.66 just four years later. Although this was not a significant change, it does provide credence for a variable parameter.

Lowe and Sen (1996) used a gravity model to analyze patient choice behavior. Specifically, they investigated how policy changes such as hospital closures and the restructuring of the insurance system such as universal health care would affect the spatiality of patient flows. In this study, they concluded that the friction of distance is lower for the more advanced services of the larger hospitals with medical schools than for the smaller community facilities. They also inferred that universal health care would steer many patients away from the larger hospitals to smaller hospitals in poorer neighborhoods. According to Lowe and Sen, this may necessitate investing more in medical education in community center settings. They also investigated how previous hospital closures led to the redistribution of patients to other facilities. The closures reduced the number of short trips but did not generate a greater number of long trips over 30 minutes. A universal health care policy could have prevented the closures of some of the hospitals.

Location-allocation models have long been used to solve various location problems including retail-based problems (Hodgson 1978; Beaumont 1980). For instance, the *p*-median model has been especially useful for finding new locations in competitive environments (Hakimi 1983), though most location-allocation models, including the *p*-median model, assign customers to the closest facility (Beaumont 1981). Store patrons, however, do not always exhibit this behavior, and there have consequently been a growing number of attempts to apply the principles of the gravity model to competitive optimal facility location problems. These models have been structured in a plethora of ways. Most retail-based models have an objective function that maximizes market share or profit. Others have applied the basic location-allocation models to maximize retail coverage or minimize consumer travel distance. Hodgson (1978, 1981) was one of the first to incorporate a gravity-type spatial interaction model into a location-allocation model. Okunuki and Okabe (2002) presented an optimal location problem that incorporated the Huff model with continuous demand, meaning that there exists a demand probability at any point. Although there have been models that only allocated demand to the nodes such as Hakimi (1990) and Ghosh et al. (1995), models that incorporate the Huff Model can be designed with continuously varying demand that depends on a node's distance from a store. To that end, Okabe and Kitamura (1996) presented the "Network Huff Model," where distance decay assumed a convex form, which varied from Hakimi (1990), where the form was concave. Other examples include Berman and Krass (1998), who modeled demand from both nodes and paths, and Okunuki and Okabe (2002), who permitted demand along links. Colome et al. (2003) made an amendment to the Maximum Capture Model (MAXCAP) by ReVelle (1986) to calculate the capture of market share with the gravity model instead of just proximity. Finally, Drezner and Drezner (2007) introduced the "gravity *p*-median model," where customers have different probabilities for competing facilities that differ from standard *p*-median problems where customers are assigned to the closet facility.

18.5 Concluding Remarks

We have presented Reilly's Law of Retail Gravitation and the Huff Model in their original forms, and have discussed the empirical tests of Reilly's law and the associated debates over parameter estimation and applicable spatial uses. Further, we have explored the proliferation of analyses using the Huff Model; four decades after its introduction, however, Huff (2003) opined that the model has been commonly misused and has not been applied to its full potential. For instance, Huff described how many analysts were failing to statistically validate their parameters and were simply arbitrarily choosing their values. Clearly, statistically validating the variables and parameters increases the accuracy of the conclusions. Optimal locations are very sensitive to varying parameters and measures of attractiveness (Drezner and Drezner 2002), and proper use of the model includes gathering empirical data from customers (origin data). Although acquiring these data may be an exigent task for academics, many retail analysts are fortunate enough to have these data at their fingertips. At the time of their landmark studies, Huff and Reilly did not have these data and relied upon empirical tests. Still, the eventual ubiquitous availability of reliable data has changed the nature in which analysts study retail trade. In modern retail trade, many retailers have great customer origin data. Much of these data are acquired through loyalty store cards (a.k.a. frequent-buyer cards). For Reilly and others in his era, the proliferation and advent of customer data loyalty cards would have greatly simplified and expedited the process of identifying retail trade areas.

Reilly/Huff modeling is well-known for its retail applications, but we also discussed a few of the other noteworthy applications of the modeling. One pronounced example is its applications for health care. Even within the literature of both retail and health care, however, there are the same debates about the parameter estimation and the accuracy of modeling applications in particular situations. It is rather obvious that although retail, health care, and other industries and public organizations can learn from each other, they cannot necessarily apply the empirical results because the scenarios may be remarkably different.

Finally, there is still room for extensive adaptation of the modeling. As we discussed, the flexibility of gravity modeling is one its advantages. These models can be customized based upon the unique factors associated with different types of organizations. The Huff Model is credited with switching the focus from the destination to the customer, and current and future adaptations of the modeling are and will be involved with accounting for the individual customer. Specifically, segmenting customers based on varying lifestyles is becoming increasingly popular for retailers. By structuring gravity models to assign greater probabilities to customers who are more likely to purchase or spend more on particular products and services, the opportunity for improving the accuracy of the modeling, especially as it relates to sales forecasting, is greater.

References

Applebaum W (1966) Methods for determining store trade areas, market penetration, and potential sales. J Mark Res 3:127–141
Beaumont, JR (1980) Spatial interaction models and the location-allocation problem. J Reg Sci 20:37–50
Bell DR, Ho TH, Tang CS (1998) Determining where to shop: fixed and variable costs of shopping. J Mark Res 35:352–370
Berkeley G (1713) Three dialogues between Hylas and Philonous. In: Mathias MB (ed). Pearson Longman, New York (2007)
Berman O, Krass D (1998) Flow intercepting spatial interaction model: a new approach to optimal location of competitive facilities. Locat Sci 6:41–65
Bruno G, Improta G (2008) Using gravity models for the evaluation of new university site locations: a case study. Comput Oper Res 35:434–444
Bucklin LP (1971) Retail gravity models and consumer choice: a theoretical and empirical critique. Econ Geogr 47:489–497
Carey HC (1858) Principles of social science. Lippincott, Philadelphia
Carrothers GAP (1956) A historical review of gravity and potential models of human interaction. J Am Inst Plan 22:94–102
Colome R, Lourenco HR, Serra D (2003) A new chance-constrained maximum capture location problem. Ann Oper Res 122:121–139
Converse PD (1949) New laws of retail gravitation. J Mark 14:379–384
Converse PD (1953) Comment of movement of retail trade in Iowa. J Mark 18:170–171
Cooper LG, Nakanishi M (1988) Market share analysis: evaluating competitive marketing effectiveness. Kluwer, Boston
Douglas E (1949) Measuring the general retail trading area: a case study. J Mark 14:46–60
Drezner T, Drezner Z (2001) A note on applying the gravity rule to the airline hub problem. J Reg Sci 41:67–73
Drezner T, Drezner Z (2002) Validating the gravity-based competitive location model using inferred attractiveness. Ann Oper Res 111:227–237
Drezner T, Drezner Z (2007) The gravity *p*-median model. Eur J Oper Res 179:1239–1251
Eppli MJ, Shilling JD (1996) How critical is a good location to a regional shopping center? J Real Estate Res 9:5–32
Fotheringham AS (1981) Spatial structure and distance decay parameters. Ann Assoc Am Geogr 71:425–436
Fotheringham AS, O'Kelly ME (1989) Spatial interaction models: formulations and applications. Kluwer, Boston
Gautschi DA (1981) Specification of patronage models for retail center choice. J Mark Res 18:162–174
Ghosh A, McLafferty S, Craig CS (1995) Multifacility retail networks. In: Drezner Z (ed) Facility location: a survey of applications and methods. Springer, New York, pp 301–330
Hakimi SL (1983) On locating new facilities in a competitive environment. Eur J Oper Res 12:29–35
Hakimi SL (1990) Locations with spatial interactions: competitive locations and games. In: Francis RL, Mirchandani PB (eds) Discrete location theory. Wiley, New York, pp 439–478
Hodgson MJ (1978) Toward more realistic allocation in location allocation models: an interaction approach. Environ Plan A 10:1273–1285
Hodgson MJ (1981) A location-allocation model maximizing consumers' welfare. Reg Stud 15:493–506
Huff DL (1962) Determination of intraurban retail trade areas. Division of Research, Graduate School of Business Administration, University of California, Los Angeles, CA
Huff DL (1963) A probabilistic analysis of shopping center trade areas. Land Econ 39:81–90
Huff DL (1964) Defining and estimating a trade area. J Mark 28:34–38

Huff DL (2003) Parameter estimation in the Huff model. ArcUser 6:34–36
Huff DL, Jenks GF (1968) A graphic interpretation of the friction of distance in gravity models. Ann Assoc Am Geogr 58:814–824
Jain AK, Mahajan V (1979) Evaluating the competitive environment in retailing using multiplicative competitive interactive models. In: Sheth J (ed) Research in marketing. JAI Press, Greenwich, pp 217–235
Lee ML, Pace RK (2005) Spatial distribution of retail sales. J Real Estate Finance Econ 31:53–69
Lowe JM, Sen A (1996) Gravity model applications in health planning: analysis of an urban hospital market. J Reg Sci 36:437–461
Luce RD (1959) Individual choice behavior. Wiley, New York
McLafferty S (1988) Predicting the effect of hospital closure on hospital utilization patterns. Soc Sci Med 27:255–262
Nakanishi M, Cooper LG (1974) Parameter estimation for a multiplicative competitive interaction model: least squares approach. J Mark Res 11:303–311
Okabe A, Kitamura M (1996) A computational method for market area analysis on a network. Geogr Anal 28:330–349
O'Kelly ME (1983) Multipurpose shopping trips and the size of retail facilities. Ann Assoc Am Geogr 73:231–239
O'Kelly ME, Miller HJ (1989) A synthesis of some market area delimitation models. Growth Change 20:14–33
Okunuki K, Okabe A (2002) Solving the Huff-based competitive location model on a network with link-based demand. Ann Oper Res 111:239–252
Reilly WJ (1929) Methods for the study of retail relationships. Research Monograph 4, Bureau of Business Research, The University of Texas, Austin
Reilly WJ (1931) The law of retail gravitation. Knickerbocker Press, New York
ReVelle C (1986) The maximum capture or "sphere of influence" location problem: Hotelling revisited on a network. J Reg Sci 26:343–358
Reynolds RB (1953) A test of the law of retail gravitation. J Mark 17:273–277
Sheppard ES (1978) Theoretical underpinnings of the gravity hypothesis. Geogr Anal 10:386–402
Thrall GI (2002) Business geography and new real estate market analysis. Oxford University Press, New York
Wagner WB (1974) An empirical test of Reilly's law of retail gravitation. Growth Change 5:30–35
Young WJ (1975) Distance decay values and shopping center size. Prof Geogr 27:304–309

Chapter 19
Voronoi Diagrams and Their Uses

Mark L. Burkey, Joy Bhadury and H. A. Eiselt

19.1 Introduction

Voronoi diagrams are a very simple geometrical construct with a large variety of applications. Simply put, the problem can be described as follows. Consider some *d*-dimensional space in which a number of given points (sometimes referred to as seeds, attractors, or generators) are located. To each seed we assign a set that includes all points that are closer to the seed it is assigned to than to any other seed. Such a set is called a Voronoi set. The collection of all Voronoi sets is then a Voronoi diagram. Voronoi diagrams can be constructed for a number of different metrics. Clearly, different metrics will lead to different measures of proximity that result in rather different Voronoi diagrams.

The first steps into the direction of this geometrical construct were taken by mathematicians in the nineteenth century. Gauss (1840) appears to have been the first to graphically represent quadratic forms in a special case of Voronoi diagrams. Ten years later it was Dirichlet (1850) who further developed that representation. In his treatise (Voronoi 1908), the Russian mathematician G.F. Voronoi (1868–1908) exploited the Lejeune-Dirichlet theorem on the reduction of positive integer quadratic forms in two dimensions and its equivalence to a configuration of hexagons in the plane. Three years later, the geographer Thiessen (1911) independently developed "Thiessen polygons" for a spatial missing data problem. As the latter contribution provides a much more obvious explanation and use of the concept, we

M. L. Burkey (✉)
School of Business and Economics, North Carolina A & T State University,
Greensboro, NC, USA
e-mail: burkeym@ncat.edu

J. Bhadury
Bryan School of Business and Economics, University of North Carolina – Greensboro,
Greensboro, NC 27402-6170, USA
e-mail: joy_bhadury@uncg.edu

H. A. Eiselt
Faculty of Business Administration, University of New Brunswick, Fredericton, NB, Canada
e-mail: haeiselt@unb.ca

H. A. Eiselt, V. Marianov (eds.), *Foundations of Location Analysis,*
International Series in Operations Research & Management Science 155,
DOI 10.1007/978-1-4419-7572-0_19, © Springer Science+Business Media, LLC 2011

have chosen to concentrate on Thiessen's work. Note that this is in no way meant to indicate that Voronoi's work is any less original.

As reported by Okabe et al. (2000), estimation of the magnitude of ore deposits in Russia during the early years of the twentieth century used the concept of Voronoi diagrams. Crystallographers such as Niggli (1927) and Wigner and Seitz (1933) rediscovered the concept and used it for their work. The latter were interested in atomic structures. Their terse description of Voronoi diagrams is this:

> If we draw lines connecting the nearest atoms and consider the planes bisecting these perpendicularly, we have every atom surrounded by a truncated octahedron.

Delauney (1934) introduced another, but intimately related, tessellation of space. Today, we know that it is the graph-theoretic dual of Voronoi diagrams. It will be described in some detail in Sect. 19.3 of this chapter. The seminal contribution by Shamos and Hoey (1975) connected the proximity concept of Voronoi diagrams with location models. The many contributions by authors such as Aurenhammer and Edelsbrunner have advanced a variety of aspects of Voronoi diagrams and have made them what they are today. Excellent surveys are those by Aurenhammer (1991) and the book by Okabe et al. (2000).

19.2 Thiessen's (1911) Contribution

Thiessen's (1911) concern deals with the computation of averages given spatial data. In particular, consider a region in which the average rainfall is to be determined. A number of reporting stations exist, but they may not be evenly distributed within the region, or some stations have failed to report for some reason. While one could simply take an average of the existing data, this may lead to significant errors. Envisage a situation in which one part of the region is basically dry, while another experiences heavy rainfall. Furthermore, assume that the reporting stations are not representative, so that an average computed on that basis will fail to paint the true picture. In today's parlance, the problem deals with missing spatial data.

Thiessen presents his ideas in the form of an example. A square region is subdivided into small squares of 4 in.^{2} each. The true rainfall amounts are shown in Fig. 19.1. The total rainfall amount is 30 on the entire 16 squares, resulting in an average of 1.875 in. per square.

1	2	3	4
1	2	3	3
1	2	2	2
1	1	1	1

Fig. 19.1 Thiessen's original example

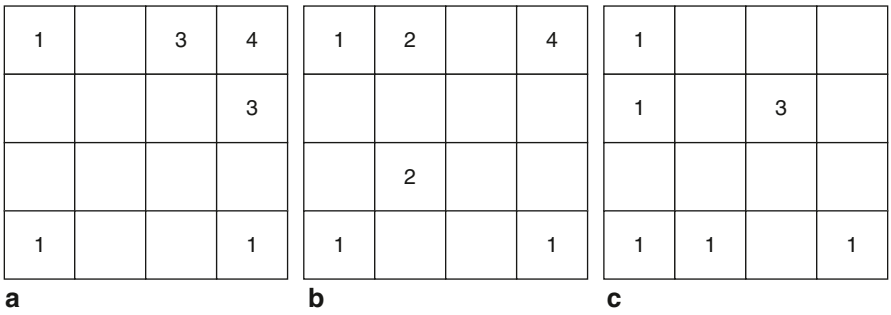


Fig. 19.2 Different scenarios with missing data

Suppose now that for some reason, not all weather stations are reporting. Figures 19.2a–c show three different scenarios, each of which relating to a different set of weather stations reporting, all based on the same true rainfall data shown in Fig. 19.1. The averages in the scenarios shown in Fig. 19.2a–c are 2.166,7, 1.833,3, and 1.333,3. Note the significant errors of the averages calculated above. Thiessen then suggests that

> The amount of rain recorded at any station should represent the amount for only that region inclosed (sic) by a line midway between the station under consideration and surrounding stations.

Based on this suggestion he constructs what geographers have long since referred to as a "Thiessen tessellation." The subdivision of space for this example as reported by Thiessen is shown in Fig. 19.3a. Here, the four regions are of size 4¼ with rainfall 1, 5 with rainfall 2, 3 with rainfall 4, and 3¾ with rainfall 1, resulting in a weighted average of 30/16 = 1.875, which happens to be the exact average.

However, closer inspection will reveal that the tessellation shown in Fig. 19.3a is not correct, following Thiessen's own description. As a matter of fact, if we were to take his example, the proper tessellation is shown in Fig. 19.3b. There is an area of 6¾ with a rainfall of 1, an area of 6⅜ with a rainfall of 2, and an area of 2⅞ with rainfall of 4, resulting in an average of 31/16 = 1.937,5, only a slight overestimate.

Published almost a decade after Thiessen's publication, Horton (1917) claims that Thiessen's method was "independently developed and has been extensively

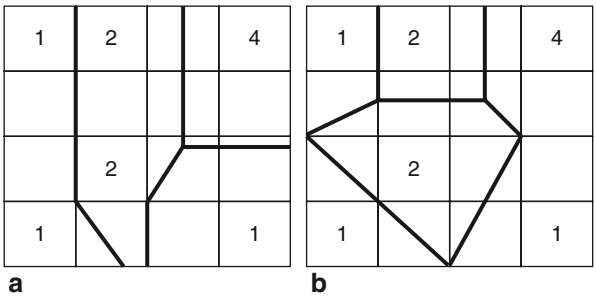


Fig. 19.3 **a** Thiessen tessellation of the original example. **b** Correct Thiessen tessellation

used by the writer." He also developed the "inclined plane method," a spatial interpolation method that determines missing data by way of weighted average of surrounding existing data. Thiessen's method was later developed further by Whitney (1929), who anticipated duals of Voronoi diagrams without formally introducing them and most likely without being aware of them. These concepts are defined and discussed in the next section.

19.3 Voronoi Diagrams and Their Properties

This section will formally introduce Voronoi diagrams and display examples for a variety of distance functions. Then some of their properties are explored including another tessellation that is closely related to Voronoi diagrams. Finally, a number of extensions of the basic concept and some applications are introduced and discussed. Even though there are no restrictions on the number of dimensions in which Voronoi diagrams can be constructed, for reasons of simplicity we restrict ourselves to two dimensions in this paper.

19.3.1 Measures of Distance

The concept of distance is central to Voronoi diagrams. While there are many types of distances such as gauges (Nickel and Puerto 2005), we will restrict ourselves to Minkowski distances, usually referred to ℓ_p distances. Given two points $A=(a_1, a_2)$ and $B=(b_1, b_2)$ in $\mathbb{R}^2$, the ℓ_p distance is formally defined as

$$d_{AB}^{(p)} = [|a_1 - b_1|^p + |a_2 - b_2|^p]^{1/p},$$

where p is a parameter chosen by the decision maker. It turns out that if $p=1$, we obtain the *Manhattan*, *rectilinear*, *rectangular*, or simply ℓ_1 distance. Formally, it is defined as

$$d_{AB}^{(1)} = |a_1 - b_1| + |a_2 - b_2|,$$

and it mimics movements that occur only parallel to the axes. The ℓ_1 distance function may be appropriate to model inner-city movements (without one-way streets), or wherever travel takes place along corridors such as on shop floors, in stores, or in a warehouse. This metric is commonly used in facility layout models; see, e.g., Francis et al. (1994).

For $p=2$, we obtain

$$d_{AB}^{(2)} = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}.$$

This is the usual *as the crow flies*, *straight line*, *Euclidean*, or simply ℓ_2 distances. Many authors (such as Thiessen in the second section of this chapter) have used Euclidean distances as a default. A related measure of distance is the *squared Euclidean distance* defined as

$$\left(d_{AB}^{(2)}\right)^2 = |a_1 - b_1|^2 + |a_2 - b_2|^2.$$

While not a Minkowski distance in the strong sense, it is a specific instance of a generalization of these distances that can be arrived at by using different exponents in the inner and outer part of the square brackets of the general formula (here: 2 and 1). Geometrically, this function does not actually define the length of a line segment but an area. However, it may be interpreted not as a distance but a cost or a disutility. The function itself has some interesting and desirable properties. First of all, it assumes that the cost or disutility of travel grows nonlinearly. For example, the disutility of commuting is more costly for the second mile than for the first mile, and more costly for the tenth mile than for the ninth. There are two reasons that economists use to justify this assumption. First, people normally get increasing disutility from higher levels of a bad thing, such as work or pollution. Secondly, there is an increasing opportunity cost of an individual's time as more and more time is spent driving. The first 15 minutes spent driving might pose only a minor inconvenience, but the second 15 minutes take the individual away from a more important activity.(Of course, the opposite might be true in the case of shipping a product. It may be less costly per mile to ship a long distance by ship or train than a short distance by truck or van. For details, see, e.g., Rydell (1967).

The general effect of varying $p \in [1, 2]$ on the distance measured between two points $A=(1, 1)$ and $B=(2, 2)$ is illustrated in Fig. 19.4.

Figure 19.5 demonstrates what happens as p decreases from 2 toward 1. In particular, decreasing p from 2 to 1 "mimics" the effect of barriers to straight line travel. Studies performed with real-world transportation data (e.g., Love and Morris 1979,

p	Distance
1.0	2.00
1.1	1.88
1.2	1.78
1.3	1.70
1.4	1.64
1.5	1.59
1.6	1.54
1.7	1.50
1.8	1.47
1.9	1.44
2.0	1.41

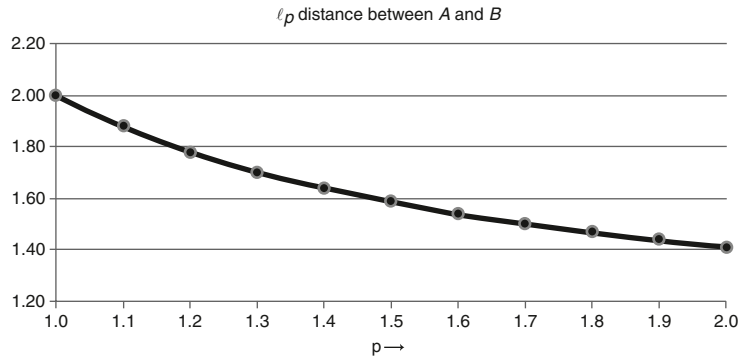


Fig. 19.4 Calculation of ℓ_p distance between $A=(1, 1)$ and $B=(2, 2)$ for $p \in [1, 2]$

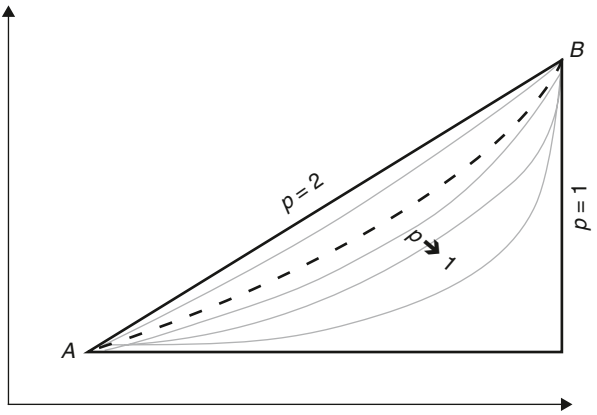


Fig. 19.5 Illustration of distance for $p \in [1, 2]$

Brimberg et al. 1994, and Fernández et al. 2002) show that variants of the ℓ_p norm with intermediate values of p between 1 and 2 best predict actual road distances.

Finally, we can let $p \to \infty$, which results in the *Chebyshev* (or *max*) distance

$$d_{AB}^{(\infty)} = \max\{|a_1 - b_1|, |a_2 - b_2|\}.$$

19.3.2 Basic Voronoi Diagrams

We are now able to formally define Voronoi diagrams. Let a set of seed points Q_1, Q_2, …, Q_n be given in d-dimensional real space $\mathbb{R}^d$. For convenience, Q_i denotes a given point as well as its location. The *convex hull* of the set of seed points is the intersection of all sets that include all seed points. To each seed point Q_i we can then assign a *Voronoi set* (or *Voronoi area* or *Voronoi cell* in $\mathbb{R}^d$) which we denote by $V^p(Q_i)$ given the ℓ_p metric. If no confusion can arise, we will refer to the area simply as $V(Q_i)$. The set $V^p(Q_i)$ consists of all points in space which are no farther from the seed point Q_i than to any of the other given seeds $Q_j, j \neq i$. As such, a Voronoi area solves the closest assignment problem of points to the seed. Formally, we define $V^p(Q_i) = \{Q\colon d^p(Q, Q_i) \leq d^p(Q, Q_j)\ \forall\ i \neq j\}$. The collection of all Voronoi sets then defines the *Voronoi diagram* VD^p or simply VD. Since each point $Q \in \mathbb{R}^d$ belongs to at least one Voronoi set, it follows that the union of all Voronoi sets spans the given space (or, alternatively, is a tessellation of the given space).

To visualize the concept, consider the simple case of only two points Q_i and Q_j in the two-dimensional Euclidean plane. Define then the bisector of Q_i and Q_j as $B^p(Q_i, Q_j) = \{Q\colon d^p(Q, Q_i) = d^p(Q, Q_j)\}$, i.e., the set of points Q whose distance to Q_i equals that to Q_j. The graphs in Fig. 19.6a–e display the bisectors for Minkowsky distances with the usual parameters $p = 1, 2, \infty$ as well those with the more unusual parameters $p = ½$ and 3.

Before further investigating the properties of bisectors in the various metrics, a special case related to the ℓ_1 metric should be mentioned. Given the case that

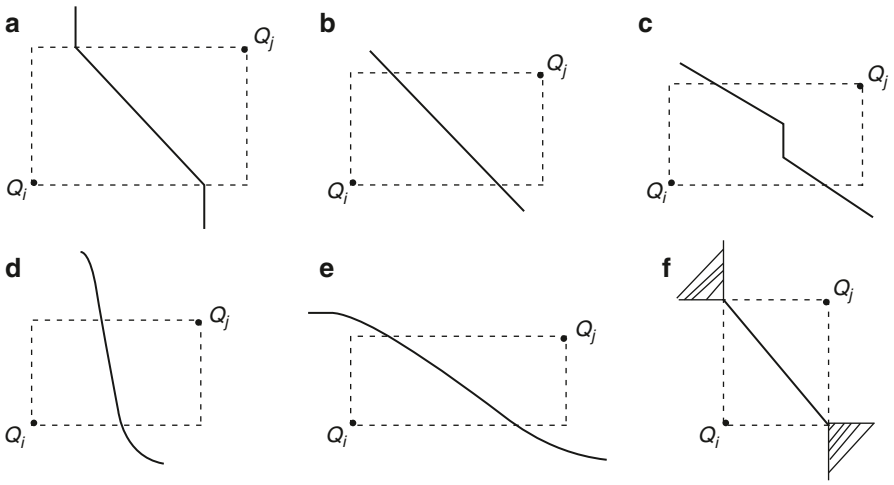


Fig. 19.6 **a** Bisector in ℓ_1 metric. **b** Bisector in ℓ_2 metric. **c** Bisector in ℓ_∞ metric. **d** Bisector in $\ell_{1/2}$ metric. **e** Bisector in ℓ_3 metric. **f** Special case of bisector in ℓ_1 metric

$|x_i - x_j| = |y_i - y_j|$ as shown in Fig. 19.6f, the bisector within the square defined by Q_i and Q_j coincides with the diagonal of that square. Outside this square, the bisector is no longer uniquely determined. Larson and Stevenson (1972) appeared to be among the first to notice this phenomenon. Actually, all points in the crosshatched area, the "plateaus," as Larson and Stevenson call them, are equidistant to Q_i and Q_j. At first glance, this non-uniqueness could be considered as an interesting, but ultimately highly unlikely anomaly as it occurs only if Q_i and Q_j are located at the corners of a square. Consider, however, a situation in which Q_i and Q_j form a rectangle with almost equal side lengths, then minor movements of either point will cause the bisector, and this the two sets it generates, to change suddenly and dramatically. Whether or not such a feature is desirable depends on the specific application.

Also note that among the ℓ_p distance functions, only ℓ_2 has a linear bisector. Furthermore, $B^p(Q_i, Q_j)$ is piecewise linear only for $p = 1$ and $p \to \infty$. In all other cases, the bisector is nonlinear.

Define now the p-halfspace $H_i^p(Q_i, Q_j) = \{Q : d^p(Q, Q_i) \leq d^p(Q, Q_j)\}$ as the set of all points that are closer to Q_i than to any of the other points $Q_j, j \neq i$, based on the ℓ_p distance. Note that the line bordering such an ℓ_p-halfspace is the bisector separating Q_i and Q_j in the ℓ_p metric. Then the i-th Voronoi set can be written as $V^p(Q_i) = \bigcap_{j \neq i} H_i^p(Q_i, Q_j)$. This allows us to state

Lemma 1: *Given an ℓ_p distance function, then $p = 2$ implies that $V^p(Q_i)$ is convex.*

The proof follows immediately from the convexity of $H_i^p(Q_i, Q_j)$, which itself is a result of the linearity of $B^p(Q_i, Q_j)$. This result also implies that $V^2(Q_i)$ is a bounded or unbounded polygon. Note that some of the p-halfspaces maybe be redundant and thus unnecessary in the actual construction of the Voronoi diagram.

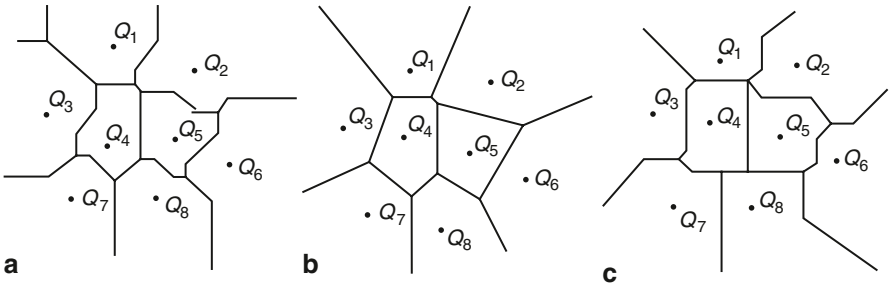


Fig. 19.7 **a** Voronoi diagram in ℓ_1 metric. **b** Voronoi diagram in ℓ_2 metric. **c** Voronoi diagram in the ℓ_∞ metric

Figure 19.7 provides an example for Voronoi diagrams for ℓ_1, ℓ_2, and ℓ_∞ metrics. (Note the plateau in the Northwest corner of Fig. 19.7a).

In general, each point in a Voronoi diagram in $\mathbb{R}^d$ which is equidistant to more than d points is commonly referred to as a *Voronoi point*. These points are the intersections of $d+1$ or more bisectors $B^p(Q_i, Q_j)$. We will call a Voronoi point degenerate, if it is equidistant to $d+2$ or more of the seeds, based on whatever metric is used. Only one Voronoi point in Fig. 19.7c is degenerate (the point that is equidistant to Q_1, Q_2, Q_4, and Q_5). In the following it will be assumed that no degenerate Voronoi points exist.

Suppose now that a Voronoi diagram for n seeds Q_1, Q_2, …, Q_n has been constructed. A *Delauney tessellation* is then constructed by connecting two seeds Q_i and Q_j, whenever the Voronoi areas $V(Q_i)$ and $V(Q_j)$ are direct neighbors, i.e., if they have points on a dividing line in common. Note that the lines that connect seeds on the convex hull are included in the Delauney tessellation (Fig. 19.8).

It must be noted that the structure of the Delauney tessellation depends on the metric used for the Voronoi diagram. In other words, if two points Q_i and Q_j generate Voronoi areas $V(Q_i)$ and $V(Q_j)$ which are adjacent in one metric, they do not have

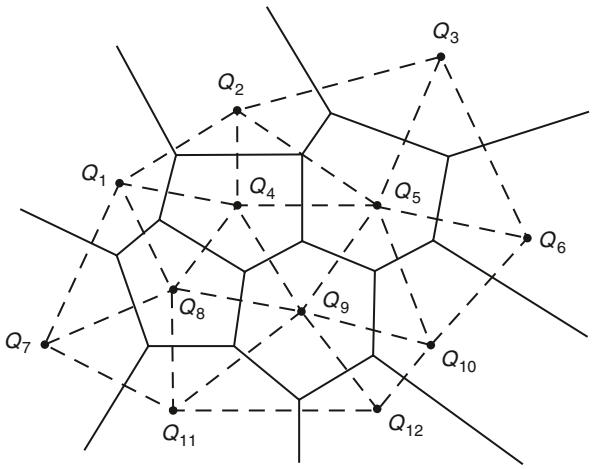


Fig. 19.8 *Solid line* Voronoi diagram for twelve seed points Q_1, Q_2, …, Q_{12}, and *broken line* Delauney tessellation

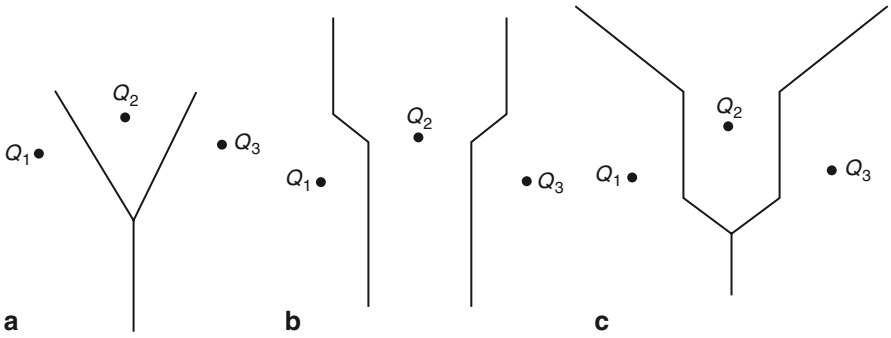


Fig. 19.9 **a** ℓ_2 metric. **b** ℓ_1 metric. **c** ℓ_∞ metric

to be adjacent in some other metric. This property has been pointed out by Eiselt and Pederzoli (1986). As an example, consider the Voronoi diagrams in Fig. 19.9 spanned by the same three points Q_1, Q_2, and Q_3.

Clearly, the Delauney tessellations of the Voronoi diagrams in the ℓ_2 and ℓ_∞ metrics contain the edges (Q_1, Q_2), (Q_2, Q_3), and (Q_1, Q_3), while the dual of the Voronoi diagram in the ℓ_1 metric includes the edges (Q_1, Q_2) and (Q_2, Q_3). In order to demonstrate that the duals of the Voronoi diagrams in the ℓ_2 and ℓ_∞ metrics may also be different, rotate the space by 45°. The shape of the $\ell\infty$ Voronoi diagram is then identical to the one displayed in Fig. 19.9b (rotated by 45°) and the dual includes the edges (Q_1, Q_2) and (Q_2, Q_3), whereas the Voronoi diagram in the Euclidean metric is unchanged.

Construction methods for Voronoi diagrams fall into two categories: divide-and-conquer techniques, and incremental methods. Typically, divide-and-conquer methods are somewhat more difficult to implement than incremental methods, but have a better worst-time complexity. Among other things, Shamos and Hoey (1975) demonstrated that a Voronoi diagram with n seed points can be computed by their divide-and-conquer method in two dimensions for the ℓ_2 metric in $O(n \log n)$ time, while the worst-case time complexity of incremental methods is $O(n^2)$. However, incremental methods appear to be easier to implement. Subsequent authors such as Hwang (1979) have extended this result to ℓ_1 and other metrics. A practical approach to the construction of Voronoi diagrams was put forward by Fortune (1987). Based on algorithmic ideas by Fortune (1987), Shute et al. (1987) constructed Delauney tessellations for ℓ_1 and ℓ_∞ metrics, in $O(n \log n)$ time. A good review of the subject is provided by Aurenhammer and Klein (1996).

19.3.3 Extensions to Basic Voronoi Diagrams and Their Applications

This section will explore some extensions to basic Voronoi diagrams and a few select applications. The choice of extensions and applications has been made in view of their applicability to location problems.

One straightforward extension of the standard Voronoi diagram allows different weights associated with the seed points, so that the tessellation is then computed based not only based on distance, but on the weights as well. Important applications of weighted Voronoi diagrams are found in the retail sector, where "attraction functions" that combine distances and weights have been used for a long time completely unrelated to Voronoi diagrams. In this context, the weights symbolize floor space, availability of parking, the general price level, friendliness of staff, and similar features that are but a few components relevant in a customer's decision making. Reilly (1931) was the first to acknowledge the problem and to introduce a single factor called "attractiveness." The probabilistic counterpart to Reilly's deterministic model was put forward by Huff (1964). In general, a customer at site i is attracted to a facility at site j governed by the function

$$a_{ij} = \frac{A_j}{d_{ij}^k},$$

where the weight A_j denotes a basic attractiveness of the facility located at site j, and k is a parameter that measures distance decay. In case $k=2$, we obtain Huff's gravity model. Given this attraction function, customers will then choose the facility they are most attracted to.

A number of problems are associated with this seemingly straightforward weighted extension of the basic model. As an example, consider a line segment of length 2 with a customer at its left and, a small facility with attractiveness "6" at the center, and a large facility with attractiveness 36 at the right end. At its location at the left end, the small and large facility exert a respective "pull" of $6/1=6$ and $36/4=9$ on the customer, who will decide to patronize the larger facility. Suppose now that the customer drives towards the large facility, but makes a stop at ½. The distances to the two stores are now ½ and 1½, respectively. Recalculating his preferences, the "pull" of the small and large store are now 24 and 16, respectively, so that the customer is now more attracted to the smaller store. This inconsistency was noted by Drezner et al. (1996), who prove that only additive selection rules are consistent. The authors also offer a consistent approximation to inconsistent multiplicative rules. In two dimensions, Voronoi diagrams based on attraction functions will have the Voronoi area of the less attractive facility surrounded by the Voronoi area of the more attractive facility, provided that the study region is sufficiently large. The above example on a one-dimensional line segment is but a cross section of a two-dimensional space.

The weighted model described above is generally known as having multiplicative weights, as the distances are multiplied by weights. This is also known as the Appolonius model. In contrast, in the additive model weights are added to the distances. This additive model was first investigated by Johnson and Mehl (1939), who examined the growth of crystals, bubbles, and cells. Both extensions result in Voronoi areas with nonlinear boundaries.

Another extension concerns the *Voronoi diagram of order k*. The idea of this concept is to tessellate the space into regions, such that each region $V^*(R)$ is associated

with a set $R \supseteq \{Q_1, Q_2, \ldots, Q_n\}$ with $|R|=k$, in a way that all points in region $V^*(R)$ are closer to all seeds in R than they are to any seed not in R. The most important orders are $k=1$ (which are the standard Voronoi diagrams discussed above), and $k=n-1$. In the latter case, suppose that the only seed not in the set R is Q_j. In this case, the region $V^*(R)$ includes all points closer to any seed than Q_j, or, to put it in other words, $V^*(R)$ includes all points from which Q_j is farther than any of the other seeds. This is why a Voronoi diagram of order $n-1$ is also referred to as the *farthest point Voronoi diagram*. Both these special cases are of particular importance in location modeling.

In order to explain just two of the concepts that link Voronoi diagrams to location modeling, we need to introduce two concepts. The first is the *smallest enclosing circle*. This is the smallest circle in the plane that includes all of the given seeds. The second is the *largest empty circle*. This is the largest circle, whose center is located inside the convex hull of the set of seeds that that does not include any of the seeds themselves. The center of the smallest enclosing circle is the solution of the 1-center (=1-minimax) problem in the two-dimensional plane with ℓ_2 distances (see Chaps. 4 and 5 of this volume for more details). On the other hand, the center of the largest empty circle is the solution of the 1-anti-center (=1-maximin) problem in $\mathbb{R}^2$ with ℓ_2 distances. The following two observations can be used to solve the 1-center and the 1-anti-center problems very efficiently with the use of Voronoi diagrams.

Lemma 2: *The center of the smallest enclosing circle is either at the center of the seeds that are farthest from each other, or at a Voronoi point of the farthest-point Voronoi diagram.*

Furthermore, we can state

Lemma 3 (Shamos and Hoey 1975): *The center of the largest empty circle is either located at a Voronoi point of the standard (1st order) Voronoi diagram, or at one of the points determined as the intersection of the boundary of a Voronoi area and the convex hull of the set the circle is to be located in.*

What follows are a number of applications, in which Voronoi diagrams play an important role. For reasons of space, this list is very selective. For a much more detailed discussion of the many applications of Voronoi diagrams, readers are referred to the authoritative works by De Berg et al. (2008) and Okabe et al. (2000) and Preparata and Shamos (1985), as well as the survey by Aurenhammer (1991).

A very difficult and compelling application of the basic Voronoi model is the incorporation of competition in Voronoi diagrams. Competitive location models date back to Hotelling's (1929) seminal work, which is described in Chap. 7 of this volume. Suppose now that customers are located in the plane with uniform density. Furthermore, a number of facilities (e.g., retail facilities) are also located in the plane. The task is now to determine the optimal location for a new retail facility, given that its objective is to maximize the number of customers that will patronize its store. Given uniform demand density, this objective is equivalent to maximizing the size of the Voronoi area associated with the new facility. As Eiselt et al. (1985) point out, this task first requires the triangulation of the Voronoi area, the determi-

nation of its size, and then its optimization. The objective is a polynomial that only holds within some nonlinearly bounded area. Okabe and Suzuki (1987) devise a nonlinear optimization method, coupled with a heuristic, for this global optimization method. Actually, their study does not simply find the optimal location for a single facility, but it repeatedly optimizes the optimal locations for one facility at a time. The research question is whether or not the process will stabilize. It turns out that while the patterns become similar to regular honeycomb patterns described by Lösch (1962), the pattern then self-destructs and no equilibrium is reached.

Another application concerns the task of drawing a topographic map from data observed in the field. A number of points in the area of interest have been measured, so that for each such point, the exact coordinates and its altitude are known, e.g., by way of GPS measurements. The problem is now to interpolate these points to enable the designer to draw contour lines on topographic maps. This could be accomplished by constructing a Voronoi diagram and the Delauney tessellation. It is then possible to interpolate the altitudes along the arcs of the dual, thus enabling the mapmaker to draw smoother contour lines than without this technique, see, e.g., Gold (1989). A good survey can be found in Okabe et al. (2000).

An application from the military field was suggested (but not solved) by Lee and Yang (1979). It deals with the surveillance and tracking of ships and it can be described as follows. The set of ships is subdivided into two classes, the red and the blue ships. Assuming that the present locations of all ships is known, the red points in the plane denote the locations of target ships, e.g., enemy submarines, while the blue ships indicate the locations of the ships in our own fleet. One problem is now to determine which of the blue ships is used to track and/or intercept the red target ships. The fact that this game is dynamic and stochastic makes it very difficult to solve. In addition, behavioral assumptions concerning the actions and moves of the red ships will have to be made. This introduces an element of fuzziness into the model. In order to initialize the game, it has been suggested to first determine the Voronoi diagram for the blue points as seeds. In this case the use of the Euclidean metric appears justified. Whenever now a red point Q is located in the Voronoi area $V(Q_i)$, it means that the target ship is closer to our own vessel Q_i than to any of our other ships, so that we will use Q_i for the chase. In case multiple target ships are present in a single Voronoi area, the authors suggest an interactive man-machine procedure.

Some further applications of Voronoi diagrams can be found in the works by Eiselt (1989), Aurenhammer (1991), and Okabe et al. (2000).

19.4 A Case Study Involving Voronoi Diagrams

To illustrate the application of Voronoi diagrams with different attraction functions and metrics, we use data from the Piedmont-Triad region of North Carolina. This region comprises the 12 counties of northern central North Carolina (out of 100 counties in the state). Included in this region are the Greensboro-High Point

Fig. 19.10 Triad region of North Carolina

and Winston-Salem Metropolitan Statistical Areas and the surrounding counties. Figure 19.10 shows the geographic location of this region. The major cities in this region are Greensboro, Winston-Salem, and High Point which are the third, fourth, and eighth most populous cities in North Carolina. These cities have approximately 188,000, 173,000 and 72,000 people. The 12 counties of the Triad Region contain 1.5 million of North Carolina's eight million people. (All population data are taken from the 2000 United States census).

The seed points comprise the centroids of the urban areas within this 12-county region as defined by the US Geological Survey's national atlas. The USGS describes these as the "Generalized footprint of built-up areas." (For these data and further information, see Nationalatlas 2008) There are 21 such areas in these 12 counties, shown in Fig. 19.11.

Note from Fig. 19.11 that several counties have several "attractors" (i.e., seed points), while two counties have none. While this may seem curious at first glance, the counties without attractors (Stokes and Caswell Co.) are large in area but sparsely populated. For example, the county seat of Stokes County, Danbury, had a population of only 108 people in the year 2000. Also, according to the 2000 census more than 70% of the workers in each of the counties had a job outside of their county of residence, therefore being "attracted" to some other location for work. The major attractors are listed in Table 19.1, along with their populations. All attractors not listed have populations under 10,000 people.

We will construct several Voronoi diagrams in the next sections, and will use the centroids of the region's 976 Census Block Groups for the year 2000 to assign households to each attractor. Then, we will compare the attraction functions based on each diagram's assignment of households to attractors.

Fig. 19.11 The 12 triad counties and 21 urban areas

Table 19.1 Large attractors in the triad region

Name	Population
Greensboro	187,800
Winston-Salem	173,238
High Point	72,096
Burlington	57,711
Lexington	21,250
Asheboro	16,606
Thomasville	15,915
Eden	15,238
Reidsville	12,183
Mount Airy	11,657

19.4.1 Voronoi Diagrams of the Piedmont Triad Region

19.4.1.1 Unweighted Voronoi Diagrams

Figures 19.12, 19.13, and 19.14 will show the tessellation of the Triad Region under unweighted Minkowski distances, i.e., attraction functions in which all base attractions A_j are equal.

While Fig. 19.12 shows the subdivision of the area by using Euclidean distances, Figs. 19.13 and 19.14 displays a similar tessellation based on ℓ_1 and $\ell_{1.5}$ distances, respectively.

Fig. 19.12 Unweighted Voronoi diagram with ℓ_2 distances

19.4.1.2 Weighted Voronoi Diagrams

Consider now weighted models with attraction functions, in which the attractors are not all equal. Borrowing from Krugman (1980) and other economists and marketing researchers (for a survey, see, e.g., Martin and Sunley 1996), we may use the population of a seed as the base attraction A_j.

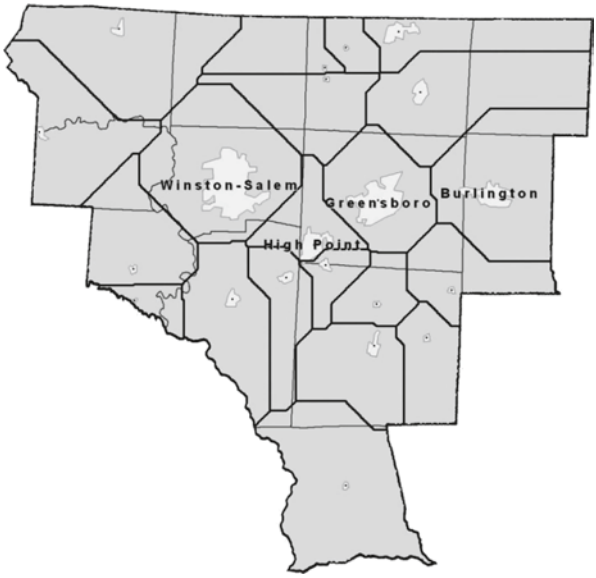


Fig. 19.13 Unweighted Voronoi diagram with ℓ_1 distance

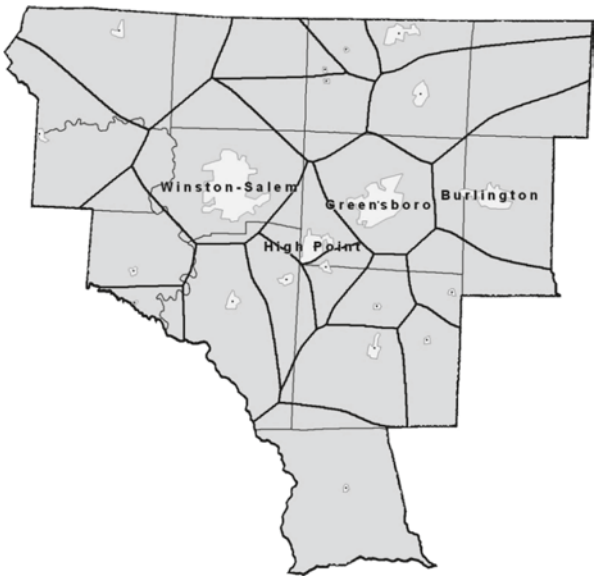


Fig. 19.14 Unweighted Voronoi diagram with $\ell_{1.5}$ distance

The resulting weighted Voronoi diagram with Euclidean distances is shown in Fig. 19.15. As is evident from Fig. 19.13, the most prominent Voronoi cells belong to Winston-Salem and Greensboro, the two most heavily populated and hence, the largest attractors in the Piedmont Triad region. Their Voronoi cells are separated by the curved line running roughly North-South between them. As is evident, these areas are nonconvex, and include all of the area not contained in a small circle around

Fig. 19.15 Weighted Voronoi diagram with ℓ_2 distances

each of the other attractors. As mentioned earlier in Sect. 19.3.3, this nonconvexity is problematic in that it implies that customers who are to the east of the Voronoi cell of Burlington would still be attracted to Greensboro rather than Burlington. Clearly this would be questionable as a predictor of reality.

As a result, the population-weighted Euclidean metric appears to give results that are at least counterintuitive and do not coincide with reality. The Voronoi diagram boundaries also cut through the limits of the known "built up areas" in many cases, assigning people living inside these areas not to the facilities in their own proximity but to one of the two main attractors, Winston-Salem and Greensboro, instead. Clearly, if the "built-up areas" are any measure of reality, then simply weighting by population gives too much weight to the attraction factor proxied by population, and not enough to costs related to distance.

There are two ways that one can change the attraction function to make the diagrams more realistic. The first would involve making the attraction function have a negative second derivative with respect to variety (e.g., population). Krugman (1980) and others have used utility functions that assume that people have a preference for variety, but that their utility increases in variety at a decreasing rate. Therefore, imposing a natural logarithm or square root function on population might be appropriate.

Alternatively, one could achieve a similar, economically more realistic result by assuming that the disutility associated with the distance to an attractor increases at an increasing rate. Applying the usual attraction function with squared Euclidean distances results in the Voronoi diagram shown in Fig. 19.16. For reference, the ℓ_2 distances are also shown in Fig. 19.16 in light gray.

While the Voronoi cells of Greensboro and Winston Salem still have very wide areas of influence, increasing the importance of distance relative to population has

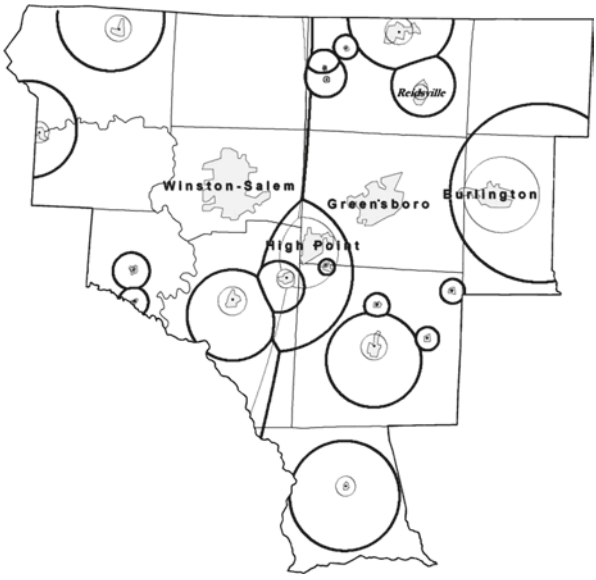


Fig. 19.16 Weighted Voronoi diagram with $\ell_2^2 = (\ell_2)^2$ distances

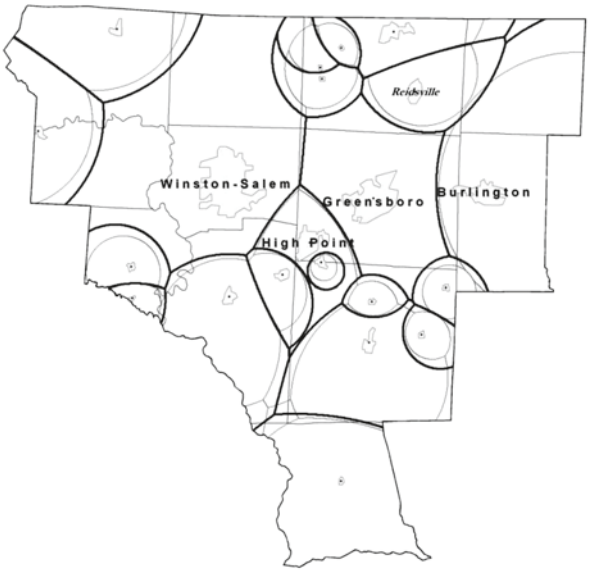


Fig. 19.17 Weighted Voronoi diagram with $(\ell_2)^3$ distances (*gray lines*) and $(\ell_2)^4$ (*black lines*)

increased the size of the areas for the smaller attractors. None of the smaller attractors are "torn apart" as they were when using only the Euclidean metric.

Figure 19.17 shows a weighted Voronoi diagram based on cubed Euclidean distances and Euclidean distances raised to the fourth power.

19.4.2 Comparing the Voronoi Diagrams

This section will focus on comparing the different attraction functions discussed in Sect. 19.3 above. In particular, we will conclude this section with an empirically-based validation of these attraction functions that will allow us to conclude which one best predicts travel behavior in the Piedmont Triad region.

As a first method of comparison among the different attraction functions, we list below in Table 19.2 several measures from each Voronoi diagram. For the four largest attractors in the Piedmont Triad (Greensboro, Winston-Salem, High Point and Burlington), Table 19.2 lists the areas and total income of their respective Voronoi cells generated under each attraction function.

As a second way to quantitatively compare the different attraction functions, we use the Herfindahl-Hirschman indices of inequality with respect to several measurements of interest. (Other potential measures include Lorenz curves and the Gini index). The measures we choose for comparison are income, population, and land area. Original references regarding these measure are Lorenz (1905), Gini (1921), Hirschman (1945, 1964), and Herfindahl (1950).

Table 19.2 Summary statistics of the top four attractors

Attractor	Unweighted ℓ_2		Unweighted ℓ_1		Unweighted $\ell_{1.5}$		Weighted ℓ_2		Weighted $(\ell_2)^2$		Weighted $(\ell_2)^3$	
	Area 000s mile^2	Income in $ millions	Area	Income	Area	Income	Area	Income	Area	Income	Area	Income
Greensboro	3,242	6,661	2,962	6,728	3,196	6,639	32,352	13,138	19,248	9,542	8,687	8,255
Winston-Salem	5,619	7,514	5,968	7,735	5,906	7,632	23,629	11,843	19,464	10,268	13,809	9,284
High Point	1,216	2,735	1,563	2,614	1,425	2,829	882	1,927	2,603	2,660	2,402	2,636
Burlington	5,183	2,682	5,278	2,736	5,448	2,724	1,244	1,784	4,725	2,599	7,170	2,848

A Herfindahl-Hirschman Index (*HHI*) is a common measure of concentration or inequality that was primarily invented to measure the concentration of production among a small number of firms.

$$HHI = \sum_i \left(\frac{x_i}{\sum_k x_k} \right)^2$$

The *HHI* is simply the sum of the squared proportions of each entity. As this measure gets closer to one, it indicates a higher concentration among fewer entities. For example, if each of ten firms produce one tenth of the industry output each, then the *HHI* equals $10(0.1)^2 = 0.1$. If instead one of the ten firms produces 90% of the industry output, and the other nine firms produce 1.1% each, then the *HHI* would increase to $(0.9)^2 + 9(0.011)^2 = 0.811089$. Thus, one can see that equal shares of n elements produce an *HHI* of $1/n$, while the index will approach the value of one, as one member has a larger, unequal share. Therefore, with 21 attractors in the Piedmont Triad region, an equal share (of area, population, etc.) would generate an *HHI* of $1/21 \approx 0.0476$.

In contrasting the results obtained using the different attraction functions, we find below (see Table 19.3 below) that for the unweighted ℓ_1, ℓ_2, and $\ell_{1.5}$ attraction functions, the area is very close to being equally divided with an *HHI* around 0.06, with the largest shares of area for any attractor of around 8 or 9%. Thus these attraction functions predict only a marginal amount of agglomeration in the Piedmont Triad region.

However, as also seen below (and is clear from the Voronoi diagram) the Voronoi diagrams weighted by the population of the seeds show significantly more agglomeration than the unweighted Voronoi diagrams with respect to all three distance measures. For example, in the population-weighted diagram, 43 and 39% of the household incomes are assigned to the largest two attractors, while in the unweighted ℓ_2 Voronoi diagram, the largest shares of income are 25 and 22%. Interestingly, the unweighted diagrams are most heavily concentrated in income, whereas the population-weighted ℓ_2 Voronoi diagram is most heavily concentrated in area, with Greensboro's Voronoi area controlling 54% of the total surface area.

The attraction function that uses squared Euclidean distance weighted by population serves as a middle path between the unweighted and population weighted diagrams, by reducing the importance of population relative to the distance, and is again most heavily concentrated in income.

This example assigns people in block groups to attractors. One feature of interest concerns the distance between a block group and the attractor for each attraction function under investigation. Suppose now that we compute the weighted average distance that people would have to travel from their home block group measured in Euclidean distances to the assigned attractor. The Voronoi diagram based on unweighted ℓ_2 distances minimizes the distances, achieving the "most efficient" assignment in terms of transportation. This will be the benchmark measure. It is

Table 19.3 Descriptive comparison measures

Attraction function	Herfindahl-Hirschman index			Avg ℓ_2 distance from block group to attractor	
	Income	Population	Area (mile^2)	Distance (mile)	(ℓ_2Ratio)
Unweighted ℓ_2	0.134,3	0.115,7	0.024,7	5.57	1.000
Unweighted ℓ_1	0.137,9	0.118,3	0.024,1	5.60	1.007
Unweighted $\ell_{1.5}$	0.136,2	0.117,4	0.024,4	5.57	1.001
Weighted ℓ_2	0.344,9	0.328,1	0.176,5	11.76	2.112
Weighted $(\ell_2)^2$	0.230,0	0.206,7	0.087,8	8.02	1.441
Weighted $(\ell_2)^3$	0.188,6	0.164,6	0.044,7	6.41	1.153

apparent that other attraction functions will assign people to attractors other than the closest. Given that, we can then calculate the ratio of the distances generated under each model to the benchmark measure. For instance, traffic planners who are interested in patterns of urban sprawl and associated excess traffic congestion, could use such a ratio as a measure of excess transportation costs when people shop or work at a job in a large city, rather than the closest city. For example, the most efficient travel assignment would predict that the average commute would be 5.57 miles each way (see Table 19.3). However, the weighted ℓ_2 metric with population as weights assigns people more heavily to larger (rather than closer) attractors, and would predict average commutes of 11.76 miles, which is 2.11 times as large as the most efficient assignment.

Below, we attempt to cross-validate how well these attraction functions predict actual human behavior. To test this, we begin by estimating the county border crossing behavior predicted by each attraction function. These predicted numbers are then checked against cross-county worker flow files from the Census Bureau for the year 2000, since this is the finest level at which this information is recorded. Refer to Table 19.4, which shows the origin and destination for each worker 16 years of age or older for people who both live and work in the 12 Triad Counties. Looking at the top row, 47,734 workers live and work in Alamance County, and 2,388 people live in Caswell County, but work in Alamance County. Conversely, only 164 people live in Alamance and work in Caswell (2nd row, 1st column). With the set of attractors we are using, there is a limit to how accurately we can model behavior since two counties had no urban areas. However, attraction functions which do better could be considered to be more appropriate, at least in terms of modeling employment behavior.

Having computed the estimates of cross-county travel as predicted by each attraction function, we compare them to the actual data above by computing two measures of error.

The first measure of error is a chi-square goodness of fit and an absolute error measure where, for each (origin, destination) pair,

$$\chi^2 = \sum (actual - predicted)^2 / (actual)$$

Table 19.4 Cross county worker flow, actual origin-destination

Work county	County of residence (origin)												
	Alamance	Caswell	Davidson	Davie	Forsyth	Guilford	Mont-gomery	Randolph	Rocking-ham	Stokes	Surry	Yadkin	Total workers
Alamance	47,734	2,388	323	25	287	4,050	16	578	503	15	13	11	55,943
Caswell	164	2,693			12	44		27	171				3,111
Davidson	129	9	40,621	521	4,136	2,982	165	2,607	96	252	101	134	51,753
Davie	19		314	7,710	902	67		11	24	58	73	541	9,719
Forsyth	418	22	11,062	5,242	119,233	7,636	36	694	870	10,259	4,316	5,504	165,292
Guilford	6,443	800	14,668	410	16,515	187,150	205	20,278	11,960	1,620	500	323	260,872
Montgomery			46		18	8	8,130	419	17				8,638
Randolph	301	12	2,540	53	392	3,984	897	38,637	73	20	14	19	46,942
Rockingham	271	844	177	28	358	1,720		143	25,523	1,360	79	14	30,517
Stokes		7	63	30	1,165	68		10	511	6,330	512	72	8,768
Surry			10	69	560	10	9	12	21	1,167	24,821	1,678	28,357
Yadkin			39	327	663	45				66	1,146	7,572	9,858
Total res. workers in triad	55,479	6,775	69,863	14.415	144,241	207,764	9,458	63,416	39,769	21,147	31,575	15,868	679,770
Work outside the triad	8,219	3,142	3,030	2,219	3,597	5,315	2,092	2,387	1,869	562	1,883	1,399	35,714

The second measure of error is the mean absolute error, which is given by the sum of absolute deviations of predicted minus actual number of people crossing between each county and divided by two.

$$Error = {}^{1}\!/_{2} \sum |actual - predicted|$$

For example, suppose that there were only two counties, and in reality 100 people commute both from county 1 to county 2 and from county 2 to county 1. If our model wrongly predicted that 99 people travel from county 1 to county 2 and 101 from county 2 to county 1, note that the model is only wrongly assigning 1 person. The term in the numerator would be computed as [|100–99|+|100–101|]=2, so to arrive at the correct number of persons wrongly assigned, we divide by 2 in the denominator. The results are shown below in Table 19.5.

The attraction function with weighted $(\ell_2)^2$ distances performs best with a 15.9% error rate. In other words, this model's predictions are 15.9% off in their predictions of actual inter-county worker flows. This is much better than a simple linear weighted model, and outperforms all of the unweighted models.

We will conclude this section by applying the standard Voronoi diagram to find the largest empty circle centered at any point in the Triad Region so as to solve the 1 maximin problem, i.e., the optimal location of an undesirable facility. Following our discussion in Sect. 19.3.3, the following algorithm will solve the problem.

Algorithm: Locate a Single Undesirable Facility

Step 1: Draw the Voronoi Diagram of points Q_1, Q_2, Q_n using the unweighted ℓ_2 metric.

Step 2: Superimpose the outer boundary of feasible locations on this Voronoi Diagram.

Step 3: Construct the Voronoi cells $V^2(Q_i)$ for all points Q_i.

Step 4: Determine $\bar{Q}_i \in V^2(Q_i)$, the point in Q_i's Voronoi cell that is farthest from Q_i.

Step 5: Determine D_i as the ℓ_2 distance between Q_i and $\bar{Q}_i$. The optimal location is the facility that determines $D^* = \max\{D_i\}$.

Table 19.5 Predictive ability of attraction functions

Attraction function and metric	Chi-square measure of error	Mean absolute error	Error rate (%)
Unweighted ℓ_2	329,501.10	130,340	19.17
Unweighted ℓ_1	307,678.50	130,440	19.19
Unweighted $\ell_{1.5}$	299,653.10	127,717	18.79
Weighted ℓ_2	956,172.30	197,863	29.11
Weighted $(\ell_2)^2$	293,826.90	108,231	15.92
Weighted $(\ell_2)^3$	299,653.10	127,716	18.79

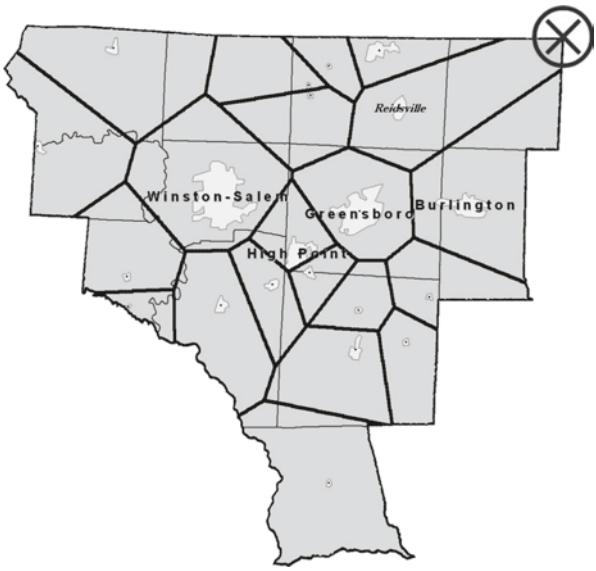


Fig. 19.18 Obnoxious single-facility location

The optimal solution is shown in Fig. 19.18 by a large "X" in the Northeast corner of the region. This location is 32.84 miles away from the nearest seed point, Reidsville.

19.5 Conclusions

This chapter has followed the steps taken by one of the first geographers to tessellate space based on the concept of "nearest neighbors." This was followed by a general introduction of Voronoi diagrams and a short survey of some of their properties. The concept was then put to work in a real-life application that examined settlement patterns and agglomerations in the Piedmont Triad Region in North Carolina. In particular, a number of weighted und unweighted distances functions were applied to the region. In addition to different Voronoi diagrams, we explored the computation of descriptive measures for the different models, thus allowing model validation.

While this chapter has focused on measurements in "geographic spaces," it should be noted that interesting work in many fields is being done which extends measures designed for geographic spaces into "conceptualized spaces." The earliest reference to mention modeling in such spaces appears to be Hotelling (1929), who discussed the location of politicians along a single-dimensional ideological spectrum. Whether it is the stand on a political issues, preferences for products, or similar issues, many of them could potentially be modeled as points in a multidimensional space; see, e.g., Aspinwall (2002), Shaw (1982), and Bower and Whitten (2000). Once this is done, voter and customer behavior may be put into some

proximity-based concept, resulting in a Voronoi diagram. Such a diagram will then allow to predict voting outcomes, market shares of products, and similar measures. Much work needs to be done, though, in order to make these concepts workable.

Acknowledgments This work was in part supported by a grant from the Natural Sciences and Engineering Research Council of Canada. This support is gratefully acknowledged. The authors would also like to thank Professor Vladimir Marianov for his assistance with the Thiessen reference and for providing some of the figures in this paper. Thanks also to our assistant #21 (Courtney Palmer) for providing some of the figures.

References

Aspinwall M (2002) Preferring Europe. Eur Union Polit 3:81–111

Aurenhammer F (1991) Voronoi diagrams—a survey of a fundamental geometric data structure. ACM Comput Surv 23:345–405

Aurenhammer F, Klein R (1996) Voronoi diagrams. http://www.pi6.fernuni-hagen.de/publ/tr198.pdf. Accessed 21 Aug 2009

Bower JA, Whitten R (2000) Sensory characteristics and consumer liking for cereal bar snack foods. J Sens Stud 15:327–345

Brimberg J, Dowling PD, Love RF (1994) The weighted one-two norm distance model: empirical validation and confidence interval estimation. Locat Sci 2:91–100

De Berg, M, Cheong O, van Kreveld M, Overmars M (2008) Computational geometry: algorithms and applications, 3rd edn. Springer, Berlin

Delauney B (1934) Sur la sphère vide. Bulleting of the academy of sciences of the U.S.S.R. Cl Sci Math Nat, Ser 7:793–800

Dirichlet GL (1850) Über die Reduktion der positiven quadratischen Formen mit drei unbestimmten ganzen Zahlen. Z Rein Angew Math 40:209–227

Drezner T, Drezner Z, Eiselt HA (1996) Consistent and inconsistent rules in competitive facility choice. J Oper Res Soc 47:1494–1503

Eiselt HA (1989) Modeling business problems with Voronoi diagrams. Can J Adm Sci 6:43–53

Eiselt HA, Pederzoli G, Sandblom C-L (1985) On the location of a new service facility in an urban area. In Bartel H (ed) Proceedings of the administrative sciences association of Canada 6/9, Montreal, pp 42–55

Fernández J, Fernández P, Pelegrin B (2002) Estimating actual distances by norm functions: comparison between the $\ell_{k,\,p,\,\theta}$-norm and the $\ell_{b1,\,b2,\,\theta}$-norm and a study about the selection of the data set. Comput Oper Res 29:609–623

Fortune SA (1987) A sweepline algorithm for Voronoi diagrams. Algorithmica 2:153–174

Francis RL, McGinnis LF, White JA (1994) Facility layout and location: an analytical approach. Prentice-Hall, Englewood Cliffs

Gauss CF (1840) Recursion der "Untersuchungen über die Eigenschaften der positive ternären quadratischen Formen" von Ludwig August Seeber. Z Rein Angew Math 20:312–320

Gini C (1921) Measurement of inequality of incomes. Econ J 31:124–126

Gold CM (1989) Surface interpolation, spatial adjacency and GIS. In: Raper J (ed) Three dimensional applications in geographical information systems. Taylor & Francis, London. http://www.voronoi.com/wiki/images/9/93/Chapter3-Surface_interpolation%2Cspatial_adjacency.pdf. Accessed 21 Aug 2009 (Chapter 3)

Herfindahl OC (1950) Concentration in the U.S. steel industry. Unpublished dissertation, Columbia University, New York

Hirschman AO (1945) National power and the structure of foreign trade. University of California Press, Berkeley, CA

Hirschman AO (1964) The paternity of an index. Am Econ Rev 54:761
Horton RE (1917) Rational study of rainfall data makes possible better estimates of water yield. Eng News-Rec 79:211–213
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
Hwang FK (1979) An $O(n \log n)$ Algorithm for rectilinear minimal spanning trees. J ACM 26:177–182
Johnson WA, Mehl RF (1939) Reaction kinetics in processes of nucleation and growth. Trans AIMME 135:416–458
Krugman P (1980) Scale economies, product differentiation, and the pattern of trade. Am Econ Rev 70:950–959
Larson RC, Stevenson KA (1972) On insensitivities in urban redistricting and facility location. Oper Res 20:595–612
Lee DT, Yang CC (1979) Location of multiple points in a planar subdivision. Inf Proc Lett 9:190–193
Lorenz MO (1905) Methods of measuring the concentration of wealth. Pub Am Statist Assoc 9:209–219
Lösch A (1962) Die räumliche Ordnung der Wirtschaft, 3rd edn. Fischer, Stuttgart (1st edn published in 1940)
Love RF, Morris JG (1979) Mathematical models of road travel distances. Manag Sci 25:130–139
Martin R, Sunley P (1996) Paul Krugman's geographical economics and its implications for regional development theory: a critical assessment. Econ Geogr 72:259–292
Nationalatlas (2008) http://nationalatlas.gov/mld/urbanap.html. Accessed 21 Aug 2009
Nickel S, Puerto J (2005) Location theory: a unified approach. Springer, Berlin
Niggli R (1927) Die topologische Strukturanalyse. Z Kristallogr, Kristall-geom, Kristallphys, Kristallchem 65:391–415
Okabe A, Suzuki A (1987) Stability of spatial competition for a large number of firms on a bounded two-dimensional space. Environ Plan A 19:1067–1082
Okabe A, Boots B, Sugihara K, Chiu S-N (2000) Spatial tessellations: concepts and applications of Voronoi diagrams, 2nd edn. Wiley, Chichester
Preparata F, Shamos MI (1985) Computational geometry. Springer, New York
Reilly WJ (1931) The law of retail gravitation, 2nd edn. Pilsbury, New York
Rydell CP (1967) A note on a location principle: between the median and the mode. J Reg Sci 7:185–192
Shamos MI, Hoey D (1975) Closest point problems. Proceedings of the 16th annual symposium on Foundations of Computer Science, pp 151–162
Shaw RW (1982) Product proliferation in characteristics space: the UK fertilizer industry. J Ind Econ 31:69–91
Shute GM, Deneen LL, Thomborson CD (1987) An $O(n \log n)$ plane-sweep algorithm for ℓ_1 and ℓ_∞ Delaunay triangulations. Technical report 87-5, computer science and mathematics, University of Minnesota Duluth
Thiessen AH (1911) Precipitation averages for large areas. Mon Weather Rev 39:1082–1084
Voronoi G (1908) Nouvelles applications des parameters continus à la théorie des forms quadratiques, deuxième mémoire, recherché sur les paralleloèdres primitifs. Z Rein Angew Math 134:198–287
Whitney EN (1929) Area rainfall estimates. Mon Weather Rev 57:462–463
Wigner E, Seitz F (1933) On the constitution of metallic sodium. Phys Rev 43:804–810

Chapter 20
Central Places: The Theories of von Thünen, Christaller, and Lösch

Kathrin Fischer

20.1 Introduction

The question of why economic activities are concentrated in certain places and not in others, why so-called "central places" exist at which an agglomeration of people and trade takes place, and where these central places are to be found, has long been a focus of spatial economists. In the nineteenth and twentieth centuries, three German scientists concentrated on that area, and the results of their research became famous and influential in Germany and all over the world. The three scholars in question are: Johann Heinrich von Thünen ("Der isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie," Teil I, 1826), Walter Christaller ("Die zentralen Orte in Süddeutschland," 1933) and August Lösch ("Die räumliche Ordnung der Wirtschaft," 1940). Von Thünen was the first to develop a theory of land use, and was praised as "one of the patron saints of econometrics" by Schumpeter (1955). Christaller founded the Theory of Central Places which, in the 1950s, was the only theory "concerning systems of cities that was at all well developed" (Berry 1964) and, especially in the 1960s and 1970s, became the major concept to be applied in regional planning in Germany. Lösch, who is described as an "extraordinary personality" by Stolper in the foreword to Lösch's book, developed the first general equilibrium concept regarding the system of locations of economic activities that had ever been presented.

The three scholars worked in diverse areas: von Thünen concentrated on agricultural land use, and proposed a location theory for agricultural products. Christaller derived a concept to explain the locations, the sizes, and the interrelation of urban settlements. Lösch refined and generalized the resulting theory of central places to a concept on the "nature of economic regions." Although their work is differently

K. Fischer (✉)
Institute for Operations Research and Information Systems (ORIS),
Hamburg University of Technology, Schwarzenbergstr. 95, 21073 Hamburg, Germany
e-mail: kathrin.fischer@tu-harburg.de

H. A. Eiselt, V. Marianov (eds.), *Foundations of Location Analysis,*
International Series in Operations Research & Management Science 155,
DOI 10.1007/978-1-4419-7572-0_20,

focused, the three researchers have an important aspect in common: in each case, an economic region is defined by the major center which is at its core. The major center has some "sphere of influence" surrounding it, and this sphere of influence economically depends upon the center.

In this paper, the major aspects of von Thünen's, Christaller's, and Lösch's theories are presented, and their interrelations as well as their impact on further developments in the field are analyzed. Moreover, some of the most important subsequent work is discussed, underlining the importance of the fundamental contributions of these three authors for spatial studies and location theory.

20.2 Classical Theories on Central Places

This chapter discusses the theories developed by von Thünen, Christaller and Lösch and the most important results regarding central places. The presentation of each of the theories is followed by a short discussion of its limitations, variations and of possible extensions.

20.2.1 Johann Heinrich von Thünen's Ring Theory

Johann Heinrich von Thünen (1783–1850) was a landowner and farmer who was interested in the mechanisms that lead to the different uses of land in different areas, in different systems of cultivation, and also in the factors that influence the fertility of the soil and the prices of the produce gained from the land. To develop explanations for the phenomena he observed on his own estate, he took an analytical approach leading to a *partial equilibrium model*.

20.2.1.1 Assumptions and Development of the Theory

In developing his approach, von Thünen (1921) considers an "isolated state" (hence the title of his work), in which there is a central city and which is surrounded by wilderness. The surface of the land is assumed to be flat and homogenous without mountains or rivers, and the soil is assumed to have the same constant *a priori* quality everywhere, leading to a standard yield. Farmers are assumed to transport their products directly to the city which is the only place of consumption. Furthermore, each farmer is assumed to behave as a *homo oeconomicus* with the goal of maximizing his profit.

Based on these assumptions, von Thünen wants to determine to which use the land should be put (the "optimal land use"), depending on the distance to the town, which has a crucial impact on the cost of transportation, and under the condition that the demand of the town has to be fulfilled ("supply model"). Early in his paper and without much preceding analysis, he observes that the differences in transportation

cost, resulting from different weights and volumes of products, will lead to a *ring structure* around the central city, where the different rings grow different products. From this seminal insight he develops his model of the "Thünen rings" for agricultural activity which has been praised as the "world's first economic model" by Hall in the introduction to von Thünen's book (1966).

A major part of the first section of von Thünen's book is devoted to an analysis of the farm price of grain. On the one hand, this price—or better: what the grain is worth at a certain distance from town—is determined by the transportation cost. If the price in the central town is known, the prices in different places around the city will amount to the difference of this price and the transportation cost. Therefore, prices will decrease with increasing distance from town, and there will be a certain limit beyond which it is not profitable to produce any grain at all. This defines the limits of the respective ring. In contrast to the majority of authors, von Thünen does not—at least at first—assume that cost of transportation per unit is proportional to the distance travelled. This is due to the fact that he does take into account the amount of food for the horses that needs to be taken on each trip, and hence the unit costs of transportation are slightly decreasing. However, later on he assumes transportation cost to be proportional to distance, e.g., in the case of butter.

On the other hand, the area on which grain is grown and the fertility of the soil are important factors which influence the yield, and therefore the cost, of grain production. The fertility depends on factors such as the use of manure or the rotation of crops. Hence, while von Thünen assumes a homogenous plane and therefore the same "inherent quality of the soil" in his development of an *intensity theory*, he nevertheless takes into account the effect of different levels of fertility.

The data which von Thünen collected on his own farm lead to the proposition that a less fertile soil which produces less grain should only be used for grain when the price is high, as otherwise no profit will be made, due to the high cost of production. In studying different cultivation systems, he found that the improved system in which seven different crops are rotated is not always better than the three-field system. Which of the systems is better depends on the grain prices, with lower prices supporting the three-field system.

In his analysis, von Thünen also takes production costs into account, which he assumes consist of a money ("town-based") part and of a "farm-based" part, calculated in terms of units of grain. It is one of the important contributions of von Thünen that he studies *all* cost aspects and their impact on the choice of location for the different products. He concludes that the product that leads to the greatest decrease in cost when it is produced close to the market and hence saves the highest amount of cost should be produced there; this is the product which leads to the highest *land rent*. Hence, the *land rent* or so-called "locational rent" is the profit which results from the land itself, after the deduction of all cost and the interest for buildings and other objects apart from the land. This rent is the same as the classical "economic rent," but, as von Thünen points out, it is different from the "land rent" in the sense of Adam Smith who does not deduct the respective interest. The land rent also represents land value, and therefore it equals the maximum amount a farmer would be ready to pay for using the land.

From this, Lösch (1962) concluded that, under the assumption of linear transportation costs, the land rent R for a specific product is determined by relation (20.1).

$$yp = yc + yfm + R \quad \Leftrightarrow \quad R = y(p - c) - fmy, \tag{20.1}$$

with the following parameters an variables:

y: yield per unit of land (in tons of product)
c: production cost per ton of product
p: market price per ton of product
f: freight rate per km and ton
m: distance to market (in km)

Therefore, the returns from a product equal the sum of the production cost (which may vary with the distance from town), the transportation cost and the land rent. In other words, the land rent is a residual, as it equals the difference between the cost resulting from the production and transportation of the product at the "marginal site," which is equal to the product's return, and the costs that have to be incurred at the place currently under consideration. (The "marginal site" is the site furthest from town where the product still has to be produced in order to satisfy the towns demand.) The land rent then results from the advantage that a farm has over the "worst" farm that still supplies the market with the same product (usually a farm with high production or transportation costs), and therefore it is a measure of *marginal productivity*: "Land rent does not spring from capital or labor, but from the fortuitous advantage one farm enjoys over the others in the quality of its soil or location" (von Thünen 1966). It can also be interpreted as an opportunity cost: if, for example, another crop were to replace the growing of grain, it should lead to a lower total cost, *including* the grain's land rent.

Consequently, the limits between the rings in which the varying uses of land take place are defined by the land rent: one ring ends and the next ring begins at the point where, for instance, the land rent for the produce of the first equals the land rent for the produce of the second. This concept of marginal productivity can be illustrated as shown in Fig. 20.1. A similar figure on milk production is provided by Lösch (1962).

Figure 20.1 illustrates the tradeoff between land rent and transportation cost. It shows how the Thünen rings result from differences in the costs of transportation (or production) and in the rent that a farmer who raises a certain crop would be ready to pay at a certain distance from the town (named the "bid-rent" by Alonso 1964). The faster the rent for a certain product diminishes, shown by the steeper respective line in Fig. 20.1, the closer to the town the product must be grown.

The Thünen rings can also be derived analytically as explained by Lösch (1962). Consider two products *I* and *II*, such that product *I* leads to a larger rent R_I than product *II*, which leads to rent R_{II}. Using relation (20.1), we can then write

$$R_I > R_{II} \Leftrightarrow y_I(p_I - c_I) - fmy_I > y_{II}(p_{II} - c_{II}) - fmy_{II} \tag{20.2}$$

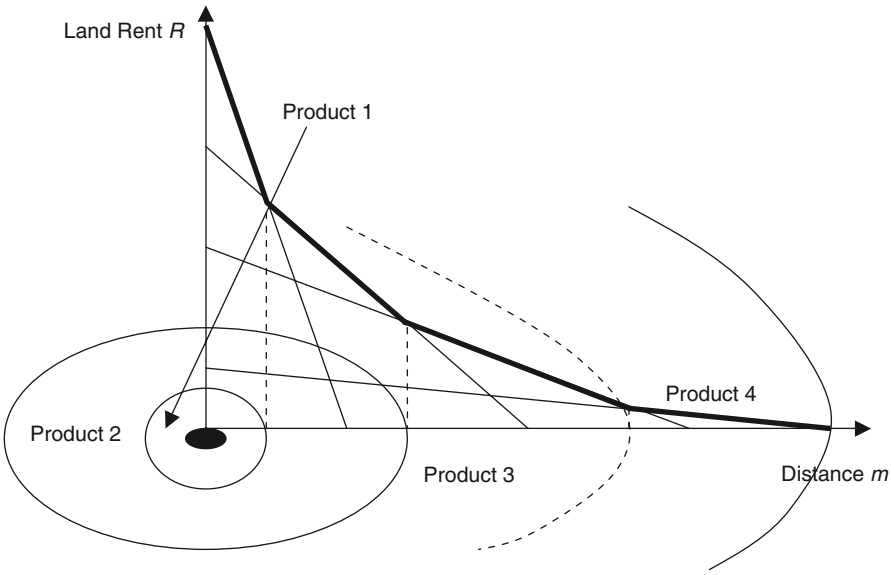


Fig. 20.1 Land rents for different products and the development of von Thünen's rings

At the town, the distance m from the market is 0, and hence this is equivalent to:

$$y_I(p_I - c_I) > y_{II}(p_{II} - c_{II}), \tag{20.2'}$$

which leads to the inequality

$$\frac{y_I(p_I - c_I)}{y_{II}(p_{II} - c_{II})} > 1 \tag{20.3}$$

Therefore, inequality (20.3) must be satisfied, if product *I* can be more profitably produced in the town than product *II*. Analogously, a condition for *II* to be more profitable at the periphery of the area can be derived. The two resulting inequalities are:

$$\frac{y_I}{y_{II}} > \frac{y_I(p_I - c_I)}{y_{II}(p_{II} - c_{II})} > 1 \tag{20.4}$$

If only one of these conditions is true, either crop *I* or crop *II* is grown exclusively as the other cannot be advantageously grown anywhere. If both conditions are fulfilled, we find that if $y_I > y_{II}$, crop *I* will be grown in the inner ring (and note that if $y_{II} > y_I$, the conditions have to be reversed and product *I* will be in the outer ring). Lösch (1954) wrote "the choice of a crop is then a function of distance." While the yield is higher for product *I*, the maximum possible profit per unit, $(p_I - c_I)$, must be smaller, resulting from the first part of condition (20.4). For product *I*, therefore, the profit per unit is absorbed more quickly by the cost of transportation than it is the

case for the other crop. In summary, at some distance from the market the two crops give the same profit, and further out in the periphery, the production of the second product will be advantageous.

20.2.1.2 Von Thünen's Rings

Von Thünen gives a very specific presentation regarding the allocation of the production of different products to the different rings around a central town. This central town is the (only) place of consumption, i.e. the place where the products are sold to the customers, whereas production takes place in the respective rings. According to von Thünen, in the first ring dairy production and intensive farming are to be found, as these products need to be brought to the market very quickly in order for them not to perish; as in von Thünen's time, cooling devices, and especially refrigeration trucks, were unheard of. Obviously, the selling price of the milk has to be so high that it is not attractive for the farmers to put the land to any other use than food production for the cows, with the exception of the production of selected products, such as strawberries, because they would not survive a long transport, and potatoes, because it would be too expensive to transport them over a long distance. A special feature of this ring is that manure is mainly bought from the city. Von Thünen calls this ring the "free cash cropping" ring (in German: "Freie Wirtschaft").

The second ring contains timber and, closer to the town, firewood production. During von Thünen's lifetime, wood was needed for heating and cooking, and as its transportation was difficult due to its weight, it was located close to the place of consumption. In the third ring, crops such as grain are found, because they are not perishable and are much easier to transport than wood. Three different "sub-rings" are defined by the different types of crop rotation: an inner ring, in which the crop alternation system is used, a middle ring with the improved system, and an outer ring with a three-field system.

The fourth (or sixth, if we count the three sub-rings mentioned above separately) and final ring should be devoted to animal farming, including, for example, the production of butter. Von Thünen concludes that butter production should take place at about 30 miles from the town, as it is not worthwhile to produce it at a shorter or longer distance. Closer to the city (except for the "Freie Wirtschaft" ring), the land rent for stock farming is negative and no stock farming will take place there, because close to the town the cost of production steeply increases and overcompensates the decrease in transportation cost. This is due to the fact that crops such as rye are less expensive at larger distances to the town due to the lower land rent, and therefore stock farming, which has a higher production but lower transportation cost than rye (due at least partly to the *consumption* of rye by those working in stock farming), can be done more efficiently further away from the town.

Stock fattening can commence far away from town but has to be finished close to it, as the animals lose too much weight on their way to the town if they have to walk long distances. Also, young cattle can be raised at the outer ring. It should be noted that "industrial" crops which extract a lot of fertility from the ground (e.g.,

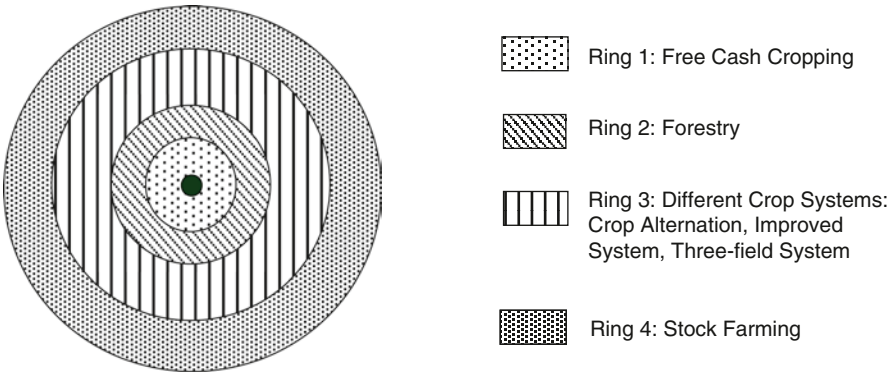


Fig. 20.2 The von Thünen rings in overview

oilseeds, such as rapeseed, tobacco, and flax), are to be found in this outmost ring, as are sheep farming and the production of wool. At about 50 miles from the town, all farming activity ends, as the land rent is too low to support it. Overall, there is a tendency of rising intensity towards town, but there are exceptions like forestry (ring 2). The concept of the rings is illustrated in Fig. 20.2.

Assuming identical costs and/or yields, von Thünen states some general rules regarding the location of different agricultural products:

(a) The higher the production cost, the farther from the market the product should be produced.
(b) The higher the yields of a product, the closer to the market it should be produced.
(c) The crop extracting the most fertility from the soil should be grown farther from the market.

While (b) agrees with the result derived by the analytic procedure, result (a) is somewhat less convincing. Von Thünen states that the lighter or more expensive good is produced farther away, as freight is not so important for such goods, whereas Lösch (1962) subsequently argues that there are cases where it makes sense to produce the cheaper good at the periphery. Moreover, in contrast to von Thünen, Lösch comes to the conclusion that the von Thünen rings are only one possible result. They will form if the economy is dynamic (in the sense that farmers react to changes in the market, such as the introduction of a new crop), whereas in a traditional economy, reversed von Thünen rings are possible as well.

In the second part of his book, von Thünen himself discusses and criticizes his own major assumptions. He states that the soil usually is not of the same quality everywhere, and he drops the assumption of only one central town. The existence of additional smaller towns leads to "sub-centers" which have their own smaller systems of rings. Of course, there are interdependencies between the different rings, as the land rents have to be equal where the borders meet.

Von Thünen also states that towns must be distributed such that their location maximizes national income. According to him, such a pattern will result if all (agri-

cultural) goods are produced at the location which leads to the lowest cost. He has got in mind some kind of spatial equilibrium structure; however, he does not elaborate on the question why and how towns should come into existence at the optimal places and in an optimal pattern.

Moreover, because towns tend to be found near rivers, von Thünen considers the influence of rivers on the regional pattern and concludes that rivers lead to zones of equal transportation costs that stretch along them, as transportation on the rivers can be assumed to be cheaper than transportation on land.

20.2.2 Christaller's Central Place Theory

Walter Christaller (1893–1969) was the first researcher who focused on systems of settlements and the hierarchy of towns, instead of studying them as single units. Nevertheless, von Thünen's work, which mainly deals with a single town, served as one of the major foundations of Christaller's studies.

Christaller's theory of "Central Places" is presented in his book "Die zentralen Orte in Süddeutschland," published in 1933 ("Central Places in Southern Germany," the English translation, appeared in 1966). Christaller is mainly interested in the laws and principles that determine the number, size, and distribution of towns, in order to explain the existing structures he observed in Southern Germany. From his point of view, these could not be explained by geographical aspects, but only through economic theory and, therefore, economic laws.

20.2.2.1 Assumptions and Basic Terms

Christaller's book comprises four parts: a theoretical foundation, the development of a method, a descriptive part on real phenomena, and a final verification of the theory. Observing that centralization around some kind of core or nucleus is one of the basic principles in nature, he states that it is the major purpose of a (market) town to be in the centre of an agricultural area. Being "central" and the notion of centrality are therefore relative notions: Christaller defines *central places* as those settlements which are important for the surrounding area because they provide it with so-called central goods. *Central goods and services* are produced at only a few central places, but are needed and consumed at many different and dispersed places. Examples are medical services, cinemas, schools, and stores. Christaller emphasizes that "centrality" is not so much about the production of goods, but that sales and services are primarily offered at central places due to the capital requirements related to establishing those services.

In order to determine the laws according to which the central places develop, Christaller makes a number of key assumptions: first, the area under study is a flat and homogeneous surface (isotropic plane), on which the population is evenly distributed. Next, all consumers have the same demand regarding the "central goods," and they all have the same income and identical purchasing power. However, those

who live further away from the central place—which in the first instance is assumed to be located in the middle of the respective area—have to use part of their budget (in terms of time and money) in order to travel to the central place, meaning they have to pay transportation costs, and therefore not all of them can spend the same amount on central goods. Finally, transportation costs are assumed to be proportional to distance, and hence customers always prefer the nearest central place (if there is a choice).

Hence, it should be noted that in contrast to von Thünen who assumes the existence of one town—the center of consumption—and dispersed production of (agricultural) goods which have to be transported to the town by the producers (farmers), Christaller builds his approach on the assumption that the consumers have to travel to the central place in order to buy the central goods. Therefore, while in von Thünen's theory a standard transportation cost function can be used to model the impact of transportation on the price of the goods, in Christaller's theory the disutility resulting from travelling plays a role. In other words, "*economic distance*" is the most important factor in determining if a place is indeed central, and this notion relates to the cost of transportation, the time a consumer has to invest in transportation, and the disutility connected with it. The economic distance leads to the *range of a good*, which is the maximum distance people are ready to travel to buy the good, but the willingness to travel to a central place will be different for different individuals, and hence the *economic distance* or range is also an individual measure. Each central place is surrounded by a so-called *sphere of influence*, which is the market area that it serves. The size of this area depends, among other things, on the price of the good and on the transportation cost. It has to be noted, however, that better roads or railways can facilitate transportation of central goods, and therefore reduce cost and "transportation resistance," which leads to a higher consumption of central goods, to an increase of the ranges, and to better developed central places.

With relation to the sphere of influence, Christaller distinguishes between *centers of high and low order*. He assumes that at a center of higher order (such as a place with a university) all the goods and services of lower order (like a school) are on offer as well, but not the other way around. This leads to a *hierarchy* of central places, where the importance of a center is not equal to the number of people living there, but depends on the intensity with which central functions are executed. This, in turn, is related to the number of central goods that are on offer and their ranges.

On the one hand, the range of a central good depends on the distribution of the population and on the order of the central place: the higher the order of the central place (and usually, the larger the place itself), the more different central goods are on offer, making the place more attractive and increasing the range of the goods. On the other hand, the amount of central goods that are consumed and hence the *importance of a central place* depend on the sphere of influence, on the number of inhabitants, and on the population density: the higher the population density, the more central goods will be consumed, and the larger the sphere of influence, the better the central place will be developed.

The characteristics of the goods are important as well: A central good that can easily be substituted will have a lower range than a good which can hardly be sub-

stituted at all. For example, bread can be bought in many stores and hence can be substituted without difficulty, while a special wedding cake has to be ordered at a specialized bakery. Moreover, it matters if the good is available only in limited amounts or if there are no limitations, if the prices are fixed or variable, and if the good is also offered at other places. With respect to the last aspect, Christaller differentiates between the *absolute range* (the distance at which people do not buy the good at all) and the *relative range* (the distance at which they prefer to buy the good from another central place).

These different aspects lead to the individual range of a good, which is an upper limit. There is, however, also a lower limit, the so-called *threshold*, which is the minimum distance from which people have to travel to the central place to buy the good in order to make offering it worthwhile and profitable. Based on range and threshold, the place can be classified as a "higher order place" if both are large, and as a "lower order place" if both are small. If the upper limit is high and the lower is small, the good can be offered in many places, and hence it is a "low order" good.

Christaller concludes that each central place will expand its market area as much as possible, and because the ranges are identical for central places of the same order, these central places have to be spaced regularly. Furthermore, there is a tendency towards more than one central place, as when people have to travel smaller distances, more people can get serviced and total consumption increases. However, those who offer the service or good have to be able to make a living from it, which requires enough customers to support them. Hence, the optimal constellation must be such that the demand of the whole population is satisfied from a minimum number of centers, and this leads to the maximum possible profit for those who offer the good. In this sense, Christaller is aiming at an *equilibrium pattern* of central places. It can be called a supply equilibrium, as it is aimed at serving the whole population.

20.2.2.2 Christaller's System of Central Places

Under the assumption that each customer is always served from the nearest location (an assumption that, to this day, is made by most location analysts, e.g., in competitive location models), Christaller develops a basic spatial pattern which the locations of the central places have to follow in order to serve the whole population with all central goods. The development of this pattern is explained as follows.

First, assuming one central good with a certain range r, and a homogeneous plane with an evenly distributed population, a simple, circular sphere of influence results, as shown in Fig. 20.3.

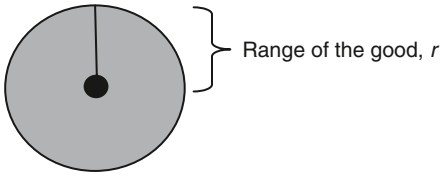


Fig. 20.3 Circular structure of the market area around the central place

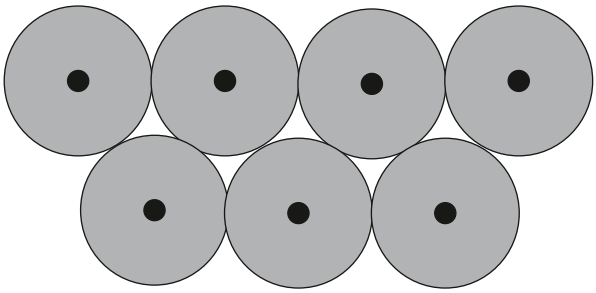


Fig. 20.4 Pattern resulting from many circular market areas

The black dot represents the central place from which all customers within the range of the good, shown by the gray area, can be served. As those people who are not in this area cannot be served from there, additional central places are needed to serve them. The resulting pattern, the so-called hexagonal circle packing, which is the densest packing of circles in the plane (Fejes 1960/1961), is shown in Fig. 20.4.

However, as is obvious from Fig. 20.4, in this pattern there are areas which are not served at all. Hence, to serve all customers, the central places have to be moved a little closer together, such that the spheres of influence overlap. Defining central places for the goods of lower order at those points where the market areas meet leads to a hexagonal pattern of central places and to market areas as illustrated in Fig. 20.5.

As stated above, if the whole population is to be served by a minimal number of central places, these places have to be in a regular pattern. When the central places are arranged in the form of equilateral triangles, as it is the case in Fig. 20.5 (note the dotted lines), the market area for each supplier reaches a maximum and the whole population is served: according to Christaller's objective, this is the optimal spatial structure of central places.

In order to determine the distance between the central places of the lowest order, the length of one of the edges of the basic triangle, ℓ, has to be found. The Pythago-

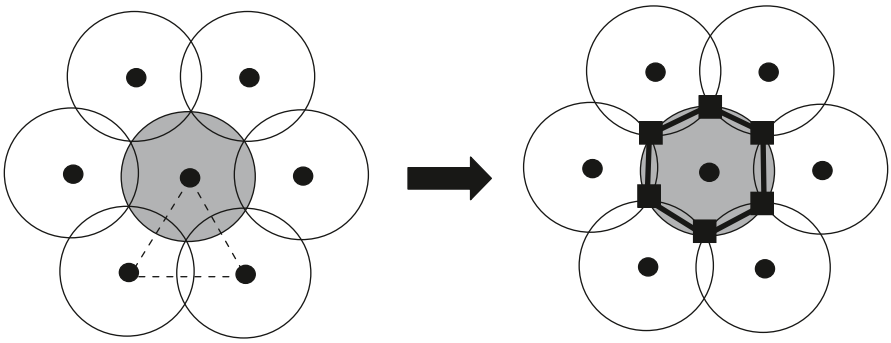


Fig. 20.5 A hexagonal market area

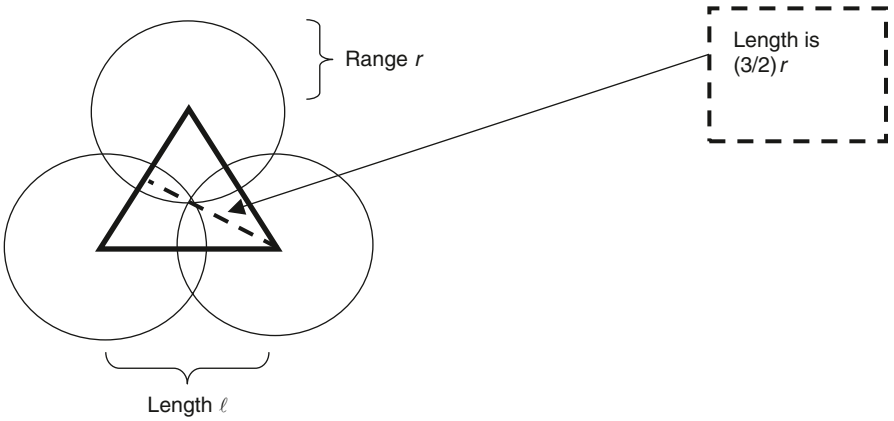


Fig. 20.6 Basic triangle of central places. (Similar to Lang 2002)

ras formula (with r being the range of the lowest order good which is assumed to be known) leads to relation (20.5).

$$\ell^2 = \left(\frac{3}{2}r\right)^2 + \left(\frac{\ell}{2}\right)^2 \Rightarrow \frac{3}{4}\ell^2 = \frac{9}{4}r^2 \Rightarrow \ell^2 = 3r^2 \Rightarrow \ell = r\sqrt{3} \qquad (20.5)$$

Figure 20.6 illustrates relation (20.5) with such a triangle.

The triangular basic shape leads to the maximum possible market areas for the central places and to the hexagonal structure that has been already described above. As there are different central goods of increasing order, a system of hexagons of different sizes results (see Fig. 20.7). The resulting location pattern is called *supply*

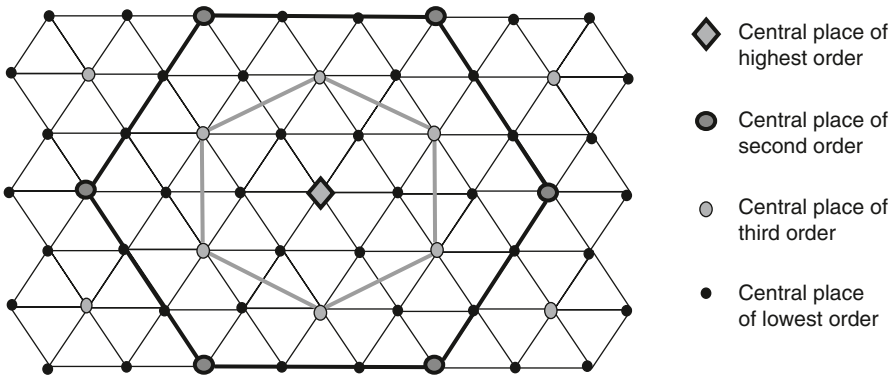


Fig. 20.7 Hierarchical structure of central places

principle, marketing principle or $K=3$ *system* because the market area of a higher order place is three times the size of the market area of the next lower order place. Moreover, three marketplaces are served with higher order goods from a place of higher order, namely the higher order place itself and one third of each of the six surrounding places of lower order. (Note that customers are assumed to patronize the closest location, allowing the demand of the lower order place to be split evenly among the three higher order places to which it has the same distance.)

The smallest central places are only "supportive places" and are to be found in the middle of the triangles shown in Fig. 20.7. The resulting structure of settlements shows certain basic principles: the lower the order, the larger the number of settlements of this order will be, while higher orders will serve larger areas.

Christaller's first main result, therefore, is that there is a *regular pattern of central places* which follows certain laws: there is an important place "in the center" (highest order), with six small places around it (lowest order). Then follows a ring of medium sized places around that and, following another ring of small places, there is a peripheral ring of medium-to-large places (second-highest order). The second result implies the *existence* of those different categories of central places, and, according to Christaller's third result, the *number* of central places of each order increases geometrically, with the lowest number of settlements for the highest order. The numbers of central places will then develop as shown in Table 20.1.

While the smallest central places offer only a few goods—Christaller estimates ten—the next larger places might offer about 40, the next 90, then 180 and 330 goods, and so forth, so the number of goods on offer increases. The importance or level of a central place is directly related to this number of goods.

Christaller's system or principle is *rational*, as it leads to an optimal use of the central places and to the smallest possible loss in the economy; the producers and salesmen make the maximum profit, and all the consumers are served. Of course, in reality there are many obstacles to this optimal pattern; therefore, for example, places of the lowest order can be missing completely. Historical development has a big influence on the existing structure of central places, too, as when one or two big central places already exist, they determine the structure of the smaller places around them. However, governments can help to establish a more efficient structure by setting up their administrative offices in the right places.

It is a special feature of Christaller's approach that he does not develop a structure based on the (existing) traffic conditions, but that he assumes the traffic conditions to result from the system of central places. He argues that the existence of central goods, for the exchange of which people have to travel, leads to the ex-

Table 20.1 Numbers of central places and market areas

Order of place	1st (highest) order	2nd order	3rd order	4th order	5th order
Number of places	1	2	6	18	54
No of market areas	1	3	9	27	81

istence of traffic and the respective infrastructure, but not the other way round. Consequently, traffic structures in the supply system will be dissatisfying because the central places of different orders usually are not to be found on straight lines. If, for example, two central places of the second-highest order (gray with bold black line in Fig. 20.7) are connected, then only two places of the lowest order are to be found on that connection (see bold black lines in Fig. 20.7). The traffic structures can be modified to include more places of different order, but in essence the supply principle does not lead to a good solution of traffic and transportation problems.

Therefore, Christaller also considers the *transportation or communication principle*, which is aimed at the realization of as much transportation as possible, (the maximization of connectivity) at the lowest possible cost (a minimal network length). This leads to a different structure of central places which can be illustrated as shown in Fig. 20.8.

According to the transportation principle, the lower order places are to be found on the edges of the hexagon instead of the corners. Therefore, each higher order place serves a total of four places of the lower order, the place itself and half of each of the six neighboring places, and hence, this principle is also called the $K=4$ *principle*. There is a larger number of central places in this "linear" traffic-oriented system, and hence the "supply principle," which means to serve all customers from the minimal number of central places, does not work here. Moreover, as more central places lead to a higher demand for central goods, demand will be higher and therefore traffic will be more intense when the transportation principle is applied.

A third principle that is discussed by Christaller is the *administrative or political principle*. Here it is necessary to find a unique allocation of some lower order places to a higher order place, such that the respective group of settlements defines an administrative district. In this principle, seven central places—one of higher order and six of lower order—are put together to build a unit, which is why the system is also called the $K=7$ *system*.

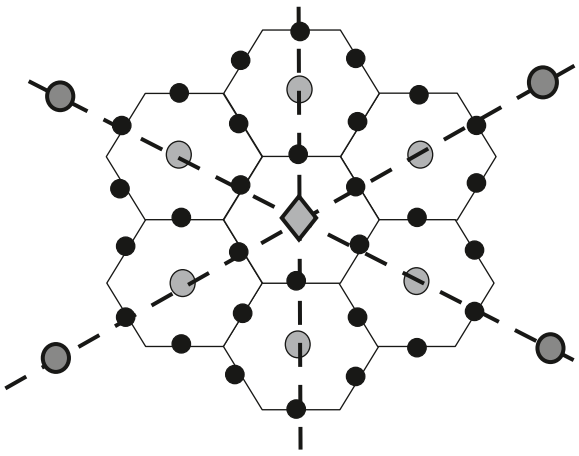


Fig. 20.8 The $K=4$ system or transportation principle

20.2.2.3 Variations and Empirical Implications of Christaller's Theory

After the development of the three principles described above which are of a static nature, Christaller proceeds with introducing dynamic aspects into his theory. He discusses changes in demand resulting from changes in the size of the population, i.e., an increasing (decreasing) number of people, which leads to higher (lower) importance of certain central places, and can even result in new central places coming into existence (or old places disappearing). Changes of prices are also considered, as are changes in the number of central goods that are on supply at certain central places. Christaller points out that the introduction of an additional central good does not only lead to higher sales of the respective good, but also to higher sales of other goods. The major advantages of agglomeration are discussed by Christaller, as are the effects of the land prices (and related rent), which usually will be higher in places of higher order than in places of lower order. He considers this to be an obstacle to too much centralization.

Furthermore, he states that technological progress will lead to cheaper production and transportation, as indeed it did in the last century. A decrease in the cost of transportation leads to more money being available for actually buying the goods. Moreover, people will be able to travel longer distances to buy the goods, thus increasing the range of the goods. Therefore, the goods can be offered in fewer and more distant central places, while still serving all customers. As a result, the larger central places will grow even more while others lose importance, and smaller central places close to a larger one may even disappear altogether. For example, if a new railroad station is built, central place(s) which now can be reached by train will increase in importance, as the "economic distance" to them decreases. Moreover, people will buy different goods at this place which in turn will lead to an increase of the range of the goods offered there, leading again to more agglomeration.

In the second part of his book, Christaller establishes a relationship between his theory of central places and the situation in Southern Germany. He applies his theory to the existing structures to judge which settlements are central and to determine their order. To do this, he first classifies different administrative, cultural, medical, entertainment, and organizational services as well as sales, crafts, and traffic services as of low, medium or high importance. For example, while a police station is of low importance, a Lower Court would be of medium and the Superior Court of major importance. To quantify the importance of a central place, he uses the number of private phone lines that exist in a place. At his time, he found there to be one phone line per 40 inhabitants on average. According to his approach, a central place is of higher importance if it has more than the "expected number" of phone lines, and of lower importance, if it has less.

In his general discussion, Christaller starts from the least important central places (the "supportive" central places) and then works his way up through *M*-places (markets), *A*-places (Amtsgericht, the Lower Court), followed by the *K*-places (Kreisstadt, small district town), the *B*-places (Bezirkshauptort, major district town), the *G*-places (Gaubezirksstadt, Superior Court), the *P*-places (Provinzialhauptort, seat of provincial government), and up to the most important *L*-places (Landeszentrale,

major central city). But in the specific analysis of Southern Germany, he defines the *L*-places which are central for a larger area first, and then works his way down to the smaller places. He calls the central place of highest order "system-building" or "system-defining," as it is the basic element of the system. Around this major place, central places of lower order are located in a regular pattern as illustrated in Fig. 20.7. This might be the reason why it is often argued that Christaller went "top down" in his analysis, while Lösch went "bottom up" (see von Böventer 1963). However, they both start their theoretical development from the smallest settlements while finding the existence of a major central place to be crucial for the resulting overall structures.

The typical distances that Christaller observes in reality are 7–9 km between each two central places of the lowest order (*M*-places), and hence the radius of the sphere of influence is 4–5 km. This is the distance that can be covered by a one-hour walk, and obviously there are many central goods for which the "critical distance" or range is about an hour. The distances between the central places should—according to the supply principle and going from the lowest to the highest level—then obey the following scheme:

$$4 - 4\sqrt{3} - 12 - 12\sqrt{3} - 36 - 36\sqrt{3} - 108 - 108\sqrt{3}, \text{ i.e.,}$$
$$4 - 7 - 12 - 21 - 36 - 62 - 108 - 187.$$

These are the theoretically correct distances. They are not always found in reality, though, and if they are not, there must be an "explanation which is due to special economic, historical or natural circumstances."

In the third part of his work, Christaller studies the five different central places of highest order in Southern Germany, *viz*., Munich, Nuremberg, Stuttgart, Frankfurt and Strassburg, and discusses the urban structures around these central places. He finds that in the case of Munich and Nuremberg, his rational system does seem to work and reality fits the central place system rather well. In the case of Stuttgart, however, the results are not that clear cut. While in most cases the structure follows the supply principle, in other cases it can be better explained by the transportation principle, as is the case for the city of Frankfurt.

In the fourth and last part of his book, Christaller concludes that his three principles indeed are "laws of distribution of central places" that are at work in different areas. He considers the supply principle to be the major distribution principle, and the transportation principle and the administrative principle to be secondary and, therefore, less important.

20.2.3 Lösch's Theory of Economic Regions

In 1940, August Lösch (1906–1945) published his book "Die räumliche Ordnung der Wirtschaft" (the English translation "The Economics of Location" appeared in 1954), in which he refined and generalized Christaller's theory of central places—

but, as he claims, without even knowing Christaller's book beforehand. Similar to Christaller, the focus of Lösch's work is on the interdependencies of locations of production and consumption and on the nature of economic regions (such as the distribution of population and cities), and not on the isolated study of one specific location or location choice.

Among other research, Lösch's book is based on the work of Palander (1935) and Ohlin (1933). He criticizes Weber's partial equilibrium theory, and, following Stolper in his foreword to Lösch's book, Lösch "…was the first to present a full general equilibrium system describing in abstract the interrelationship of all locations." However, in contrast to Christaller, whose aim it is to find a way to supply the whole area with a minimum number of marketplaces—i.e., this is essentially some kind of minimum covering problem as it was later introduced by Toregas et al. (1971)—Lösch concentrates on the effects of competition which lead to the smallest market areas possible. Hence, essentially his objective is to maximize the number of independent economic units and, therefore, of locations.

20.2.3.1 Introduction

In Part I of his book, Lösch discusses previous work in the area (such as the work by von Thünen, discussed in Sect. 20.2.1 of this chapter), and lays the foundation of his own work. Lösch's basic assumption is that each location is chosen such that utility is "as great as possible." For an industrial location, this leads to the "location of the greatest nominal profit." If demand is completely inelastic, as assumed by Weber (1909), this is the point of minimum transportation cost. This point can be determined by different means, especially geometrically by using Weber's isodapanes. These are lines of identical total freight per unit, in which costs for the transportation of raw materials as well as costs for the transportation of the final products are taken into account. Production costs must also be considered, as Lösch points out. (As Isard (1956) demonstrated, the point of minimum transportation costs remains optimal with respect to profit maximization, if the production function coefficients are fixed.)

However, it has to be taken into account that usually demand depends on the price and the location, and that the three aspects are therefore interdependent. In other words, it cannot be assumed that the demand is independent of the location, because the market area depends on the location, and so market area and demand will change with it. Lösch points out that in this situation, isodapanes are of no use at all, and that the only possibility to find the best location is a "trial and error approach."

While his predecessors such as Weber (1909) only concentrate on parts of the system instead of considering the system as a whole, Lösch presents an integrated analysis that is one of his major results, *viz*., the "general equilibrium in space." This general equilibrium results from two forces: the maximization of individual advantages (utility) and the maximization of the number of independent economic units. From these, Lösch develops five conditions which define the equilibrium.

According to condition 1, the location of each individual (be they farmer, entrepreneur, or customer) must be as advantageous as possible, meaning it has to lead to the highest possible profit or utility. Moreover, there are three conditions which make sure that the number of enterprises reaches its maximum. Condition 2 states that there must be so many locations that the entire space is covered. According to condition 3, all abnormal profits must disappear, leaving prices equal to costs. Condition 4 states that all areas of supply, production, and sales must be as small as possible, or more entrepreneurs would come onto the market. Finally, condition 5 implies that "at the boundaries of economic areas it must be a matter of indifference to which of two neighboring locations they belong."

These five conditions and the resulting types of equations define "the size and limits of market areas, the situation of production locations within them and within the entire area, and the f.o.b. prices." The resulting system of equations, however, cannot be solved in general terms. As Stolper in the foreword to Lösch's book puts it, the theory is "too all-inclusive to be applicable." Moreover, in the subsequent discussion, Lösch points out that the best location for the producers does not have to be optimal for the consumers, and that the structures in industrial production which lead to the existence of cities are different from those in agriculture, as the latter are much more dispersed.

20.2.3.2 Lösch's Theory of Economic Regions

Lösch states in the preface to his work that "Parts II and III are the kernel of the whole book". The discussion below will mostly concentrate on Part II which contains Lösch's development of economic regions and relates to Christaller's system of central places. An overview of the content of Part I was already given above in Sect. 20.2.3.1, and a brief overview of the contents of Parts III and IV can be found below, at the end of Sect. 20.2.3.3.

Part II of Lösch's book is dedicated to *economic regions*. In his analysis, Lösch assumes that raw materials are evenly distributed, that the whole area is homogeneous and that there exist only regularly distributed farms. If now any of these farmers starts to produce a good such as beer, this good will be bought by other farmers, but only by those who are not too far away. Hence, it can be assumed that demand decreases with increasing distance, and furthermore that only those who live within the necessary shipping distance ("Versendungsreichweite") will buy the product from the respective supplier at all. (Note that Lösch does not consider the range or the maximum distance people would travel to buy the product. Instead, he only considers the threshold, i.e., the distance that has to be covered to render production worthwhile, as in contrast to Christaller he concentrates on minimum sized market areas).

Due to the homogeneity assumption, Lösch's first approach leads to circular market areas as does Christaller's. But also under Lösch's assumptions, this structure cannot be optimal because parts of the plane are left unused (in contradiction to his conditions 2 and 4). Therefore, the circles have to be reduced to hexagons or

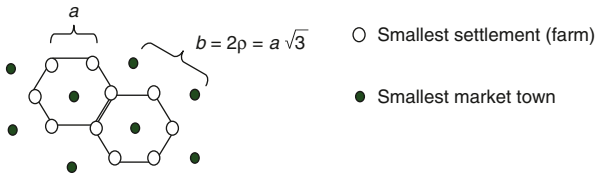


Fig. 20.9 Honeycomb scattering, smallest market areas

honeycombs, which are completely enclosed in the circles and cover a somewhat smaller area. Those hexagons will have the minimal size necessary to support the living of the suppliers, in accordance with condition 4, and will therefore allow for the maximum number of independent enterprises. The size of the hexagons belonging to any specific product can be described by the radius of the inscribed circle, ρ, which depends on the production cost and the demand. If now the smallest settlements (farms) have a distance of *a,* and their areas are regular hexagons, they will be found in a form of "honeycomb scattering" as shown in Fig. 20.9.

The distance between the smallest market towns, *b,* corresponds to the diameter of the circle inscribed in the hexagon, 2ρ (where ρ is expressed in freight costs and *b* is expressed in kilometers). Finally, the furthest distance at which the good must be sold to make its production worthwhile is called *nV*. This corresponds to Christaller's threshold.

The smallest possible value for *nV* is *a* (if we assume production to take place in one of the settlements) and the smallest number of settlements served is three, as in Christaller's model (each market town serves 1/3 of each of the six settlements surrounding it, and it serves itself). The distance between two market towns, *b*, is also the same as in Christaller's model, i.e., $b = a\sqrt{3}$. However, in contrast to Christaller, Lösch argues that there could be products that still have a threshold of $nV = a$, but for which the number of settlements served is not three, but four. This is illustrated in Fig. 20.10.

Each of the small settlements is served from two market places, so a total of four settlements is served from each place that offers goods of order 2, and for such a good, the distance between two places offering it is $b = 2a(= a\sqrt{4})$. In other words, the market area is bigger than in the case of the first good, but the necessary shipping distance remains the same, as only a larger fraction of the same settlements is served. This is a general result: with the increasing order of the goods, the

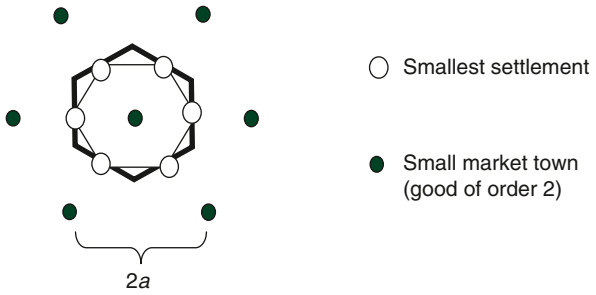


Fig. 20.10 Second smallest market areas

Table 20.2 The ten smallest economic areas

Area #	1	2	3	4	5	6	7	8	9	10
n	3	4	7	9	12	13	16	19	21	25
b	$a\sqrt{3}$	$a\sqrt{4}$	$a\sqrt{7}$	$a\sqrt{9}$	$a\sqrt{12}$	$a\sqrt{13}$	$a\sqrt{16}$	$a\sqrt{19}$	$a\sqrt{21}$	$a\sqrt{25}$
nV	a	a	a	$a\sqrt{3}$	$2a$	$a\sqrt{3}$	$2a$	$2a$	$a\sqrt{7}$	$a\sqrt{7}$

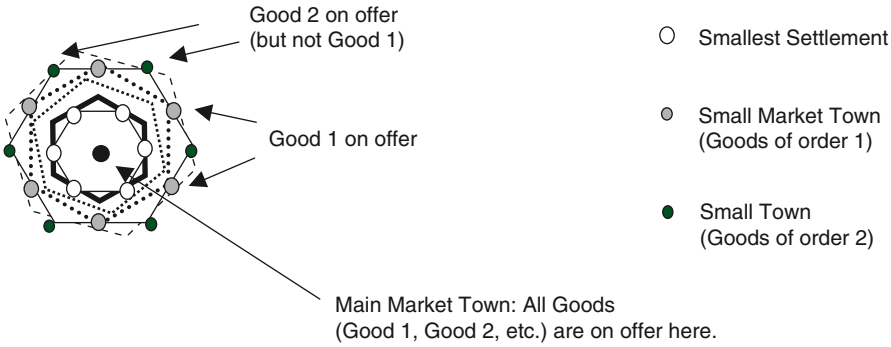


Fig. 20.11 The six smallest market areas

distance b between the market places and the number of places served increases monotonously, whereas the necessary shipping distance nV does not increase in each step, as shown in Table 20.2. The resulting system of hexagonal market areas is presented in Fig. 20.11. Only the first six hexagons are given here (for a similar presentation, see also Lang 2002); the development of the areas of higher order proceeds analogously.

Table 20.2 summarizes the development for the 10 smallest possible economic areas (though they do not all have to exist) and shows the number of settlements served n, the distances between the different centers b, and the necessary shipping distance nV.

A comparison with Christaller's concept shows that the first and the fourth hexagon are the same as in Christaller's model, while the second and third market areas are different. In contrast to Christaller's approach, as can be seen from the development described above and as is illustrated in Fig. 20.12, Lösch does not assume that each good of lower order is on offer in all places of higher order. The only exception is the most central place (the town of the highest order) that offers all the goods. However, the smaller centers specialize in different goods, and they are therefore not in a strict hierarchical order.

According to Lösch, not all possible market areas as they are illustrated above have to actually exist. On the one hand, the resulting market area might be too small to make producing and selling a specific good worthwhile and, on the other hand, the splitting of settlements between central places is not a stable arrangement. In his opinion, it is therefore especially likely for the market areas 3, 6 and 8 to be established, as in these constellations no splitting is necessary.

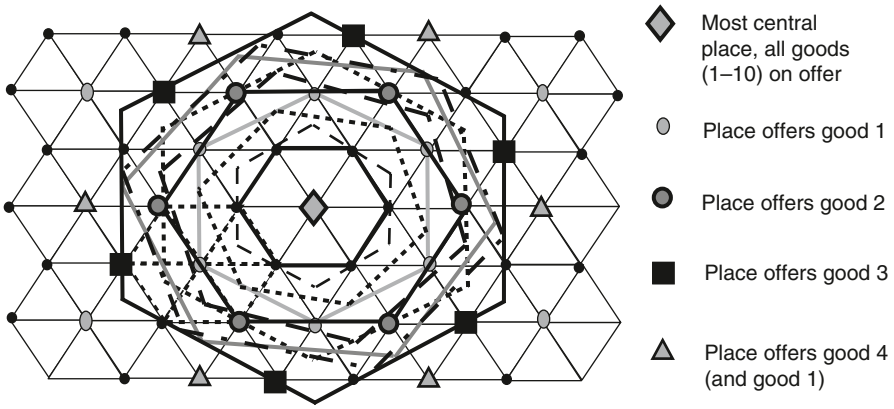


Fig. 20.12 Lösch's system of market areas

As can be seen from Fig. 20.12, there are areas in which more central places are to be found, and places in which there are less of them. Specifically, there is an "empty" ring without central places around the main central place, and there are sections with more central places in some kind of "wheel structure" around the main place. This structure is illustrated in Fig. 20.13.

In the gray spokes, we find places at which more than one good or class of goods is on offer, and these are central places or sites that, as in Christaller's model, are to be found in a regular pattern. The resulting structure for each of the goods (or better: for each of the sizes of market areas, as goods with the same size of market

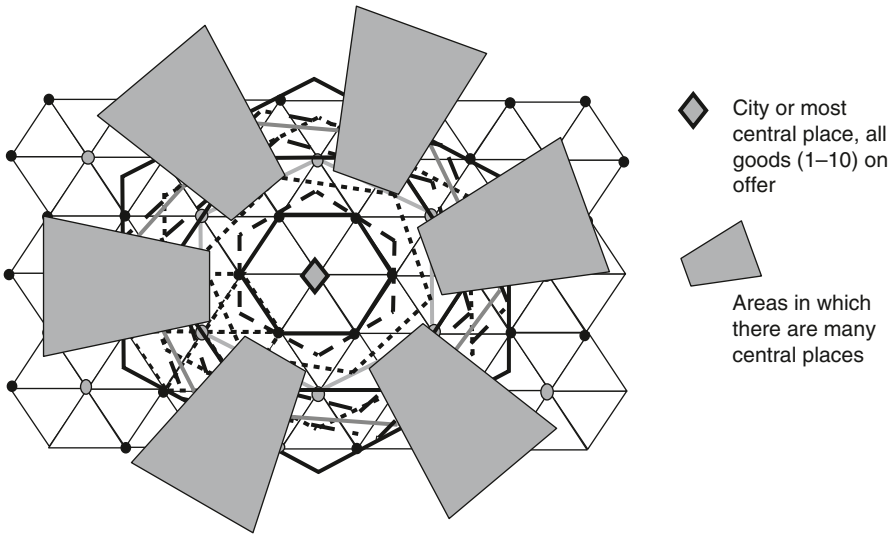


Fig. 20.13 Lösch's system with "city-richer" and "city-poorer" sectors

area will belong to the same class and hence to the same network) is a *honeycomb network*. The market areas of a specific size are to be found adjacent to each other, and the networks of the different sizes cover the whole area. They can be "thrown over our plane at will," but they should have one center in common (the major city), and they should be arranged in such a way that sectors with more and sectors with fewer market towns result. Due to the existence of those "city-richer" and "city-poorer" sectors, there are some bigger and more intensively used traffic lines which mostly leave the central city and lead out from there to the sectors, especially to those that contain more central places. The resulting pattern resembles a cobweb, with the major city as the midpoint. The central city is also the midpoint of a system of rings of economic activity, which is the industrial equivalent to von Thünen's agricultural rings.

Obviously, there will not only be one such system with one most central (major) city, but there will be many of them that are adjacent to each other and that in themselves have a honeycomb structure. In this way, Lösch develops a system of regular economic regions which with their network structure "form…an organic whole". The complete system consists of three different stages: simple market areas, regional networks, and regional systems.

20.2.3.3 Special Cases, Variations and Empirical Results

Lösch examines some typical regional systems, namely those where each area has the size of K regions of the next smaller size. For example, a 3-*system* would contain settlements and areas of the type 1 as the smallest units, then towns of type 4 would be the next larger size (as type 2 and 3 cannot cover three areas of the type 1, they do not exist in a 3-system), followed by type 11 on the next stage which contains three areas of type 4 (again, types 5–10 do not exist), followed by type 30 and type 77. The resulting system has the same structure as developed by Christaller (see his $K=3$ *system* as illustrated in Fig. 20.7). According to Lösch, it has the advantage of having a clearer structure than the full system, but the disadvantage of being less economical, as in this system the areas of many goods are larger than they actually need to be.

Lösch also criticizes Christaller for his claim that the $K=3$ system was most economical. However, this is merely a question of definition, as Christaller assumes the supply principle, and his objective is therefore to supply the whole population from as few places as possible. Obviously, this must lead to a result different from Lösch's who assumes that the number of "independent existences" (suppliers) should be maximized. But nevertheless, Christaller's theory of central places can be viewed as a special case (or better, the three systems can be viewed as three special cases) of Lösch's more generalized theory.

After the presentation of his general theory, Lösch varies his assumptions and studies economic differences (price, product differentiation, the importance of the freight rate) and natural differences (productivity, accessibility, human and political differences) and their impact on his model and its results. It turns out that they

lead to modifications of the theoretical structure in reality, allowing irregular and overlapping market shapes and less regular networks.

Part III of the book is devoted to the specifics of trade. Here, Lösch relates the location problem to other important areas, such as the choice of occupation. Again, he takes a marginal view and develops conditions for an equilibrium in which each individual, e.g., a worker or an entrepreneur, realizes the highest utility possible.

In the fourth and final part of his work, Lösch studies the spatial structure of locations in the United States. For example, among bank branches in Iowa he finds a uniform distribution which agrees with his theoretical results of a regular network structure, and therefore he concludes that "the spatial distribution of most non-agricultural enterprises corresponds very well, after all, with our theoretical model." Moreover, he finds that a "regular distribution of towns throughout the world is extraordinarily common."

20.3 Assessment of the Classical Contributions

As already stated in the introduction, the contributions by von Thünen, Christaller, and Lösch are pioneering and seminal, serving as the basis of many subsequent developments in different fields, including regional science, geography, economic theory, and location theory. Their work is highly interdisciplinary: the three books are named among the "path-breaking books in regional science" (Waldorf 2004) by the members of the Regional Science Association International (RSAI), while at the same time their authors—first and foremost von Thünen, but also Christaller and Lösch—are considered to be very important contributors to economic theory.

20.3.1 Von Thünen's Contribution

Von Thünen's work is the first publication in which—in Part I more implicitly, in Part II explicitly—*marginal analysis* is used to analyze and model an economic problem, namely optimal agricultural land use and the prices resulting from it. Based on the principles of arbitrage and marginal productivity, von Thünen develops his concept of the land rent and an equilibrium concept for land use. As these problems necessarily involve the existence (and location) of central towns as centers of consumption and the question of where to locate which agricultural activities, his contribution is a seminal part of the literature on spatial analysis. At the same time, it is an important contribution to the field of economics: in his "Foundations of Economic Analysis," Samuelson (1983) mentions J.H von Thünen alongside Leon Walras, J.S. Mill, and Adam Smith as one of the most important economists ever. Von Thünen's outstanding work has clearly had a major impact in more than one scientific field, and, as Fujita (2000) points out, his views and insights on modern aspects like agglomeration, are well ahead of his time.

Perhaps the most important contribution of von Thünen is the introduction of transportation costs into economic theory and the extensive analysis of their effects on land use and prices. Von Thünen was among the first to introduce the dimension of space into economic modeling, and in this way he prepared the ground for the many publications in location analysis, regional sciences and spatial economics that followed.

Although von Thünen's model has often been criticized for its simplifying assumptions—the homogenous plane, the single city, the static approach—it still remains one of the most important models of agricultural regional structures, and it is still studied and discussed, applied and modified by economists as well as geographers. This is, as Block and Dupuis (2001) assert, probably due to the important theoretical contribution, but also to its simplicity and to the fact that it is empirically relevant. For example, Rutherford et al. (1967) find a slightly modified ring (actually a belt) structure for agricultural activities in Australia, around and near Sydney.

The major criticism regarding von Thünen's model concerns his assumptions, especially regarding the uniformity of the plane. That he does not take into account the obviously uneven distribution of natural resources, renders his theory less relevant and applicable, at least at first sight. Authors such as von Böventer (1963) point out, though, that many of von Thünen's results remain true, albeit possibly in modified form, if this assumption is dropped.

While von Thünen mainly concentrates on "advantages of site" (i.e. the location), according to Lösch there are different factors which lead to different rings for different products: besides the advantages of site, there are advantages of source (like the quality of the soil) and advantages of scale (where larger amounts lead to lower cost per unit). Lösch also emphasizes that, as in industrial production, there is a tendency to maximize the number of producers, allowing each farm to be only the size necessary to support a family, and to maximize rent. This is an aspect that von Thünen does not study.

Moreover, von Thünen's theory does not explain how cities emerge, as it takes the central city as given. However, in the second part of his work, von Thünen relaxes the assumption of only one major city and even mentions a system of cities of different sizes covering and serving the whole state. Therefore, he can be viewed as a true predecessor for Christaller's central place theory. Finally, it should be noted that von Thünen already elaborates on aspects of industrial location, a theory which was to be formulated in detail no less than 80 years later by Weber (1909).

20.3.2 The Contributions by Christaller and Lösch

Christaller's central place theory is pioneering in the studies of economic regions because it does not concentrate on individual locations in isolation, but instead takes into account the interrelations of different economic activities and their locations. He is the first author to develop a complete system and hierarchy of urban

settlements. It is especially remarkable that the structure of this system is not determined by traffic conditions or other existing structures, but solely by the goods which are on offer at specific places and consumed at many others, the central goods. At the core of his theory is the idea that the goods and their consumption define and shape the economic landscape, an idea which is simultaneously new, simple, and utterly convincing. By taking into account that consumers have only a limited budget and can buy only goods relatively close to their home place, and that suppliers have to be able to sell enough of their product to make its production worthwhile, Christaller develops an equilibrium system of central places and the respective market areas which ensures the cost-minimal provision of all customers with all goods.

Central place theory became fundamental for subsequent developments in different fields of sciences, such as in urban systems research (Coffey 1998) or urban economics. Wang (1999) remarks that central place theory does "treat cities as spaceless points" in order to analyze the structure of a system of cities. Hence, the major contribution of central place theory is to explain the existence of urban centers and of their hierarchical order. As Christaller's own analysis of Southern-Germany and other empirical studies show, the patterns he developed are not to be found in reality in this pure hexagonal form, but nevertheless in many cases central place theory describes the locations of towns and trade activity rather well.

Christaller makes the basic assumptions that the plane is homogeneous and that each customer will patronize the closest location. The former has often been criticized, even by Christaller himself, as obviously physical features like mountains or rivers also have an effect on central place locations and the respective market areas. If, however, the assumption were dropped, a modified structure would result, but still Christaller's basic result of a hierarchy of urban settlements remains true. The second assumption has developed into a standard assumption in location analysis, and it is also common to split the demand of a customer location between two facilities if both are at the same distance from the customer (Plastria 2001), e.g., in competitive location theory (for an introduction and overview, see Eiselt and Laporte 1989). In gravity models (Huff 1964) this assumption is modified to take into account aspects of attractiveness, for example of different stores.

In fact, closest center choice does not describe actual behavior, because people tend to go to places where they can satisfy different needs at the same time. In other words, multi-purpose shopping will usually take place (Eaton and Lipsey 1982), in contrast to the single purpose travel assumed by Christaller in his basic model. For this reason, most trips are made to higher order places instead of places of lower order. The same tendency can be observed regarding medical treatment: While medical services of different order are offered in different places, and hence the medical sector is organized hierarchically and fits Christaller's assumptions very well, people tend to go to a larger regional hospital instead of requiring treatment at a smaller and more local clinic, even if they have got only a minor health problem. This behavior, in turn, increases the size of high order places and leads to an even stronger hierarchy, as Christaller himself actually anticipated; however, this devel-

opment is not part of his basic theory but only of the subsequent discussion, see Sect. 20.2.2.3 above.

Moreover, with increasing industrial production and less expensive transportation, physical proximity of producers and customers became less important than Christaller assumed. Finally, customers usually are not uniformly distributed, but, mainly due to historical reasons, centers with higher numbers of inhabitants usually already exist.

Nevertheless, central place theory had a big impact on actual regional planning decisions in Germany, both during and after World War II. As Rössler (1989) reports, Christaller was involved in the development of Hitler's "General Plan of the East" at Himmler's "Planning and Soil Office." This was a plan to reconfigure the geography in Eastern countries such as Poland, where millions of inhabitants were forced to leave their homes and were relocated or deported to enable the setup of a new hexagonal structure of settlements.

After the war, in the 1960s and 1970s, central place theory remained the most important theoretical concept for German regional planning activities, as these activities concentrated on supplying the whole population with the necessary goods within a preset distance (or travel time). In applications to the planning of German regional structures, a maximum of three to four hierarchical stages of towns and settlements are considered. The concept was much criticized during the 1980s and 1990s due to its suggested lack of flexibility and because it was said not to take into account modern ideas of "sustainable development." However, it is still useful as a basic concept for political planning and activities in the areas of spatial and regional decision making, even if in adapted form (Blotevogel 2002). It has also been used for settlement planning in other countries, especially for the polders in the Netherlands (Yoshio 2006).

The major contribution of Lösch is the generalization of Christaller's hierarchically structured approach to a more flexible system of central places. In contrast to Christaller, Lösch's approach leads to different types of places which specialize in different goods, but are not in a strict hierarchical order. "Smaller" centers can also serve "larger" centers in his system, which is more realistic than Christaller's concept. However, the resulting pattern of locations is less regular than the one which is produced by Christaller's theory, and therefore it is more difficult to evaluate it empirically (Lang 2002). This is probably why most scientists, and especially those with an empirical background, focus more on Christaller's original theory and less on Lösch's generalized approach.

In contrast to Christaller, Lösch discusses different reasons for the existence of towns which are to be found in large individual enterprises, in agglomeration (because of advantages of larger numbers in sales and procurement), in advantages of certain sites in terms of natural conditions and the structure of the population, and in the fact that competitors will come into the market until there are no remaining rents and pure competition is reached. With this discussion, Lösch sets the stage for a theoretical foundation of the existence of towns, as it was developed later on by other authors.

20.4 The Impact of the Work of von Thünen, Christaller, and Lösch

As alluded to above, the work by von Thünen, Christaller and Lösch had a major impact on the subsequent development in different areas of economic theory. Some of the most important contributions which are based on their theories are presented in this section, in order to illustrate the significance of their seminal work.

20.4.1 The Impact of von Thünen's Work

Soon after it was published, many German economists such as Hermann (1832), Schüz (1843), and Roscher (1854), built upon von Thünen's work. Even now, von Thünen is very famous in Germany. There is even a "Thünengesellschaft" that, among other activities, publishes Thünen-Jahrbücher. Predöhl (1928), who integrates the problem of location choice into production theory, uses von Thünen's work as a founding pillar of his theory; however, while von Thünen concentrates on agricultural production and location, Predöhl's focus is on industrial activities.

Internationally, von Thünen's work was only little known until Isard (1956) and a little later Chisholm (1962) discussed it in their books and until, in 1966, an English translation was published. This was the same year as the translation of Christaller's book, which had appeared in its original German version more than 100 years after von Thünen's book!

Spatial pricing models as introduced, for example, by Beckmann (1952, 1968) can be said to have their origin in von Thünen's work, as they explicitly take into account the impact of transportation cost on the price structure, an idea which was first developed by von Thünen.

Von Böventer (1963) presents a common framework for agricultural and urban location theory. In his discussion, von Thünen's rings as well as Christaller's (and Lösch's) central place theory play an important role as basic models of economic theory. Therefore, it should be noted that agricultural and urban land use theories are based on the same ideas and have common roots, mainly in von Thünen's work, whereas industrial location theory goes back to different sources (Launhardt 1885 and Weber 1909).

Alonso studies the structure of cities in his book "Location and land use" (1964). Based on von Thünen's model for agricultural structures, he develops a monocentric city model with one central business district (*CBD*) in the middle, surrounded by a residential region. (An earlier contribution along similar lines goes back to Burgess (1923) who, mainly on a empirical basis, studies the structure of the City of Chicago.) Under the assumption of individual utility maximization, Alonso develops a system of resulting bid rents (land rents), i.e., the prices of the land at different distances to the center. Analogously to von Thünen's theory for agricultural land use, due to the cost of transportation the land rents for urban use decrease with

increasing distance to the centre, which leads to decreasing land use intensity, e.g. to taller buildings in the *CBD* and to smaller buildings with less floors in the outskirts. Depending on their intensity of land use and thus their productivity, industrial activities can be found in even larger distance from the city center in a ring surrounding the residential areas, if they require much space, or closer to the center, in a second ring around the retail and service area.

Alonso's book is the pioneer work in urban economics and location analysis, and monocentricity—a concept that corresponds to von Thünen's assumption of a single, centrally located city—for a long while remained a basic assumption in urban economics which was also used by Mills (1967), Muth (1969) and others.

Sinclair (1967) argues that von Thünen's theory and results, especially the decreasing intensity of land use at larger distances from the market, were still valid for underdeveloped areas, while it was not true for the industrialized parts of the world, as here the most important factor was urban expansion. This "urban sprawl" leads to the reversed pattern, i.e., to rings of increasingly intensive land use with growing distance from the city.

Krugman (1991), the founder of the "New Economic Geography" (*NEG*), a rather new branch of spatial economics, builds upon von Thünen's work, as do Fujita and Thisse (2002) in their work on agglomeration. The aim of the *NEG* is to "explain the formation of a large variety of economic agglomeration in geographical space, using a general equilibrium framework" (Fujita and Mori 2005). It is the general equilibrium modeling approach—following, in a way, the spirit of Lösch, but going much further in modeling the market mechanisms—that characterizes the "New Economic Geography" and distinguishes it from traditional Economic Geography as represented by von Thünen and Christaller. The explanation for the formation of centers (i.e., regions in which economic activities concentrate) and cities given by *NEG* is mainly based on increasing returns to scale and, therefore, imperfect competition, and on the existence of transportation costs. One major result is Krugman's "core-periphery model" according to which two rather similar economic regions can develop differently, due to a small advantage one of them has got, e.g., in terms of costs: one of them develops into an industrial agglomerated "core" and the other into non-industrialized periphery.

If there are different industries with differing scale economies or transportation costs, it can be shown that there is a tendency towards a hierarchical structure as it was already developed by Christaller. Based on von Thünen's, Christaller's and Lösch's results, Fujita et al. (1999) develop an integrated model of the economy, consisting of an industrial core and an agricultural periphery, and provide an explanation of the formation of cities and systems of cities in which, as in Krugman's approach, the importance of imperfect competition is emphasized.

Finally, it should be noted that von Thünen's ring model is still discussed and applied today, e.g., to explain the location of milk production (Block and Dupuis 2001). In the tradition of Alonso, the "Concentric Zone Model," which is based on Thünen's rings, is used with respect to urban structures to study the location of different economic activities in an urban setting as described above, or, for example,

to analyze the residential locations of different income groups assuming a circular city (de Bartolome and Ross 2007).

20.4.2 The Impact of Christaller's and Lösch's Work

Central place theory has acted as a foundation for many contributions regarding systems of cities. Christaller's seminal work was followed by many publications, one of them being the work by Lösch (1962). Christaller himself refined his concept, extended it to a European scale, and discussed the different aspects that influence the locations of agricultural activities (which will usually be dispersed in the "sphere of influence" of a central place), of industrial activities (which are to be found close to a centre or on a traffic line connecting central places) and of the central places (markets) themselves (Christaller 1950).

The ideas of Christaller were introduced to the English speaking world by Ullman in his paper "A theory of location for cities" (1941). (Note that, by publishing some major results in an English journal, Lösch (1938) reached a bigger readership at a slightly earlier stage.) Later on, these ideas are discussed by Isard (1956) who combines central place theory with Weber's results on production location, with market area theory and with von Thünen's results on agricultural location to derive the first unified and generalized location principle. A little later, a quantitative model for systems of central places is presented by Berry (1964).

The two aspects of Christaller's and Lösch's theories which influenced the subsequent literature most are (a) the hierarchy of locations or market places and (b) the hexagonal structure of locations and market areas. Some contributions to both areas are discussed below.

20.4.2.1 Hierarchy of Locations

Beckmann (1958) and Parr (1969) are interested in the sizes of the cities on different levels of hierarchy in Christaller's central place system. Based on the assumptions that the size of a city is proportional to the population it serves and that each city of a certain order has a certain number of "satellite" cities of the next lower order, Beckmann develops a system of multipliers by which the respective city sizes can be found. Beckmann and McPherson (1970) generalize the approach such that the number of "satellites," and hence the relation between the sizes of the market areas, is allowed to change from level to level. In other words, they modify Christaller's hexagonal structure in a way similar to Lösch's approach. Central place theory is therefore now often linked to the question of city sizes and urban growth (see, e.g., Nourse 1978). However, as Burns and Hfaly (1978) point out, centrality is not only related to population size, but should primarily be measured in economic units such as occupation and related incomes.

Von Böventer (1963) combines von Thünen's and Christaller's theories to develop "a hierarchy of villages within a ring formation for the commodities" or, if more than one town is taken into account, "hierarchies of agricultural villages within systems of interrelated Thünen rings." He emphasizes that in their pure form, von Thünen's theory can be mainly applied to the primary (agricultural) sector, Christaller's theory to the tertiary (services) sector and Lösch's theory to the secondary (manufacturing and production) sector. Regarding central place theory, this view is supported by Wyckoff (1989), who finds that it mostly held true for the service sector in Colorado at the end of the nineteenth century.

Beavon and Mabin (1975) restate and clarify some aspects of Lösch's theory, especially with respect to the development of the system of market areas of different hierarchical order. They emphasize that the "city-rich" and "city-poor" sectors are a constraint, and not a result, of Lösch's system. Moreover, they argue in favor of its use as a theory for urban development, i.e., as a concept which can represent and explain what they call the "internal tertiary structure" of a city.

In their work on spatial competition among shopping-centers, Eaton and Lipsey (1982) assume multi-purpose shopping behavior on the side of the customers and, based on this assumption, develop a hierarchy of shopping centers similar to Christaller's hierarchy of central places. In their seminal study, they concentrate on a one-dimensional market and on only two goods, and they develop important insights on agglomeration effects. Empirical studies regarding the attractiveness, growth and decline of shopping centers based on Christaller's central place theory and on the results of Eaton and Lipsey are presented, e.g., by Ryan et al. (1990) and Dennis et al. (2002).

20.4.2.2 Hexagonal Structure

Isard (1956) drops the assumption of a uniformly distributed population and concludes that the size (and shape) of a market area depend on the population density, and therefore will vary. A similar result is achieved by Rushton (1972). Friedmann (1961) continues the development of central place theory, concentrating on aspects of political, cultural, and social authority due to which the surrounding regions depend on the respective major center, and on the sub-centers which take care of some subordinate services. Again, the approach leads to an irregular structure of the resulting regions and centers. Allen and Sanglier (1979) study the influence of the introduction of new goods and services by simulation. Also their results do not show a regular pattern, but different irregular results. On the other hand, in his empirical study of the current German structures, Lang (2002) finds rather regular patterns of cities, but no obvious hexagonal structure.

Eaton and Lipsey (1975, 1976) point out that in a competitive environment, i.e., under the assumption of free market entry, no hexagonal market structure has to result in a two-dimensional market, given a uniform distribution of customers. While the hexagonal structure is the "planner's solution" that minimizes total transportation cost, and thus is efficient (Beckmann 1968), according to Eaton and Lipsey

it is not the solution that results from profit-maximizing behavior of the firms, at least when up to 19 firms are studied. However, Okabe and Suzuki (1987) come to the conclusion that for an even larger number of firms (up to 256 firms in their numerical tests), the "quasi-global equilibrium" configuration that results from spatial competition, will be similar to the hexagonal structure resulting from social planning. In an iterative procedure, Okabe and Suzuki apply Voronoi polygons in order to determine the market areas the competing firms can achieve from their current locations, then the firms relocate in turn if they can increase their market area by the respective move, and so on. The procedure results in a near-hexagonal structure; however, it should be noted that the two structures do not agree completely, but are only similar, and even in the case where the simulation is started with the socially optimal pattern, this pattern is actually destroyed during the process.

Drezner and Zemel (1992) examine the sequential competitive location problem in the plane in which two competitors can open multiple facilities each. Under the assumption of a uniform distribution of customers, the first competitor wants to choose his locations such that the second competitor is prevented from capturing too much of the market. The authors show that under these circumstances the hexagonal, honeycomb pattern is the best location structure for the first competitor to defend his market area.

Okabe et al. (1997) study systems of successively "inclusive" and "exclusive" hierarchical facilities. Here, an "inclusive" type of hierarchy is one, in which the facilities of higher order offer all services of lower order as well, and which occurs, e.g., in the medical sector. In the "exclusive" hierarchy, not all services of lower order are offered at each central place. As a result, the planning problem consists of the decision about the hierarchical structure of the facilities *and* about their spatial configuration in areas with uniform customer distribution. As it turns out, the regular triangular lattice as used by Christaller and Lösch is a basic feature of the solution for each stage of the "exclusive" problem. From the solutions of the different hierarchical stages, a solution of the "inclusive" problem can be derived which closely resembles the hexagonal structure suggested by Christaller.

Finally, Suzuki and Okabe (1995) show that a hexagonal structure also results for the continuous p-center problem in which the maximum distance from a user to his closest facility is to be minimized. Hence, the basic hexagonal structure that was first developed and studied by Christaller and Lösch is a characteristic feature of many different spatial planning situations which are of major interest to researchers and scientists up to this day. Moreover, central place theory is discussed and applied also in special areas of research such as sport tourism; see, e.g., Daniels (2007).

20.5 Future Research Directions

As the classical publications, as well as most of the work that followed, concentrate on a homogeneous plane with a uniform or regular customer distribution, future research might focus on the modification of these assumptions, such as non-uniform

demand and/or forbidden regions. Especially in the field of hierarchical and competitive location, this would lead to new and interesting insights which are more closely related to reality.

With the increasing globalization of all economic activities, global structures and systems of cities, and locations are increasingly the focus of economists and regional scientists. Therefore, central place theory and the theories building upon it can be useful to derive theoretical insights regarding the future development of these global structures, such as the pattern of future "global central places." Moreover, the ring concept that was originally developed by von Thünen and modified by Alonso for urban structures could be used to explain and perhaps also forecast the future development of those huge cities: for example, residential areas can be expected to be found farther and farther away from the city centers, while the centers' predominant functions are to host commercial activities.

In general, urban structures and their developments have to be studied to be able to plan and (politically) direct their growth and development. In particular, the development of "medium-sized cities" will be of interest, as it has not received as much attention as large cities have. It can be expected that, especially in Germany, central place theory will remain an important supporting tool for political decision making in regional planning, specifically with respect to the development of infrastructure in certain areas and with respect to the allocation of financial incentives to certain branches of industry.

Changes in transportation infrastructure have an important impact on both large scale and small scale planning and need to be taken into account in the future development of models in both areas. Here, the concept of "economic distances" as developed by Christaller may lead to further important insights, as the actual distances are less and less important, while the importance of "felt distances" increases. Due to the possibility of getting to literally every point on earth within a day or two, and of getting information from everywhere within seconds, some services do not have to be offered locally at all, but can be received by the customers even over a very long distance. It would be interesting to examine how the structure of the system of central places changes due to these developments.

Finally, environmental and ecological issues could also be incorporated in the respective approaches. This change of the planning objective will most likely lead to modified ring or network structures.

References

Allen PM, Sanglier M (1979) A dynamic model of growth in a central place system. Geogr Anal 11:256–272

Alonso W (1964) Location and land use. Harvard University Press, Cambridge

Beavon KSO, Mabin AS (1975) The Lösch system of market areas: derivation and extension. Geogr Anal 7:131–151

Beckmann MJ (1952) A continuous model of transportation. Econometrica 20:642–660

Beckmann MJ (1958) City hierarchies and the distribution of city size. Econ Dev Cult Change 6:243–248
Beckmann MJ (1968). Location theory. Random House, New York
Beckmann MJ, McPherson J (1970) City size distributions in a central place hierarchy: an alternative approach. J Reg Sci 10:25–33
Berry B (1964) Cities as systems within systems of cities. Pap Reg Sci 13:146–163
Block D, Dupuis EM (2001) Making the country work for the city: von Thünen's ideas in geography, agricultural economics and the sociology of agriculture. Am J Econ Sociol 60:79–98
Blotevogel HH (2002) Fortentwicklung des Zentrale-Orte-Konzepts. Forschung- und Sitzungsberichte ARL, 217. Akademie für Raumforschung und Landesplanung, Hannover, Germany
Burgess EW (1923) The growth of the city: an introduction to a research project. Publ Am Sociol Soc 18:85–97
Burns RS, Hfaly RG (1978) The metropolitan hierarchy of occupations—an economic interpretation of central place theory. Reg Sci Urban Econ 8:381–393
Chisholm M (1962) Rural settlement and land use, Hutchinson, London
Christaller W (1933) Die zentralen Orte in Süddeutschland, Fischer, Jena
Christaller W (1950) Das Grundgerüst der räumlichen Ordnung in Europa: Die Systeme der europäischen zentralen Orte. Frankfurter Geographische Hefte 24(1)
Coffey WJ (1998) Urban systems research: an overview. Can J Reg Sci 21:327–364
Daniels MJ (2007) Central place theory and sport tourism impacts. Ann Tourism Res 34:332–347
De Bartolome CAM, Ross SL (2007) Community income distributions in a metropolitan area. J Urban Econ 61:496–518
Dennis C, Marsland D, Cockett T (2002) Central place practice: shopping centre attractiveness measures, hinterland boundaries and the UK retail hierarchy. J Retailing Consum Serv 9:185–199
Drezner Z, Zemel E (1992) Competitive location in the plane. Ann Oper Res 40:173–193
Eaton BC, Lipsey RG (1975) The principle of minimum differentiation reconsidered: some new developments in the theory of spatial competition. Rev Econ Stud 42:27–49
Eaton BC, Lipsey RG (1976) The non-uniqueness of equilibrium in the Lösch location model. Am Econ Rev 66:77–93
Eaton BC, Lipsey RG (1982) An economic theory of central places. Econ J 92:56–72
Eiselt HA, Laporte G (1989) Competitive spatial models. Eur J Oper Res 39:231–242
Fejes Toth L (1960/1961) On the stability of a circle packing. Ann Univ Sci Budap, Sect Math 3–4:63–66
Friedmann J (1961) Cities in social transformation. Comp Stud Soc Hist 4:86–103
Fujita M (2000) von Thuenen and the new economic geography. Discussion Paper no. 521, Kyoto Institute of Economic Research. Kyoto, Japan
Fujita M, Mori T (2005) Frontiers of the new economic geography. Discussion Paper 27, Institute of Developing Economies, JETRO, Japan
Fujita M, Thisse J-F (2002) Economics of agglomeration: cities, industrial location and regional growth. Cambridge University Press, Cambridge
Fujita M, Krugman P, Venables A (1999) The spatial economy: Cities, regions and international trade. MIT Press, Cambridge
Hermann FBW (1832) Staatswirthschaftliche Untersuchungen über Vermögen, Wirthschaft, Productivität der Arbeiten, Kapital, Preis, Gewinn, Einkommen und Verbrauch. Anton Weber'sche Buchhandlung, München, Germany
Huff DL (1964) Defining and estimating a trading area. J Mark 28:34–38
Isard W (1956) Location and space-economy: a general theory relating to industrial location, market areas, land use, trade and urban structure. MIT Press, Cambridge
Krugman P (1991) Geography and trade. MIT Press, Cambridge
Lang B (2002) Die Untergliederung der Bundesrepublik Deutschland in strukturierte Wirtschaftsregionen. Peter Lang Verlag, Frankfurt, Germany
Launhardt W (1885) Mathematische Begründung der Volkswirtschaftslehre. Wilhelm Engelmann, Leipzig

Lösch A (1938) The nature of economic regions. South Econ J 5:71–78
Lösch A (1954) The economics of location. Translated by W.H. Woglom. Yale University Press, New Haven
Lösch A (1962) Die räumliche Ordnung der Wirtschaft, 3rd edn. Fischer, Stuttgart (1st edn published in 1940)
Mills ES (1967) An aggregative model of resource allocation in a metropolitan area. Am Econ Rev 57:197–210
Muth R (1969) Cities and housing. University of Chicago Press, Chicago
Nourse HO (1978) Equivalence of central place and economic base theories of urban growth. J Urban Econ 5:543–549
Ohlin B (1933) Interregional and international trade. Harvard University Press, Cambridge
Okabe A, Suzuki A (1987) Stability in spatial competition for a large number of firms on a bounded two-dimensional space. Environ Plan A 19:1067–1082
Okabe A, Okunuki K-I, Suzuki T (1997) A computational method for optimizing the hierarchy and spatial configuration of successively inclusive facilities on a continuous plane. Locat Sci 5:255–268
Palander T (1935) Beiträge zur Standorttheorie. Almqvist & Wiksell, Uppsala, Sweden
Parr JB (1969) City hierarchies and the distribution of city size—a reconsideration of Beckmann's contribution. J Reg Sci 9:239–253
Plastria F (2001) Static competitive facility location: an overview of optimisation approaches. Eur J Oper Res 129:461–470
Predöhl A (1928) The theory of location in its relation to general economics. J Polit Econ 36:371–390
Rössler M (1989) Applied geography and area research in Nazi society: central place theory and planning, 1933 to 1945. Environ Plan D 7:419–431
Roscher W (1854) Die Grundlagen der Nationalökonomie. J.G. Cotta'scher Verlag, Stuttgart, Germany
Rushton G (1972) Map transformations of point patterns: central place patterns in areas of variable population density. Pap Reg Sci 28:111–129
Rutherford J, Logan MI, Missen GJ (1967) New viewpoints in economic geography. Harrap, Sydney
Ryan DL, von Hohenbalken B, West DS (1990) An econometric-spatial analysis of the growth and decline of shopping centres. Reg Sci Urban Econ 20:313–326
Samuelson P (1983) Foundations of economic analysis. Harvard University Press, Cambridge
Schüz CWCh (1843) Grundsätze der National-Oeconomie. Druck und Verlag von C.F. Osiander, Tübingen, Germany
Schumpeter JA (1955) History of economic analysis. Oxford University Press, New York
Sinclair, R (1967) von Thünen and urban sprawl. Ann Assoc Am Geogr 47:72–87
Suzuki A, Okabe A (1995) Using Voronoi diagrams. In: Drezner Z (ed) Facility location—a survey of applications and methods. Springer-Verlag, New York, 103–118
Toregas C, Swain R, ReVelle C, Bergman L (1971) The location of emergency service facilities. Oper Res 19:1363–1373
Ullman E (1941) A theory of location for cities. Am J Sociol 46:853–864
von Böventer E (1963) Towards a united theory of spatial economic structure. Pap Reg Sci Assoc 10:163–187
von Thünen JH (1921) Der isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie, 2nd edn. Fischer, Jena, Germany. (The 1st edition was published in 1826)
von Thünen JH (1966) von Thünen's isolated state. Translated by Wartenberg CM. Pergamon Press, Oxford
Waldorf BS (2004) Path-breaking books in regional science. Pap Reg Sci 83:59–89
Wang F (1999) Modeling a central place system with interurban transport costs and complex rural hinterlands. Reg Sci Urban Econ 29:381–409
Weber A (1909) Über den Standort der Industrien. 1. Teil: Reine Theorie des Standorts. Mohr Siebeck Verlag, Tübingen, Germany

Wyckoff W (1989) Central place theory and the location of services in Colorado in 1899. Soc Sci J 26:383–398
Yoshio, S. (2006) Planning on Settlement Location in the Ijsselmeerpolders and Central Place Theory. Geogr Rev Jpn 79:566–587

Index

H. A. Eiselt, V. Marianov (eds.), *Foundations of Location Analysis,*
International Series in Operations Research & Management Science 155,
DOI 10.1007/978-1-4419-7572-0, © Springer Science+Business Media, LLC 2011